Conference on New Techniques and Technologies for official Statistics

NTTS 2023

Charlemagne Building, Brussels, 7–9 March 2023

Boulevard Charlemagne 52, 1000 Brussels

Book of abstracts

'New Techniques and Technologies for Statistics' (NTTS) is an international biennial scientific conference series organised by Eurostat on new techniques and methods for official statistics, and the impact of new technologies on statistical collection, production and dissemination systems. **NTTS 2023** will take place in **Brussels on 7-9 of March 2023**, with satellite events on the 6 and 10 of March. The programme of the satellite events can be found at: https://ec.europa.eu/eurostat/cros/content/NTTS2023_en#section-76

The purpose of NTTS conferences is both to allow the presentation of results from currently on-going research and innovation projects in official statistics and to stimulate and facilitate the preparation of new innovative projects, with the aim of improving and enhancing the quality and usefulness of official statistics, as well as to prepare activities related to research in statistics. The target audience includes the official statistics R&D community at large: colleagues from statistical institutes, academia, European institutions and actors from the private sphere.

Web Data (JENK1M.1) **10**

A Web-Intelligence information system to support the production of EGR statistics *11*

Improving the quality of the job vacancy survey in Poland using administrative and online data *16*

USING WEBDATA TO DERIVE THE ECONOMIC ACTIVITY OF BUSINESSES *21*

Sampling (MANS 1M.1) **26**

Random Forest algorithm to adjust for Census population over-counts *27*

Combining probability and non-probability samples on an aggregated level *33*

Keywords: categorical data, combined estimator, selectivity, data integration *33*

Data Sources (MANS 1M.2) **38**

How to identify private households in French surveys? *39*

Administrative Data assisting Data Collection through Machine Learning *45*

Data Integration (JENK 1M.2) **50**

Quality measure to evaluate statistical matching methods *51*

Data linking from different sources – key to better statistics *56*

POST01: Poster session **58**

Exploring the Italian ecoregions: institutional fragmentation and socio-demographic characteristics .. *59*

Estimation of hidden populations using single-source capture recapture models with singleRsource package *64*

Protection of linked tables with a suppressive approach. Method and Use Cases. *67*

Case studies in using Satellite Earth Observation for National Statistics: the GAUSS project *71*

Application of an Adaptive Survey Design on the Italian Population Census *75*

A New Way to Learn Data Science: The Funathon, a Non-competitive Hackathon *80*

Visualising the pathways in and out of employment during COVID-19 in Estonia *84*

Web scraping (GASP 1A.1) **88**

Web scraped data in consumer price indexes *89*

Use of data obtained by web scraping for statistical purposes *94*

Can real estate web data augment official statistics? *98*

Innovative data collection in Social Statistics (JENK 1A.1) **102**

What is the business case for a smart household budget survey? Experiences from the field in six countries *103*

CRŒSS: A cross-domain data collection platform for the ESS *108*

Data Visualization (MANS 1A.1) **110**

An innovation approach of the data dissemination and data visualisation in official statistics. *111*

Territorial Economic Data viewer (TEDv). Data integration and visualisation of multiple EU funds at regional level *112*

restatapi: an R package to search and retrieve data from the changing Eurostat dissemination chain *116*

Smart data (MANS1A.2) **120**

Smart Mobility Surveys: Current App Developments, Methodological Perspectives, and Respondent Interaction *121*

Challenges in the Smart Household Budget Survey: how to involve & motivate respondents *125*

Advances in Eye-Tracking and Cognitive Interviewing Methodology: Dos, Don'ts, and Decisions *131*

Demography (JENK 1A.2) **138**

Avoiding the avoidable: Simulating long-term population changes in Germany without avoidable mortality *139*

Estimation of mode effects in the French population census *144*

Geographic Infrastructure for the 2021 Population and Housing Census *148*

From Classifications to Artificial Intelligence: innovative services of the Publications Office (GASP1A.3) **152**

A successful collaboration on classifications and what is in it for the statistical community *153*

Machine Learning Support for Repository Services — Reception, Cellar, and Semi-automated Metadata Automation *158*

Confidentiality (JENK1A.3) **164**

Data linkage with a search engine: identifying individuals in a population register *165*

Synthetic Dataset Generation for Anonymised Machine Learning Applications *169*

R/Julia/Python (MANS1A.3) **173**

Julia as a software for Official Statistics and Social Sciences *174*

New Scientist IT Infrastructure *179*

An R package for automatically generating candidate correspondence tables between classifications *183*

New skills in Symbolic Data Analysis for Official statistics (GASP2M.1) **188**

An illustration of the use of the measures *189*

s-concordance and s-discordance in applications *189*

Exactly mergeable summaries *191*

Citizen Science (JENK2M.1) **192**

To what extent do data privacy concerns and digital distrust act as barriers to smartphone sensor data collection in general population surveys and what can be done to mitigate them? 193

Eurostatistics - from PDF format to interactive web visualisation using R 196

Nowcasting (MANS2M.1)..200

GDP-Flash estimates: a quality assessment..201

Density forecasts of inflation: a quantile regressions forest approach...205

Analyses of the connection and the interaction between retail products markets and cash-holding ratio through panics using dynamic factor..216

European Master in Official Statistics (EMOS) (GASP2M.2)..220

Estimating Spatial Distribution of the NEET Rate in Italy by Small Area Methods 221

Measuring competitiveness: a new composite indicator for Italian municipalities..............................230

Interactions within a multi-layer EU inter-bank network..239

Online Data (JENK2M.2)..245

Development of an OJA gold standard to classify occupation...246

URL finding methodology...251

National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption 255

Data Literacy (MANS2M.2)...261

Creating, testing and maintaining a functional library for statistical methods....................................262

POST02: Poster session 265

Evaluation of Some Methods of Detecting Outliers...266

Pre and post COVID-19 related statistics analysis for identifying future opportunities based on historical dynamics...267

Unsupervised ranking and categorisation of companies using web scraping and machine learning....272

Integrating multimode survey data with VTL 276

Modelling local level housing demand based on the Multi-sectoral Regional Microsimulation Model (‘MikroSim’)..279

Non-traditional data and methods for identifying patterns of inequality to digital access in the EU ...284

Natural Language Processing to automate data coding 288

Maritime mobility statistics using open data...293

Statistical Data Triplification : The case of semi-automatic generation of RDF triples from relational databases...298

Local population projections with Bayesian hierarchical models 303

Data circularity for the circular economy..307

Statistics during pandemics (GASP2A.1) **309**

SORS decision making support system in COVID-19 emergency situations *310*

The Effect of Non-pharmaceutical Policy Interventions on Population Mobility During the COVID-19 Pandemic *313*

The behaviour of alternative trend-cycle decomposition methods during the pandemic *318*

Equality (JENK2A.1) **323**

Promoting the collection and use of equality data: measuring 'racial or ethnic origin' *324*

Beyond the gender pay gap *332*

New statistics on discrimination and gender based violence at EU level *336*

Machine learning 1 (MANS2A.1) **342**

Early provision of economic short-term indicators using Machine Learning *343*

Machine learning methods to support a big data-based household budget survey *349*

Machine learning and wealth measurement : an experiment on housing wealth of French households *354*

Social statistics (GASP2A.2) **360**

Subjective poverty in the European Union: the EU-SILC approach *361*

Chatbots to 'talk' with public data: the case of the Luxembourg Income Study (LIS) *365*

Adding a new mode to the mix: Implementation and evaluation of CAWI in EU-SILC *370*

Tracing respondents' footpaths – How user journey analysis can offer new insight into survey experience *374*

High frequency Data (JENK2A.2) **379**

COVID-19 daily number of diagnoses: an ARIMA analysis *380*

Time disaggregation in the Labour Force Survey using statistical learning and state-space models *384*

Temporal Disaggregation of the Services Producer Price Index *389*

Machine Learning 2 (MANS2A.2) **393**

Measuring the Italian Sentiment on economy: a Deep Learning application to the sentiment analysis *394*

An Experimental Study on Using NLP to Summarize Biotechnology Activities in Türkiye in a few Words *399*

Fuzzy Clustering based Autocoding Methods for Family Income and Expenditure Surveys *404*

MNO data (GASP2A.3) **409**

Estimation of the length of stay of foreigners in Poland using mobile big data *410*

Use of mobile network data for official statistics *414*

Synthetic mobile networks data generation 418

Generation of Synthetic Data from actual data of Mobile Network Operators (MNO) through Generative Adversarial Networks (GANs): motivations, techniques and preliminary results 423

Transport Statistics (JENK2A.3) **428**

Port Visits Using New Data Sources 429

TranStat - application of maritime data in the traffic intensity and transportation volume 432

Using Open-source Data to Estimate Mobility Statistics 437

Blockchain and Artificial Intelligence (MANS2A.3) **441**

Exploring the potential of blockchain to support innovation in European statistics. An interactive workshop by the European Blockchain Services Infrastructure 442

webAI - A new type of web-based indicators for economic analyses: The example of sustainability 445

The usage of blockchain technology for official statistics 449

Microsimulation (GASP3M.1) **452**

A microsimulation model for population projections in official statistics 453

A combined methodology SimulSTAT for public data indicators by using statistical and simulation means for improving decision-making in administration policies 458

A confidentiality concept for a simulation data centre 463

Privately Held Data (JENK3M.1) **468**

(Re-)using privately held data : experimental statistics on online collaborative economy platforms ... 469

Practical Privacy-Aware Data Linkage and Statistical. Aggregation based on Privacy Enhancing Techniques 473

Towards a Taxonomy for Business-to-Government Data Sharing 478

Agriculture Statistics (MANS3M.1) **483**

Strategy to Modernise Agricultural Statistics: New Pathways for the Future 484

Strengthen the statistical capacities in the developing country to close the agricultural data gap 488

Communication and dissemination (GASP3M.2) **493**

The release package of the Eurostat regional yearbook 2022 – Something for everyone? 494

Visualization and communication of statistics 498

Environmental Statistics (JENK3M.2) **504**

Satellite altimetry to water level monitoring: Italian and Finnish case studies 505

Degree of urbanization in Catalonia: high-resolution grids and commuting times 506

Remote sensing application in official statistics.Urban green index for Romanian cities. 511

Energy data spaces and statistics (MANS3M.2) **516**

POST03: Poster session **518**

Eurostat's GISCO services at your disposal *519*

From a glossary to a controlled vocabulary: a preliminary Istat experience *522*

An open source data science platform to foster innovative and reproducible statistical projects *527*

Mapping circular economy EU funded projects using text mining *532*

Integrating Big Data in the traditional survey of maritime transport statistics: open challenges *537*

How well-behaved are revisions to quarterly fiscal data in the euro area? *542*

An Android app for continuous census management *546*

Bridging independent taxonomies using AI: an application to skills mismatch using PIAAC, ESCO Skills and Online Job Ads *550*

Use of multilevel grids to release protected grid data from the French 2021 Census *555*

Hedonic House Price Index for Poland based on listings *559*

Geospatial Data Store - centralization of spatial data in Statistics Poland *564*

Business statistics (GASP3A.1) **566**

The Italian Social Mood on Economy Index: sub-indices focused on business *567*

Business register improvements: a balance between search, scrape and 3rd party web data *571*

Innovating Business Data Collection: System-to-System Data Communication applied to an agricultural survey *576*

Geospatial statistics (JENK3A.1) **581**

Towards standardised geospatial statistics *582*

New and improved tools for spatial statistics *584*

GSGF Europe – opportunity to re-assess national geospatial and statistical data integration practices *587*

Price data (MANS3A.1) **588**

Evaluating energy prices and costs impacts on households and industry's costs *589*

Price imputation methods based on statistical algorithms *593*

Automatic classification for price statistics using machine learning methods *595*

Geospatial statistics 2 (JENK3A.2) **600**

Wasserstein Distances for Spatial Statistics: The Spatial-KWD library *601*

On the Application of the Fourier-based Distance to Spatial Statistics *606*

Exploring Spatial and Demographic Official Statistics on Personal Insolvency *610*

Geocoding millions of addresses in a reproducible manner for Big Data climate risk analysis 615

Energy Statistics (MANS3A.2)...**620**

Energy consumption decomposition analysis using European official statistics – methodology and input data..621

Y-0 ESTIMATES OF SUPPLY SIDE OF ENERGY BALANCES AND KEY INDICATORS 626

Import dependency by origin indicator...628

Evaluating energy prices and costs impacts on households and industry's costs.................................631

On line Job Advertisment (GASP3A.3)...**635**

Towards statistics on skills: considerations on using WIH OJA data...636

Changes of skill requirements across jobs and labour markets: evidence from web vacancies 641

*Demand for digital skills: experimenting with information extraction from online job advertisements*645

Online Job Advertisements: the Italian Case Study..650

GISCO (JENK3A.3)..**654**

New spatial snow statistics in Finland using Earth observation and modelling 655

Innovative in situ and other Earth Observation platforms for air quality official statistics: the GAUSS project...659

Web cartography for gridded statistics – the Gridviz library..664

The Polish Use Case in Project GAUSS – extent and quality of green areas Smart Statistics based on EO data..669

Time series analysis (MANS3A.3) **673**

Online job ads' time series: A novel approach 674

Fixing the model for the seasonal component: INE's revision policy..679

The complexity of the turning points detection in 2022 684

Testing for cointegration in STAR models: A simulation-based study..688

JDemetra+ 3.0: New (R) tools for (high-frequency) time series analysis..693

Web Data (JENK1M.1)

Session Chair: **Roeland Beerten** *(National Bank of Belgium)*

A Web-Intelligence information system to support the production of EGR statistics

Antonio Laureti Palma, Ioannis Sopranidis, Fernando Reis (Eurostat), Alexandros Bitoulas*,* Anca Maria Kiss, Gabriele Marconi *(Sogeti)*

Improving the quality of the job vacancy survey in Poland using administrative and online data

Maciej Beręsewicz *(Poznań University of Economics and Business)*, Marek Wydmuch *(Poznań University of Technology), Herman Cherniaiev, Robert Pater (University of Information Technology and Management in Rzeszow).*

Using Webdata to derive the economic activity of businesses

Manveer Mangat *(Statistics Austria)*, Alexander Kowarik *(Statistics Austria)*

A Web-Intelligence information system to support the production of EGR statistics

Keywords: Eurogroups Register, Multinational Enterprises, Web Scraping, API

Introduction

Economic globalisation creates new opportunities for businesses to organise their production chains using multinational enterprise (MNE) groups. In this context, the highly volatile evolution of the dynamics of such global organizations makes it extremely challenging for the statistical business registers to support the production of macroeconomic and business statistics with the highest possible quality.

The EuroGroups Register (EGR)

The EGR (Euro Groups Register) is the European statistical register on MNE groups created by the European Statistical System and managed by Eurostat. It receives input data from the National Statistical Institutes (NSIs) of the European Member States and EFTA countries and a commercial data provider, consolidates them and makes it available for statistical purposes [1].

In EGR, about 10% of the MNE groups account for 90% of the employment. This clearly indicates a high polarisation of the MNE groups recorded in EGR, with few and very large ones dominating. To ensure better timeliness and higher accuracy for a limited set of these large MNE groups that impact significantly the European statistics, while at the same time accepting only a good coverage and level of accuracy for the smaller MNE groups, the European Statistical System agreed on a new strategy. The so-called “two-tier approach” refers to the split of the EGR population in two groups, a “top-tier” (largest, most important and most complex MNE groups) and a “bulk-tier” (the rest of MNE groups). This can allow for two different data quality management processes for the update of the two sub-populations.

The “top-tier” group, which has a significant impact on economic and business statistics, needs to be monitored regularly and to be updated frequently in order to provide users with highly accurate and up-to-date statistical information. To this end, new public sources from the World Wide Web (the Web), can help interpreting any changes or modifications that may occur in enterprise groups.

Use of innovative ways to improve the quality of EGR data

Within the EGR innovation process, different activities are foreseen for improving the quality of EGR information. One of these activities, is to use public data sources and provide them to the users as options to compare and update missing or erroneous information.

Such data sources could be the official web site of an MNE, its annual reports, specialized public group registers, or other easily accessible public source of information.

Web intelligence is a relatively new area of scientific research that makes use of big data tools for extracting and exploring information from the Web. To explore these capabilities and answer to a smart approach for supporting EGR in its innovation process, Eurostat carried out a “proof-of-concept” study (hereafter “the study”) on ‘Smart Data for MNEs’ [2].

This abstract presents the work done so far in this direction: the proof-of-concept study to obtain public information from the web, an assessment of the quality of these results, as well as ongoing and future work towards the implementation of the new Web-Intelligence based information system which can support the data quality improvement on multinational enterprise groups.

Methods: Web-intelligence approach

In order to explore different ways to improve the data quality in EGR, Eurostat has experimented with the use of public information from the web. This required specific web intelligence (web scraping and API querying) on a list of selected sources, and the implementation of a final dataset, along with a data viewer to display the information sought in an efficient view.

Web Intelligence was focused on a selected number of MNE groups (about 200 MNE groups) mainly operating in the EU and EFTA countries and included some whose headquarters is based outside the European Union. The web-intelligence process was carried out in two phases: in the first phase (discovering phase), public and open sources for MNE groups were investigated and in the second one (implementation phase), an information database was built integrating all available scraped data.

Discovery phase

The discovery phase identified the public sources that could be used, looking at information on the control structure of the groups, their global group heads, the country of global decision centres, the main activity codes, the consolidated persons employed, turnover and assets.

A big pool of **web sources** were analysed (landscaping), to assess the relevance and quality of the available information, as well as any technical limitations imposed by the source on data retrieval. Out of these, seven web sources were finally selected, namely ***Wikipedia, Wikidata, DBPedia, GLEIF, Open Corporates, Open PermID*** and ***EDGAR***.

Sources such as GLEIF, Wikipedia, Wikidata and DBpedia, provided reference dates and some sort of time series for specific variables and were thus more useful in the analysis of quantitative data such as number of employees or turnover. For qualitative data, the best structured information came from GLEIF and OpenPermID.

Implementation phase

The information database (DB) was created on the basis of EGR key variables at group level. The DB implementation process was articulated for each key variable and was based on the three standard steps: extraction of each piece of information from the web sources, transformation of the information to reconcile it with the EGR variables, loading of the reconciled information into the DB.

The data were extracted either through **web scraping**, or via an **API** (whenever available). The source code [3] was written in *R and Python*.

The transformation step first required a variable reconciliation with the EGR database. This step included the record linkage of the extracted information with the proper EGR group identifier.

The loading step needed first a quality evaluation of the data source for each variable considered and then the proper loading process into the information database.

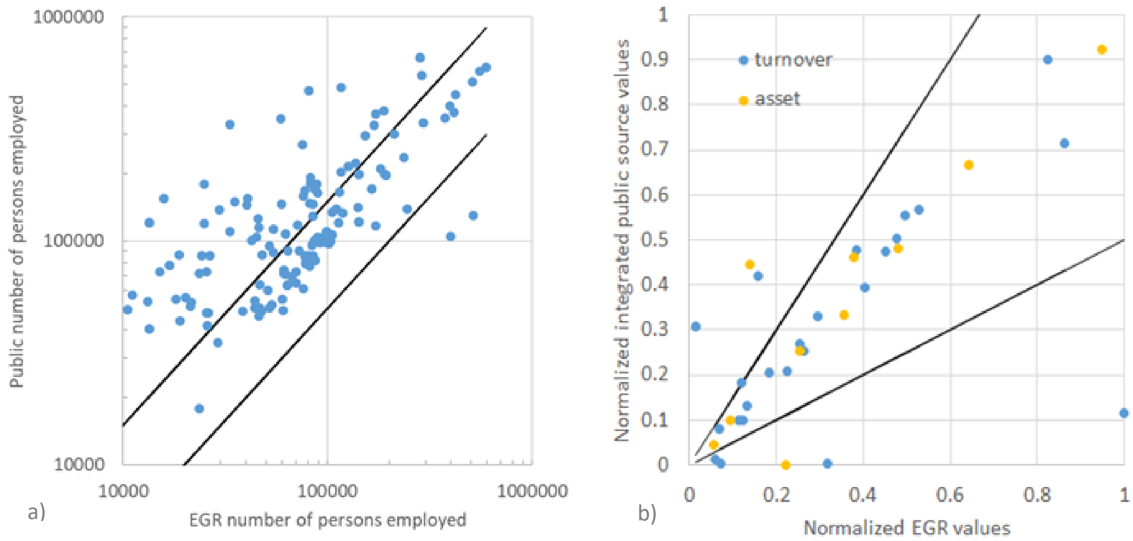
Quality evaluation of the data sources

In the quality evaluation of the loading step, each value of each transformed variable was compared with the information in EGR to measure the quality level and then a ranking order of each variable source loading.

To define a quality indicator for our variables, we first defined a relative difference of the variable between the public variable value and the EGR variable value. We then defined the variable quality indicator as the number of occurrences of the public variable with a relative difference in the range -1/2 to ½, divided by the total number of variable occurrences, i.e. all occurrences of the public source with relative differences in the interval [-1/2, 1/2] contribute to the numerator of the variable quality indicator.

The final information database was built on the base of an integration process which contained the information at group level from both the public sources and the official data from EGR.

Figure 1: a) Log-log scatter plot of number of persons employed, EGR versus public source; b) Normalized scatter plot of turnover and asset, EGR versus public source



Results

The comparison of the public data versus the official EGR data shows that results are positive for variables *Country of the Ultimate Controlling Unit (UCI)*, *Turnover* and *Assets*.

Figure 1a shows the scatter plot of *number of persons employed* as recorded in EGR and as collected from the public sources, for each MNE group. Figure 1b shows normalized values for *Turnover* and *Asset* according to EGR and to public sources. Each dot represents a group and the two straight lines identify the range boundaries of the quality indicator. From figure 1a it is evident that the scraped values overestimate employment compared to EGR values. It is not easy to interpret this upward shift. One possible reason could be that EGR employment is accurate for the MNE groups' parts inside EU (because data are provided by NSIs), but is not for the parts outside the EU (where only commercial sources are used and coverage is partial).

In figure 1b, it is possible to recognize an acceptable fitting of the integrated public sources with EGR for both variables, which reflects that the quality indicator of the integrated sources is high. In these cases there are not systematic data shifts but instead, points outside the quality range lines are randomly distributed.

Data visualization

Under the proof-of-concept study, the final information database was made available also by a means of a dashboard. The dashboard, which served as an illustrative example, displayed various information from different web sources, summarised together in a unified view, giving a general overview of the MNE groups. Screenshots of this dashboard can be found in the proof-of-concept study [2].

Conclusions

Based on the analysis of the results, this study concludes that public sources can be used as additional source of information for the support of the users in improving the quality of the data of the MNE groups. Moreover, public sources can be taken into account when complementing EGR missing information on MNE groups, but however their contribution needs to be precisely qualified.

With regard to most of the attributes of a MNE group, the gain seems to be positive for the country of the ultimate controlling unit (UCI), turnover and assets. The conclusion on employment is more moderate and further analysis is needed to understand the employment gap, which could be attributed to the lack of coverage of information from outside Europe in EGR. If confirmed, the employment data from the public sources could well complement any missing data in EGR for the countries outside the EU.

Further work

The study and the quality analysis carried out in this paper are based on a limited number of MNE groups only. Following the positive assessment of the data acquired from the proof-of-concept study, Eurostat will start the implementation of a web data collection in its own Data Platform, hosted under its Web Intelligence Hub [4], to extend the coverage and verify the possibility to implement the results for the purpose of the production of EGR data**.**

References

[1] Eurogroups register - Statistical business registers - Eurostat (europa.eu)

[2] Smart Data for Multinational enterprises (MNEs) – using open source data to obtain information on Multinational enterprises — 2021 edition - Products Statistical working papers - Eurostat (europa.eu)

[3] https://github.com/eurostat/Smartdata4MNEs

[4] https://ec.europa.eu/eurostat/cros/content/trusted-smart-statistics-%E2%80%93-web-intelligence-hub_en

Improving the quality of the job vacancy survey in Poland using administrative and online data

Keywords: non-probability data, measurement error, automated classification of occupations, validation survey

1. Introduction

The standard way of measuring the number of job vacancies is to use a job vacancy survey (JVS), where a certain number of companies is sampled and asked to report the number of their vacancies. In Poland, a JVS called *the Demand for Labour survey* is conducted on a quarterly basis, with 100k entities (legal and local units) sampled at the beginning of the year and asked to report the number of employees and vacancies at the end of each quarter. The survey suffers from non-response and non-contact, which results in an overall response rate of around 60%.

While the JVS is the most popular source of information about vacancies, there are other (non-statistical) sources that can be utilised for this purpose. One type of such sources are administrative data, which mainly include job offers submitted to labour offices or vacancies in government agencies or government-owned companies. The second category are online sources containing job ads placed on recruitment or classified portals. These sources are currently being used by Cedefop [1] to produce statistics about the demand for specific occupations, skills and competencies. The methodology for using administrative and online sources for vacancy statistics is still being developed [2].

In this study we focus on ways of improving the Polish JVS by linking it with administrative and online data. While a lot of research concentrates on correcting non-response, we mainly deal with the measurement error due to under-reporting or mis-reporting of vacancies (e.g. reporting zero vacancies while actually having them). In order to tackle this problem we obtained data from the Central Jobs Database, which contains all vacancies submitted to public employment offices (PEOs) in Poland and the country's leading online portal Pracuj.pl. In order to obtain information about occupations from job descriptions in the ads, we developed a hierarchical classifier based on natural language processing.

The paper is structured as follows. In section 2 we briefly discuss data sources, the classifier and training, the validation sample of administrative and online data and the overall procedure to improve the JVS. In section 3 we present selected results of the hierarchical classification, measurement error and corrected estimates of the number of vacancies in Poland.

2. Methods

2.1. Data sources

In our study we used JVS data for the period from 2015 to 2021, which included the population frame and samples. The sampling frame consists of about 650k entities, both legal and local units (e.g. the company's headquarter and its regional subsidiaries are treated as separate units). The sampling frame is prepared once a year (e.g. the frame for 2018 was prepared at the end of 2017) and is used for the whole year. At the beginning of the year 100k units are sampled and asked to report on a quarterly basis.

Next, we used the Central Jobs Database (CBOP), which is an administrative record including all vacancies submitted to PEOs. Submitted data are verified by PEO staff and classified into 6-digits occupation codes (about 2.9k codes; 4-digit codes are equivalent to ISCO, 6-digit codes partially overlap with ESCO) according to the Polish classification. Administrative data suffer from under-coverage (certain units do not submit vacancies) and over-coverage (out-of-scope units, outdated vacancies). CBOP records are available through an API and contain detailed descriptions of ads. To study these errors, a validation sample was drawn.

Finally, we use online data from Pracuj.pl, which is a leading recruitment website in Poland. We obtained up-to-date and historical data using web-scraping. Because entity identifiers (REGON/NIP) were available, we were able to identify population units. Like CBOP, Pracuj.pl suffers from converge errors.

2.2. Hierarchical classification of occupations

In order to obtain information about occupations in online job advertisements, we developed a classifier which utilised the hierarchical structure of occupation codes (e.g. code 2 - specialists, 25 - Information and communication technology specialists, …, 251402 - Mobile application developer). The classifier is based on machine learning algorithms for NLP (BERT) capable of handling the Polish language (herBERT, [3]) and takes into account a hierarchical loss function. In order to train the model, we sampled 3000 ads for hand coding by 3 experts (PEO employees) and other data sources, such as official descriptions, the thesaurus of Statistics Poland or thousands of CBOP ads classified by PEO staff.

2.3. Validation samples

In order to assess over-coverage in CBOP and Pracuj.pl we conducted two validation studies at the end of 2021Q3. We sampled only those ads of entities included the JVS sample for 2021 and that meet JVS definition of vacancy (e.g. only contracts of employment). The first study was conducted by employees of the Statistical Office in Bydgoszcz, Poland, the second by two Bachelor students at Poznań University of Economics and Business.

2.4. Overall procedure to improve JVS

The overall procedure to improve the Polish JVS is as follows (this is only mentioned to put the results of this study in context): (1) Link JVS (frame) with CBOP and Pracuj.pl at entity level; (2) Correct CBOP and Pracuj.pl for over-coverage; (3) Edit JVS based on corrected CBOP and pracuj.pl data; (4) Impute missing data for non-respondents and due to the editing procedure (5) Calibrate sampling weights; (6) Obtain estimates (and bootstrap variance estimate by repeating steps 2-5).

3. Results

3.1. Hierarchical classification algorithm

The classifier was trained on a dataset containing 64k records and was validated using 7.5k job advertisements. Model accuracy measures are presented in Table 1 for 1-, 2-, 4-, and 6-digit occupation codes. ACC/Recall@1 refers to the first, most likely, occupation; Recall@2 refers to top 2 occupations, and recall@3 to top 3. According to Table 1, the model has an accuracy of 80% for 1-digit codes, 62% for 6-digit codes, and the use of top-3 predictions ensures an accuracy of 96% for 1-digit codes and 80% for 6-digit codes. Note that this is reported for all validated ads and there was a situation when experts suggested different 6-digit codes based on a given job description. In this table each row of the validation sample is treated as a separate ad.

Table 1. Results of the classification algorithm

Measure	1 digit	2 digits	4 digits	6 digits
ACC/Recall@1	0.80	0.77	0.72	0.62
Recall@2	0.93	0.89	0.85	0.75
Recall@3	0.96	0.93	0.89	0.80

Source: own elaboration

3.2. Results of the validation studies

In this section we present results of two validation studies regarding overcoverage and underreporting by linking with the JVS survey for the end of the 2021Q3. We note that for CBOP outdated rate was 12% while for Pracuj.pl 2%. For those validation sample units linked with JVS respondents 54% did not reported vacancies but had them registered at CBOP and Pracuj.pl this share was higher and equal to 76%. This indicate that under-reporting (measurement error) is significant problem for the JVS.

Table 2. Overcoverage as a share of up-to-date vacancies/ads and under-reporting in JVS based on validation samples for 2021Q3

Source		Up-to-date vacancies		Vacancies not reported by the units linked with JVS respondents
		Yes	No	
CBOP (n=835)	m	5 363	749	2 607
	%	88	12	54%

Pracuj.pl (n=506)	m	492 ads (1119 vacancies)	12 ads (unknown no. of vacancies)	720
	%	98	2	76%

Source: own elaboration

3.3. Estimation results

Finally, we compare corrected estimates of vacancies using administrative and online data. According to our study, the current JVS estimates of vacancies in Poland are underestimated by around 20-40% (depending on the quarter). The proposed estimator takes into account imputed values for non-respondents, imputed values after the editing process (correction for underreporting) and new calibration weights that account for respondents and non-respondents with imputed vacancies.

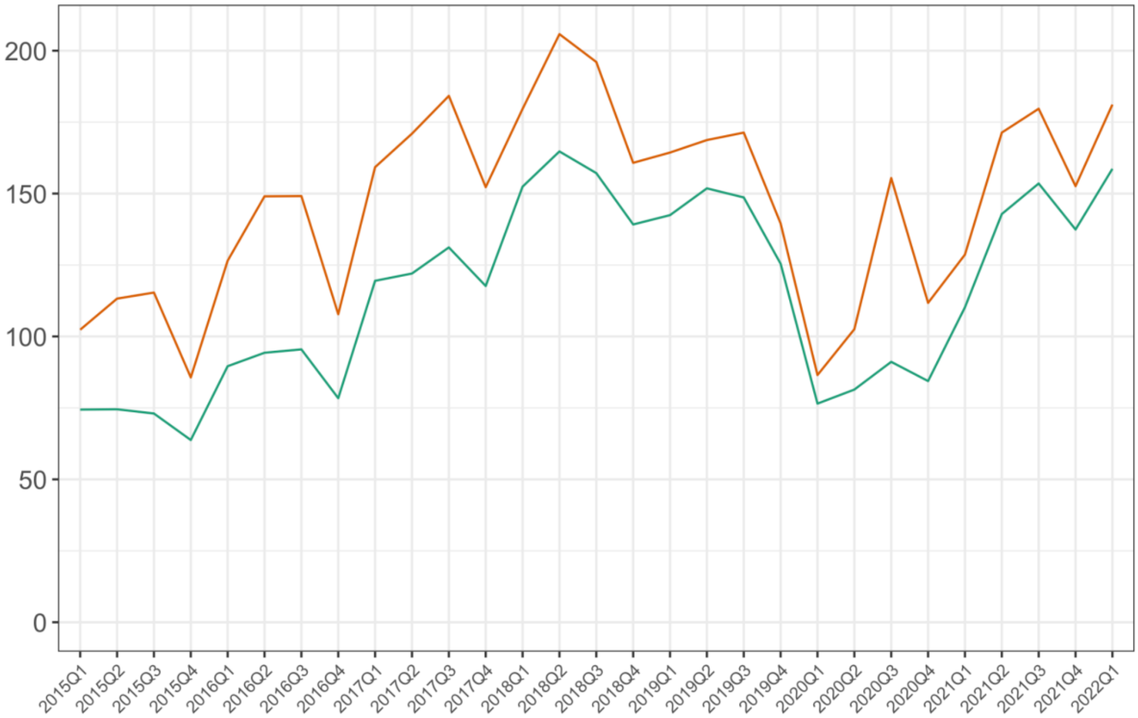


Figure 1. Comparison of the estimated number of vacancies (in thousands) before (green) and after (orange) the correction

4. Conclusions

In this paper we present results of the study aimed at improving the quality of the Polish JVS. For this purpose we used administrative and online data and showed that the JVS suffers from high measurement error due to under-reporting. We developed a classification algorithm based on NLP, which can be used to obtain occupation codes and impute the number of vacancies at the level of 2-digit occupation codes. Further studies are needed to take into account misclassification errors due to automatic classification and possible linkage errors due to insufficient information about entities.

References

[1] Siebel & Nerguisian (2021). The European Centre for the Development of Vocational Training (Cedefop).

[2] Beręsewicz & Pater (2021). Inferring job vacancies from online job advertisements. Eurostat

[3] Mroczkowski, Rybak, Wróblewska, and Gawlik (2021). HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

USING WEBDATA TO DERIVE THE ECONOMIC ACTIVITY OF BUSINESSES

Keywords: Business Statistics; Machine Learning; Natural Language Processing; New Data Sources; Text Mining

INTRODUCTION

The economic activity of businesses (NACE) is often a key characteristic for the production of most business statistics. The NACE code is determined for all units in the population and stored in the statistical business register. It is important that misclassification in NACE codes are avoided, since they can lead to seriously biased business statistics estimates. Determining and maintaining the main economic activity of a business is a difficult classification task that often requires manual editing. Accordingly, this leads to the swift depletion of time resources, especially for national statistical institutes with a large business population size.

Due to an increasing web presence of businesses, the use of data on the web is starting to become a viable data source to help classify the economic activity of businesses. One project within the ESSNet Web Intelligence Network, which started in April 2021, aims to develop automatic procedures to support manual editing of business registers

NACE codes.

In this paper, the current state of an automated procedure to predict the economic activity in terms of NACE 2 codes is presented. First, each business contained in a statistical business register is linked to a webpage that contains its unique identifier (such as VAT or CRN). If the unique identifiers are not available on the webpage, a model based on string matching techniques is employed to facilitate the linking procedure between such webpages and businesses. Once the data set of linked businesses and webpages is established, the webpages are scraped and processed using natural language processing techniques. Finally, using a neural network along with a feature selection method the most likely NACE 2 codes are predicted.

METHODS

Data Set

Maintaining the NACE code in the statistical business register (SBR) or even improving its quality needs a considerable amount of resources and a lot of manual editing. To alleviate the manual editors, a classification procedure, which uses text data from enterprise websites to predict the most likely NACE codes, is highly desirable. The web data used for this purpose is the data collected during the ICT cycles 2019, 2020 and 2021. During these cycles, the websites are linked to one or more enterprises using a URL finding procedure. The considered URL-Finder links part of the enterprises deterministically (if the unique identifier on a webpage matches the unique identifier of an enterprise in the SBR, the corresponding enterprise and URL are linked) and the remaining ones through a model based on string matching (e.g. when the address and name on a webpage matches the address and name of an enterprise in the

SBR, the corresponding enterprise and URL are linked). Once the data set of linked businesses and webpages is established, the webpages are scraped:

Pre-processing

For the NACE classification, the text scraped on the landing page as well as certain subpages, which contain specific keywords (e.g. “enterprise”, “company”, “home”, “about us”) in the link or link description names, that might provide some information about the enterprises, is used. Only the text elements in addition to the link names are kept and the rest of the scraped html code is discarded. Subsequently the text is processed by applying the following steps: Transform each word with the German morphological lexicon[1], remove all digits and punctuations, remove characters not part of the German dictionary, remove German stop words, apply lemmatization using the German version of the hunspell dictionary, and aply stemming using the German version of the hunspell dictionary.

Feature Selection

After applying the pre-processing steps, the scraped text contains 1 014 678 different words. To select a balanced set of features, the feature selection procedure proposed by [1], where a global and a local feature selection score function is combined, is used. As a global score function, the Gini Index (GI), Distinguishing Feature Selector (DFS) and Information Gain (IG) is used, and for the local score function the Odds Ratio (OR) is used. This feature selection is applied on the texts grouped by the NACE 2 codes.

For each NACE code, up to 200 and 500 words, denoted by W-200 and W-500 respectively, are established. Note that in practice one would apply this strategy to the training set, but since the procedure is rather time consuming, it was used for the whole labelled set.

Model Specification

To conduct the NACE classification a neural network model is employed. Modern neural network software is efficiently implemented and is designed to handle thousands of features. In addition, pre-trained word embedding are used.

Two different specifications for the neural network are considered. The first network (label : “wide” in Figure 3.1 and 3.2.) only consists of feed forward layers and uses the one hot encoded W-200 words from the webpages weighted by the term frequency inverse document frequency transformation as inputs. The second network (label: “wide and deep” in Figure 3.1 and 3.2.) enhances the aforementioned neural network by additionally providing it withW-500 words which are transformed using pre-trained word embeddings from fastText. This additional structure consists of multiple convolutional filters applied to the word embeddings. The results from the feed forward and convolutional layers are concatenated in a penultimate layer and

[1] http://www.danielnaber.de/morphologie/

then supplied to a final softmax layer. The R-Package keras [2] and the tensorflow software [3] are used to run the models.

RESULTS

To test both model specifications as well as the different global selection scores GI, DFS or IG, a 40-fold cross-validation procedure is applied. Thereby 80% of the data is used as the training set, 10% as the validation set and the remaining 10% as the test set.

The models are trained on all the available data (2019, 2020 and 2021) but the results presented here are limited to the enterprises, which were part of the 2021 ICT population. Figure 3.1 shows the distribution of accuracy, F1 score and top-5 accuracy for each of the methods used over all cross validation runs.

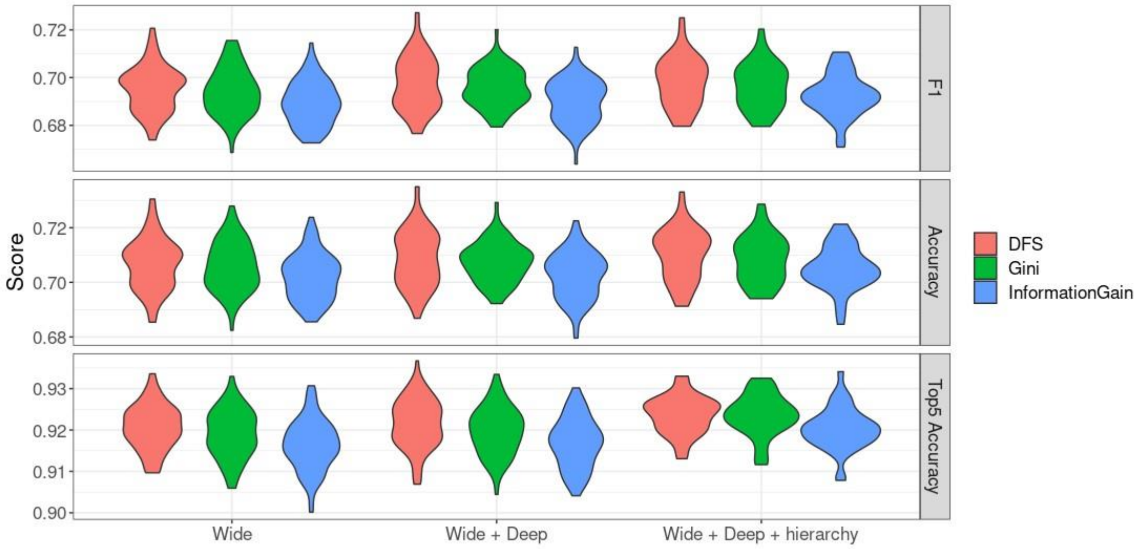


Figure 3.1. Distribution of accuracy, F1 score and top-5 accuracy for each of the methods after cross validation runs

The label Wide + Deep + hierarchy refers to applying the cross-validation first for predicting the NACE 1 level and using the predicted probabilities for the NACE 1 category as predictors for predicting the NACE 2 level.

From 3.1 it is evident that there are hardly any differences between the model settings and feature selection score. Looking at the NACE 2 level prediction accuracy we obtain similar results. Figure 3.2 shows the average accuracy (y-axis) by NACE 2 digits (x-axis) for each model specification and feature selection score. The figure is split into three panels which indicate the number of enterprises from the ICT 2021 for which a website was found by each NACE 2 digit.

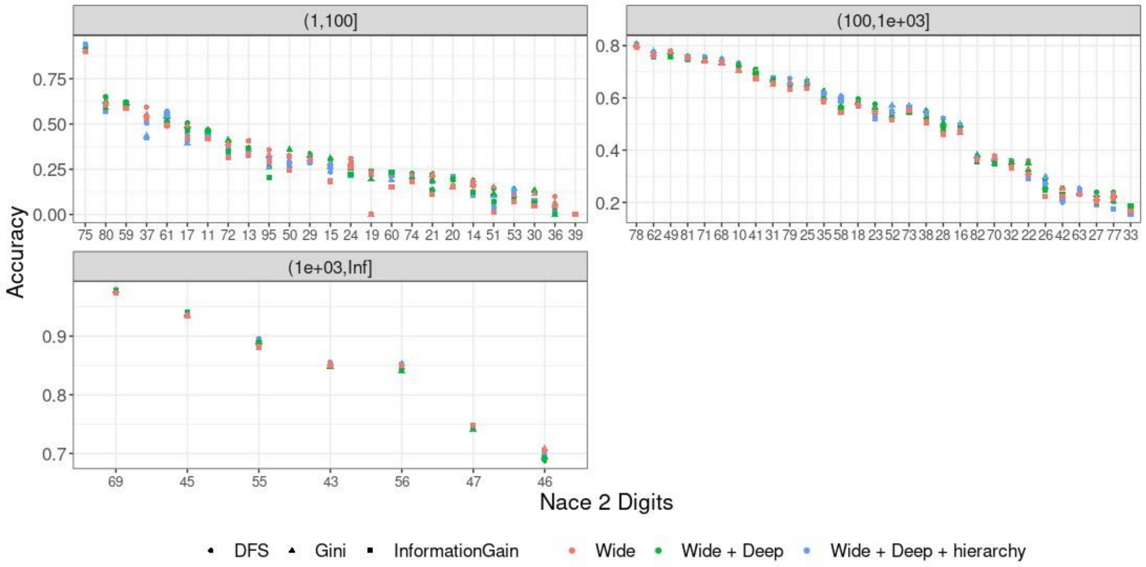


Figure 3.2: Average accuracy (y-axis) by NACE 2 digits (codes) (x-axis) for each model specification and feature selection score. The panels split the NACE 2 codes by number of enterprises available in the training data.

CONCLUSIONS

The economic activity of businesses (NACE codes) is a key characteristic often used to classify the output of economic statistics into subpopulations of different industries. Misclassification of NACE codes can lead to seriously biased output. Therefore, many national statistical institutes have a procedure to manually check and edit the NACE codes the most influential businesses, which is very time consuming. Business websites are an important source when editing the NACE codes. In this paper preliminary results on automatically predicting NACE codes using text-mining models is presented. The procedure starts with identifying webpages of businesses in a statistical business register and scraping relevant text from these webpages. Afterwards the text is processed using natural language processing techniques and finally used in a neural network to predict the most likely NACE codes. The preliminary results show regardless of the feature selection method and neural network used, similar performance results are obtained.

The next steps entail improving the NACE classification method. One possible way to facilitate the improvement of the NACE classification models is to improve the URL linking procedure. The currently regarded training data only consist of deterministically linked pairs of websites and enterprises which has the apparent disadvantage that it is biased towards larger enterprises, as they tend to exhibit a higher level of compliance in regards to implementing legal regulations (e.g., disclosing unique identifiers on their webpage). By employing string similarity techniques - opposed to string matching techniques – one could potentially enhance the existing training data set and possibly mitigate an existing bias. Using the (hopefully) balanced training data set, enhanced by the pairs of websites and enterprises linked by the new model based on string similarity techniques, a potential next step could be to predict the NACE 3 codes of the enterprises.

REFERENCES

[1] Uysal, A. K. (2016). “An Improved Global Feature Selection Scheme for Text Classification.” Expert Sys. Appl. 43 (C): 82-92

[2] Allaire, J.J., and Chollet, F. (2019). “Keras: R Interface to ’Keras’.” https://CRAN.Rproject.org/package=keras.

[3] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C. Corrado, G.S. et al. (2015). “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.” http://tensorflow.org/.

Sampling (MANS 1M.1)

Session Chair: **Cristiano Tessitore** (*Eurostat)*

Random Forest algorithm to adjust for Census population over-counts

Violeta Calian *(Statistics Iceland),* Margherita Zuppardo *(Statistics Iceland)*

The End of Random Sampling

Gary Brown *(Office for National Statistics-ONS)*

Combining probability and non-probability samples on an aggregated level

Ton de Waal *(Statistics Netherlands-CBS),* Sofia Villalobos Aliste *(Utrecht Universtity)*; Sander Scholtus *(Statistics Netherlands)*;

Random Forest algorithm to adjust for Census population over-counts

Keywords: Census, machine learning, R language, open-source code, non-response

1. Introduction

One of the most important purposes of the 2021 digital Icelandic census is to provide an accurate description of the population residing in the country. The standard way to identify this population is to directly consult public registers on the long-term residence status of everyone in the country. However, it is a known problem that many people, of both Icelandic and foreign origin, do not notify the authorities when moving overseas, despite it being a legal requirement.

This phenomenon cannot be ignored in the context of the census without incurring into significant biases in the estimates of demographic and social attributes such as age and income distribution, fertility and mortality rates or employment and education profiles [1].

Fortunately, most individuals that are present in the country leave a track of their presence in public registers by, for instance, paying taxes and being employed in the country, attending school, buying real estate property or changing their address within the country. Other demographic attributes such as gender, age, or country of origin also provide some indication on the likelihood of each person of migrating for a significant period of time. We refer to all these indicators as ‘signs of life’ (SOL). This is essentially a binary classification problem, where the status of each individual has to be coded as ‘in’ or ‘out’ of the country according to certain predictors.

Since the previous edition of the Census in 2011, there has been an explosion of new resources in the field of machine learning and many open-source packages which are designed to solve this exact class of problems have become available. Statistics Iceland took advantage of this development and trained and optimized a classification algorithm to estimate the true population for the 2021 Census. Such techniques had not been previously applied to similar matters in the field of demography, to our knowledge. Other methods were used instead [2], based on more ad-hoc scores defined over signs of life.

In the following, we describe the process solving the overcounting problem by applying classification statistical methods to the total Census population. First, we provide a brief description of the dataset used to train the algorithms. Next, we show how several most widely known machine learning algorithms can be applied to our problem and motivate the choice of one of them, the Random Forest algorithm, based on a set of performance measures. Later, we describe the way that tuning the parameters of the Random Forest affects the performance of the method and justify the choice of optimum parameters. We finally give a brief overview of how the application of this model changed the Census statistical results.

2. Methods

2.1. Building a training dataset

The first step in building a training dataset is to identify a group of individuals that can be placed with certainty in or out of the country at a given time. For this purpose, we used the Icelandic Labor Force Survey (LFS) data for the years between 2014-2018. Of the 17710 individuals, 18 years old and over, which were sampled in this survey, only 537 were declared as ‘out’ of the country, making our dataset rather unbalanced.

Identifying the most important predictors (SOL) of presence is a challenge in itself. On the one hand SOL need to be readily available for all individuals in the public registers when the model is eventually applied to the total population for predicting the status of true presence/absence. For example, banking information and car ownership were excluded from the model due to this very reason, but may be significant and could be included in the future.

On the other hand, information in the public registers needs to be re-coded in order to maximize the importance of predictors in the final model. For example, we found that having children in the public school system or being registered as a student did not have a big importance in the final model as such. Hence, we decided to use these variables in a more parsimonious way, describing how many people in the family unit are in the school system and this new variable proved to have a higher impact.

Our final dataset has a total of 20 predictor columns, of which 9 are binary variables and 11 numeric ones. A full list of predictors may be provided on request.

2.2. Performance metrics

We used the open-source R package ‘Caret’ [3] to train a variety of algorithms in a unified notation. In order to choose the final algorithm, we first had to decide which performance metrics were the most relevant for our problem. These can be best defined as a function of the elements of the confusion matrix (CM) shown below

	TRUE IN	TRUE OUT
P IN	TP	FP
P OUT	FN	TN

ACCURACY	$\frac{TN + TP}{N\_tot}$
SPECIFICITY	$\frac{TP}{TP + FN}$
SENSITIVITY	$\frac{TN}{TN+FP}$
POP. ERROR	$\frac{\lvert FP - FN \rvert}{N_{tot}}$

The measure we selected were: accuracy, sensitivity and specificity, as defined in the table above, which are standard and widely used for binary classification. We split the our dataset into 70% of the rows for training and the metrics were evaluated on the remaining 30% of the rows for testing.

In addition, since our problem is applied to Census data, we used the 'population error' as a fourth metric. This is the percentage of the total population which is either over- or under-estimated. This accounts for the fact that the total number of people in the country can be kept close to the true number if the 'wrong' estimates compensate each other.

Note also that sensitivity and specificity often compete when tuning a model. For our problem, we want to be on the safer side of keeping the specificity higher thus allowing a slight over-count of the population, even though this means lowering the model's sensitivity.

3. Results

Table 1. Performance of different algorithms

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Population error (%)
Register data	96.7	0.0	100.0	3.3
Logistic regression	96.8	14.2	99.6	2.6
Decision tree	97.0	16.4	99.8	2.3
Neural network	96.9	17.5	99.6	1.6
AdaBoost	95.1	26.3	97.5	0.01

Random forest (untuned)	97.0	25.1	99.5	2.1
Optimized RF (final model)	96.0	48.0	98.0	0.04
Latest results (revised data)	96.7	56.0	98.2	0.2

3.1. Performance of different algorithms

The table above shows how different algorithms performed under our metrics of choice. Among those, we decided to exploit the Random Forest algorithm available in the RandomForest package for R [4]

3.2. Random forest optimization

Figure 2 shows how the Random Forest probability cutoff values affect our selected performance measures. While the specificity and the accuracy stay rather high when increasing the cutoff value (of the probability which defines the status as in/out of the country) above a certain value, the specificity and the population error are most affected by this tuning. For this reason, we choose the cutoff parameter corresponding to specificity above 98% while optimizing the other metrics. This led to a cutoff value of about 20%. The resulting metrics are shown in Table 1. Tuning of the other parameters, specifically the stratification proportion (sampsize) needed due to the un-balanced in/out samples, allowed further improvement as shown in the last row of Table 1.

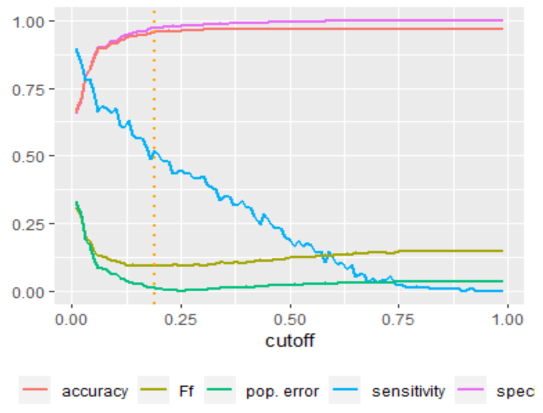


Figure 2. How cutoff tuning affects the performance of the Random forest. The optimal cutoff is chosen as the one allowing for the highest sensitivity while keeping specificity above 98%.

A similar method was applied to the foreign citizenship population and to the total register population, separately. There were 7100 individuals predicted to be out of the country, of which 3770 foreign citizens. This is in the range that we would expect based on the data on delayed de-registration from previous years.

Conclusions

The Machine Learning methods described in this paper allowed us to adjust the Icelandic population for the errors due to the individuals who emigrated without informing the national registers. Our final model of choice, based on the Random Forest algorithm, allowed a sensitivity of above 50%, while keeping specificity and population error at ‘safe’ levels.

3.3. Future development

Although we are satisfied with the results of our work, we would like to improve the performance of our model in the future. This could be done by improving the training data, by using the information about late deregistration that will occur in the next months after the reference Census date. This method can also be used routinely, to adjust the total population count of individuals in Iceland.

This will also allow us to use our model to predict non-response in public surveys. By sampling individuals from the population that is predicted as ‘in’, it will save time and resources, and lower the non-response rates.

References

[1] A. Monti, S. Drefahl, E. Mussino, J. Härkönen, Over-coverage in population registers leads to bias in demographic estimates, Population Studies (2019), 1-19.

[2] E. Maasing, E.-M. Tiit, M. Vahi, Residency index – a tool for measuring the population size, Acta et Commentationes Universitatis Tartuensis De Mathematica (2017), 129-139

[3] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software (2008) 1-26.

[4] Leo Breiman, Adele Cutler, Andy Liaw and Matthew Wiener. Breiman and Cutler's Random Forests for Classification and Regression. URL https://cran.r-project.org/web/packages/randomForest/index.html

Keywords: Survey/administrative/big data, data science, quality, statistics@2023

1. INTRODUCTION

Statistics are no longer the monopoly of National Statistical Institutes (NSI). Data are everywhere, and real-time statistics are becoming ubiquitous. The NSI tradition of publishing retrospective statistics, based on data collected solely by in-house surveys, is being overturned in favour of publishing faster indicators based on multi-sourced data from wherever they fall. Quality is addressed through an increased focus on data governance, access improved through government acts and data sharing agreements, collation eased through new data science wrangling techniques, and finally analysis accelerated through game-changing technology.

This presentation explores the options for producing statistics@2023, analysing the strengths and weaknesses of different approaches, illustrating their benefits, and the pitfalls that could result from reliance on a single solution. The talk will finish with a cautionary tale from 2021/22, as experienced by the UK Office for National Statistics – a saga of data sharing, congruence, targeted sampling, and … statistics.

2. METHODS

This presentation will review methods for estimating statistics from survey, administrative and big data – and combinations of the three. The methods will take into account the provenance of the data, and the data generating process where applicable. Formal quality measures for the resulting statistics will not be covered, but quality will be a key dimension of the assessment of strengths and weaknesses.

3. RESULTS

The new results presented will be the outcome of the saga experienced by the UK Office for National Statistics in 2021/22 – the methods concerning which were published in Robinson (2022), with the results due in November 2022 in a pre-scheduled release. Legacy results from a range of countries will be presented to illustrate the pros and cons of the various approaches reviewed for statistics@2023.

4. CONCLUSIONS

The conclusion of the presentation will be to neither trust one method, nor a pre-determined suite of methods, to solve all issues. Statistics are complex, and the best techniques to produce them are equally complex. New technology is not a panacea, and cannot replace the human qualities of curiosity, caution and compromise.

REFERENCES

Robinson, D. (2022) “Comparison of ONS business enterprise research and development statistics with HMRC research and development tax credits data”, ONS website, 29/09/2022.

Combining probability and non-probability samples on an aggregated level

Keywords: categorical data, combined estimator, selectivity, data integration

1. Introduction

Probability surveys are experiencing important drawbacks nowadays: participation rates are decreasing, and the respondent burden is increasing, which could yield less accurate estimates, needing a big investment of resources in order to benefit from the feature of unbiasedness that makes them so desirable. Non-probability samples like administrative records are having a rise in popularity due to their convenience and low costs. Unfortunately, non-probability samples are often selective and, since the inclusion probability in such samples is unknown, estimators based on such non-probability samples are generally biased. Wiśniowski et al. (2020) proposed a Bayesian approach where the non-probability sample is used to improve estimates based on the probability sample. In this paper, we examine a frequentist method that combines estimates from a probability and a non-probability sample on an aggregated level. Our method does not require any data on the level of the individual units.

We assume that estimates for proportions of a categorical target variable y per category of a categorical background variable x are available from both a relatively small probability sample with size $n^{(P)}$ and from a (possibly) selective non-probability sample, e.g. an administrative dataset, with size $n^{(NP)}$. Since the sampling design for the probability sample is known, the estimator based on the probability sample is usually unbiased. We assume that this is indeed the case in our situation. However, the sampling variance of this estimator is usually quite large. The estimator based on the non-probability sample is very likely to be biased, but its variance is generally quite small. The proposed method combines the estimator for the probability sample with the estimator for the non-probability sample by constructing a weighted sum of both estimators. The used weight aims to minimize the mean squared error (MSE) of the combined estimator.

2. Proposed Estimator

The proposed method is inspired by a small area estimation approach. We will therefore refer to the categories of background variable x as domains. We will denote these domains by k $(k = 1, \ldots, K)$. The categories of the target variable y are denoted as c $(c = 1, \ldots, C)$. We assume that estimates for proportions Z_{kc} $(k = 1, \ldots, K; c = 1, \ldots, C)$ are available for both samples. We will denote the estimator for proportion Z_{kc} from the probability sample by $\hat{Z}_{kc}^{(P)}$, and the corresponding estimator from the non-probability sample by $\hat{Z}_{kc}^{(NP)}$. For simplicity, we assume that the probability sample is drawn by simple random sampling.

We use the observed proportions $\hat{Z}_{kc}^{(P)}$ for domain k and category c in the probability sample, respectively $\hat{Z}_{kc}^{(NP)}$ in the non-probability sample, as estimators for the true population proportion Z_{kc}. For the probability sample, the estimator $\hat{Z}_{kc}^{(P)}$ is unbiased. The variance estimate of $\hat{Z}_{kc}^{(P)}$ is given by $\hat{Z}_{kc}^{(P)}\left(1-\hat{Z}_{kc}^{(P)}\right)/\left(n^{(P)}-1\right)$. We want to construct a combined estimator of the form

$$\widehat{D}_{kc} = W_{kc}\hat{Z}_{kc}^{(P)} + (1-W_{kc})\hat{Z}_{kc}^{(NP)} \qquad (1)$$

where W_{kc} is a weight between zero and one. If the bias and variance of $\hat{Z}_{kc}^{(NP)}$ were known, we could compute its MSE and find the weight W_{kc} for which the MSE of $\widehat{D}_{kc}$ given by (1) is minimum. Those optimal weights would be given by (Särndal et al., 1992)

$$W_{kc} = MSE\left(\hat{Z}_{kc}^{(NP)}\right)/\left(MSE\left(\hat{Z}_{kc}^{(P)}\right)+MSE\left(\hat{Z}_{kc}^{(NP)}\right)\right). \quad (2)$$

Technically, the design-based sampling variance of $\hat{Z}_{kc}^{(NP)}$ is unknown/undefined. Nevertheless, assuming that the sample is large, a reasonable variance estimate might be $\hat{Z}_{kc}^{(NP)}\left(1-\hat{Z}_{kc}^{(NP)}\right)/\left(n^{(NP)}-1\right)$. The bias of $\hat{Z}_{kc}^{(NP)}$ cannot be estimated from the non-probability sample only, and we have to rely on some model assumptions. We introduce the notation $b_{kc} = E_d\left(\hat{Z}_{kc}^{(NP)}\right) - Z_{kc}$, where E_d denotes the expectation under the (unknown) sampling design of the non-probability sample. The bias b_{kc} is unknown in practice. However, note that within each domain k we have $\sum_{c=1}^{C} b_{kc} = 0$, since the estimated proportions in each domain add up to one. Stated differently, if a certain category has too many in the non-probability sample, there also has to be a category with too few units in this sample. Therefore, we assume a model such that b_{kc} is distributed as a random variable with mean $E_b(b_{kc}) = \beta_c$, i.e. the expected bias in category c $(c = 1, \dots, C)$ is assumed to be constant across domains, $\sum_{c=1}^{C}\beta_c = 0$, and $Var_b(b_{kc}) = E_b((b_{kc}-\beta_c)^2) = \sigma^2$. Here the subscript b indicates that expectations are calculated under our model for b_{kc}.

We can calculate the expected MSEs (EMSEs) for under the above model for b_{kc}. We can derive

$$EMSE\left(\hat{Z}_{kc}^{(P)}\right) = \frac{1}{n_k^{(P)}}\left(\frac{n_k^{(NP)}}{n_k^{(NP)}-1}v_{kc} + \beta_c\left[2E_bE_d\left(\hat{Z}_{kc}^{(NP)}\right)-1\right] - \beta_c^2 - \sigma^2\right) \qquad (3)$$

$$EMSE\left(\hat{Z}_{kc}^{(NP)}\right) = \beta_c^2 + \sigma^2 + \frac{v_{kc}}{n_k^{(NP)}-1} \qquad (4)$$

where $n_k^{(NP)}$ is the size of the non-probability sample in domain k and $v_{kc} = E_bE_d\left(\hat{Z}_{kc}^{(NP)}\left(1-\hat{Z}_{kc}^{(NP)}\right)\right)$. An unbiased estimator for v_{kc} is $\hat{Z}_{kc}^{(NP)}\left(1-\hat{Z}_{kc}^{(NP)}\right)$; similarly, $E_bE_d\left(\hat{Z}_{kc}^{(NP)}\right)$ is estimated by $\hat{Z}_{kc}^{(NP)}$. Ordinary least squares estimates for β_c $(c = 1, \dots, C)$, respectively σ^2 are: $\hat{\beta}_c = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{Z}_{kc}^{(NP)} - \hat{Z}_{kc}^{(P)}\right)$ and $\hat{\sigma}^2 = \frac{1}{(K-1)C}\sum_{k=1}^{K}\sum_{c=1}^{C}\left(\hat{Z}_{kc}^{(NP)} - \hat{Z}_{kc}^{(P)}\right)^2 - \frac{K}{(K-1)C}\sum_{c=1}^{C}\hat{\beta}_c^2$. Plugging these estimates into (3) and (4) leads to estimates $\widehat{EMSE}\left(\hat{Z}_{kc}^{(P)}\right)$ and $\widehat{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)$ for $EMSE\left(\hat{Z}_{kc}^{(P)}\right)$, respectively $EMSE\left(\hat{Z}_{kc}^{(NP)}\right)$. We now use the estimates $\widehat{EMSE}\left(\hat{Z}_{kc}^{(P)}\right)$ and $\widehat{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)$ to calculate weights $\widehat{W}_{kc} = \widehat{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)/\left(\widehat{EMSE}\left(\hat{Z}_{kc}^{(P)}\right)+\widehat{EMSE}\left(\hat{Z}_{kc}^{(NP)}\right)\right)$, and we plug these weights into (1) and obtain our combined estimator. For details, we refer to Villalobos-Aliste (2022).

3. Simulation study

To assess the proposed approach, we simulated a population and repeatedly drew two datasets (a probability sample and a non-probability one), and applied our proposed estimator to those datasets. Our population consists of $N = 100{,}000$ units.

We considered one, four, ten, and 15 domains, three, five, eight, and 15 categories, and a first scenario where all these categories are of equal size in each domain, and a second scenario where categories have unequal sizes. We simulated scenarios with sample size per domain for a probability sample of $n^{(P)} \in \{10, 100, 400, 900\}$. For the non-probability samples, sample sizes per domain were $n^{(NP)} \in \{100, 1000, 2000, 6000\}$. We also created two levels of selectivity for the non-probability sample: one scenario where selectivity is weak and one scenario where selectivity is severe. The details on how selectivity was created and other details on the simulation study can be found in Villalobos-Aliste (2022).

We considered a full factorial design, giving rise to 1024 scenarios of different simulation conditions, and drew $R = 1000$ simulations for each of them. For each simulation we calculated the estimates for $\hat{Z}_{kc}^{(P)}$, $\hat{Z}_{kc}^{(NP)}$ and $\widehat{D}_{kc}$.

4. Evaluation criteria

To evaluate the results we first compute the root mean squared error (RMSE) over the R simulations. The RMSE for a domain k and category c is computed as $RMSE_{kc} = \sqrt{\sum_{r=1}^{R}\left(Z_{kc} - \hat{Q}_{kc,r}\right)^2/R}$ with $\hat{Q}_{kc,r}$ an estimate for domain k and category c in simulation r $(r = 1, \dots, R)$. Depending on which estimator we want to evaluate $\hat{Q}_{kc,r}$ is equal to $\hat{Z}_{kc,r}^{(P)}$, $\hat{Z}_{kc,r}^{(NP)}$ or $\widehat{D}_{kc,r}$. We combine the $RMSE_{kc}$ into one performance measure per domain $ARMSE_k$ defined by $ARMSE_k = \sum_{c=1}^{C} RMSE_{kc}/C$. Finally, we combine the $ARMSE_k$ into one overall performance measure by taking the mean of the values by domain: $MARMSE_k = \sum_{k=1}^{K} ARMSE_{kc}/K$.

We also assess the bias of $\hat{Z}_{kc,r}^{(NP)}$ and $\widehat{D}_{kc,r}$ by means of $MAB = \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{c=1}^{C}\left|Z_{kc} - \hat{Q}_{kc,r}\right|/RKC$.

5. Results

Table 1 shows the percentage of times out of a total of 256 simulation conditions that the combined estimator shows a lower MARMSE than the estimator based on the probability sample, the estimator based on the non-probability sample, and both estimators, for different selectivity levels and different category sizes. The 256 simulation conditions differ with respect to the number of domains, the number of categories, and sample sizes.

Table 1. Percentage of combined estimators with lower MARMSE

Selectivity	Size of categories	better than $\hat{Z}_{kc}^{(P)}$	better than $\hat{Z}_{kc}^{(NP)}$	better than both
Weak	Equal	94	64	58
	Unequal	92	64	57
Severe	Equal	83	84	67

	Unequal	80	85	65

In our simulation we also found that when $W_{kc} \geq 0.7$, the combined estimator always had the lowest MARMSE of the three estimators. When $W_{kc} \leq 0.6$, the estimator based on the non-probability sample always performed best. When $0.6 < W_{kc} < 0.7$, it depended on the specific scenario which of the three estimators performed best.

The percentage of times in which the mean absolute bias of the combined estimator was equal to or lower than the mean absolute bias of the estimator based on the non-probability sample was close to 100% for the situation of weak selectivity (99% when the categories are of equal size in each domain and 98% when the categories have unequal sizes). For the case of severe selectivity this percentage was equal to 100%.

6. Conclusions

In this paper, we have proposed and evaluated a method to combine estimates based on a probability sample with estimates based on a non-probability sample with the aim of minimizing the MSE of the combined estimator of proportions in a contingency table. The method models and estimates the bias of the estimator for the non-probability sample, which allows us to estimate the EMSE under the posited model. The method also estimates the EMSE for the estimator for the probability sample, and combines both estimators by minimizing the EMSE.

Through a simulation study, we showed that the combined estimator can indeed lead to a reduction of the MSE, rather than merely a reduction of the EMSE under the posited model. This evaluation indicates that using our proposed combined estimator is better than using either the estimator for the non-probability sample or the estimator for the non-probability sample when the weight W_{kc} is relatively high (in our case, when $W_{kc} \geq 0.7$).

A major advantage of the proposed method is that it is not necessary to link the two samples at the level of individual observations, or even use any individual observations. It is also not important whether the non-probability sample and the probability sample overlap or not. Moreover, the method is quite robust, in the sense that the EMSE of the combined estimator is always less than or equal to the lowest EMSE of the estimators for the non-probability sample and the probability sample, and – more importantly – the actual MSE of the combined estimator is never higher than the highest MSE of the estimators for the non-probability sample and probability sample. Finally, the combined estimator is very easy to implement in software like R since it requires very few lines of code. For many more details on the proposed method, including an application to a real dataset from Statistics Netherlands, we refer to Villalobos-Aliste (2022).

References

[1] Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A. and Blom, A.G. (2020), Integrating Probability and Nonprobability Samples for Survey Inference. Journal of Survey Statistics and Methodology, 8, pp. 120-147.

[2] Villalobos-Aliste, S. (2022), Combining Probability and Non-Probability Samples for Estimation. Master thesis, Utrecht University.

[2] Särndal, C-E, Swensson, B. and Wretman, J.H. (1992), Model Assisted Survey Sampling. New York: Springer.

Data Sources (MANS 1M.2)

Session Chair: **Sofie De Broe** *(Sciensano / University of Maastricht)*

SORS - Integrated data collection system
Branko Josipovic *(Statistical Office of the Republic of Serbia)*, Nebojsa Tolic *(Statistical Office of the Republic of Serbia)*

How to identify private households in french surveys?
Lea E Tholozan (Insee)*; Helene Chaput (Insee)

Administrative Data assisting Data Collection through Machine Learning
Sonia P Quaresma (INE)*; Alvaro Combo (INE); Pedro Guerreiro (INE); Carlos Valente (INE); Paula Marques (INE)

How to identify private households in French surveys?

Keywords: households, private household, household member, surveys, incomes, expenses, dwelling, French, European, IESS regulation.

INTRODUCTION

In French household surveys, the concept of household is approached by two different definitions:

- The **household** in the broad sense, which is composed of all persons living together in the same dwelling, regardless of their family or economic relationships.
- The **private household**[2], which is a person living alone or a group of people who live together and share household income[1] and expenses[3], which is called "making a common budget". For instance, a dwelling with three room-mates contains only one household in the broad sense, but three private households (each consisting of one person).

The use of the private household definition is essential to correctly assess the standard of living of the population. Indeed, living in a shared apartment with two separate budgets does not have the same impact on the standard of living as sharing income and expenses as a couple without child for instance. Moreover, the IESS regulation[2][3][4] recommends that the scope of European household surveys should focus specifically on private households and not on households in the broad sense.

In most French household surveys, it is at the beginning of the questionnaire, in the "household outline" module, that the number of private households in the dwelling is identified. To do so, the groups of people who make a common budget and budget separately from the other inhabitants are identified. However, "budgeting separately" is not easily understood by the respondents, who often equate it with having separate bank accounts or giving allowance to a child, which is not what we want to identify.

In order to avoid these misunderstandings, additional questions were added to check for contentious cases of separate or joint budget reporting. This "household outline" module is embedded in the French SILC[5] (Statistics on Income and Living Conditions) and in 2021, only 1,04% of the dwellings surveyed were composed of more than one private household.

[2] "sharing household income" means contributing to the private household income or benefiting from the private household income, or both.

[3] "household expenses" means expenses incurred by private household members in relation to providing themselves with the essentials of living. They include house-related expenses (namely rent, house or apartment charges and housing insurance) as well as other expenses related to daily life, encompassing needs such as food; clothes; sanitary products; furniture, equipment and utensils; commuting and other transport; medical care and insurance; education and training; leisure and sports activities; and holidays.

As part of the redesign of the household surveys and their adaptation to self-administered questionnaire, Insee first decided to remove these control questions which made the questionnaire more cumbersome for the respondents and to add a control question for room-mates who would report on a common budget.

The French LFS[6] (Labour Force Survey) – which until then was based on the broader sense of household – integrated this new "household outline" module in 2021. This allowed us to assess the impact of removing these control questions on the number of private households identified.

This first year of collection of the new module revealed much higher reporting of separate budgets: in 2021, 6.9% of the dwellings surveyed in the French LFS consisted of more than one private household.

The purpose of this paper is to explain the causes of this increase and to propose a new formulation of the private household identification module to avoid this break in the series while continuing to properly identify each kind of household.

METHODS

The SILC questionnaire integrates the former version of the "household outline" module. The module begins by asking if any of the residents of the dwelling have a separate budget. If so, the respondent is asked to list the members of each household. Based on the family relationships between the members of each private household, contentious cases[4] are identified for which two control questions are asked to ensure that the respondent has understood the definition of a separate budget.

This questionnaire is rather long and difficult for the respondent, therefore, it is not suitable for a selfadministered survey. Nevertheless, it ensures that private households are properly identified.

For its integration in the French LFS in 2021, this module has been simplified and shortened in order to be understandable by respondents answering by internet, without the help of an interviewer. The module begins by asking if there are people in housing who budget separately and recalls some definitions of what it means to "budget separately." The respondent then reports the list of individuals who constitute each of the private households. If two roommates are on a joint budget, a control question appears. It reminds the respondent that, *a priori*, two room-mates are budgeting separately because they only share the expenses for housing, and then asks him to confirm his declaration of the household outlines.

Nevertheless, this new version of the questionnaire, although simpler, does not probably fulfill its main function, namely to identify private households as defined in the IESS regulation. Indeed, the respondent is burdened with the understanding of the concept of "separate budget" even though its statistical definition is far from its common meaning. This partly

[4] Three contentious cases are identified: couples who report being on separate budgets, families where a child under age 25 is reported on a separate budget from his or her parents, and families where a child over age 25 is reported on a joint budget with the parents.

explains the dramatic increase of the proportion of dwellings composed of several private households in the French LFS compared to the French SILC in that same year.

First, this paper will compare the French SILC questionnaire with the questionnaires of other European countries. The ways in which each country has adapted the identification of private households to its cultural realities might help us find new applications for the French questionnaire.

Then, we will present a comparison of the family composition of the dwellings and private households surveyed by SILC and LFS in 2021. This will provide insight into which type of family structure tends to overreport separate budgets.

Finally, we will propose a new questionnaire for the "household outline" module. We will test it on the 2021 French LFS data as well as in the 2023 French SILC Focus Group. Hopefully, the results will be conclusive and provide a similar proportion of private households compared to the previous version included in the SILC questionnaire. If so, the module will be first embedded in French LFS 2024 and then proposed for all other French household surveys.

RESULTS

Review of some of the SILC questionnaires:

We did a quick review of SILC questionnaires from other European countries that were either in French or that we were able to translate. Some questionnaires do not seek to go into detail about the identification of private households and remain at the level of a broader household concept. This is the case for SILC questionnaires of the UK[7], Ireland[8], Belgium[9] or Spain[10]. Italy[11], on the other hand, translates the concept of private household by the term “family” (*famiglia*), therefore excluding other inhabitants for economic reasons[5] (tenants, employers, servants, etc.).

First results from 2021 LFS and SILC collections:

These first results come from the 2021 LFS and SILC collections. Both of these surveys are representative of almost the same population. They then allow us to compare the two formulations of the “household outline” question module. LFS included a version without control questions, while SILC included the former version, with control questions to compensate for a misunderstanding of the term “separate budget”.

Table 1. Distribution of the number of private households in the surveyed dwellings in French LFS and SILC in 2021.

Percentage of private households in the dwellings	French LFS	French SILC
One private household	93,14%	99,29%
Two private households	5,90%	0,87%

[5] *“Vive in questa abitazione esclusivamente per motivi economici come ad esempio nel caso di collaboratori domestici, babysitter, affittuari oppure datore di lavoro?”*

More than three private households	0,96%	0,17%

Source: LFS and SILC surveys 2021, Insee
Coverage: Ordinary dwellings (i.e. excluding collective households and institutions) in France excluding Mayotte

In 2021, according to SILC (see table 1), 99,3% of the dwellings are occupied by only one private household. This proportion is much lower according to LFS (93,1%).

Table 2. Distribution of respondents in a couple with another dwelling resident in the same or different private households.

Person in a couple with another private household member	French LFS	French SILC
Yes	95,6%	99,8%
No	4,4%	0,16%
Total	200 970	16 000

Source: LFS and SILC surveys 2021, Insee
Coverage: persons in a couple with another dwelling resident, ordinary dwellings (i.e. excluding collective households and institutions) in France excluding Mayotte

In 2021, 200 970 respondents of the French LFS were in a couple with a cohabitant and among them, 4,4 % declared a separate budget from their partner (table 2). In comparison, among the 16 000 respondents of the French SILC in a couple with a cohabitant, only 0,16% declared a separate budget.

Table 3. Among dwellings with multiple households, distribution of the family structure of private households by family structure of all cohabitants in the dwelling.

Family structures of housing consisting of several private households	Family structures of the private households					
	One person household	Lone parent with child(ren)	Couple without any child	Couple with child(ren)	Other types of family structure	Total
Lone parent with child(ren)	10,57%	1,83%	0%	0%	0,01%	2 622 individuals (12,41%)
Couple without any child	20,54%	0%	0%	0%	0%	4 338 (20,54%)
Couple with child(ren)	15,44%	18,25%	7,99%	4,58%	0,07%	9 788 (46,34%)
Other types of family structure	11,03%	2,41%	1,69%	2,38%	3,21%	4 376 (20,72%)
Total	12 162 (57,57%)	4 753 (22,50%)	2 044 (9,68%)	1 471 (6,96%)	694 (3,29%)	21 124 (100%)

Source: LFS and SILC surveys 2021, Insee

Coverage: Ordinary dwellings (i.e. excluding collective households and institutions) with multiple private households.

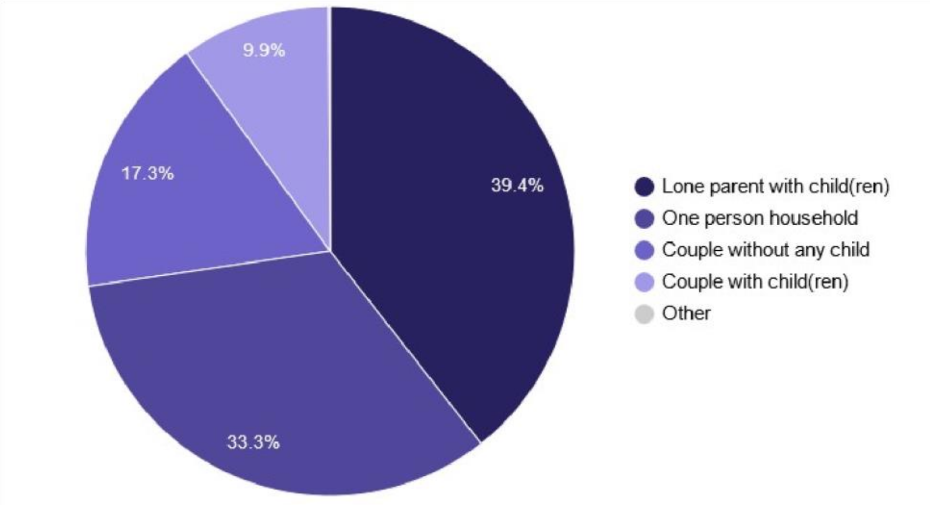


Figure 1. Distribution of family structures of private households that compose dwellings inhabited by a couple with child(ren) reporting separate budgets.

In 2021, according to LFS, 21,000 people live in multi-household housing (table 3). Of these, 21% are couples without children who are therefore counted in one private household each.
On another hand, 46% of these dwellings are inhabited by couples with children, and among them 39% (figure 1) declare a separate budget between the two partners, which leads to the counting of more than 3,850 lone parents who actually live with their partner.

cONCLUSIONS

This over-reporting of separate budgets seen in the new French LFS, compared to SILC, can be explained by the fact that the respondents do not understand the term "separate budget" as we seek to identify it statistically, but rather by its common definition, *i.e.* "having separate accounts", "managing one's own revenue" or "managing allowance independently" etc. It will therefore be necessary to chose carefully the vocabulary writing the final proposal for the "household outlines" module.

Therefore, the over-reporting of separate budgets comes mainly from couples with or without children, which causes an over-representation of single parents in the French LFS. Children who are close to or into young adulthood (17 to 25 years old) but still live in their family home are also often reported as having a separate budget from their parents, even though they usually remain a rather important part of their parents' budget and expenses

We can conclude that couples with or without children under 25, as well as single parents with one or more children under 25 years old are – in the vast majority of cases – part of a single private household. Conversely, in dwellings occupied by room-mates, usually the later each constitute a one-person private household.

This is why the final “household outlines” module will not be presented when the dwelling is composed of one of these types of relationships structures. Inhabitants will be grouped directly into the correct household type(s) based on the identified relationships (family, roommates or other non-family relationship).

For all other “non-classical” family structures, the module will be proposed with only two questions. The first question will be introduced by a definition of the private household and then ask whether some inhabitants budget separately from others, by having separate expenses, especially food expenses. If so, we will ask to identify which inhabitant belongs to which private household.

As a result, this questionnaire will be asked of far fewer people. It will also be shorter and take into account potential misunderstandings of the term “separate budget” directly within the first question. It will yet to be tested.

rEFERENCES

[1] Methodological guidelines and description of EU-SILC Target variables (2021), 35-36

[2] Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples, amending Regulations (EC) No 808/2004, (EC) No 452/2008 and (EC) No 1338/2008 of the European Parliament and of the Council, and repealing Regulation (EC) No 1177/2003 of the European Parliament and of the Council and Council Regulation (EC) No 577/98 (Text with EEA relevance)

[3] Commission Delegated Regulation (EU) 2020/257 of 16 December 2019 supplementing Regulation (EU) 2019/1700 of the European Parliament and of the Council by specifying the number and the title of the variables for the labour force domain (Text with EEA relevance)

[4] Commission Delegated Regulation (EU) 2020/258 of 16 December 2019 supplementing Regulation (EU) 2019/1700 of the European Parliament and of the Council by specifying the number and the titles of the variables for the income and living conditions domain (Text with EEA relevance)

[5] French SILC Questionnaire 2021 (Questionnaire SRCV 2021)

[6] French LFS Questionnaire 2021 (Questionnaire EEC 2021)

[7] UK SILC Questionnaire 2017

[8] Ir ish SILC Questionnaire 2020

[9] Belg ian SILC Questionnaire 2020

[10] Switz SILC Questionnaire 2020

[11] Spanish SILC Questionnaire 2020

[12] Italian SILC Questionnaire 2020

Administrative Data assisting Data Collection through Machine Learning

Keywords: Machine Learning, Administrative Data, Income Variable, Data combination

Introduction

Most variables collected for the labour force survey questionnaire are mandatory. However, income is an exception. It is only collected for workers on account of a third party and the variable is imputed whenever it is associated with a non-response rate greater than 5%.

Even when an answer is obtained it must be validated as the value recorded should only pertain the monthly income (take home) pay from the main job. Currently the strategy for selecting the respondents for value confirmation consists in investigating values below a “low threshold” or above a “high threshold”. This approach is very simplistic, and doesn’t consider that different: occupations, education levels, regions and so on may influence the income value.

The Labour Force Survey questionnaire includes several other variables that characterize the respondent; thus, a study was conducted to ascertain their informativeness regarding the ability to predict the income. Albeit the use of several regression methods the results were not very satisfactory.

In order to improve the results, administrative data containing earnings information was used. The administrative data is itself object of several treatments briefly described in the current paper [1] in section 2, as well as the treatment of the survey data and the subsequent combining of data. This improved our rsq results from 0.441 to 0.909.

Section 3 describes the machine learning algorithms used and the results obtained, while section 4 presents the conclusions and discusses further work.

Data Sets

For the purpose of predicting the income of an individual, two approaches were followed. Initially the other variables from the Labour Force Survey (LFS) were considered to establish a model.

Due to the poor results achieved a second data set was considered to enrich the survey data. The administrative data (DMR - SS) from the social security possessing earnings information was therefore treated and linked to the LFS data: These steps are explained in the current section.

Labour Force Survey (LFS) Data

Labour force data came from a sample survey. The sampling base was extracted from the National Housing File (FNA) and consisted of main residence family accommodation. It was aimed at residents in family accommodations who, in the reference week, lived in that accommodation, considering that it was their main residence. The information characterizing individuals in relation to the labour market was collected only for residents aged 16 to 89 and a stratified and multistage sampling scheme was followed.

The sample is of the panel type with a rotation scheme in which the housings remain in the sample for six consecutive quarters. The total sample is divided into six subsamples (rotations) and each quarter each subsample is replaced by another after having been observed six times. A subsampling strategy (wave approach) was adopted taking advantage of the organization of the LFS in rotations. This feature makes it possible to build a microdata annual base composed of four subsamples corresponding to the rotation that first enters the sample (new rotation) in each quarter but asking some of the questions to on only 1/6 of the quarterly sample (new rotation). This methodology makes it possible to reduce the statistical burden on families and, consequently, the interview time, as they only respond to the entire questionnaire in one of the six inquiry quarters (in the quarter in which they first enter the sample). This happens for some of the variables including the income so only data from the first rotation was used.

Due to the changes in the survey methodology [2] only data from January 2021 to February 2022 was used for this study, around 30000 households each quarter.

Monthly Employees Earnings – Social Security “DMR-SS”

A second dataset was selected coming from an administrative source, social security. This contains the Monthly Employees Earnings – Social Security “DMR-SS” and was already processed and enriched comprising information stratified by age, region, occupation and education level groups. Part of the problem with this source was its completeness. For this reason, data from 2019 was used for the current study.

To overcome the issue of completeness several versions of the information for the same reference period are requested, to accommodate revisions in the original data into the statistical process. Since 2014, Statistics Portugal receives monthly information from the Social Security (SS) regarding the monthly remuneration statements (DMR) that are filled in by national companies. For the same month, four versions of the information are delivered, received one month apart, due to the possibility of companies correcting the information provided. With a view to producing official statistics, it is intended to identify, in the first version received, the companies that tend to make corrections and the missing (companies that do not deliver the declaration) and carry out an imputation exercise in order to obtain provisional values on salaries, closer to the definitive values (which may only be known three months later) than those received from the SS. The current procedure assumes that the latest version (4th version) is the correct one, and the imputation exercise is not carried. A dataset with the history of each company during the previous year is built that tracks through twelve yes or no variables how significant are the corrections a company has made. Based on 2 criteria the companies subject to imputation are detected. The criteria are three or more corrections have been made or the Support Vector Machine identifies the company as prone to corrections.

After these treatments the data is stratified to allow combining data at an aggregated level instead of unit level [3]. Combining this data is important as it not only increases the various application of the data source but will also stimulate the extraction of new information. Especially for this study example combining survey and administrative data, will increase the informativeness of the variables and it might provide ways to get a grip on the income composition of the surveyed population.

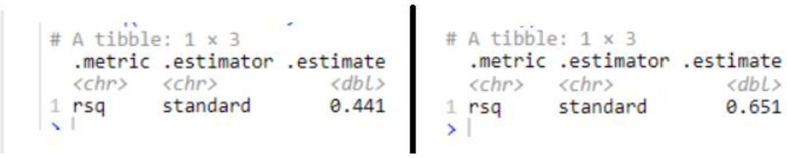
The combination of data is therefore conducted at an aggregated level, and it falls under the concept of data ‘fusion’; i.e. integration of multiple sources data to produce synthetic data that is more informative than the original, or indirect usage category [4].

No direct identifiers (for example, social insurance numbers, names of persons, etc) were available thus it was not possible to precisely identify the statistical units and link them. Instead, key variables were used as there are a set of variables that, when considered together, can be used to identify units [5]. The combination of the datasets was performed using an iterative method where matching was tried using a space of five key variables (gender, profession, education level, region and age) followed by different combinations of reduced dimensions of the initial space.

Methods and Results

The preliminary analysis of the income variable shown a non-normal distribution which may affect the model performance (as indeed was confirmed by the results). To minimize this effect a logarithm of the variable was used instead.

An initial dataset exploratory modelling was performed using linear regression through glmnet R method. The default penalty of 1 was not used as some variables are always to be included unpenalized in the model, such as the demographic variables gender and age. Also, we had no prior knowledge or preference over the variables so applying separate penalty factors to each coefficient did not make sense.

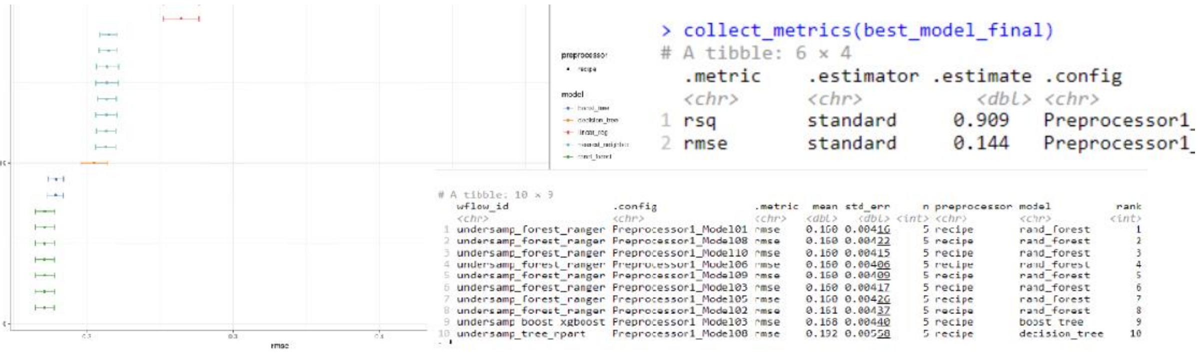


```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.441
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rsq     standard       0.651
>
```

Figure 1. Results using directly the variable (left) and its logarithm (right)

The first results confirmed our choice regarding the logarithm of the variable and after some usual treatments (dummy variables for classes, etc) an efficient grid search via racing with ANOVA models was conducted.



```
> collect_metrics(best_model_final)
# A tibble: 6 × 4
  .metric    .estimator .estimate .config
  <chr>      <chr>          <dbl> <chr>
1 rsq        standard       0.909 Preprocessor1_
2 rmse       standard       0.144 Preprocessor1_
```

```
# A tibble: 10 × 9
   wflow_id                .config              .metric  mean std_err     n preprocessor model           rank
   <chr>                   <chr>                <chr>   <dbl>   <dbl> <int> <chr>        <chr>          <int>
 1 undersamp_forest_ranger Preprocessor1_Model01 rmse    0.160 0.00416     5 recipe       rand_forest        1
 2 undersamp_forest_ranger Preprocessor1_Model08 rmse    0.160 0.00422     5 recipe       rand_forest        2
 3 undersamp_forest_ranger Preprocessor1_Model10 rmse    0.160 0.00415     5 recipe       rand_forest        3
 4 undersamp_forest_ranger Preprocessor1_Model06 rmse    0.160 0.00406     5 recipe       rand_forest        4
 5 undersamp_forest_ranger Preprocessor1_Model09 rmse    0.160 0.00409     5 recipe       rand_forest        5
 6 undersamp_forest_ranger Preprocessor1_Model03 rmse    0.160 0.00417     5 recipe       rand_forest        6
 7 undersamp_forest_ranger Preprocessor1_Model05 rmse    0.160 0.00426     5 recipe       rand_forest        7
 8 undersamp_forest_ranger Preprocessor1_Model02 rmse    0.161 0.00437     5 recipe       rand_forest        8
 9 undersamp_boost_xgboost Preprocessor1_Model03 rmse    0.168 0.00440     5 recipe       boost_tree         9
10 undersamp_tree_rpart    Preprocessor1_Model08 rmse    0.192 0.00558     5 recipe       decision_tree     10
```

Figure 2. Results show Random Forests wins consistently the race

The best results were consistently achieved with random forest, when declaring an interaction between the variables Education and National Classification of Occupations (CNP at two digits) as would be expected. The interaction between Education and NACE provided slightly less good results probably due to its broader scope.

The plot confronting the real value (V2430) and the predicted value shows that resources are being unnecessarily used investigating the set of values inside the box just because they are below the lower threshold (first dashline – 300 €) although these are expectable values. On the other hand, values not expected (on the ellipses) should be inspected.

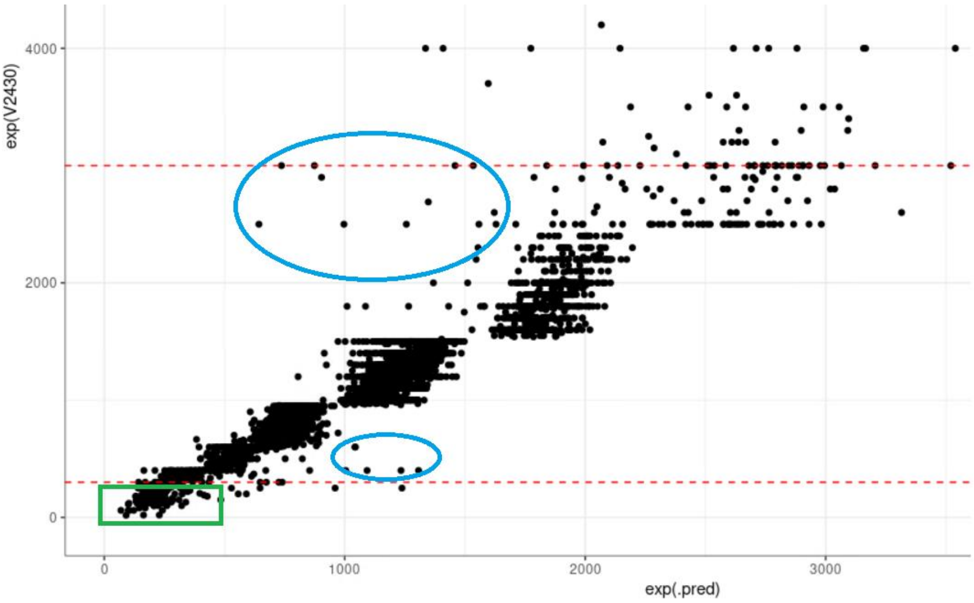


Figure 3. Results confronting the value (V2430) and the prediction

Conclusions

To decide which observed income values were suspicious and could be targeted for further investigation we used regression algorithms to predict the values and computed the distance to the real values. The results obtained were very relevant showing us that we were spending resources scrutinizing values that were expectable just because they were very high or very low. Simultaneously there’s a set of values that albeit being in the middle-income range are unexpected and may now be detected and looked into.

Despite the encouraging results achieved, we would like to start collecting data on the income values investigation and from there, identify which were found suspicious and investigated, which were corrected and the corrected values. Although this information was for operational reasons reported it was not recorded in an organized and systematic way making it impossible to treat this case using classification algorithms. With labelled data it will be possible in a future study to do so.

References

[5] F. Santos, 2019, Support Document on the Mehodology of Detection and Imputation of Data from Monthly Remuneration Statements (DMR-SS), INE Statisitcal Methods

[6] Methodological Document Labour Force Survey (2021 series)

[7] P. Daas, J. Maślankowski, D. Salgado, S. Quaresma, T. Tuoto, L. Consiglio, G. Brancato, P. Righi, M. Six, A. Kowarik, Methodology and Quality, Deliverable K9: Revised version of the methodological report (2020) ESSnet Big Data II

[8] Deliverable A1: Usage of Administrative Data Sources for Statistical Purposes (2015) MIAD - Methodologies for an Integrated Use of Administrative Data

[9] M. Templ, A. Kowarik, and B. Meindl, Statistical Disclosure Control for Micro-Data Usingthe RPackage sdcMicro, Journal of Statistical Software (2015), Volume 67, Issue 4

Data Integration (JENK 1M.2)

Session Chair: **Mojca Bavdaž** *(University of Ljubiana)*

Intergration of social administrative data: The case of Turkish education-labour force database
Turgay Altun *(Turkish Statistical Institute- TURKSTAT)*

Quality measure to evaluate statistical matching methods
Arnout van Delden (Statistics Netherlands)*; Francesca Goudie (Statistics Netherlands); Ton de Waal (Statistics Netherlands)

Data linking from different sources - key to better statistics
Paweł Murawski *(Statistics Poland)*

Quality measure to evaluate statistical matching methods

Keywords: bias, variance, random hot deck, distance hot deck

Introduction

Sample surveys are a crucial way for National Statistical Institutes (NSIs) to collect data, where participants answer questions related to variables of interest. However, sometimes one is interested to estimate a contingency table that is based on variables which have been collected in two different surveys. The number of units in the overlap of two such surveys is often too small to reliably estimate the contingency table. This overlap is small especially for NSIs which have a sampling coordination system that aims to reduce response burden, by reducing the probability that the same unit participates in more than one survey.

Rather than directly using the overlap, one can apply a set of methods referred to as 'Statistical Matching' (SM) to estimate the joint relationship between variables observed in different samples. This estimate is obtained by matching different, but similar, units to each other. The similarity between the units is based on common background variables that are present in both data sets, denoted by X. Let Y be the categorical target variable in sample A, of size n_A, with T_Y categories and with observations $y_1, \dots, y_{n_A}$. Let Z be the target variable in sample B, of size n_B, with T_Z categories and with observations $z_1, \dots, z_{n_B}$. One then obtains the statistical matching situation of Figure 1, columns 1–4.

Table 1. Statistical matching situation

Sample	Y	Z	X	Y*
A	✓		✓	✓
B		✓	✓	✓

In statistical matching, since one is not relying on the overlap only, one cannot estimate the joint (Y, Z) distribution directly. One either needs to make an assumption, use certain restrictions, make use of a small overlap or of a third data set (see D'Oratio et al, 2006 for an overview). In the current paper we consider the situation that there is a small overlap between sample A and B. Furthermore, we use the commonly used conditional independence assumption (CIA) that the distribution of Y and Z are independent given the auxiliary variables X. Unfortunately one cannot know in practice whether this assumption really holds for the data set at hand. If the assumption does not hold, the estimated contingency table will be biased. Therefore, NSI's seldom publish output based on statistical matching. An exception is when a proxy target variable Y^* is available for both samples (see Figure 1 column 5), see D'Oratio et al. [1] and Kim et al. [2]. Such a variable might be obtained from register data that can be linked at unit level to both samples. Still, NSI's would like to have an estimate of the bias of the table before they decide that the output can be published. To our knowledge, such a quality indicator does not yet exist. Current quality estimators for contingency tables provide a so-called uncertainty range (using Frechét-bounds) with respect to the unknown true population parameters that does not depend on the actual statistical matching method applied, see Manski [3], Rässler [4] and references in D'Oratio et al, ([1], p.98-100).

In the current paper we propose an estimation procedure for the bias and variance of a table produced by statistical matching and we evaluate how well that procedure works.

Proposed quality estimation procedure

In addition to the notation that has already been introduced, let C denote the overlap which contains all the variables Y, Z, X and Y^*. Let the total number of observations across the datasets be defined as $n = n_A + n_B$. Further one realisation of a sample A and B from the population is referred to as sample pair s, and denoted by $\mathrm{A}^{(s)}$ and $\mathrm{B}^{(s)}$. In practice one has only a single pair of samples A and B, in a simulation study one can draw multiple sample pairs s, see section 3. Let p_{ij} denote the true cell proportion in the population for category i of Y and category j of Z, and let $\tilde{p}_{ij}^{(m,s)}$denote the estimated proportion by applying SM method m to sample pair s. Further let $\hat{p}_{ij}^{(s)}$ be the estimated proportion given by the cell proportions in the overlap $\mathrm{A}^{(s)} \cap \mathrm{B}^{(s)}$.

For each sample pair s, a large number of bootstrap sample pairs r ($r = 1, \dots, R$) is drawn of size n_A from $\mathrm{A}^{(s)}$ and n_B from $\mathrm{B}^{(s)}$. For each bootstrap sample pair r statistical matching method m is applied. Let $\hat{p}_{ij}^{(r,s)}$ be the corresponding cell proportion in the overlap $A^{(r,s)} \cap B^{(r,s)}$. Finally, let $\tilde{p}_{ij}^{(m,r,s)}$ be the estimated proportion by applying statistical method m using $A^{(r,s)}$ and $B^{(r,s)}$.

The true (unknown) bias and variance of $\tilde{p}_{ij}^{(m)}$ are given by

$$B\left(\tilde{p}_{ij}^{(m)}\right) = \mathbb{E}_s\left(\tilde{p}_{ij}^{(m,s)}\right) - p_{ij}, \text{ and} \tag{1}$$

$$V\left(\tilde{p}_{ij}^{(m)}\right) = \mathbb{E}_s\left(\tilde{p}_{ij}^{(m,s)} - \mathbb{E}_s\left(\tilde{p}_{ij}^{(m,s)}\right)\right)^2 \tag{2}$$

where $\mathbb{E}$ stand for the expectation over all possible sample pairs s from the population and over any stochastic aspects of the SM method.

Their estimated counterparts, using the bootstrap samples, for a given sample pair *s* are:

$$\hat{B}\left(\tilde{p}_{ij}^{(m,s)}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{p}_{ij}^{(m,r,s)}\right) - \hat{p}_{ij}^{(s)}, \text{ and} \tag{3}$$

$$\hat{V}\left(\tilde{p}_{ij}^{(m,s)}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{p}_{ij}^{(m,r,s)} - \bar{\tilde{p}}_{ij}^{(m,s,\bullet)}\right) - \hat{p}_{ij}, \bar{\tilde{p}}_{ij}^{(m,s,\bullet)} = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{p}_{ij}^{(m,r,s)}\right). \tag{4}$$

Simulation study

We compared the true versus the estimated bias and variance in a simulation study, taken from a large real sample: 2016 Public Health Monitor in the Netherlands [5]. In the simulations the sample data represented the population, with target variable Y being income in five categories (quintiles) and variable Z General Health in three categories (bad, moderate, good). Four matching variables X were chosen Age (14 categories in 5 year age-groups), Ethnicity (13 categories), Sex and Education Level (4 categories). We selected the target variables and their auxiliary variables such that their association (Cramer's V) was not too low, and that the number of missings was limited. We obtained 421,226 complete cases.

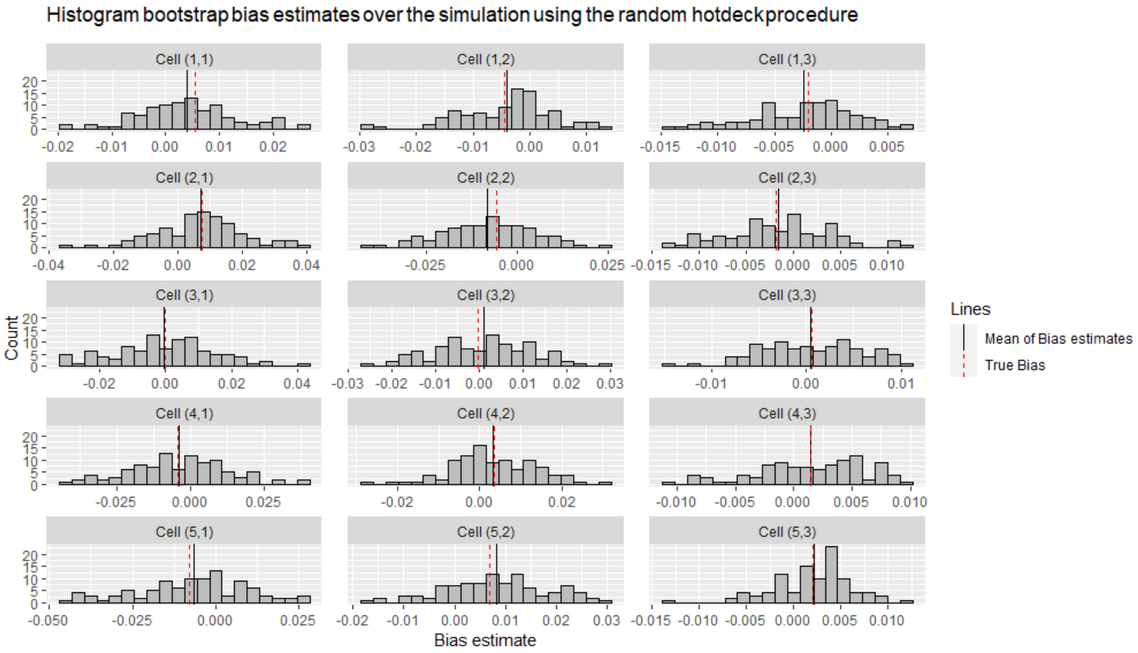
A synthetic proxy variable Y^* was created by adding misclassifications to Y. Each category of Y had the same probability to be misclassified and all observed categories were equally likely given that a category was misclassified. We used a misclassification probability of 0.3 and of 0.475, resulting in a Cramers' V between Y and $Y^*$of 0.6 and 0.4 respectively.

In the simulation study, we drew $S = 100$ replicates of sample A and B from the population data, with n_A and n_B 10500 persons (2.5% sample from the population). The 100 sample pairs were used to estimate the true bias and variance of the cell proportions (eq. (1) and (2) but now based on S sample pairs rather than taking the expectation). Within each sample pair s, $R = 200$ bootstrap sample pairs were drawn to estimate the bias (eq. 3) and the variance (eq. 4), so we obtain a distribution of 100 of those estimates. $S = 100$ and $R = 200$ were shown to yield stable estimates.

Table 2. Contingency table based on the complete cases, used for our simulations[6]

	General Health		
Income quintiles	Good	Okay	Bad
0-20%	0.046	0.027	0.008
20-40%	0.113	0.067	0.015
40-60%	0.156	0.055	0.010
60-80%	0.189	0.047	0.007
80-100%	0.217	0.038	0.005

We tested the proposed quality measures for two statistical matching methods: random hot deck and distance hot deck. For the random hot deck method, donor units in sample A were matched with recipients in sample B when their categories of the matching variables were exactly identical. If no units could be found, a variable was dropped (in fixed order). When multiple units were found, randomly one was chosen. With the distance hot deck method a distance was computed between units in A and B according to the Gowers distance (Gower [6]) and the unit with the shortest distance was selected as donor. This method was used with the help of the function NND.hotdeck in the package StatMatch in R.



[6] We did not use the weights of the health monitor, nor did we impute missing values. Therefore, the true distribution in the Dutch population might differ somewhat from this table.

Figure 2. Distribution of the true and the estimated bias per cell of the contingency table (income quintile, health category) after applying a random hot deck statistical matching.

Results

The true bias was more evenly distributed over the different cells of the contingency table for the random hotdeck method than for the distance hotdeck method, but differences were limited (not shown). Differences between the estimated and true bias by the two methods were comparable, here we only present results of the random hotdeck (with a Cramers' V of 0.6). The estimated bias (eq. 3) as averaged over the $S = 100$ replicates, was quite close to the true bias, but the estimated bias for different samples could vary considerably, see Figure 1.

The true variances were very small for both the random hotdeck and the distance hotdeck method, see Figure 2. The values correspond to a standard error of about 5% of the mean value. The estimated variances (eq. 4), as averaged over the $S = 100$ replicates, were quite close to their true values, the individual values for a given sample pair s could be further away from the true value, see Figure 2.

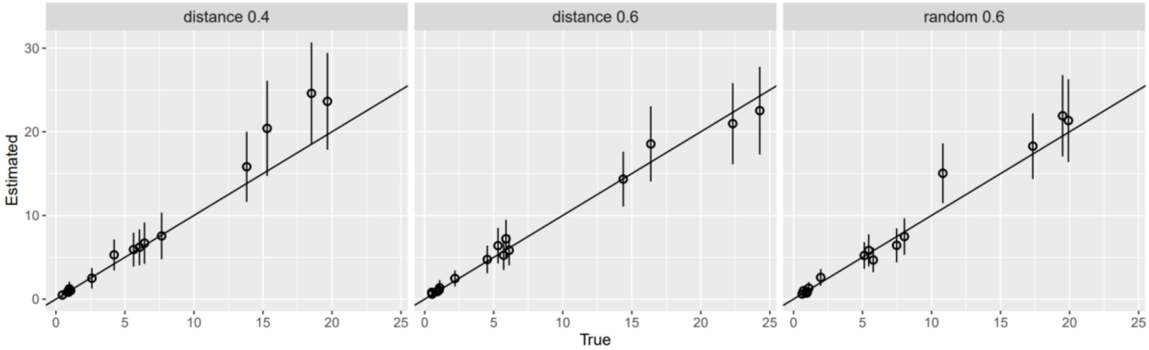


Figure 3. True versus estimated variance (in 10^{-6}) for the 15 cells of the contingency table. Vertical lines represent 2 × standard deviation over the $S = 100$ replicates.

Conclusions

We propose a method to evaluate the bias and variance of output after statistical matching. To our knowledge such a method does not exist yet. Averaged over 100 samples A and B, the bias and variance estimates were quite close to their true values for both random and distance hot deck. In a practical SM situation, one will only have a single pair of samples A and B. Bias and variance estimates based on a single pair of samples were further away from their true values. Although the first results are promising, it would be good to test the method under more conditions: matching methods, sampling sizes, overlap sizes and more associations between Y, Z, X and Y^*.

References

[1] D'Orazio, M., Zio, M.D. and Scanu, M. (2006). Statistical Matching: Theory and Practice. Chichester, UK: Wiley.

[2] Kim, J.K., Berg, E. and Park, T. (2016). Statistical matching using fractional Imputation. Survey Methodology 42, 19-40.

[3] Manski, C.F. (1995) Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press.

[4] Rässler, S. (2002) ¨ Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York: Springer-Verlag.
[5] Public Health Monitor 2016 of the Community Health Services, Statistics Netherlands and the National Institute for Public Health and the Environment
[6] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27 (4), 857–871.

Data linking from different sources – key to better statistics

Keywords: data linking, data integration, data quality, statistical research,

Introduction

Data users expect high quality data as soon as possible and at the lowest possible level of territorial aggregation (xy coordinates are becoming a standard for combining address data.) In order to meet these expectations, public statistics are gathering more and more data from administrative and non-administrative sources in order to improve the quality of the data provided. There is a slow but constant process of refusal to participate in statistical surveys across Europe. Economic considerations and the associated limited financial resources are also of paramount importance. Regardless of this fact, the questionnaire research will remain the main source of data for the purposes of statistical surveys for a long time to come. To obtain the best information, in particular to describe complex and multifaceted phenomena, it is necessary to have data of the highest quality, complete and available in a short time. Usually, access to one source meeting the presented criteria is impossible or very limited. However, high-quality data with all the necessary variables in its scope can be obtained by combining data from many sources, including in particular questionnaire surveys, administrative and non-administrative data, and Big Data.

Data quality – methods of data quality testing

The quality of the input data is most important for the future integration process. Getting to know the data - generally called data profiling - should precede the join stage. It also allows to choose the optimal algorithm. Methods of data quality testing, including IT instruments, used by Statistics Poland units will be presented.

Data quality – methods of improving data quality, including the main problems

Most of the NSO has developed own methods to improve the quality of data - especially address data. The methods of improving data quality, both statistical and non-statistical, used by Statistics Poland units will be presented.

Methods

Selection of the best method depends on many factors - the quality and the manipulations (transformation) carried out on them are crucial. The IT environment in which work related to data integration is performed will be presented. The methods that are used, distinguishing between the quality of the data and the source of their origin.

Integrated Microdata System

The Integrated Microdata System, which was built in Statistics Poland in order to integrate data from multiple sources and based on them, to implement multi-faceted research on socio-economic phenomena, will be presented. These studies are characterized by high quality, short study implementation time and the greatest possible reduction of costs and, to a large extent, reduction of the burden on respondents.

Data integration methods

Deterministic methods

Probabilistic data linking was possible for most of the datasets that were used in the research. Only for a part of the population and some data sources it was necessary to use other methods of combining data. During the presentation, all data sources used in the 2021 census will be indicated along with the results of integration data for individual data sets.

Probabilistic methods

For entities that did not have a unique join key in a given set, it was necessary to use probabilistic data linking methods. During the presentation, methods of combining data that were used in the 2021 census will be presented along with the results of the integration of these data.

Results

As examples, the results of data integration from multiple sources will be presented in order to build the population of entities specified for participation in Census 2021.

The second example will be the integration of data from different sources for the determination of the Census 2021.

Conclusions

Data/records linkage techniques can help identify the same person, administrative unit and so on. Prior knowledge of whether a linkage is feasible is very important. The development of methods and IT tools seems to be key importance given the growing requirements of data recipients. Therefore, the presentation of unique methods of data processing for the purpose of their integration, which are used in Statistics Poland, together with a detailed description of the methods of combining data and its results, may contribute to the broadening of knowledge in this field in the European and global statistical community.

POST01: Poster session

Exploring the Italian ecoregions: institutional fragmentation and socio-demographic characteristics
Raffaella Succi *(National Institute of Statistics–ISTAT)*, Marija Mamolo *(National Institute of Statistics–ISTAT)*

Estimation of hidden populations using single-source capture recapture models with singleRsource package
Piotr Chlebicki, Maciej Beręsewicz *(Poznań University of Economics and Business)*

Machine Learning for Estimating Commuting Population
Loredana Di Consiglio *(National Institute of Statistics–ISTAT)*

Protection of linked tables with a suppressive approach. Method and Use Cases
Julien Jamme (*Insee*)*; Nathanael Rastout (*Insee*); Clément Guillo (*Insee*)

Case studies in using Satellite Earth Observation for National Statistics: the GAUSS project
Phillip Harwood *(Evenflow)*, Orestis Speyer *(National Observatory of Athens),* Ewa Panek *(Institute of Geodesy and Cartography)*

Bootstrap procedure for variance estimation of small area estimates with application to job statistics
Djalel-Eddine Meskaldji *(Swiss Federal Statistical Office)*

Application of an Adaptive Survey Design on the Italian Population Census
Claudia De Vitiis (ISTAT)*; Stefano Falorsi (ISTAT); Francesca Inglese (ISTAT); Paolo Righi (ISTAT); Marco Dionisio Terribili (ISTAT); Alessio Guandalini (ISTAT)

A New Way to Learn Data Science: The Funathon, a Non-competitive Hackathon
Marie-Pierre de Bellefon (*Insee*)*; Romain Lesur (*Insee*)

Visualising the pathways in and out of employment during COVID-19 in Estonia
Kadri Rootalu *(Statistics Estonia)*

Towards an Open Source ETL Statistical System
Cristiano Tessitore (*Eurostat*), Marius Felecan (*Gopa Luxembourg*)

Exploring the Italian ecoregions: institutional fragmentation and socio-demographic characteristics

Keywords: Italian ecoregions, institutional fragmentation, socio-demographic characteristics of the ecoregions

INTRODUCTION

Research on sustainability has to recognise that society is embedded in the global ecological system, which integrates all the biological components of the planet, their interrelations and the interactions with the other physical components. Human ecology, studying the interactions of humans with their environments, focuses on understanding the role and managing the impact of human behaviour on the natural environment. However, in order to face the challenges that are proper to sustainability, research needs an interdisciplinary approach in addressing the strong mutual interactions between humans and the environment and a perspective which sees human society as acting within the biosphere.

From the ecological perspective, ecoregions represent worldwide an important tool for orienting sustainable development policies. Each ecoregion covers a relatively large area of land or water with a similar biodiversity of flora, fauna and ecosystems. Therefore, they represent a valuable framework for the implementation of environmental-related policies and strategies oriented to biodiversity conservation. However, a holistic approach to sustainability requires to include in this perspective also the societal component, supporting the vision suggested with “The SDGs wedding cake” [1].

The present research builds on the recent official statistics published by Istat which classify Italian Municipalities according to territorial ecoregions ([2]; [3]). From a statistical point of view, this new classification binds together the natural environment and the society: it allows to combine the rich information on ecological characteristics of ecoregions, including climate, geomorphology and biodiversity, and the information normally collected at municipal level concerning socio-demographic and economic characteristics.

Firstly, the paper aims to explore the institutional fragmentation characterising the Italian ecoregions and their nested hierarchical levels. Environmental-related strategies of biodiversity conservation should be tailored for areas with a homogeneous ecological character. Thus, such an approach has to overcome the classical administrative boundaries, such as those of the municipal territory itself or higher administrative units. However, the larger is the number of institutional actors in search of agreement, the more difficult it is to reach consensus for the implementation of policies [4].

In Italy, the need to adopt an eco-systemic approach to land use planning with the aim to focus on areas which are homogeneous in terms of ecological and socio-economic features (beyond the administrative boundaries), has been central to a scientific and cultural debate. The final report of the project “Verso la Strategia Nazionale per la Biodiversità”, a collaboration between the former Ministry of Environment, Territory and Water resources Protection and the WWF-Italy, summarises the complexity of such an approach and the possible problematic issues concerned [5].

Secondly, the paper presents some main socio-demographic characteristics of the Italian ecoregions. The degree of socio-demographic heterogeneity within ecoregions (and its hierarchical levels) should be considered when discussing strategies oriented to environmental conservation. Socio-demographic characteristics are relevant to understand the target of environment-oriented policies at the societal level.

METHODS

We use the Istat classification of Italian Municipalities (i.e. LAUs) according to terrestrial ecoregions ([3]). The classification is the result of a collaboration between Istat and the Interuniversity Research Center "Biodiversity, Ecosystem Services and Sustainability" (CIRBISES), Department of Environmental biology (DBA) of the Sapienza University of Rome. Since 2020 these data represent official statistics.

Italian ecoregions are organised in four nested hierarchical levels, i.e. Divisions, Provinces, Sections and Subsections. This nested hierarchical classification splits the land “into increasingly homogeneous units, based on specific combination of the climate, bio-geographical, geomorphological and hydrographic features that determine presence and distribution of species, communities and ecosystems” ([3]).

To evaluate the institutional fragmentation of an ecoregional area we consider: the absolute number of administrative units belonging to the ecoregion; the number of parts of administrative regions at NUTS-2 level; the average administrative unit area in km2. We also consider the number of municipalities with less than 5.000 inhabitants, units that have a low administrative capacity, insufficient financial resources and reduced managerial capacity.

We use Istat census socio-demographic variables at municipality level and combine them with the ecoregional classification of municipalities. Our preliminary analysis considers the last year for which the ecoregional classification of municipalities is available (1 January 2020), and consequently the census data on 31 December 2019. In this preliminary analysis we focus on the percentage of national territory covered by ecoregions, the total amount of population by ecoregion and the population aged 65+ in the ecoregions.

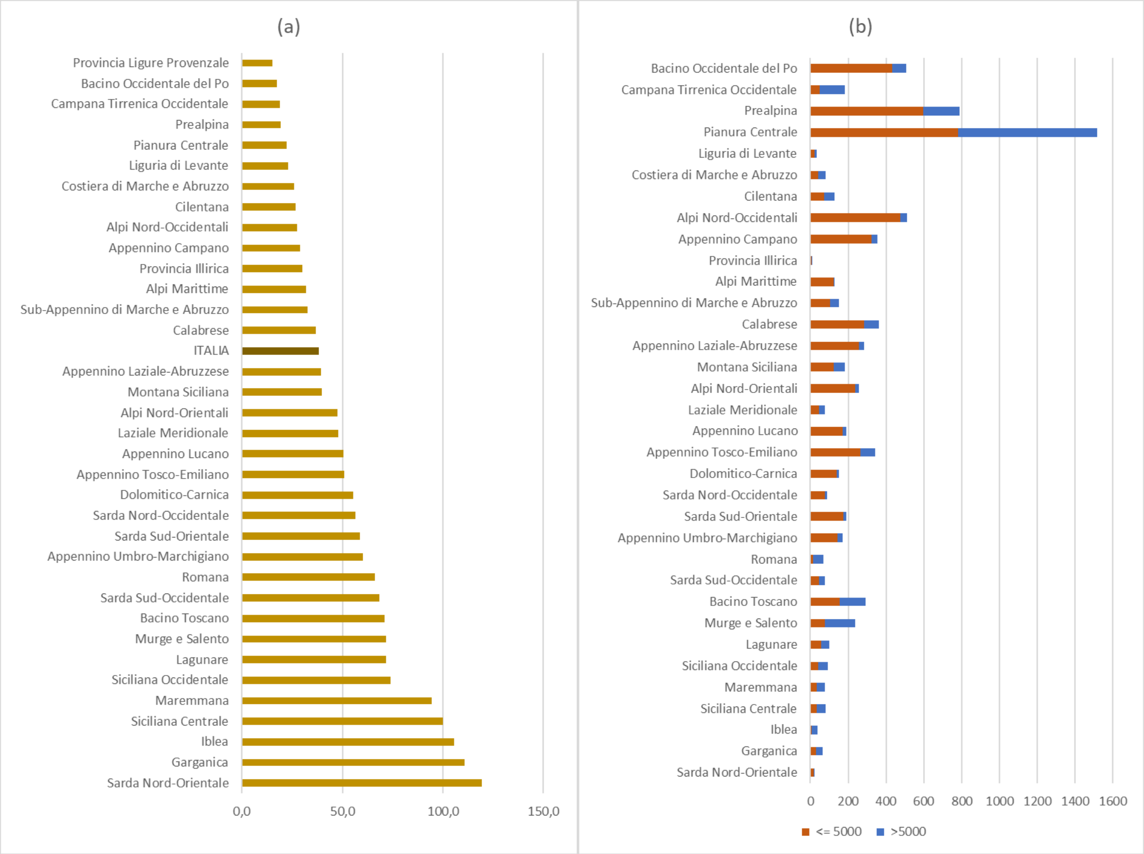
RESULTS

On 1 January 2020, in Italy there are 20 NUTS-2 regions, 196 NUTS-3 regions and 7.903 municipalities (LAU). To capture the ecological peculiarities of the country, the municipalities have been classified according to 35 ecological subsections. The institutional fragmentation is already evident considering that at the national level the 7903 municipalities have an average area of 38,2 km2, corresponding to a square grid of side 6 km. However, the phenomenon is not homogeneous across the ecoregional areas: we observe that in 14 subsections out of 35 the average municipality area is lower than the national mean (see Figure 1, a). The first three ecoregional subsections with the lowest average area are Provincia Ligure Provenzale (15,1 km2), Bacino Occidentale del Po (17,6 km2) and Campana Tirrenica Occidentale (18,9 km2).

Among these 14 subsections we find those with the highest number of municipalities, i.e. decision centres: Pianura Centrale with 1516 municipalities, Prealpina with 791 municipalities and Bacino Occidentale del Po with 508 municipalities (see Figure 1, b). Another element of frailty is the high number of municipalities with less than 5.000 inhabitants (at national level 5.496 in total, 69,5%), that characterize the Alpine, Prealpine and Apennine ecoregions.

Moreover, if we consider the NUTS-2 regional level, we observe that the subsections are split in 85 portions of land with an average of 3.353,7 km2, corresponding to a square grid of side 60 km.

Figure 1. Italian ecoregional subsections by a) average municipality area (km2) and b) n. of municipalities with inhabitants <= 5.000 and with inhabitants > 5.000 - 2020



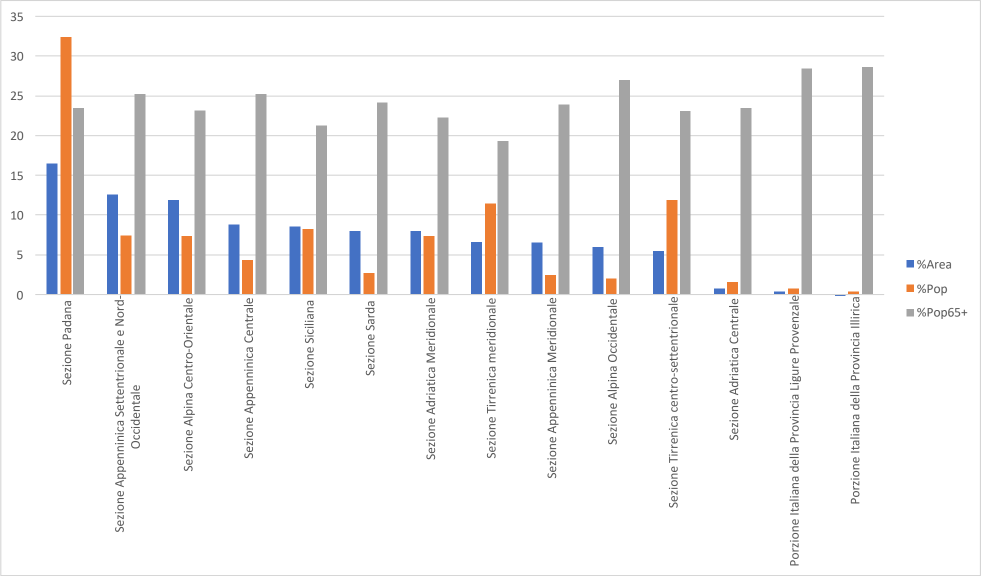
Source: own elaboration on Istat, Classification of Italian Municipalities according to Ecoregions (2020)

From a socio-demographic point of view, the Italian ecoregions are heterogeneous. The section Padana represents 32,4% of the total population, even though it covers 16,5% of the territory (see Figure 2, the sections are ordered by % of covered area). With a slightly lower percentage of covered area (about 12%), the sections Alpina Centro-Orientale and Appenninica Settentrionale/Nord-Occidentale represent respectively about 7% of the total population. The ratio between the percentage of total population and the area covered by the ecological sections is about double also in the sections Tirrenica CentroSettentrionale and Tirrenica Meridionale.

The age composition of the population shows a high percentage of people aged 65+ across all ecoregional sections. Besides the higher percentages in the small areas covered by the parts of Provincia Illirica and Ligure Provenzale (about 28%), a higher percentage of people aged 65+ is found in the sections Alpina Occidentale (26,9%), Apenninica Settentrionale/Nord-Occidentale and Centrale (about 25%) and subsection Sarda (24%).

A further division of the territory into subsections confirms that the Pianura Centrale, which covers 11,1% of the Italian territory, includes the highest percentage of the total population (25,6%). The percentage of the population aged 65+ ranges from a minimum of 17,8% in the subsection Campana Tirrenica-Occidentale to 28,7% in the subsection Alpi Marittime.

Figure 2. Area, Population and People aged 65+ in the Italian ecoregional sections (percent) - 2020.



Source: own elaboration on Istat, Classification of Italian Municipalities according to Ecoregions (2020)

CONCLUSIONS

In the next steps of our analysis, we aim at studying the impact of the administrative fragmentation on the land consumption considering the ecoregions as the reference unit and combining in the analysis the territorial classification of municipalities by the degree of urbanization. Furthermore, we intend to explore the socio-demographic heterogeneity within the ecoregions, at different levels. Since ecoregions and their further division into smaller units cover different parts of various administrative regions (at NUTS-2 level), we would like to put into evidence similarities and differences between areas which are part of the same ecoregion but belong to different NUTS-2 administrative regions; and, on the other hand, the similarities and differences between areas which are part of the same NUTS-2 administrative region but belong to different ecoregions.

REFERENCES

[1] Stockholm Resilence Centre, “A new way of viewing the Sustainable Development Goals and how they are all linked to food” (2016), accessed October 2022, https://www.stockholmresilience.org/research/research-news/2016-06-14-the-sdgswedding-cake.html

[2] C. Blasi, G. Capotorti, R. Copiz, D. Guida, B. Mollo, D. Smiraglia L. Zavattero Classification and mapping of the ecoregions of Italy, Plant Biosystems, 148:6, 12551345 (2014)

[3] Istat, Classificazione dei Comuni secondo le Ecoregioni d’Italia - Nota metodologica, (2020).

[4] Bartolini, D., "Administrative fragmentation and economic performance of OECD TL2 regions", *OECD Journal: Economic Studies*, vol. 2016/1, https://doi.org/10.1787/eco_studies-2016-5jg318w59m6h (2017).

[5] MATTM, “Ecoregioni, biodiversità e governo del territorio - la pianificazione d’area vasta come strumento di applicazione dell’approccio ecosistemico”, rapporto finale del progetto “Verso la Strategia Nazionale per la Biodiversità” (2009).

Estimation of hidden populations using single-source capture recapture models with singleRsource package

Keywords: population size estimation, hard-to-reach population, truncated regression, the dark number

INTRODUCTION

Population size estimation is an important issue in official statistics, social sciences and natural sciences [1]. One way to tackle this problem is by applying capture-recapture methods, which can be classified depending on the number of sources used, i.e. one source or two and more sources.

In this study we focus on the first group of methods, i.e. single-source capture-recapture (SSCR). SSCR models assume that observed counts follow truncated count distributions (e.g. zero-truncated Poisson, one-inflated zero-truncated geometric) and this assumption is used to estimate missing (hidden) zero counts. The literature includes applications of SSCR methods for estimating the number of irregular migrants, home violence cases or homeless people [1].

In the paper we focus on presenting the new R [2] package called *singleRsource*, which implements state-of-the-art SSCR models with user-friendly functions. The package is currently under development and has not been submitted to CRAN yet.

The structure of the paper is as follows. In section 2 we briefly present SSCR models implemented in the package. Section 3 presents an overview of the package with selected examples from the literature. Paper ends with conclusions and references.

METHODS

SSCR models are based on the assumption that a given person may be observed multiple times in a given data source (e.g. police records). For instance, migrants working without a valid work permit may be approached and verified multiple times during the year by border guards or labour inspectors. Thus, the resulting counts are zero-truncated, as the population outside a given register is not observed. In the literature, the number of zeros is called *the dark number* [3].

In order to estimate the size of a population, distributional assumptions about observed counts are postulated. The most common are zero-truncated Poisson, negative binomial or geometric distributions, and when covariates are available, a regression model corresponding to the postulated distribution can be applied. Recently, much attention has been devoted to studying one-inflation in observed count data, which results either from individuals exiting the population (e.g. irregular migrants expelled from the country) or a change in their behaviour (e.g. learning how to avoid being the subject of scrutiny) [4].

In the package we implemented the following SSCR models:

- Chao’s regression,
- Zelterman’s regression,
- zero-truncated and zero-one truncated Poisson, negative binomial and geometric regression,
- one-inflated zero-truncated and zero-truncated one-inflated Poisson, negative binomial and geometric regression (extended by the authors),
- pseudo-hurdle Poisson and geometric regression (proposed by the authors).

The package implements analytical and bootstrap variance estimators of the population size estimator as well as functions to assess the fit and quality of the model (e.g. leave-one-out diagnostics, rootograms).

It should be noted that there are several R packages that make it possible to fit zero-truncated regression models (e.g. countreg [5], VGAM [6]) but our package offers many other models, including those proposed by the authors, and is designed for population size estimation problems.

RESULTS

The main function of the package is `estimate_popsize`, which can be used to specify the regression formula, dataset, model, type of variance and estimation methods, among other things. The code below presents an application of the package for the Dutch study about irregular migrants presented in the paper [7].

```
ModelPo <- estimate_popsize(formula = capture ~ ., ## regression formula
                            data = netherlandsimmigrant, ## dataset
                            pop.var = "analytic", ## analytical variance
                            model = "ztpoisson", ## zero-truncated Poisson
                            method = "robust") ## IWLS algorithm
```

results may be obtained using `summary(ModelPo)`.

```
#> estimate_popsize(formula = capture ~ ., data = netherlandsimmigrant,
#>     model = "ztpoisson", method = "robust", pop.var = "analytic")
#>
#> Pearson Residuals:
#>      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
#> -0.488779 -0.486085 -0.297859  0.002075 -0.210439 13.921578
#>
#> Coefficients:
#>                      Estimate Std. Error z value P(>|z|)
#> (Intercept)            -2.317      0.449   -5.16 2.5e-07 ***
#> gender                  0.397      0.163    2.44 1.5e-02   *
#> age                     0.975      0.408    2.39 1.7e-02   *
#> reason                  0.011      0.162    0.07 9.5e-01
#> nationAsia             -1.092      0.302   -3.62 2.9e-04 ***
#> nationNorth Africa      0.190      0.194    0.98 3.3e-01
#> nationRest of Africa   -0.911      0.301   -3.03 2.5e-03  **
#> nationSurinam          -2.337      1.014   -2.31 2.1e-02   *
#> nationTurkey           -1.675      0.603   -2.78 5.5e-03  **
#> -----------------------
#> Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 ' '
#>
#> AIC: 1714.896
#> BIC: 1764.747
#> Deviance: 1128.549
#>
#> Log-likelihood: -848.4481 on 1871 Degrees of freedom
#> Number of iterations: 8
#> -----------------------
#> Population size estimation results:
```

```
#> Point estimate 12691.45
#> Observed proportion: 14.8% (N obs = 1880)
#> Std. Error 2809.508
#> 95% CI for the population size:
#>              lowerBound upperBound
#> Studentized    7184.917   18197.99
#> Logtransform   8430.749   19723.38
#> 95% CI for the share of observed population:
#>              lowerBound upperBound
#> Studentized   10.330814   26.16592
#> Logtransform   9.531836   22.29932
```

The first part of the results is similar to the `lm/glm` output, while the most important output is presented in the `population size estimation results`. This section includes a point estimate (here 12.7k), the share of the observed population (14.8%), analytical standard error (2.8k) and two confidence intervals. During the conference we plan to show more models and results based on new datasets.

CONCLUSIONS

In the paper we introduce the *singleRsource* R package for estimating SSCR models. The package implements state-of-the-art models as well as some new models proposed by the authors. The software is prepared for users interested in estimating the size of populations, particularly those that are hard-to-reach or for which information is only available from one source and dual/multiple system estimation cannot be utilised.

REFERENCES

[1] Böhning, Bunge , & van der Heijden (Eds.). (2018). Capture-recapture methods for the social and medical sciences. CRC Press.

[2] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[3] Cruyff, van Dijk & van der Heijden (2017). The challenge of counting victims of human trafficking: Not on the record: A multiple systems estimation of the numbers of human trafficking victims in the Netherlands in 2010–2015 by year, age, gender, and type of exploitation. Chance, 30(3), 41-49.

[4] Godwin & Böhning (2017). Estimation of the population size by using the one-inflated positive Poisson model. Journal of the Royal Statistical Society. Series C: Applied Statistics, 66(2), 425–448.

[5] Zeileis, Kleiber & Jackman (2008) Regression Models for Count Data in R. Journal of Statistical Software, 27 (8), 1-25.

[6] Yee, Stoklosa & Huggins (2015). The VGAM Package for Capture-Recapture Data Using the Conditional Likelihood. Journal of Statistical Software, 65(5), 1-33.

[7] van der Heijden, Bustami, Cruyff, Engbersen & van Houwelingen (2003). Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modelling, 3(4), 305-322.

Protection of linked tables with a suppressive approach. Method and Use Cases.

Keywords: SDC, suppressive methods, linked tables, R, Tau-Argus

INTRODUCTION

When a NSI releases tabular data from one source, the number of tables which are released can be very great. The more tables there are, the more potential links there are between them, and the more difficult it is to handle confidentiality issues.

We consider that the NSI uses a two-stage suppressive method (primary and secondary suppression), with the second one designed as the solution of an optimization problem [1]. We also assume that the software Tau-Argus is the best tool to apply this optimization problem on a table of at most 4 dimensions. Finally, we assume that none of the released tables crosses more than 4 dimensions.

In this context, the difficulty is to deal with the number of linked tables in each planned data release. Most of the releases contain more than 10 tables, and there are more than a hundred in the most complex ones.

In this paper, we propose, firstly, to describe some use cases to show where complexity lies and how to handle it to capture all the links between the tables. Secondly, we present an algorithm that can protect an indefinite number of linked tables. The main types of links (margins, additivity, non-nested hierarchies) are handled. The algorithm is implemented in an R package called *rtauargus* and we provide some results to compare it with the native way to proceed with the Tau-Argus software. At last, we present the results of the cell protection processes applied on the use cases.

METHODS

Protect linked tables with an iterative process

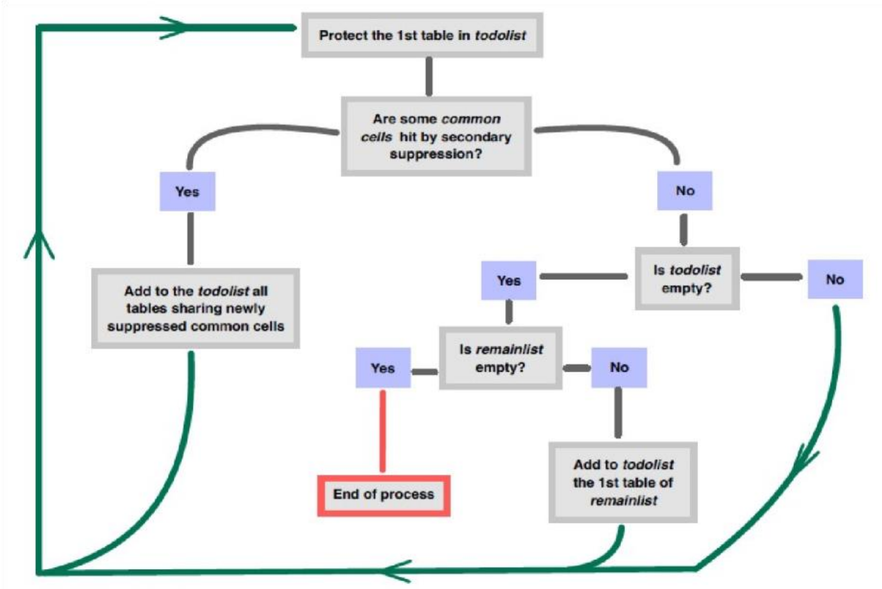
Let’s assume that the sensitive cells, regarding the confidentiality rules, have been detected previously. To protect a large set of linked tables, the algorithm works with two lists. The first one, called *todolist*, is a list of tables which have to be protected. The second one, called *remainlist*, is the list of original tables which haven’t yet been protected at all. We initialize the *todolist* with the only first table of the set and the *remainlist* with all the others, except the ones where none of the cells are risky regarding the confidentiality rule (no primary suppression).

The Figure 1 shows how the algorithm proceeds to protect all tables till the suppression is stable. Firstly, the secondary suppression on the table in the *todolist* is done with TauArgus. Then, if some new common cells have been hit by secondary suppression, all the tables of set sharing these common cells are added to the *todolist* and the process is continued by protecting the next table on the *todolist*. Otherwise, that is to say if none of the common cells has been hit, the next step of the process depends on the state of the two lists. If the *todolist* isn't empty, the process can carry on, otherwise the *todolist* is filled with the first table in the *remainlist*. But, if this latter is also empty, then the process can stop.

Thus, the secondary suppression is processed on each table with sensitive cells at least once and can be processed several times on the same table, if several times some common cells in it are suppressed while processing other tables.

The first advantage of this algorithm is to be able to handle any number of tables. It could be longer but not impossible in theory. In addition, we are sure that it will end. In the worst case, the process will end with the suppression of all common cells. But, in our intensive experience with real data (OFATS, IFATS, CIS, ICT and other several european business statistics surveys), the stability is quite fast: 2 or 3 iterations mostly to a maximum of 5.

A prior merging of all tables of the set makes it much easier to find the common cells between all the tables. Actually, the search is only a filter on the merged table thanks to some boolean variables indicating to which tables a cell belongs.



Protect the 1st table in todolist
Are some common cells hit by secondary suppression?
Yes
No
Add to the todolist all tables sharing newly suppressed common cells
Yes
Is todolist empty?
No
Yes
Is remainlist empty?
No
End of process
Add to todolist the 1st table of remainlist

Figure 1: Algorithm implemented in rtauargus::tab_multi_manager() to protect multiple linked tables

An implementation in the R package *rtauargus*

The package *rtauargus* is a light-weight package to interface R and Tau-Argus [2]. The secondary suppression is always done with Tau-Argus. The implementation in R can be useful to ease reproducibility and integration in a production process.

The *rtauargus::tab_multi_manager()*, especially, implements the algorithm presented above. Even to protect very large set of tables, the code to write is very parsimonious [3].

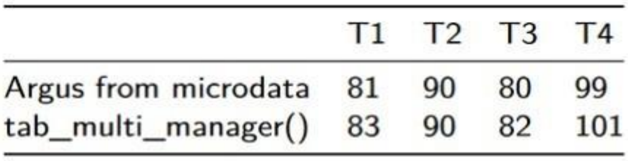
About automatically building the set of tables - Perspectives

As the protection of the large sets of tables is now an easy task, the main difficulty to handle this kind of sets is to build this set from the dissemination scheme. Currently, this task is very empirical because the scheme is presented very differently from one case to another. Our idea would be to find which kind of metadata could represent all kind of diffusion plan, metadata from which the set of tables to protect would be automatically built. In the results section, we present one attempt to do this from the ICT survey's dissemination requested by Eurostat.

RESULTS

A comparison with native implementation in Tau-Argus

Tau-Argus proposes already a way to deal with a set of linked tables, but only with few tables (max to 5 or 6). This limitation was the main reason to develop our own solution to handle larger sets. Thus, even if we can't compare the results of our algorithm with TauArgus on very large sets of tables, we can compare them on small sets (3 to 5 tables). The table 1 shows encouraging results. Even if the over-suppression from our mechanism is expected, it is not so important on our different tests. The paper will provide more systematic tests.

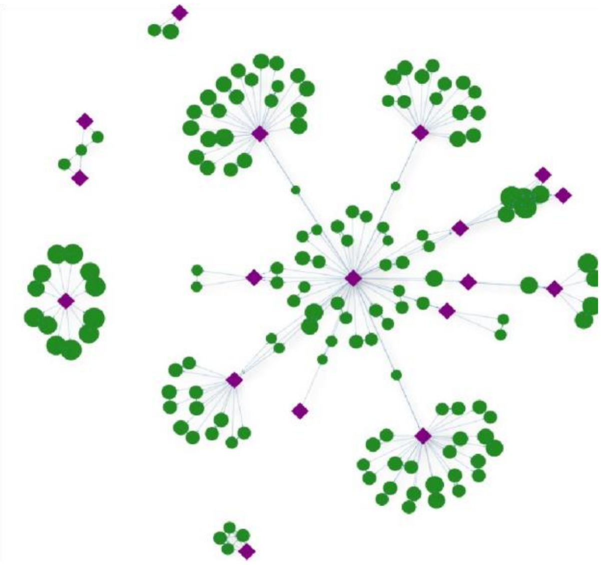


	T1	T2	T3	T4
Argus from microdata	81	90	80	99
tab_multi_manager()	83	90	82	101

Table 1: Number of suppressed cells between Tau-Argus and rtauargus::tab_multi_manager() function on one set of 4 tables

Very large sets of tables, up to 200 tables, have been tested with success. Nevertheless, an alternative way to initialize the process has proven to be more interesting in terms of suppressions in some special cases when the effects of protection intervals need to be lowered.

ICT survey's dissemination: an attempt to draw a general framework of metadata



A graph representation of some existing links has been drawn to split the dissemination plan into independent subplans. These sub-plans are the different connected components of the overall graph. And for each sub-plan, the graph lets us to build the set of linked tables which have to be protected. Even if the graph doesn't represent all the dimensions of the complexity of the plans – some non-nested hierarchies don't appear here – the technique appears very efficient to automate the preparation steps.

But, the try has to be converted. To do this, the mechanism has to be tested on other diffusion plans. *Figure 2: ICT survey - graph of the links between breakdowns' categories in the frequency tables dissemination required by Eurostat.*

cONCLUSIONS

The implementation in the *rtauargus* package of a simple algorithm to protect large set of tables is very effective to consider all the links during the protection process, and the code to do this is quite short and easy to handle. The next step would be to find a framework to represent an original dissemination plan into metadata that can be easily translated into independent sets of tables and each set being able to be dealt with *rtauargus*. The graph representation used for the ICT survey's dissemination could be a step towards this framework.

rEFERENCES

[1] M. Fischetti, J.J. Salazar, *Solving the Cell Suppression Problem on Tabular Data with Linear Constraints*, Management Science, Vol.47, 7, 2001, pp. 1008-1027

[2] P.-Y. Berrard, J. Jamme, N. Rastout, J. Pointet, F. Beroud, *The rtauargus package*, https://inseefrlab.github.io/rtauargus/

[3] J. Jamme, N. Rastout, *rtauargus, A simple way to protect multiple linked tables with R and Tau-Argus*, presentation at the Workshop of the SDC User Group, Paris, September, 2022, *https://github.com/sdcTools/UserSupport/blob/master/docs/W2022/S1_P2_01_Jamme_Rastout_rtauargus.pdf*.

Case studies in using Satellite Earth Observation for National Statistics: the GAUSS project

Keywords: New data sources, geospatial statistics, earth observation, air quality, snow, water, green areas.

Introduction

In recent years there has been increased interest in the concept of “Smart Statistics”, whereby traditional sources of official statistics (survey and administrative data) are complemented by information from sensors (including satellite imaging, sensors deployed on the roads and in the cities, environmental sensors), from tracking devices (such as data from mobile telephones or GPS), behavioural data (e.g. data from online searches, online page views) or social applications data (comments on social media, etc.). These data sources can provide entirely new insights into social and economic trends to drive public policy making.

Satellite Earth Observation (EO) data are an important source of smart statistics, especially when combined with other available data sources. Organisations such as the United Nations Statistical office (UNSTAT), the European Statistical Office (EUROSTAT) as well as many national statistical institutes/offices (NSIs/NSOs) and supporting organisations are currently seeking to incorporate satellite imagery and other space-based data sources into their operational workflows. The aim is to take advantage of the many new sources of EO data, especially the free data available through the EU’s Copernicus programme.

The GAUSS (Generating Advanced Update of Earth Observation for Smart Statistics) [1] project has been funded by the European Space Agency to provide specific demonstrations of how EO data can be used in such workflows, usually in combination with key in situ data. It provides case studies in the fields of air quality statistics reporting, green indicators for natural capital reporting and water statistics reporting, working closely with national agencies to ensure that the outputs are really meeting their needs. It will also develop best practices for an interested user community to support further development of such technologies and solutions beyond the scope of this project.

Methods

Air Quality Statistics

In Greece, the project is working together with ELSTAT to develop high resolution air quality statistics for key atmospheric pollutants, working at Local Administrative Unit level rather than the currently available coarse regional statistics. The aim is to provide statistics responding to the Air Quality Directive, meeting the requirements for reporting through the EOINET Central Data Repository [2].

These statistics are derived through a fusion of satellite data (Sentinel 5P TROPOMI) and reanalysis models (from the Copernicus Atmospheric Monitoring Service) with in situ measurements coming from both the national air pollution monitoring network and from a third-party provider (the RI-PANACEA research network [3]). These are combined using open methodologies and software to give measurements of key pollutants ($PM_{2.5}$, PM_5, NO_2, O_3) on a 1km grid, which are then aggregated to local authority level to meet reporting requirements.

Although developed for Greece, the methods are expected to be transferable to any European country. To confirm this, the project will deploy the methods in Poland, using local providers of in-situ data.

Snow Statistics

Statistics on the extent and conditions of winter snow are important information for northern countries, with such data being used for risk assessment (e.g. from snow melt) and for planning transport, logistics and power. In Finland, statistics on Snow Water Equivalent (SWE) are traditionally provided to the NSO (Statistics Finland) by the Finnish Environment Agency, SYKE, on a specified day, 16th of March, each year [4].

Within this use case, we aim at delivering (1) Average, minimum and maximum snow load on a monthly basis for provinces (to be defined), (2) Anomaly from past twenty years situation, and (3) possible trends using EO data shown in the figure below. In addition, we provide the (4) yearly (2001-2022 at the moment) information on the first snow-free day (referred to as Melt-off day, MoD) derived from EO-based Fractional snow cover maps in 500m resolution over Europe and for Finland in particular. Together with information on snow accumulation, it is possible to estimate the (5) length of the snow period for each of these years and calculate possible trends.

This product has several advantages over traditional snow statistics products, including improved granularity and timeliness. If adopted by Statistics Finland, they will give new information for use by researchers, decision makers and the general public.

Hydrological Drought

The development of a drought index based on water level can be a good indicator useful for hydrological resources managing actions of decision makers. Since the water level measures are not quite well spread in the territory, here we carried on a feasibility study of using satellite data for water level measures retrieving.

Satellite measurements may be a valid alternative to in situ measurements where instruments are rare, missing, discontinuous, malfunctioning, abandoned or difficult to access. Increasing water level measures in the territory, with the use of altimeter data, will open wide new possibilities to have a capillary hydrological monitoring system essential for forecasting, modelling and alert systems set-up. Altimeter data has been used initially for studying the ocean topography, recent studied are exploring the possibility of exploit it also for obtaining rivers water level [5].

We will first see the feasibility of using satellite data for retrieving water level data both for river and lake in Finland case study. The second step of the research is the definition of a Drought Index based on water level measures.

Quality of Urban Green Areas

In Poland, Statistics Poland is interested in improving the quality and timeliness of regional well-being statistics. In particular, they are interested in gathering information on the extent and quality of vegetation at a commune level (LAU). At the national level, this information is currently inhomogeneous, outdated and often with gaps, because it is produced individually by local governments, and at large intervals. This makes it difficult to combine the quality-of-life indicators such as health status with information about the impact of green areas.

The project is developing a service that will provide this information based on data from the European Sentinel-2 spacecraft, together with climate data from the European Centre for Medium-Range Weather Forecasts (ECMWF). The Polish Database of Topographic Objects (BDOT10k) [6] is used to determine the form of green areas, especially in urban agglomerations. The annual statistics generated on green areas are complemented with a raster preview for the entire territory of Poland. Thanks to this, the End-User can monitor the green areas created by the system and modify the procedure of delimiting the areas at every stage of data processing.

These indicators provide information necessary to fulfil the organization of the Public Services Monitoring System [7], providing local government entities, businesses, and the public with information necessary to evaluate services provided at the local level.

Results

For all case studies, the project will provide the stakeholder NSOs with outputs that can be directly used for reporting. Wherever possible, additional graphical user interfaces will be provided to allow users to investigate the statistics further, supporting understanding of the processes used to generate the output. All software and methods used are openly available for transparency. This is essential to build confidence in the quality and reliability of the statistics, prior to their adoption by stakeholders.

Conclusions

The case studies described above have shown how Earth Observation data can be used to meet key needs of National Statistical Agencies. Using the results of these case studies, the project will elaborate a future roadmap with recommendations for further developments integrating Earth Observation with Smart Statistics. This will take into consideration not just the technical issues remaining to be addressed, but also the operational and regulatory barriers to increased adoption of such products in official statistics. To help identify these barriers, the project is supported by a steering group on which key statistical agencies will be represented. This will also allow the project to exploit synergies with other initiatives in this area.

In conclusion, EO data has the potential to meet many key needs of the European statistical officers. By using key case studies to explore the practical barriers to increased adoption, the GAUSS project aims to define a pathway towards real operational use of such data in official statistics.

References

[1] https://eo4smartstats.com

[2] https://cdr.eionet.europa.eu/help/aqd

[3] https://panacea-ri.gr/?lang=en

[4] https://www.stat.fi/en/statistics/ilmatila

[5] *Remote Sens.* **2021**, *13*(21), 4456; **https://doi.org/10.3390/rs13214456**

[6] *BDOT10k. (2022). Retrieved 14 October 2022, from https://bdot10k.geoportal.gov.pl/*

[7] *SMUP. (2022). Retrieved 14 October 2022, from https://smup.gov.pl/ p.gov.pl*

Application of an Adaptive Survey Design on the Italian Population Census

Keywords: Adaptive designs, Response propensity, Target groups, Optimization.

Introduction

The aim of this paper is to present the results of the experimentation of the use of the responsive-adaptive design approach [1], [2] for the post-21 Population Census in Italy. Starting from October 2018, the Italian Population Census is carried out through an annual survey based on a sample of 2,850 municipalities and 1,500,000 households. The information is collected using a sequential mixed mode CAWI/CAPI, where the CAPI interviews are used as a follow-up of CAWI non-respondent households. The permanent census yearly provides estimates representing the entire population and all the 7,900 Italian municipalities by means of an indirect estimation process integrating information from both the sample surveys and administrative sources.

For the new post-21 cycle of the Permanent Census of Population and Housing (PCPH), budget cuts are expected with an impact on the household sample size each year. For this reason, the Italian National Statistical Institute (ISTAT) launched a project aimed at proposing and studying more efficient survey designs for PCPH, based on a reduction of the number of CAPI interviews which impacts most on the cost of the survey. The basic idea is to cope with the expected budget reduction by means of a survey design exploiting as much as possible the CAWI responses (about 45% in the previous surveys), considering that the budget could be not enough to interview all the CAWI non-respondents, but only a sub-sample of them. To this end, we test the Adaptive Surveys Design approach proposed by van Berkel et al. [3], intending to control non-response bias by optimizing response balance for subpopulations.

We evaluate the effectiveness of this design in an experimental setting by considering different scenarios defined by alternative sampling strategies and, moreover, by verifying the robustness of the response model adopted for identifying the target groups.

The general goals of the implementation of this method in this survey context are: 1) determining in advance the CAPI sampling fractions per group to select a random sample of CAWI non-respondent households, with the aim of improving quality of the estimates; 2) preserving as much as possible the same level of quality of the estimates already disseminated for the first cycle (2018-2021), paying off the reduction of CAPI sample size by increasing the CAWI sample size.

In the next paragraphs, we describe the chosen ASD approach, the implementation of the method and the experiments, and we present some results, conclusions and further development.

Adaptive survey design: implementation and simulation study

The chosen ASD approach

The basic idea of Adaptive Survey Design (ASD) approach proposed by van Berkel et al. [3] is to reduce the coefficient of variation of response propensities among the relevant target groups, CV_ρ, for reducing the non-response bias. A crucial role is played by the auxiliary variables used for identifying the target groups with different response propensities. A lower CV_ρ implies a smaller non-response bias on these variables, of course. Moreover, it implies a smaller non-response bias on survey variables before any weighting adjustment. CV_ρ is the quality indicator suggested for taking into account the solution of the optimisation problem. This indicator is estimated on the target groups in the population *N*. In each target group g $(g = 1, \dots, G)$, with population size N_g and $n_g = nN_g \; / \; N$ (i.e. proportional allocation), the total response probability is calculated assuming that all people have the same CAWI response probability $p_{cawi,g}$, the same probability $p_{elig,g}$ of being eligible for CAPI follow-up and the same CAPI response probability $p_{capi,g}$. Then, the total response probability in the group g is

$$p_g = p_{cawi,g}{+}p_{elig,g} f_{capi,g} p_{capi,g}$$

where $f_{capi,g}$ is the CAPI sampling fraction in group g, the unknown quantity to be determined. The CAPI sampling fraction in each group is determined by performing an optimisation problem that aims to minimise CV_ρ under specific constraints. Starting from p_g, it is possible to estimate the mean response propensity, the population variance of the response propensities and therefore CV_ρ.

Implementation steps

The van Berkel et al. [3] approach has been applied on the 2018 PCPH survey data enriched by linking additional variables from administrative sources, considered as a population. The first step for ASD implementation is the definition of the response model and target groups. To understand both which variables influence the most, and which units can be considered similar to others in terms of a computer-assisted response propensity, CAWI response models have been studied through a logit with household covariates: type of household, citizenship, highest educational level in the household, region (NUTS 2), type of municipalities. The logit model has worthy goodness of fit, with almost all the regression coefficients significantly higher than zero and a concordance ratio of 68.7%. To generate homogeneous groups in terms of CAWI response, a classification tree (CART) was used considering as auxiliary variables the highest educational level in the household, citizenship (Italian or not Italian) and region (NUTS 2). As the CAWI response rates vary widely by geographical area (North, Centre and South of Italy), the CART subgroups (5 final nodes) are crossed with the geographical areas for defining the final 12 target groups.

The second step is the implementation of the optimization problem. The CAPI sampling fractions, $\boldsymbol{f_{capi,g}}$ $(\boldsymbol{g = 1, \dots, 12})$, are obtained as the result of the optimisation problem that minimize the $\boldsymbol{CV_\rho}$ under the budget constraints - the maximum overall sample size $(\boldsymbol{n})$ and the overall number of CAPI interviews $(\boldsymbol{n_{capi}})$ and the minimum number of respondents $(\boldsymbol{r_{tot}})$. The objective function and the constraints are expressed as:

$$\min_{p_g}\left(\sqrt{\left[\sum_{g=1}^{G} p_g^2 \, F_g - \left[\sum_{g=1}^{G} p_g F_g\right]^2\right]}\Bigg/\sum_{g=1}^{G} p_g F_g\right) = CV_\rho$$

$$\begin{cases} n = n_{cawi} \\ n_{capi} = C \\ r_{tot} \geq R \end{cases}$$

The first constraint implies that, in the first attempt, all the sample is recruited for the CAWI mode. The second and the third constraints can be explicitly written as

$$n_{cawi}\sum_{g=1}^{G}F_g\left(1-p_{cawi,g}\right)p_{elig,g}f_{capi,g}p_{capi,g}=C$$

$$n_{cawi}\sum_{g=1}^{G}\left[F_g\,p_{cawi,g}+F_g\left(1-p_{cawi,g}\right)p_{elig,g}f_{capi,g}p_{capi,g}\right]\geq R$$

where $\boldsymbol{n_{cawi,g}=n\,F_g}$ is the sample size in the group $\boldsymbol{g}$.

The Italian population counts around 60 million individuals and 25 million households, the sample size for the list component of the PCPH is set to around 1 million households ($\boldsymbol{n=n_{cawi}}$ =1,082,340). Furthermore, the overall number of CAPI interviews is around 250,000 ($\boldsymbol{n_{capi}=C}$ =271,000), while the minimum number of respondent should be greater than 650,000 ($\boldsymbol{r_{tot}}\geq\boldsymbol{R}$ =650,000). The optimization problem is solved using the R-package Alabama [4].

Under the ASD approach, the target groups with higher CAWI non-response rates have a higher CAPI sampling fraction. Moreover, by definition, the total response rates of the target groups are balanced and close to one another. In fact, the $\boldsymbol{CV_p}$ is approximatively 0 under the optimal solution, while is equal to 0.103 when a constant CAPI sampling fraction is considered. Then, under the ASD approach, a lower non-response bias can be expected.

Simulation study

To assess the impact of the implemented approach on some of the main estimates produced by the survey, a simulation study is carried out. These simulations, evaluated first the trade-off between bias and sampling error under different ASD scenarios and estimators (direct and calibrated) and, then, under different model specifications.

In both cases, a Monte Carlo simulation (R=500 replications) is performed. The aim of the simulation is to shed light on the sampling errors of the obtained estimates too, besides their non-response bias. To handle a lighter simulation, the actual sample size of the PCPH (1,082,340 households) was reduced to 10,082. This does not impact on the optimal CAPI sampling fractions (f_{capi} optimal) that remain the same. Furthermore, to make easier the comparisons in all the scenarios the number of the respondent is always equal to 7,962.

In the first simulation study, four different sampling strategies have been compared: i. ASD with constant f_{capi} ; ii. ASD with optimal f_{capi}; iii. only CAWI interviews; iv. no target groups with constant f_{capi}. Under strategies i.-iii., a stratified one-stage sample is drawn from the 2018 PCHP theoretical sample list. This sample coincides with the CAWI component of the survey. The strata are the target groups crossed with the provinces and SR/NSR municipalities (SR=1,0) in the original PCHP sample design. Then, respondents are identified by applying the observed CAWI response rates. For strategies i. and ii., non-respondents and eligible households are stratified using the same strata, that is considering the target groups. However, just a fraction of them is contacted for the CAPI interviews based on the constant or optimal CAPI sampling fractions respectively (Table 2 - f_{capi} constant and f_{capi} optimal). Finally, the respondents are identified by applying, in both cases, the CAPI response rates (p_{capi}). Under scenario iii., the survey is composed of only CAWI interviews. Then, for equalizing the comparisons and obtaining the same number of respondents for the other scenarios, larger sample size is considered. Finally, under the strategy iv., the CAWI non-respondents and eligible households are stratified by provinces and SR/NSR municipalities in the original PCHP sample design. So, in this case, the target group are not considered. Then, the constant CAPI sampling fractions were applied in each stratum (f_{capi} constant).

In the second simulation study, the results obtained using different response models are compared to understand what happens when the auxiliary variables that explain the response mechanism are not available or the true model is not defined. Three different CAWI response models are simulated to assess the performance of the ASD approach in the empirical context and to test the robustness of the results to misspecification of the model. The CAWI response model considered are: i. Reference model described above is the true model through which the CAWI response is generated; ii. Model considering only the five Italian geographical areas; iii. No groups, the case in which no auxiliary information is available.

Main results and conclusions

For the sake of brevity, only the most interesting results and those related to the estimates of the unemployment rate are reported. Comparing the different scenarios and the two estimators, it emerges that the optimal CAPI sampling fraction provides always less biased estimates, at least for domains above the NUTS1 level. Under the NUTS2 levels, the relative bias is similar to the case in which the constant CAPI sampling fraction is used. Instead, concerning the other two scenarios it is always more convenient. When using the calibration estimator, the relative bias is mitigated and the values are closer among the four scenarios. Those exceptions increase for the estimates at NUTS3 and for unplanned domains (SAE), since auxiliary variables for that domains were not available in the CAWI-response model. In terms of NRMSE, for the optimal CAPI sampling fraction it is slightly higher, due to its more variability in sampling weights. Once again, calibration works for mitigating the differences among the scenarios.

Identifying the true response model, and of course, the true target groups, is always possible to reduce the bias of the estimates even up to the point of completely abating it. This happens for larger domains and for domains above the level at which the model is built (in this case regional). For domains below that level the bias is not null but still small (in this case regional). For domains below the bias is not null but still small. Considering the other two response models, it is more difficult to face the bias due to the lack of auxiliary information. However, the trend of the bias is the same. It increases as the estimation domain is small. Anyway, when using the calibration estimator, the results improve under all the scenarios and they are closer to each other. It is known, in fact, that calibration is a simple way of introducing auxiliary information in the estimates. Therefore, useful auxiliary information not included for defining the response mode can be added at the estimation stage. In terms of NRMSE, the reference model does not have performances so distant from other models (ii., iii.). This is due to the higher variability in sampling weights imposed by using a model that defines a larger number of target groups. However, this is not so serious and, anyway, the ASD approach aims to reduce the bias and not the sampling variance. Also in this case, calibration works for mitigating the differences among the scenarios.

Further studies are needed to extend the application of this method on a more complex setting, closer to the real case framework with two-stage sampling selection and a minimum number of CAPI interviews to be assigned to each selected municipality. It is important also to verify the properties of the approach even in a multipurpose survey as the PCPH is. Another aspect is the maintenance of the accuracy of the direct estimates at the municipality level, as ASD can act in the preservation of the quality of the indirect estimates at the municipality level.

References

[1] J.M.Brick, R.Tourangeau, “Responsive Survey Designs for Reducing Nonresponse Bias”. *J. Off. Stat*. vol. 33, pp. 735-752, 2017.

[2] R.M.Groves, S.G. Heeringa, “Responsive design for household surveys: tools for actively controlling survey errors and costs”. *J. R. Stat. Soc. Ser. A Stat. Soc*. vol. 169, pp. 439--457, 2006.

[3] K.van Berkel, S.van der Doef, B. Schouten, “Implementing Adaptive Survey Design with an Application to the Dutch Health Survey”. *J. Off. Stat.* vol. 36, pp. 609-629, 2020.

[4] R. Varadhan, Alabama: “Constrained Nonlinear Optimization”. R package version 2015. 3-1 https://CRAN.R-project.org/package=alabama, 2015.

A New Way to Learn Data Science: The Funathon, a Non-competitive Hackathon

Keywords: training, hackathon, data-science, official statistics, innovation

Introduction

Data science innovation for statistical production and studies is a crucial activity for maintaining high quality, meaningful and relevant statistics, and is pledge of productivity gains. In 2018, our national statistical institute (NSI) set up an innovation lab in data science in order to monitor and disseminate innovative statistical methods and explore new data sources through experimental projects carried out in partnership with other units of the official statistics service. One of the key missions of this innovation lab is to disseminate data science methods, promote their use and train statisticians and data scientists in various bodies of the national official statistical system. Indeed, the innovation lab explores new methodologies or sets up some experimental projects, but the ones who know exactly what is needed by their unit, and who will in fine be in charge of the generalisation of the experimentations are agents belonging to the production units. It is therefore crucial that as many agents as possible are trained or at least familiarised with data science techniques.

The innovation lab, composed of nine data scientists, used to set up some traditional classroom training sessions: during one to three days, a trainer teaches to a handful of trainees some data science or machine learning methods. This training format has proven to have some inefficiencies as regards with the learning of data science: trainees do not practice enough after the course and quickly forget what they have learned. Moreover, our NSI employs a dozen hundreds of statisticians and classroom trainings do not easily scale up. Lastly, due to the Covid-19 pandemic, classroom trainings have been suspended.

To deal with this situation, a threefold strategy has been experimented for data science training: create introductive notebooks and make them available for e-learning on a data science infrastructure, with the support of trainers who exchange regularly with the trainees; provide online master class for advanced topics; lastly, organize a non-competitive hackathon named “Funathon” to make sure everyone can become familiar with data science methods, at least once a year.

This abstract first describes the objectives of the Funathon, the organisational framework and the methodological choices we made; then it goes through the feedback and lessons learned during the first two sessions that happened in 2021 and 2022. Lastly, it concludes with the areas for improvement envisaged for the following sessions.

What is the Funathon?

Goals and targeted audience of the Funathon

The Funathon is a virtual two-days event whose aim is to be a training session. Its objective is to give participants the opportunity to train and improve their data science skills. Every employee of the NSI or other national statistical authorities can participate. The only prerequisite is to be familiar with R or Python. It is almost identical to a virtual hackathon except there is no competition among the teams. The format of a hackathon has many advantages. Such an event creates emulation among participants, encourage them to be innovative and enables team work because some goals have to be achieved in a short period of time.

Unlike a usual hackathon, we have chosen to organize a non-competitive event because we have observed that many of our colleagues consider themselves not skilled enough to participate to a hackathon: statisticians working in our NSI have different professional backgrounds and positions and most of them do not feel confident enough to use new data science techniques and methods. Conversely, the non-competitive dimension of the Funathon appears to be an opportunity to learn data science without any stake, with a good team spirit, many training materials available to the participants and a continuous support from experimented data scientists.

2.2 The framework and the methodological choices

The Funathon has been conducted twice in 2021 and 2022. Two preparatory meetings were organized for the participants: one to present the concept of the Funathon and the proposed subjects, the other to learn how the data science platform works. Participating to the event was administratively registered as an official training action, which allowed agents to take time out from their usual professional obligations and officially dedicate two days to the subject. A large and anticipated communication allowed to disseminate information in all units. For each edition, one specific topic had been chosen: global warming in 2022 and Airbnb data in 2021. The organizing team prepared a few subjects which allowed the issue to be approached from a variety of angles and levels of difficulty. For each subject, two notebooks (in Python and R) gave detailed information on how to start, and ideas on how to go further.

The Funathon in itself lasted two days. The participants were invited to use our private cloud data science platform, GitHub or GitLab to share code and work on a collaborative way. It was a virtual event, but teams were usually in the same room. The organizing team maintained a permanent contact with teams using Zoom and an instant messaging system to support all the participants. To add some training content, the 2022 edition included three masterclasses. During one hour, one of the organizers presented a general data science topic which can be of interest for all participants (“use of ElasticSearch”, “use of satellite imagery” and “why learn Python when you already know R?”)

At the end of the Funathon, a moment of sharing allowed every team to show what they did if they wished to.

Feedback of the two first editions

Illustration with the subjects of the 2022 edition

The theme of the 2022 edition was environment and climate change. The nine subjects were about: textual analysis of responses to a national poll collecting citizens' ideas on climate change initiatives, use of FastText to classify answers to this national poll, collection and use of Twitter data using ElasticSearch, characterisation, localisation and mapping of dwellings with good energy performance, attribution of a business registration number to polluting companies using ElasticSearch, prevision of pollution based on meteorological data using machine learning models, simulation of the impact of rising water levels, data storytelling about land use thanks to Quarto, land use detection thanks to the analysis of satellite imagery, detection of land cover (or bird species) by convolutional networks.

Success and difficulties

In both editions, 150 persons participated, grouped into some thirty teams. According to the satisfaction survey, 96 % were happy and 93 % wanted to participate again next year.

According to the participants, this allowed them to demystify some data science methods and better understand their interest and the diversity of uses. One of them explained for instance: "we have progressed collectively, lastingly, in the techniques of manipulation, control and production of data, but also in our ways of doing things. The work organisation methods using the new tools and sources are very effective: sharing code with several people using Git, by small iterative touches, numerous round-trips when producing statistics on large volumes of data."

According to the organizing team, this success is based mainly on four points: (I) flexibility in registration, while also being an official training session leaves time to participants to recruit other colleagues by word of mouth and constitute teams with various profiles and competencies. (II) A lot of communication is crucial, as most people are afraid to participate (hence the name). (III) The main topic of the Funathon seems to be important, at least at first, for motivating participants. (IV) Keeping a constant contact during the event allows agents from all initial technical levels to improve their skills and to have the satisfaction of seeing a concrete outcome to their work. This assistance constantly available is also very reassuring for participants.

The main difficulties were: (I) the use of the data science platform and mainly the use of Git were difficult for some participants. Onboarding on these two points is very important. (II) For the organizers, it is a challenge to get right the difficulties of subjects, between the need to avoid that participants just have to follow the script, and the need that the Funathon remains accessible to a beginner. (III) The production of the notebook and the preparatory work represents a large workload that must be anticipated.

Conclusions

The success of the Funathon owes much to the collaboration between multiple entities of official statistical services. The unit in charge of agents’ training was a great help for communicating about the event, encouraging agents to dare to participate and officialize their participation as part of their training plan. Lastly, three different units proposed some subjects in the 2022 Funathon, which contributed to the diversity and the richness of the topics and methods covered.

As ways of improvement for the next editions, the audience of the Funathon could still be expanded and the support could be even more personalized. Cooperating with other statistical institutes or international organizations concerned with data science issues would also broaden and enrich the scope of ideas of subjects, masterclasses and pedagogical material.

Visualising the pathways in and out of employment during COVID-19 in Estonia

Keywords: Labour market mobility, COVID-19, sequence analysis, trajectories

Introduction

The coronavirus crisis was a major economic challenge for all European countries, including Estonia. The restrictions, closures and social distancing rules meant that many businesses were closed either temporarily or permanently [1]. Especially during the first wave of COVID-19 employment fell most dramatically in accommodation and food services activities [2].

The situation immediately raised the question what happened to these workers. Did they remain unemployed, or did they find jobs in other sectors? Did it make more sense to wait until the crisis ends and go back to the original sector?

The aim of the current paper is to look at possibilities to visualise and model the pathways in and out of employment on the example of the accommodation and food services sector.

Methods

Data

For the current analyses data from the Estonian Employment Register was used. We built a database that indicated the main working relationship (as people may work for multiple enterprises at the same time) in the end of each month from February 2020 to August 2022. We analysed only persons whose main job was in the sector of accommodation and food services as of 29. February 2020. We determined whether they were employed (had a record in the register) or not.

As visualisation and modelling techniques we used group-based trajectory modelling and sequence analysis.

Group-based trajectory modelling

Group-based trajectory modelling is a statistical methodology that allows to analyse the evolution of an outcome over time [3]. These models can be seen as a specialised application finite mixture models [4]. In current analyses a logit link for estimating the probability of being employed in the end of the months. We only use data of individuals who had at least one status change in the period from March 2020 to August 2022. This leads us to having 15547 observations.

Sequence analysis

Sequence analysis is a technique that allows to analyse and visualise individual categorical states in sequences [5]. Additionally, the individual sequences may be compared and distances between persons may be calculated. On top of it usually clustering is used to identify similar sequences. Here we also use data of individuals who had at least one status change in the period. For clustering Ward method is used.

Results

To look at pathways out of employment and back to work we used the two modelling techniques described above. The results are presented first for the group-based trajectory models and then for the sequence analysis model. Our main aim is to compare the visualisations provided by both models.

Group-based trajectory modelling

The group-based trajectory modelling technique provided several solutions with fit statistics (for example AIC, BIC) on about the same level. The solution that was most easily interpretable had 7 trajectories starting from March 2020 and ending in August 2022. Still, the presented trajectories were similar in many of the solutions.

Three of the seven trajectories show people who lost their jobs during the first wave of covid. One of these trajectories (No 7 on the Figure 1) consisted of persons who remained unemployed. Trajectory no 6 shows persons who bounced back quickly and almost all of them were employed again in a year. Trajectory no 2 shows persons who were between these two groups: more than half of them were employed again in the end of the period (August 2022).

The trajectories no 1 and 5 included persons, who kept their jobs in spring 2020, but lost them later: either in winter-spring 2021, when a new and bigger COVID-19 wave hit or autumn 2021. They remained unemployed for at least a year, some of them went back to work in spring-summer 2022.

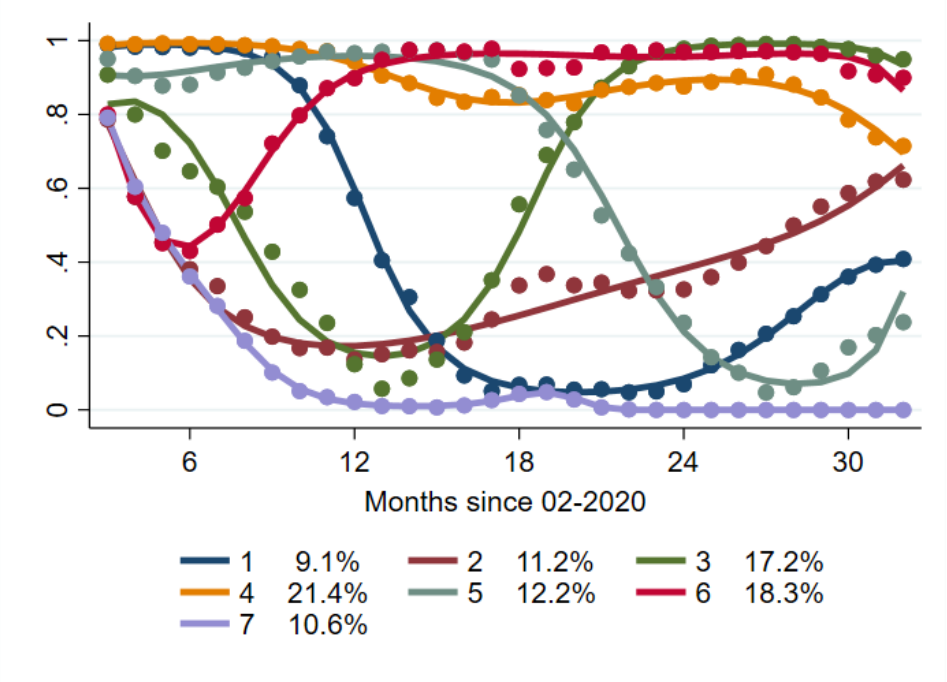


Figure 4. Proportion of working during the months following February 2020 by trajectory

It is also possible to use aforementioned trajectories in models both as dependent as well as independent variables.

Sequence analysis

The sequence analysis technique allows us to visualise either the working-nonworking states but also other characteristics of the working relation, like the sector or occupation. The solutions with binary outcome (0 not working and 1 working) clustered to 7 groups is shown on Figure 2. Purple bars show working episodes, grey bars unemployment. Y-axis shows single persons.

First of the clusters consists of people who lost their jobs and mainly did not come back to employment. Third and fourth clusters also have people who lost their jobs during the first wave of COVID-19 but returned sooner (in 3-6 months, cluster 4) or later (in a year, cluster 3). Clusters 5 to 7 show persons who lost the jobs at a slower pace and stayed unemployed or joined the workforce again in about a year (cluster 6). The biggest cluster (no 2) is made of persons who found new jobs almost immediately after the job loss. This cluster also clarifies the situation with the trajectory 4 where at any time most of the people are employed.

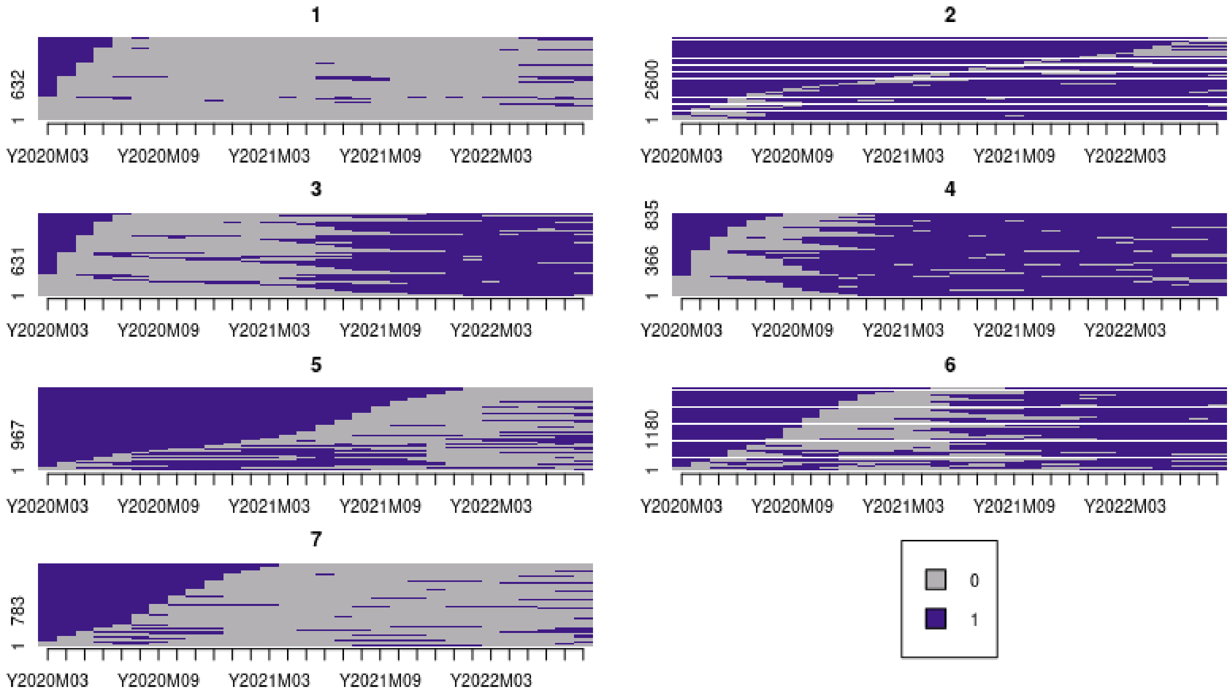


Figure 2. Individual trajectories of working (1) and not working (0) in 7 clusters from March 2020 to August 2022.

In addition to the results presented above it is also possible to look at different sectors where the persons work (not shown as graphs). Generally, persons working in the accommodation and food services sector prefer to also find their new jobs in the same sector. When they change sectors, they more often go to retail (3% were working in this sector in six months' time) or to other service sectors.

Conclusions

Both, trajectory analysis as well as sequence analysis can be used to visualise employment pathways. With the technique of group-based trajectory analysis an overall picture of groups of similarly acting people can be shown. These are aggregated from all the persons in the group.

With the technique of sequence analysis individual trajectories can be analysed. For a better understanding of underlying principles, the trajectories can also be grouped to clusters. At the same time the visualisation of sequences is more complex than with the trajectory analysis and not so easy to follow.

References

[10] Eurofound and European Commission Joint Research Centre (2021), What just happened? COVID-19 lockdowns and change in the labour market, Publications Office of the European Union, Luxembourg.

[11] Eesti Pank (2020), Labour Market Review 1/2020, https://www.eestipank.ee/en/publication/labour-market-review/2020/labour-market-review-12020.

[12] Nagin, D.S. (2005), Group-Based Modeling of Development. Cambridge, Harvard University Press.

[13] Nagin, D.S. (2014), Group-Based Trajectory Modeling: An Overview. Annals of Nutrition and Metabolism, 65, 205-210.

[14] Gabadinho, A., G. Ritschard, N.S. Müller and M. Studer (2011), Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software, 40(4), 1-37.

Web scraping (GASP 1A.1)

Session Chair: **Fernando Reis** *(Eurostat)*

Web scraped data in consumer price indexes
Peter Knížat *(Statistical Office of the Slovak Republic)*

Use of data obtained by web scraping for statistical purposes
Branka Raicevic *(Statistical Office of Montenegro)*

Can real estate web data augment official statistics?
Klaudia Peszat (Statistics Poland)*; Dominik Dabrowski (Statistics Poland); Dominika Nowak (Statistics Poland)

Web scraped data in consumer price indexes

Keywords: data web scraping, data processing, consumer price index, Jevons, GEKS-Jevons, Hedonic regression, Time-product dummy regression.

Introduction

The changes in the consumer behaviour, a consumer prefers to purchase some products through internet, calls for revisiting of the traditional data collection and integration of the automated online data collection, also called data web scraping, of products' prices in the consumer price index methodology by National Statistical Institutes (NSIs).

In this study, we outline the procedure for processing web scraped data, various methods for calculating consumer price indexes and its application on the real data.

Methods

In this section, we show the procedure for processing web scraped data and the consumer price index formulae, respectively.

Data Processing

We observe prices $[p_i^d]_r$ for product items *i = 1, …, N* in daily time periods *d = 0, …, D* that can be sold by different retailers *r = 1, …, R*. Nowadays, the online purchases by consumers are carried out usually through a web platform that compares the product items' offers from different retailers.

In the first step, we have to determine a single price of individual product items for each day. Table 1 lists the options.

Table 3. Options for determining the daily price of individual product items

Option	Formula	Narrative
Minimum Value	$p_i^d = min\left([p_i^d]_r\right)$	A consumer selects the cheapest price offer.
Arithmetic Average	$p_i^d = \frac{1}{R}\sum_{r=1}^{R}[p_i^d]_r$	All given price offers are included. Skewed towards outlier values.
Geometric Average	$p_i^d = \prod_{r=1}^{R}\left([p_i^d]_r\right)^{\frac{1}{R}}$	All given price offers are included. It balances the effect of outlier values.
Median Value	$p_i^d = \text{midpoint}\left([p_i^d]_r\right)$	Excludes the effect of all other price offers.

In the next step, we have to determine a single price of individual product items for each month *t = 1, …T*, that is, the monthly product item's price, p_i^t. Here, we average the daily time prices for each product item in a given month *t*. Similarly, we can use arithmetic or geometric average, or median value.

Consumer Price Index

The web scraped data contains the prices of product items with no information on the product items' sales or quantities sold. Hence, we can only choose indexes which belong to a family of the unweighted.

The fixed-base Jevons index, with a sample of product items defined in the base period *0*, can be defined as [1]

$$P_J^{0,t} = \prod_{i=1}^{N} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{N}}, \ \ t = 1, \dots, T \qquad (1)$$

In (1), we compare the prices of product items p_i^t in the current time period *t* with the prices of product items p_i^0 in the base period *0*.

We follow with defining the multilateral indexes that compare the prices of product items between more than two time periods simultaneously.

The GEKS-Jevons index, which uses the Jevons index as its elementary building block, can be defined as [1]

$$I_{GEKS-J}^{0,t} = \prod_{\substack{l=0 \\ l \neq t}}^{T} \left(P_J^{0,l} \times P_J^{l,t} \right)^{\frac{1}{T}}, \ \ t = 0, \dots, T \qquad (2)$$

In (2), *T* refers to the number of time periods and P_J's are the individual Jevons indexes calculated between two time periods.

We observe a number of character parameter variables z_{ik} with $k = 1, \dots, K$ for each product item *i*. Then, the hedonic regression model can be defined as [2]

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T} \partial^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t, \ \ t = 0, \dots, T \qquad (3)$$

In (3), D_i^t is a dummy variable that takes the value of *1* if the observation relates to the product item *i* observed in the time period *t* and *0* otherwise. The parameters ∂^0, ∂^t, and β_k are estimated by using the ordinary least squares method. The random errors ε_i^t of the regression model are assumed to be normally distributed with mean *0* and constant variance σ. There is no restriction of matched-item samples across time periods.

The hedonic index for each *t*, noting that a dummy variable for *t = 0* is omitted in the regression, is [2]

$$I_H = \exp\left(\hat{\partial}^t \right), \mathrm{t} = 1, \dots, \mathrm{T} \qquad (4)$$

In the case of missing character parameter variables, the time-product dummy regression model can be defined as [3]

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T} \partial^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t, \quad t = 0, \dots, T \qquad (5)$$

In (5), D_i is a dummy variable that takes the value of *1* if the observation relates to the product item *i* and *0* otherwise, a dummy for the arbitrary product item is not included in order to identify the model. Other parameters, its estimation, and variables are defined similarly as in (3).

The time-product dummy index for each *t*, noting that a dummy variable for *t = 0* is omitted in the regression, is [3]

$$I_{TPD} = \exp(\, \hat{\partial}^t), \mathrm{t} = 1, \dots, \mathrm{T} \qquad (6)$$

Results

In this section, we apply the theoretical framework from the previous section. We have been scraping the prices of the following product categories: washing machines, dryers and dishwashers.

Data Processing

First, we demonstrate the results of applying data processing steps on the real data. We select two product items from the product category washing machines. Its daily prices, as defined in Table 1, are displayed in Figure 1.

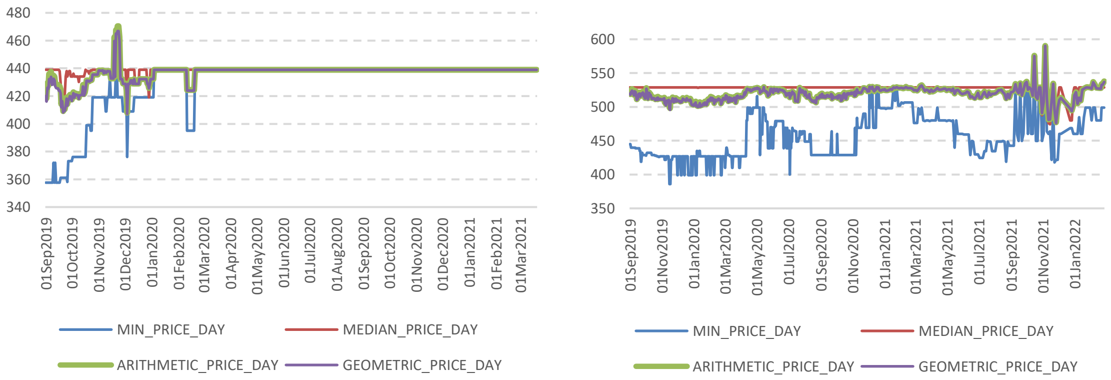


Figure 5. Product items – daily prices

Figure 1 shows that using different methods can lead to different daily prices. The first product item went to the clearance sale, however, its price remained constant before being sold out.

In the next step, we apply the averaging to create monthly prices, as described in 2.1.

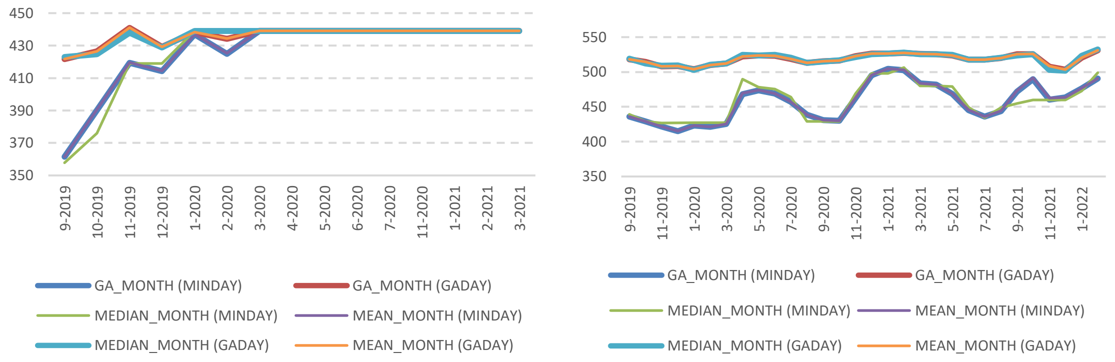


Figure 6. Product items – monthly prices

In Figure 2, for the daily price, we selected minimum and geometric average for comparison. Similar trends are shown in other product items' prices.

A careful consideration must be given for determining the choice of daily and monthly *averages* since it could lead to fundamentally different results.

Consumer Price Indexes

Next, we calculate the consumer price indexes for product categories washing machines and dryers, respectively, using a 13-month time window and select geometric averages for both day and month prices.

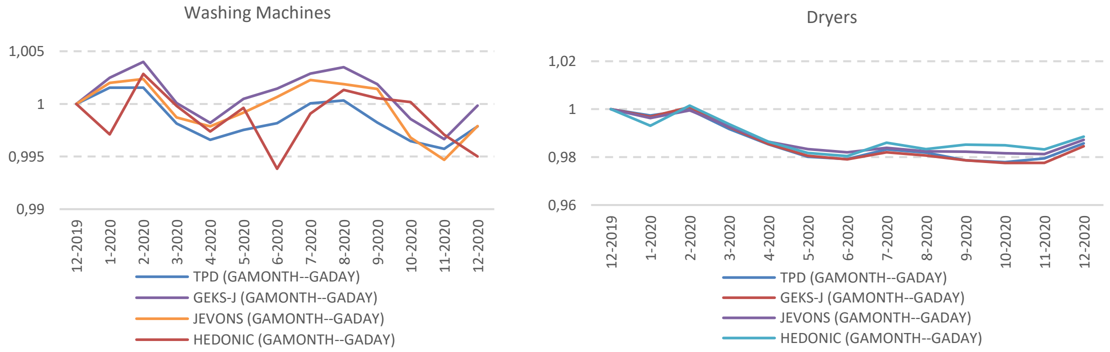


Figure 7. Consumer Price Indexes

The indexes for washing machines are more volatile that could be a result of higher attrition rate of product items in monthly samples. The hedonic index shows an irregular behaviour, which could be caused by a large amount of missing data in the product items' character variables.

Conclusions

The aim of this study is to demonstrate a complexity of product items' prices that are scraped daily from the websites. The processing of web scraped price data might lead to different, sometimes contradictory, resulting consumer price indexes.

Much work is required before the web scraped data can be implemented within the official price statistics. The recently published guide [1] deals with transaction data, which are of simpler structure, consists of sales and quantities sold for each product item in a given time period, and are normally processed for a particular retailer.

References

[15] Eurostat, Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Manuals and Guidelines (2022), Luxembourg: Publication Office of the European Union, ISBN 978-92-76-44354-4.

[16] J. de Haan, Rolling year time dummy indexes and the choice of splicing method, Unpublished draft, 29 April 2015.

[17] J. de Haan and R. Hendriks, Online data, fixed effects and the construction of high-frequency price indexes, Paper presented at the 2013 Economic Measurement Group, Sydney, Australia, November 2013.

Use of data obtained by web scraping for statistical purposes

Keywords: web sraping, collection if data, Python

Introduction

Web scraping is the automatic collection of data from the Internet. This method of data collection is increasingly used by national statistical institutes to reduce response burden, to speed up the data collection process, to obtain new indicators, to examine variables and population characteristics.

These days we use the internet for business, education, shopping, sports, travel, fun and many other tasks. Web become a big data base knowledge, information about economy and life in general.

By using web scraping tools from the Internet, it is possible to get more complete data. The periodicity of data collection by web scraping is less than through manual search. The resulting data set contains much more information than a manual site search.

A web data source is a data source that we can use by querying over Internet protocols. Web scraping is the process of automatically downloading data from the Internet without human intervention. The program that makes this possible is called "scraper".

Methods

To access the websites, the following conditions must be met:

- Legal framework for the use of data obtained by web scraping
- Adherence to certain principles when accessing the website – Netiqutte

The process of collecting data from the Internet involves the following steps

1. *Analysis of potential websites*

The statistician responsible for the survay for which the use of data obtained by web scraping is planned, analyzes all potential web sites from which data could be collected. The analysis should determine whether the website provides all the necessary information for statistical research, their reliability and quality.

2. *Analysis of the selected website in technical terms*

This type of analysis is the responsibility of the person who will use the web scraping tool to collect the required data from the website. In this phase of the analysis of a website or several websites, the technical characteristics are examined from the point of view of "scraping" and data availability.

3. *Communication with the website owner*

It is not necessary to inform the website owner about the use of data during the testing of web scraping tools and analysis of scrapped data. When starting to use the data obtained by web scraping for the production of official statistics, it is necessary to organize a meeting with the owner of the website and inform him that the data from his website will be used for statistical purposes.

4. *Collecting data from the Internet*

There are two completely different ways to collect data from the Internet:

- Collection of several items from one or two websites (eg collection of product names, product descriptions and prices of technical equipment from the website of one retail chain);
- Collecting one item from multiple different websites (e.g. collecting airline ticket prices from multiple different sites).

5. *Input control of data collected from the Internet*

The control of data obtained by web scraping is performed before processing them for the monthly consumer price index. The quality of the collected data is controlled: file size, number of prices obtained, number of categories of goods collected, very high and very low prices of goods, duplicate control and the like.

6. *Recoding data obtained by web scraping*

Web scraping data usually contains the product name and price (e.g. statistical survey Consumer price index). All products from the data set obtained by web scraping must be assigned an appropriate code in accordance with the COICOP classification. When a set of data is obtained for the first time using web scraping, data analysis is approached. The product name is used to associate the product name from the scrapped data set with the COICOP classification. The connection is made at the level of 6 digits of the COICOP classification.

Results

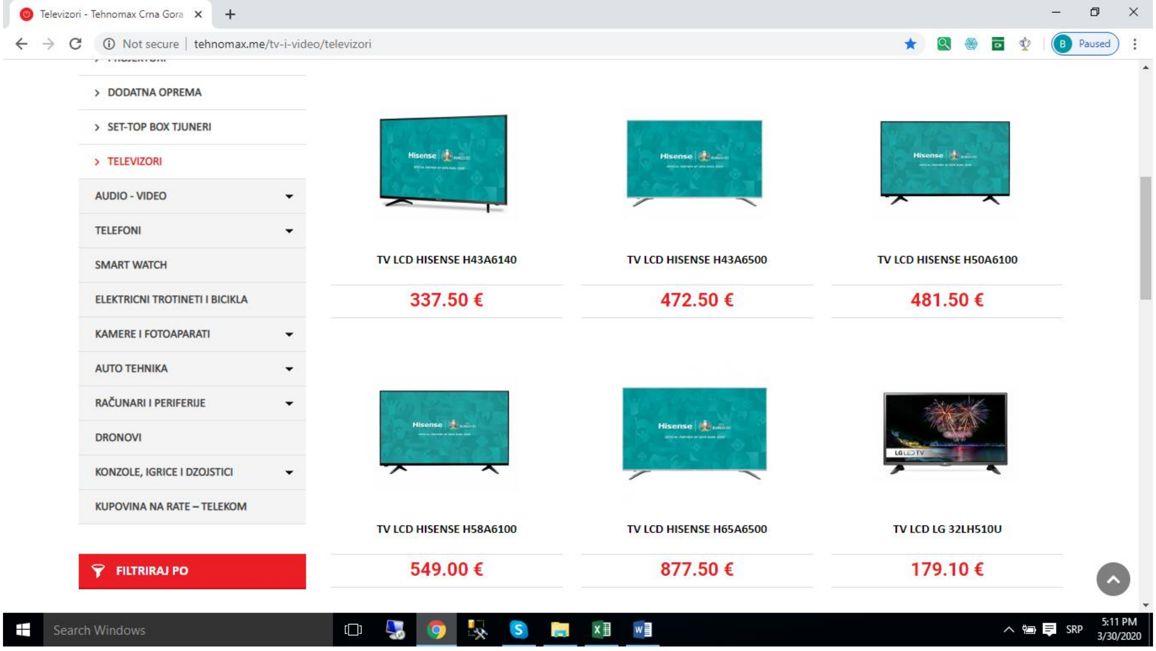
IT tool for web scraping

Web scraping is a technique for automatically accessing and collecting large amounts of data - information from the site, which can save a huge amount of time and effort to collect data.

Beautiful Soup is a Python library for collecting data from HTML and XML files. The Python programming language is open source, does not require the purchase of a license, is installed free of charge and is used by the National Statistical Institutes to collect data from the Internet.

The following is an example of the layout of the Tehnomax website, which is a potential website for collecting data on the prices of technical goods, as well as an excel spreadsheet that is the result of executing web scraping code written in the Python programming language.

Figure 1 shows the website of Tehnomax, which lists the prices of TVs sold by this company.



TV LCD HISENSE H43A6140 337.50 €

TV LCD HISENSE H43A6500 472.50 €

TV LCD HISENSE H50A6100 481.50 €

TV LCD HISENSE H58A6100 549.00 €

TV LCD HISENSE H65A6500 877.50 €

TV LCD LG 32LH510U 179.10 €

Figure 1

From this website it is necessary to scrape the information about the name and price of the TV.

In order to do this, it is first necessary to analyze the background of this page, ie the code that is in the background of this website. Based on the analysis of this code, the code is written in the Python programming language for scraping the page where the data on TV prices are located.

The table in Table 1 represents an excel file that is the result of executing code in the Python programming language.

Name	price
TV LCD HISENSE H43A6140	337.50 €
TV LCD HISENSE H43A6500	472.50 €
TV LCD HISENSE H50A6100	481.50 €
TV LCD HISENSE H58A6100	549.00 €
TV LCD HISENSE H65A6500	877.50 €
TV LCD LG 32LH510U	179.10 €

Table 1

Data from the excel file can be imported into the MS SQL Server database and and can be used for further processing.

Conclusions

Statistical institutes that use web scraping to collect data compared data collected in the traditional way and data collected from the Internet.

The price index produced on the basis of data from the Internet is lower, because the seasonal pattern of data from the Internet is more even than the price index produced with data collected in stores.

Comparing both methods shows that the difference in prices collected at the same time is not large. It can be said that the trend is the same.

The data collected using web scraping are more complete and of better quality. The workload of employees on data collection is reduced and employees can do more data analysis and thus improve quality. Using web scraping reduces the cost of research, shortens the time of collection and processing. It is important to point out that in this way the workload of reporting units is reduced.

References

[1] ten Bosch, O., Windmeijer, D., *On the Use of Internet Robots for official Statistics,* UNECE MSIS conference, Dublin, 2014

[2] ten Bosch, O., Windmeijer, D., van Delden, A., and van den Heuvel, G., Web scraping meets survey design: combining forces, BIGSURV18 conference, Barcelona, 2018

[3] Daas, P. and J. Burger (2015). *Profiling big data to assess their selectivity*. Paper presented at the conference New Techniques and Technologies for Statistics (NTTS), Brussels 2015.

[4] Stateva G., ten Bosch O., Maslankowski J., Righi A., Sannapieco M., Greenaway M., Swier N., Jansson I., *Legas aspects related to Web scraping of Enterprise Web Sites,* ESSnet Big Data Work Package 2, 2017.

Can real estate web data augment official statistics?

Keywords: web data, internet data, data quality, Web Intelligence Hub, Web Intelligence Network, real estate

Introduction

Internet data have demonstrated high potential in the modernization of the official statistical production. The experiences of many National Statistical Institutes (NSIs) and European endeavours (such as ESSnet Big Data I and II projects) have provided evidence of the relevant role of this data sources in producing new and augmenting existing statistics. In order to enhance broader application of web data in the regular production, Eurostat embarked on the development of the European platform - Web Intelligence Hub (WIH), which is expected to provide an architecture and services for web data collection, processing and analysis in compliance with the quality requirements of official statistics. The WIH is being elaborated in a participatory manner with the consortium of European NSIs under the ESSnet Web Intelligence Network (WIN) project[7].

The WIN looks into new domains of web data in order to explore the possibility of their future integration with the WIH. One of these domains is the data on real estate advertisements, with a particular focus on their application to monitor real estate market[8]. The attempts to use the data from real estate portals to augment real estate statistics appear promising [1][2], with the acknowledgment that web data not going to substitute traditional data sources, but complement them.

The paper presents the approach taken by Statistics Poland to the use of web data in real estate statistics. The goal is to verify the potential of web data to foresee the trends and market shifts resulting from, *inter alia*, crisis events (such as the COVID-19 pandemic or the outbreak of the war in the Ukraine) in a timelier manner, as well as to augment the real estate statistics by more granular data (on the lowest possible territorial level) and provide new indicators. The latter refers to the qualitative information, such as availability of parking space, security and other amenities, in order to measure the standard of the available real estate and provide valuable information for the hedonic models of prices and indices of quality of life.

New data sources, including web data, differ significantly from the traditional collection and processing methods. Therefore methodological and quality issues are of particularly importance. The article discusses the selected methodological challenges related to the application of web data sources in the real estate statistics.

[7] https://ec.europa.eu/eurostat/cros/WIN_en

[8] https://ec.europa.eu/eurostat/cros/content/issue-6-exploring-potential-new-data-source-real-estate_en

Methods

The work on web data on real estate began with the exploration and assessment of new data sources. To this end, the criteria of the selection of web pages were listed, based on the *Minimal guidelines and recommendations for implementation of web scraped data* [3], literature review and experiences of other NSIs in this fields.

Mainly the criteria covered:

- the level of completeness of the information provided publicly;
- evaluation of the type of data download availability;
- mechanisms of navigating on the page and page's structure.

As a result, real estate portals with country-wide coverage were selected as most relevant. Initially, ten portals with advertisements for sale and rent were reviewed, but due to the non-fulfilment of the criteria referring to data availability, timeliness and their usefulness (e.g. lack of mandatory variable, dynamic structure of the website etc.) only four were selected to further work. Methodological and data quality issues encountered during the exploration phase are briefly described below.

Population and coverage

One of the main challenges of the carried out work refers to the population of the observed objects. Compared to census or administrative registers, web data do not provide information about the entire population of real estate premises, but only those that are intended for sale or rent (with no guarantee of transaction completion). Therefore, under- or overcoverage of certain types of apartments is a common problem. So is the redundancy of the data resulting from the duplication of offers among different real estate portals, which can lead to significant representation errors.

What is more, web data as a source requires constant monitoring of its stability over time. It might be the case that an increase or decrease in the total number of advertisements or selected groups of apartments is not the result of structural changes in the market, but decision made by the portal owners (such as the closure of the activity, ownership link with another web page, limitation or extension of portal regional or domain scope) or its users (decline of interest in a given portal).

Existing traditional data sources can support the analyses of web data with the information on the structure of the housing stock. However, it should always be kept in mind that due to different target populations (transactions versus offers) and time-lag between the moment of data dissemination and the actual transactions, the structure of real estate data acquired from the web can differ significantly from the one of administrative or census one. Hence, the analysis of web data should be preceded by the quality evaluation, correcting for the representation errors and selection bias.

Quality of data

Real estate data acquired from the web bring multiple challenges regarding quality. First of all, statisticians do not have any control on the data and metadata presented on the portals, thus, there is a high probability of the errors in the datasets.

Additional difficulties appear when acquiring the data from more than one web page. In this case each dataset can differ in the categories of the variables obtained. For example, in one portal all the apartments with 8 and more rooms may be placed in the category „more than 5 rooms" while some portals may present the information as it is. Hence, the integration of multiple sources requires further adjustments of categories between the sources, as well as with official classifications, wherever possible.

Other common errors encountered in the web data which need to be taken care of, are:

- different values of data referring to the same information obtained from multiple page sections,
- the offer assigned to erroneous real estate category,
- multi-property offers.

The above-mentioned types of errors may be observed on the basis of micro-data exploration and detection of outliers. Preliminary analyses have demonstrated that most of such types of errors occur close to extreme observations, thus application of data cleaning methods dealing with outliers should bring good results.

Another type of errors that occur are incorrect data inserted by the offerors. They may be considered as random observational errors, as they are likely to be a result of a user's mistake or sometimes may be inserted on purpose to make the offer listed higher on the web page (e.g. by setting very low price or providing inaccurate information about the localization of the apartment). A useful approach to cope with such errors, is the application of predefined quality templates for each variable in a dataset, in order to track the error values and to standardize them. The quality templates may be prepared on the basis of the registers, e.g. by using distributions of particular variables on some level of aggregation (there is no possibility to match the observations one to one), or by analyzing the microdata of acquired datasets and setting the rules on its basis.

Results

In the process of data cleaning, the assumption was not to find the perfect match of the apartments in the corresponding register, but to create quality templates based on variable distributions to reduce the number of outliers that were incorrectly classified into a given set or have incorrect information inserted.

The distribution of the basic raw variables (price in PLN and area in m^2) acquired from the web data presented a high level of skewness (see Figure 1). This is due to the fact that the apartments offered on web portals strongly differ regionally, but also the maximum values were inserted wrongly by the offerors. Thus, a standardized data cleaning rules and procedures to deal with representation errors are to be employed. They will be followed by generation of the first experimental indicators for Poland.

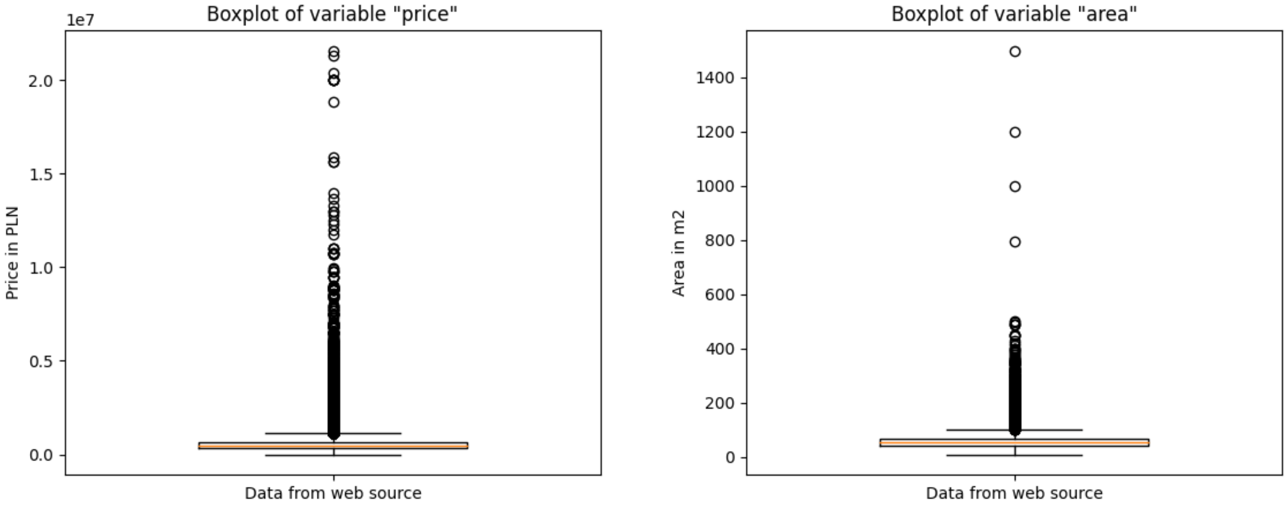


Figure 8. The examples of distribution of the basic raw variables from real estate portals

Conclusions

Web data appear to have potential to augment official real estate statistics. However, in-depth methodological work is required to ensure quality standards acceptable by official statistics. Concurrently, it should be stressed that statistics generated with web data are to play a different role than statistical information based on traditional sources. In the case of web data, it is often difficult to achieve the same quality parameters as in traditional statistics, but web data may well serve to generate flash estimates, be used to nowcasting, or predict new trends and unexpected event in a timelier manner than traditional data. The potential of web data for official statistics has been discerned by European official statistics with its practical emanation in the Web Intelligent Hub and Web Intelligence Network. In this vein, on-going work and joint efforts on the exploration of new domains are necessary, in order to stand up to move from the experimental statistics to official production.

References

[1] Beręsewicz M., 2016, *Internet data sources for real estate market statistics*, Doctoral dissertation, Poznań University of Economics and Business.

[2] Statistics Netherlands, *Indicatoren bestaande woningen in verkoop*, available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/indicatoren-bestaande-woningen-in-verkoop, accessed October 2022.

[3] *Minimal guidelines and recommendations for implementation*, available at: https://ec.europa.eu/eurostat/cros/system/files/deliverable_4_1_minimal_guidelines_and_recommendations_for_implementation_essnet_tss_win.pdf, accessed October 2022.

Innovative data collection in Social Statistics (JENK 1A.1)

Session Chair: **Eniel Ninka** *(Eurostat)*

What is the business case for a smart household budget survey? Experiences from the field in six countries
Barry Schouten *(Statistics Netherlands-CBS),* Jelmer de Groot *(Statistics Netherlands-CBS)*

CRŒSS: A cross-domain data collection platform for the ESS
Joeri Minnen *(hbits CV - spin-off VUB),* Kelly Sabbe *(STATBEL),* Jerome Olsen *(DESTATIS)*

What is the business case for a smart household budget survey? Experiences from the field in six countries

SUMMARY: Household budget surveys satisfy two of the three measurement criteria for smart surveys: they are time-consuming and cognitively burdensome, and they request information that people may not know or recall. For this reason, the introduction of smart features has been the topic of Eurostatfunded innovation projects since 2018. From a methodological viewpoint, the focus has been on effective recruitment strategies, the amount of respondent interaction, in particular as an adjustment step to receipt processing, and in (re)training of machine learning routines. However, a positive business case requires more than sound methodology. In this paper, we discuss the business case for going smart all aspects: IT, methodology, logistics and legal requirements. Considerations are based on field tests in multiple countries within the European Statistical System (ESS).

Keywords: Sensor data; Trusted smart statistics; Data collection innovation

BACKGROUND

From a technical perspective, smart surveys may employ one or more of the following features: 1) indevice processing and storage, 2) employment of internal mobile device sensors, 3) linkage to external sensor systems, 4) linkage to public online data, 5) data donation through the respondent, and 6) data donation through the NSI, i.e. linkage under consent. For households budget surveys, there may be various smart features. Statistics Netherlands performed a MOSCOW prioritization based on usability testing:

- In-device processing:
    - MUST: Semi-open product data entry, i.e. employment of daily language product lists and advanced stringmatching/search functions;
    - SHOULD: Personalized insights on expenditures; ○ MUST: Receipt text OCR-NLP; ○ COULD: Receipt product classification;
- Internal sensors ○ MUST: Receipt scanning through the mobile device camera;
    - COULD: Location tracking to assist respondents in recalling shopping/spending locations;
- Public data ○ MUST: EAN/GTIN product description data to assist in classification and product price plausibility checks;
- Data donation ○ MUST: Uploading of digital receipts;
    - COULD: Bank transactions data through the PSD2 legislation and procedures;

Part of the features has been explored in detail in a series of field tests. From a methodological point of view, smart surveys tend to reduce response burden and data outcomes are of higher quality, because of the automated measurements. For the household budget survey, this holds for the diary function that make it easier for respondents to track their expenses. And because a smartphone is often carried with them throughout the day, respondents are more prone to put an expense they made into the app. Next to that, the OCR receipt scanning automatically puts the bought items into the diary directly. And as a third advantage for the respondent, the insights page shows where one spends their money on. So with this smart, improved data collection method, the household budget survey was made more attractive to respondents. Next to the development of a new data collection tool, it is not very well known what data collection strategy fits best when using an app. It might be that an interviewer can be of added value to the respondent, or that only a letter might be enough to get people enthusiastic to participate. Not only in terms of quantity, but also in terms of data quality.

BUSINESS CASE

A smart household budget survey may have a positive business case for going smart(er), if the following criteria are met:

- Relatively high and balanced household response and completion rates can be guaranteed
- Reduced response burden by automated measurements
- Household expenditure data quality is improved, i.e. in diversity and in completeness
- In a multi-mode setting, HBS statistics remain comparable across relevant population subgroups
- Smart options are maintainable and relatively stable for the next five to ten years
- The back-office can be embedded in existing case management systems
- Respondent privacy can be protected and security follows standard best practices

The Household Budget Survey is an ESS-wide mandatory survey. Going smart in the ESS setting means:

- Smart options can be made configurable across countries
- Common denominator across countries is sufficiently large

In order to have a positive business case all smart survey levels, IT architecture, methodology, logistics, and legal should have a go.

EXPERIENCES

In the reference list to this abstract, we included a series of deliverables that were produced over the last year in two Eurostat funded projects: ESSnet Smart Surveys. They handle one or more aspects of the business case criteria. It would go too far to discuss all experiences in detail. We only selectively show three results.

The household budget survey app was fielded in six countries. In three countries, it followed exactly the same format: Luxemburg, Netherlands and Spain. Recruitment modes were randomized across samples to investigate the role of interviewers in recruiting but also motivating respondents. These are the registration, activity and completion rates:

- Registration – activity – completion rates ES ○ F2F only (N = 290): 29.4%, 26.5%, 22.4% ○ F2F/tel (N = 143): 25.4%, 21.5%, 18.8% ○ Non-interviewer (N = 433): 11.3%, 10.9%, 8.3%

- Registration – activity – completion rates LU ○ Interviewer (N = 881): 30.1%, 22.9%, 18.3% ○ Helpdesk (N = 884): 28.2%, 21.4%, 17.6%
- Registration – activity – completion rates NL ○ Interviewer (N = 685): 25.6%, 23.8%, 20.3% ○ Non-interviewer (N = 800): 15.9%, 11.6%, 9.6%
    - ○ The various rates show clearly the added value of interviewer assistance in recruitment and also completion. Rates are approximately twice as high.

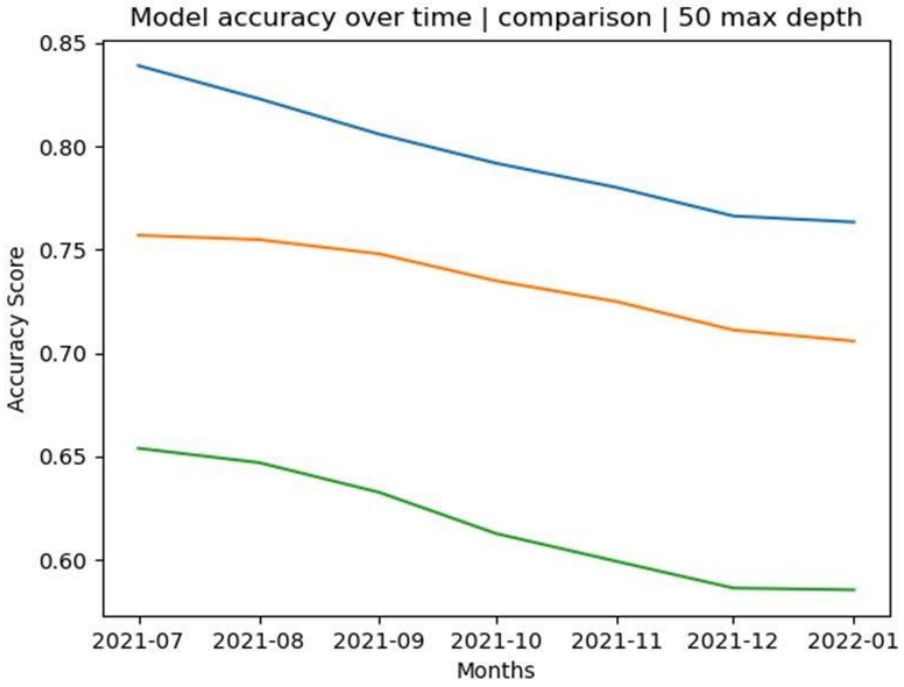
An important decision is the extent to which respondents are involved in receipt processing. In order to evaluate, in field tests respondents were randomized across different editing conditions. Table 1 shows a number of quality statistics for the groups with editing options and without editing options. When allowing for editing, the proportion of receipts with almost correct amounts is 50% higher than when it is not allowed. For the difference between true and real numbers of products, there is not much gain. However, respondents remove quite a few zero amounts when they can edit. A zero amount is provided by the app by default if it cannot find a price or the price falls below a certain OCR performance threshold.

Table 1: In-app respondent behavior for samples with and without options to edit products and prices extracted from receipts.

	With editing	Without editing
Correct amount	7.8%	2.7%
Difference amount <= 1 Euro	63.9%	11.1%
Correct number of products	27.6%	16.1%
Difference number of products <=1	43.5%	45.5%
No zero amounts	38.0%	18.5%

Figure 1 shows the accuracy of machine learning routines that were trained on different amounts of historic data, ranging from one month to one year. At the start the accuracy of a model trained on a year of data is around 85%, whereas for one month it is only 65%. When the model is not updated and kept constant the accuracy drops around 1 to 1.5% per month. Hence, performance decays very quickly in time and a good amount of data is needed to train.

Figure 1: Accuracy of machine learning for product classification as a function of time when one month (green), six months (orange) and a full year (blue) is used as train data.



CONCLUSION

From the experiences in the various smaller tests and larger field test, we draw the following conclusions about the business case:

- Response and completion rates: Unclear, rates vary greatly across countries. This may be in part due to experimental setup in these countries. Also some relevant subgroups still relatively low response;
- Reduced response burden: smart features were implemented to reduce response burden, but the difference with non-smart measurements need further investigation
- Household expenditure data quality: Data are plausible and resemble regular HBS data according to experts. A decision still is to be made where to do larger/recurring expenditures;
- Comparability across relevant population subgroups: To date still unknown as some subgroups will prefer non-smart modes;
- Maintenance and stability smart features: To some extent still open as machine learning routines for product classification require constant updating;
- Back-office and case management systems: This seems to work fine. Interoperable services have been prepared by several countries during field tests;
- Respondent privacy: Receipt scanning can be organized such that respondents can crop images and delete irrelevant, but potentially sensitive, parts;

In the ESS setting:

- Configurable across countries: Most country-specific features can be made configurable. The most extensive part is product classification where countries have different sources to create train data;
- Common denominator across countries: There is sufficient common ground to go to shared services on household budget surveys;

So where to go? The weak parts still are effective push-to-app/smart strategies, active/online learning of receipt machine learning routines, and comparability across important population subgroups. These require attention and dedicated experimental studies across multiple countries.

REFERENCES

Akkermans, J., Rodenburg, E., De Groot, J., Schouten, B., Martin Bernia, E., Balsa Criado, V., Horcajo Garcia, T., Gauche, C.,
Osier, G. (2022), The role of interviewers, deliverable 2.4, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

Akkermans, J., Van Hoek, S., Rodenburg, E., Schouten, B., Ménesi, É. Zanatyné, M., Attila, H., Parikka, T., Vrabič Kek, B., Grahonja, Č., Kebe, N. (2022), Field test analyses, deliverable 3.3, ESTAT project @HBS2, Statistics Netherlands

De Groot, J., Oerlemans, T., Rodenburg, e., Schouten, B. Clara Lope Mariscal, A., Martin Bernia, E., Poch, J., Horcajo Garcia, T., Balsa Criado, V., Gauche, C., Osier, G. (2022), Smart Survey pilots. The Consumption study, deliverable 2.1, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

Elevelt, A., Rodenburg, E., Akkermans, J., Schouten, B. (2022), Respondent engagement through personalized feedback and in-app editing, deliverable 2.3, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

Elevelt, A., Schouten, B. (2022), Bank transactions data donation, deliverable 4.1, ESTAT project @HBS2, Statistics Netherlands

Luiten, A., Lusyne, P., Schouten, B. (2021), Shareability of smart surveys in the ESS, deliverable 2.5, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

Schouten, B. (2022), Other smart features, deliverable 4.2, ESTAT project @HBS2, Statistics Netherlands

Schouten, B., De Wolf, N. and Oerlemans, T. (2022), Household Budget Survey app documentation, deliverable 2.1, ESTAT project @HBS2, Statistics Netherlands

Van Hoek, S., De Wolf, N., Van den Heuvel, G., Bos, J., Schouten, B. (2022), Receipt processing, deliverable 2.3, ESTAT project @HBS2, Statistics Netherlands

CRŒSS: A cross-domain data collection platform for the ESS

Keywords: TUS, HBS, platform, ESS, GSBPM, mobile, web, shareability

1. Introduction

The main responsibility of Eurostat is to provide reliable and comparable statistical information to the institutions of the European Union (EU) and its users. An important instrument to achieve this is through the promotion of – as far as possible – harmonised statistical methods across its member states.

TUS and HBS are two data collections in social statistics, where innovations are partially funded by grants under the European Statistical Programme. Since 2016 consecutive grants focused on Innovative Tools and Sources, of which the transition to an online data collection strategy is an important part. In a second step the ambition is to include smart data sources to the collection strategy.

This presentation is about the CROESS project which was a multi-beneficiary project with Statbel as coordinator, Destatis as beneficiary and hbits CV as subcontractor (spin-off of Vrije Universiteit Brussel).

2. Methods

The CRŒSS-project introduces a cross-domain data collection platform for TUS and HBS that is easy to share amongst countries. On the one hand, it means the availability for a researcher/statistician of a back-office to design and build a study/survey, and to collect data. On the other hand, it means the availability of a mobile (iOS and Android) and web application for the respondent in order to participate to the study/survey. Through the specific setup of the MOTUS data collection platform, the same back-office and front-offices can be used for TUS and HBS studies, and among different countries.

In addition, the CRŒSS project introduces new ways of data collection. This introduction itself is innovative through the inclusion of microservices. For example, a microservice processes paper and online ticket from shops for the benefit of HBS. Another microservice, making use of the sensors of the smartphone of the respondent, collects and analyses geolocation data for the benefit of TUS. The future of social statistics is insightful and flexible on one side, but also rigorous and comparable on the other side. The MOTUS data collection platform is able to merge these requirements into one environment.

3. Results

Within the CRŒSS-project the MOTUS data collection platform was upscaled in order to serve as a data collection platform not only for TUS, but also for HBS. Unique is the reuse of the already available research component/building blocks, and so the availability for NSI to have one platform for both TUS and HBS. On top, new components were developed specifically for HBS. For instance, the creation of an import function for COICOP classifications in the back-office. Furthermore, user-oriented microservices, namely receipt scanning, use of geolocation, and R script implementations were created and linked to the MOTUS platform. Towards the respondent a new user interface UI/UX for TUS and HBS apps (mobile and web) was evaluated and tested.

Through the CRŒSS-project, MOTUS became easier to share. Its use within different countries is now better supported and its need for comparability is boosted. Information is better structured and the data availability is more immediate. The benefits of the CRŒSS-project will be available for all European Countries, as well as for the EU-bodies.

Data Visualization (MANS 1A.1)

Session Chair: **Susanne Taillemite** *(Eurostat)*

An innovation approach of the data dissemination and data visualisation in official statistics
Sokrat Palushi *(NTT DATA),* Elma Çali *(Albanian Institute of Statistics)*

Territorial Economic Data viewer (TEDv). Data integration and visualisation of multiple EU funds at regional level
Anabela Santos *(EC - Joint Research Centre)*

restatapi: an R package to search and retrieve data from the changing Eurostat dissemination chain
Mátyás Mészáros *(Eurostat)*

An innovation approach of the data dissemination and data visualisation in official statistics.

Keywords: Dissemination System, Visualisation, Metadata, SDMX, Business Intelligence, Standardization, Power BI.

Regardless of progress made over the last decade, there are still large differences in the capabilities of statistical systems. Many countries still have insufficient tools and infrastructure to produce high-quality data. There is a growing need to modernize the dissemination and visualization of statistical information. A statistical data visualisation system is a necessary instrument in reaching a goal by creating a central data hub, integrating new data sources and statistical output. A statistical data visualisation system helps the statisticians to integrate data from multiple sources to serve as a basis for statistical analysis. Consequently, data collected from various sources and stored in the same databases can be directly visualized. Thus, implementing a cutting-edge dissemination system would be an essential process to improve data visualisation.

Key terminologies related to SDMX BI System are BI models where statisticians produce micro and macro data in each phase of production and Statistical Data and Metadata eXchange. SDMX helps data and metadata to be integrated and processed in order of visualisation and decision-making. SDMX is an ISO standard that ensures efficient data sharing across similar organizations like EUROSTAT, OECD, World Bank, UNICEF, etc. Therefore, this standard offers a format for data and metadata sharing, information models, and IT infrastructure. As a result, the data are more comparable, meaningful and usable.

On the other hand, Power BI provides the clarity of data according to a user-oriented perspective. Power BI also increases awareness of the accuracy and quality of data for statistical producers. This is the reason BI's role in data dissemination is very important. This paper will analyse with concrete examples of how the national Data Sharing and Visualisation System improve the quality of the official statistics system, and it is helpful for harmonization of the statistics production with EUROSTAT standards.

Territorial Economic Data viewer (TEDv). Data integration and visualisation of multiple EU funds at regional level

Keywords: Data integration; Territorial data visualisation; Policy monitoring.

INTRODUCTION

Access to relevant and timely data helps to support policy-decision making and to improve the effectiveness of policy intervention. However, policy analysts face several challenges to find the right data and to integrate them in an accurate reporting-system. Different data sources with different taxonomies, unstructured data and lack of time and resource are some the main bottlenecks policy experts are confronted.

The present paper aims to describe methods and techniques for data integration and visualisation, applied to Research & Innovation (R&I) funds and the case study of the Territorial Economic Data viewer (TEDv). The TEDv is the result of six main data-related activities:

- Data research: identifying needs, following data trends;
- Data collection: data gathering from different sources, combining micro and macro-level data
- Database construction: cleaning, harmonisation and enrichment
- Data analysis: producing derivate indicators
- Data visualisation: generating maps, graphs, interactive reporting and tailor-made statistics
- Data sharing: data import for policy decision making and research

The TEDv includes information from three different R&I funding programmes with different objectives (Table 1): (i) European Structural and Investment Funds (ESIF) under the thematic objective 1 (TO1 – R&I); (ii) Horizon 2020 (H2020) and; (iii) Recovery and Resilience Facility (RRF) under the thematic area R&I. Data from ESIF-TO1 and H2020 refer to the programming period 2014-2020 and include the EU contribution of the decided/selected projects or operations. RRF data comprises the estimated costs with R&Irelated expenditures in the period 2021-2026.

Table 1. Objective of ESIF, H2020 and RRF

Fund	Objective
European Structural and Investment Funds (ESIF)	To correct territorial inequalities, to enhance investment in job creation and a sustainable and healthy European economy and environment. Research and Innovation (R&I) is one of the main focus area – Thematic Objective 1
Horizon 2020 (H2020)	Financial instrument aiming to drive economic growth and create jobs. It supports research and innovation with an emphasis on excellent science, industrial leadership and tackling societal challenges

Recovery and Resilience Facility (RRF)	Temporary recovery instrument, aiming to mitigate the economic and social impact of the coronavirus pandemic. To make European economies more sustainable, resilient and better prepared for the green and digital transitions. R&I plays a key role in the recovery and twin transition.

DATA AND METHODS

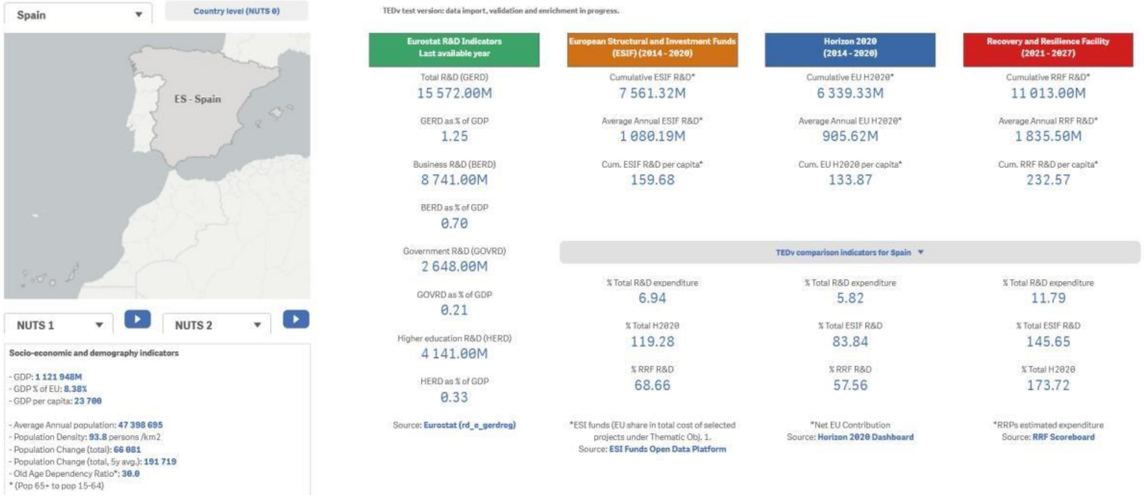
The TEDv uses essentially six main source of data:

- EUROSTAT (macro-level data) for socio-economic and demographic indicators, as well as, Research & Development (R&D) statistics. When needed we used the NUTS converter [1] to allow that all the statistical information are reported in the NUTS version 2021;
- COHESION OPEN DATA PLATFORM (macro-level data) to estimate the cumulative amount of EU funding share of total eligible costs decided until 2021 under the ESIF-TO1. When needed we used the NUTS converter [1] to allow that all the statistical information are reported in the NUTS version 2021;
- Horizon Dashboard (micro-level data) to extract the total H2020 funding allocated to each region in the programming period 2014-2020;
- Recovery and Resilience Scoreboard (macro-level data) to extract the cost estimated for R&I expenditures;
- JRC-WIFO (micro-level) database [2] for sectoring the values of ESIF-TO1 by NACE section;
- ORBIS (micro-level data) to fill gaps concerning NACE codes of the ERDF beneficiaries in the JRC-WIFO database.

Once all the statistical data are expressed in a common taxonomy (NUTS version 2021 and NACE codes), the Qlik application is used for data integration and visualisation.

RESULTS

Figure 1 and Figure 2 show an example of the data visualisation for Spain and the Andalucía region NUTS 2 level achieved with the TEDv. The tool provides a comprehensive territorial data visualization (at NUTS levels 0, 1 and 2) of EU Research and Innovation (R&I) funding and beyond.



Socio-economic and demography indicators

- GDP: **1 121 948M**
- GDP % of EU: **8.38%**
- GDP per capita: **23 700**

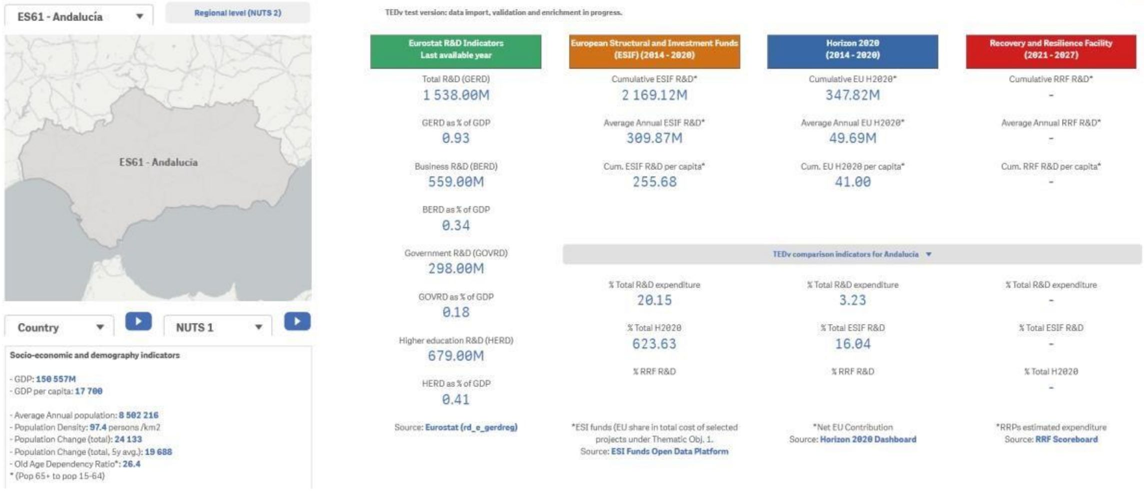
- Average Annual population: **47 398 695**
- Population Density: **93.8** persons /km2
- Population Change (total): **66 081**
- Population Change (total, 5y avg.): **191 719**
- Old Age Dependency Ratio*: **30.0**

* (Pop 65+ to pop 15-64)

TEDv test version: data import, validation and enrichment in progress.

Eurostat R&D Indicators Last available year	European Structural and Investment Funds (ESIF) (2014 - 2020)	Horizon 2020 (2014 - 2020)	Recovery and Resilience Facility (2021 - 2027)
Total R&D (GERD) 15 572.00M	Cumulative ESIF R&D* 7 561.32M	Cumulative EU H2020* 6 339.33M	Cumulative RRF R&D* 11 013.00M
GERD as % of GDP 1.25	Average Annual ESIF R&D* 1 080.19M	Average Annual EU H2020* 905.62M	Average Annual RRF R&D* 1 835.50M
Business R&D (BERD) 8 741.00M	Cum. ESIF R&D per capita* 159.68	Cum. EU H2020 per capita* 133.87	Cum. RRF R&D per capita* 232.57
BERD as % of GDP 0.70			
Government R&D (GOVRD) 2 648.00M	TEDv comparison indicators for Spain		
GOVRD as % of GDP 0.21	% Total R&D expenditure 6.94	% Total R&D expenditure 5.82	% Total R&D expenditure 11.79
Higher education R&D (HERD) 4 141.00M	% Total H2020 119.28	% Total ESIF R&D 83.84	% Total ESIF R&D 145.65
HERD as % of GDP 0.33	% RRF R&D 68.66	% RRF R&D 57.56	% Total H2020 173.72
Source: **Eurostat (rd_e_gerdreg)**	*ESI funds (EU share in total cost of selected projects under Thematic Obj. 1. Source: **ESI Funds Open Data Platform**	*Net EU Contribution Source: **Horizon 2020 Dashboard**	*RRPs estimated expenditure Source: **RRF Scoreboard**

Figure 1. Example of data visualisation for Spain



Socio-economic and demography indicators

- GDP: **150 557M**
- GDP per capita: **17 700**

- Average Annual population: **8 502 216**
- Population Density: **97.4** persons /km2
- Population Change (total): **24 133**
- Population Change (total, 5y avg.): **19 688**
- Old Age Dependency Ratio*: **26.4**

* (Pop 65+ to pop 15-64)

TEDv test version: data import, validation and enrichment in progress.

Eurostat R&D Indicators Last available year	European Structural and Investment Funds (ESIF) (2014 - 2020)	Horizon 2020 (2014 - 2020)	Recovery and Resilience Facility (2021 - 2027)
Total R&D (GERD) 1 538.00M	Cumulative ESIF R&D* 2 169.12M	Cumulative EU H2020* 347.82M	Cumulative RRF R&D* -
GERD as % of GDP 0.93	Average Annual ESIF R&D* 309.87M	Average Annual EU H2020* 49.69M	Average Annual RRF R&D* -
Business R&D (BERD) 559.00M	Cum. ESIF R&D per capita* 255.68	Cum. EU H2020 per capita* 41.00	Cum. RRF R&D per capita* -
BERD as % of GDP 0.34			
Government R&D (GOVRD) 298.00M	TEDv comparison indicators for Andalucia		
GOVRD as % of GDP 0.18	% Total R&D expenditure 20.15	% Total R&D expenditure 3.23	% Total R&D expenditure -
Higher education R&D (HERD) 679.00M	% Total H2020 623.63	% Total ESIF R&D 16.04	% Total ESIF R&D -
HERD as % of GDP 0.41	% RRF R&D	% RRF R&D	% Total H2020 -
Source: **Eurostat (rd_e_gerdreg)**	*ESI funds (EU share in total cost of selected projects under Thematic Obj. 1. Source: **ESI Funds Open Data Platform**	*Net EU Contribution Source: **Horizon 2020 Dashboard**	*RRPs estimated expenditure Source: **RRF Scoreboard**

Figure 2. Example of data visualisation for Andalucía (Spain)

TEDv also reports derived cross-funding indicators showing i.e. the average contribution of different R&I funds over total R&D and the relative importance of different R&I funds (see tables at the bottom right in Figure 1 and Figure 2). For instance, it shows that on average ESIF-TO1 contributes to around 7% of total R&D expenditure in the Spain (Figure 1) whereas in the Andalucía region it funds on average more than 20% of total R&D in the region (Figure 2). Estimated R&I-related costs financed by RRF turn out to be around 1.5 times the size of ESIF-TO1 and 1.7 times that of H2020 in Spain (Fig. 1).

cONCLUSIONS

The TEDv is the first available tool which combines statistical territorial information of different EU funding programmes in a single and coherent framework – mainly thanks to the methodological effort on territorial and sectorial/thematic allocations (via taxonomy conversions). In turn, this allows to compare the size of different EU funding and to show the contribution of them for total R&D expenditures. Such kind of statistical information can be particular useful for policy-makers, since it allow to compare the relative position of a territory with respect to country / EU averages, as well as to other EU regions.

rEFERENCES

[1] Batista e Silva, F.; Attardo, C.; Beri, M; Bucciarelli, G.; Dijkstra, L. (2020). NUTS Converter: description of the tool and calculation method. European Commission, Joint Research Centre. https://urban.jrc.ec.europa.eu/nutsconverter

[2] Bachtrögler, J., Arnold, E., Doussineau, M., and Reschenhofer, P. (2021). *UPDATE: Dataset of projects co-funded by the ERDF during the multi-annual financial framework 2014-2020*, JRC125008.

restatapi: an R package to search and retrieve data from the changing Eurostat dissemination chain

Keywords: dissemination, current and future API, R

Introduction

Changing dissemination chain

Eurostat, the statistical office of the European Union, disseminates majority of the data through the Eurostat database. Next to the navigation tree and data browser, Eurostat provides access to the datasets through web services using Application Programming Interface (API). In the context of the renovation of the dissemination chain, a project was initiated as part of the Eurostat's Methodological Network to provide users and developers with interfaces to Eurostat Representation State Transfer (REST) API implemented in various programming languages.

The aim of the development was to prepare ready-to-use interfaces available in different languages as soon as the new services are launched. These new interfaces can decrease the impact of the anticipated potential service disruption for those users/developers that are relying on the existing interfaces to the current API, and which would require them to adapt the existing interfaces or develop new ones. In this way, external users and developers benefit from the reduced transition time.

Creation of a NEW R package

Reasons for a new package

One of the planned interface was to provide access to the Eurostat database through the REST API in R, as R widely used by statisticians, methodologists and data scientists worldwide. Several statistical offices and international organizations have similar R packages to facilitate the access to their data from R through APIs. There are a few R packages to retrieve data from the Eurostat database as well, but none of them was developed with the involvement of Eurostat.

To fill this gap a new package was created in Eurostat taking into account the expected changes in the dissemination change and the possible future user needs for an improved user experience.

Similar packages and their main features

The *eurodata* package uses both the old and new API and very efficient as it uses partially C++ code and the *data.table* package, which is based on C++ as well. On the other hand it has 8 direct and 15 indirect dependencies and has limited search functionality.

The *eurostat* package has a wide range of options to retrieve and filter data. The package is caching data locally in case there are similar queries. The disadvantage of the package that it contains hard coded URLs in many of the files to the current APIs and has even larger dependencies (16 direct and 70 indirect dependencies) than the *eurodata* package.

The *rdbnomics* package can be also used to retrieve Eurostat data, but it is not using directly the Eurostat API, as it queries the dbnomics database, which contains historical/revised data from Eurostat regularly downloaded using the bulk download facility. It has only limited number of dependencies (3), but the search is not user friendly as the user has to know the exact code of the series and dimensions and it is not possible to filter based on time values.

The *RJSDMX* package uses Java and the SDMX API to retrieve data series. It has only 2 dependencies but one of them is *rJava* which is dependent on external Java installation that is not available in all workplaces due to security reasons. The search is similar to *rdbnomics*. It is not user friendly as the user has to know the exact code of the series and dimensions, but this package can filter data based on time values. Finally, it has also hard coded URL of the current API in the Java code.

Development principles

Based on the review of the already existing packages in this field, the main guiding principle for the development of the new package were the following ones:

- External configuration file to smoothly handle the change in the API
- Minimised number of dependencies on other packages
- Efficiently handle large datasets
- Provide flexible filtering options

Features of the new package

External configuration to handle the change in the API

The newly created package uses a JSON configuration file, which is located in the source code on GitHub. The file is installed locally as well with the package in case there are internet access problem to the configuration file. When the package is loaded it tries to load first from the internet the configuration and only if it fails uses the locally installed version. The URLs and other parameters of the current and the new API is stored in this JSON file together with a value that defines which set of configuration values should be used as default. The other functions of the package are extracting the URLs and parameters based on this setting from the configuration values. This way it is possible to switch the default settings without reinstalling the package only changing the default value on GitHub. Of course if someone wants, (s)he can override this default option, and load specific settings from the local file.

• Minimised number of dependencies on other packages

The package has only 3 dependencies to handle 3 types of files. The *rjson* package is used to handle the configuration file and in case future developments the JSON web service of Eurostat can be used. The *xml2* package handles the filtered datasets retrieved from the SDMX endpoint.

Finally, the *data.table* package is used to process whole datasets retrieved in TSV format from the bulk download facility.

• Efficiently handle large datasets

Two techniques are used to handle large datasets efficiently.

First, the *data.table* package is used, which uses less memory and reads, processes and writes faster tab separated value files then the standard data frame in base R. As well this package is used when the whole data set is downloaded and the filters are applied locally instead of using the SDMX endpoint for data filtering.

Second, parallel computing is applied to process SDMX files, e.g. to extract information from the Table of Contents of the datasets, or extract data from the SDMX response of data filter query. Also caching is used for whole datasets, so it does not have to download twice the same data, and the filter can be applied to the locally cached data. By default the memory is used as cache. The cache can be also disk, but in that case the processing time is increasing.

• Provide flexible filtering options

The package implements 3 types of filters:

1. standard filter for filtering through dimensions/concepts.
2. date filter to filter through time
3. frequency filter to filter by the frequency of the data

These filters can be applied individually or in any kind of combinations.

For standard filters there are several options to define them. If someone knows the exact codes and the order of the concepts then he can use as it is created with the query builder. If someone does not know the codes but familiar with the name of the concepts, then the filter can be created as a named list where the values can be any string, which can be matched exactly or partially. Finally, if someone does not know any code or concept name, then (s)he can search using single or multiple strings or regular expressions to look for them in the codes and optionally in the description of the codes. For the last 2 cases the user even does not have to know the order of the concepts, the code will take care of it.

For date filter there are many options to define it. The date filter can be a simple string or a vector. The date can be defined as day, month or year(s), or can be expressed as a range or as a string including the larger or smaller sign and a given date. These various options can be used individually or combined in a vector without being in chronological order. The code will create the union of all the defined time constraints.

The third filter provides the possibility to filter by frequency of the data. The package can handle multiple frequencies and can incorporate frequency filter defined in the standard filters as well.

As mentioned above, these filters are generally applied to query the SDMX web service, but the filters can also be applied locally on previously cached datasets. In some cases when the query response contains more than 30 thousand observations then the SDMX web service provides asynchronous delayed response. In this case the package automatically makes the filtering on the local computer not to wait for the SDMX response. It is also possible to force this local filtering. A good example for this is the data of the Time Use Survey, where the SDMX web service after filtering does not provide any values when the measurement is in time values, but the local filter provides that data.

Finally, there is option to filter out values supressed and flagged due to confidentiality and include other flags and/or add labels to dimensions for better understanding the values.

Conclusions

The ***restatapi*** package provides data scientist and the general public with a tool that can smoothly handle the changes in the dissemination API of Eurostat. It also provides flexible search options for persons who are not familiar with the specific codes and classifications used in the datasets. The package uses out the multi-core processors of present days computers to process the data in parallel on multiple threads in an efficient manner.

Ares for future development

The current version of the package could be extended by an additional options providing time-out threshold for the API calls and it would create a better user experience if improved error handling implemented with more detailed debug messages when the API call fails due to various reasons.

Smart data (MANS1A.2)

Session Chair: **Michelle Jouvenal** *(National Institute of Statistics–ISTAT)*

Smart Mobility Surveys: Current App Developments, Methodological Perspectives, and Respondent Interaction
Yvonne Gootzen, Jonas Klingwort, Mike Vollebregt, Barry Schouten *(Statistics Netherlands-CBS)*

Challenges in the Smart Household Budget Survey: how to involve & motivate respondents
Jelmer de Groot *(Statistics Netherlands-CBS)*

Advances in Eye-Tracking and Cognitive Interviewing Methodology: Dos, Don’ts, and Decisions
Mátyás Gerencsér *(Hungarian Central Statistical Office)**; Anna Sára Ligeti (Hungarian Central Statistical Office); Ferenc Mújdricza (Hungarian Central Statistical Office); Linda Mohay (Hungarian Central Statistical Office)

Smart Mobility Surveys: Current App Developments, Methodological Perspectives, and Respondent Interaction

Keywords: sensor data, GPS data, respondent tracking, time use survey, smartphone survey

1 Introduction

The need for (trusted) smart surveys in official statistics is increasing. This is caused by the fact that diary surveys, such as Time Use Surveys or Household Budget Surveys, are burdensome for respondents as they collect data in the form of a (digital) diary [1]. Additional drawbacks of such surveys are underreporting, memory errors, and nonresponse, leading to a loss in data quality. In the finished 'ESSnet Smart Surveys' pilot studies on four official statistics themes were conducted: Consumption, Time use, Health, Living conditions and environment. These studies used smartphones or wearables to measure the respondent's behavior. These sensor-based measurements provide more accurate and timely data and lower the costs for official statistics. This paper presents research on a recently developed state-of-the-art smartphone app on mobility behavior in the Netherlands and adds to the current research initiatives. The app is developed by Statistics Netherlands and the Ministry of Infrastructure and Water Management. The most complex question in app-assisted passenger mobility surveys is how to set the right balance in active-passive data collection. This amounts to specific questions such as what are the boundaries of automated travel statistics derivation? Does active and passive data collection yield comparable results? How willing and competent are respondents to adjust errors and/or missing data? How do these depend on their background? And what is the role of the device and underlying sensor technology?

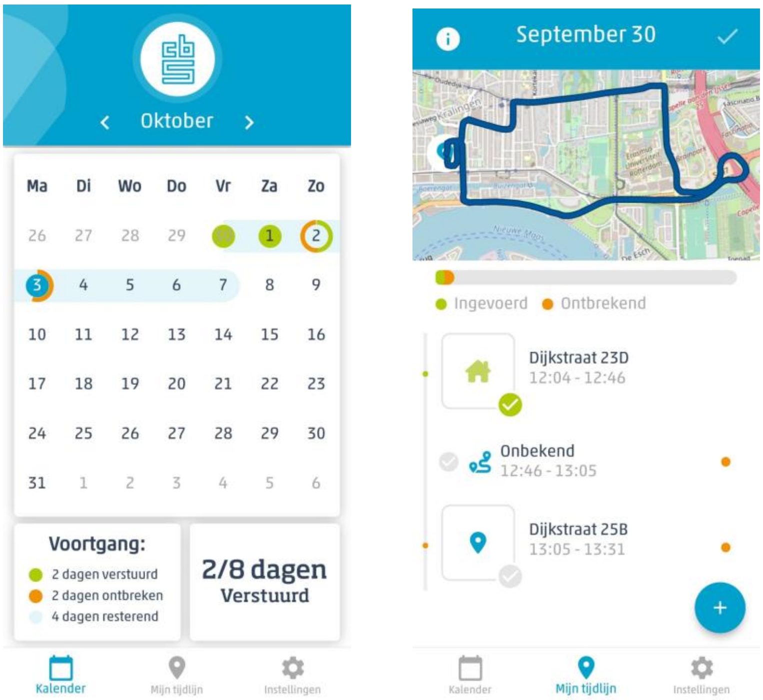
Regarding these questions, we will present lessons learned concerning app development, methodology for data validation of sensor data and respondent data, and investigations of how and what actions respondents perform.

2 Smart Mobility Survey – App Development

The use of sensor technology is of central importance in smart survey data collection on mobility behavior. Most smartphones have built-in sensors that measure motion and orientation. However, using these sensors efficiently and effectively is not straightforward. Factors such as battery management, software development kit, brand, and sensor type play essential roles concerning data quality. Relevant data quality aspects include the frequency of measured data points, accuracy, and the raw GPS location data completeness. For example, collecting high-frequency data will harm the battery run-time. Moreover, different smartphone brands provide different data quality: low-priced brands are more likely to have inferior-quality data.

The developed app can automatically select a balanced software configuration that allows high-quality data (high accuracy and complete GPS data) and maintain the battery life for a sufficient amount of time. This decreases the likelihood that a respondent disables/uninstalls the app due to experiencing a negative impact on battery run-time. The collected raw data is clustered into time periods. The clustered periods are classified into stationary or movement events using an algorithm (for details, see section 4). These classified periods will be shown to the respondents.

Figure 1 shows an example of the app. The left panel shows the digital app diary, which shows the respondent the period of interest, how many days need to be filled in, or how many have already been filled in or initiated. The right panel shows the recorded GPS data and timestamp to the respondent and the classified states. These classifications can be approved, edited or deleted by the respondent.



(a) Digital app diary

(b) GPS location data and state classification

Figure 1: Example screenshots of the developed mobility app. The digital app diary (left) and the recorded GPS location data and state classifications (right) are shown.

3 Data

To test, understand, and improve the app's functionalities, several field tests were conducted starting in May 2022 and ending in early 2023. These included small-scale internal tests and a large-scale study based on a random sample of the general population. The smaller tests aimed at technical improvements of the app. The large test was based on lessons learned in the small tests and aimed at app usability and understanding the respondent interaction. Within the smaller tests, an experiment was carried out where respondents both used the app and logged their mobility activities (stationary or movement, with timestamps) in a traditional paper diary. Here, differences between sensor technologies were analyzed, and simulation studies were carried out to identify the optimal parameter choice for the app algorithm. This algorithm is a core app component because it classifies the recorded GPS data into stationary or movement events. Within the app, the classifications are displayed to the respondents so that they can approve or modify them. If these classifications are not correct due to an inaccurate choice of parameters, respondents will likely doubt the usefulness of the app and their participation. The results of the experiment will be addressed in the following section.

4 Methodological perspectives

Before the recorded GPS data could be analyzed, several (pre-)processing steps were required. First, an accuracy filter was to be applied to select GPS data with accurate measurements. Second, a median filter was applied to drop suspicious data points that deviate strongly from the routes taken. Finally, the algorithm was applied to transform the GPS data in *events*, which are in the form of time periods annotated by a description of the expected activity (stationary or movement) [2]. Two main parameters affect the algorithm output: distance and time parameters. Their effect on two target variables was studied in a simulation study: the total number of events and the number of switches per hour. The diary, which had overlapping periods with the sensor data, was used to validate the output. In this study, the respondent data (paper diary) was kept constant, while the GPS data classifications varied depending on input parameters. Figure 2 shows a selection of the simulation study results. Concerning the total number of moves, the app data underestimates the target (the red bar is always larger than the green). This holds for different parameter choices and different sensor technologies. This result was also found for different parameter choices (not shown). However, concerning the number of switches per hour, the app data seems to pick up more switches, as reported in the diary. Again, this holds for different parameter choices and different sensor technologies. According to the classifications, it takes about 2 hours for a switch (changing from movement to stationary or vice versa).

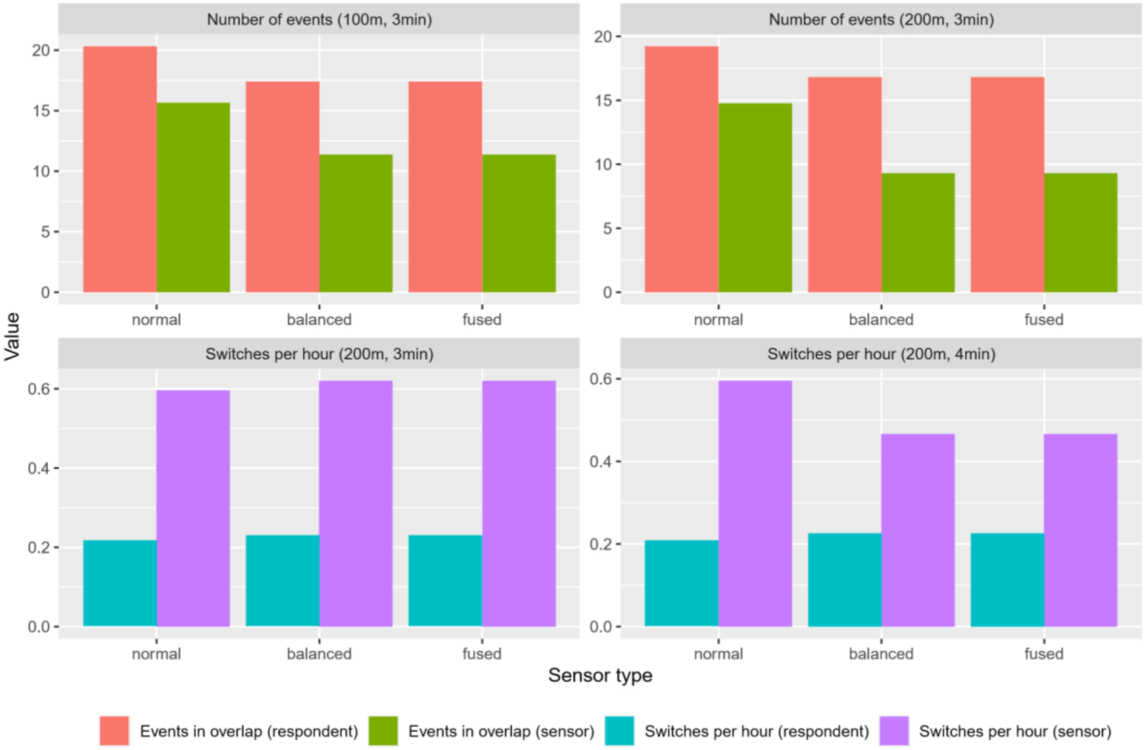


Figure 2: Simulation study results on two target variables: total number of events and number of switches per hour. The x-scale shows different sensor technologies and the y-scale the estimated values.

A different recording frequency likely explains the differences between sensor technologies. The normal sensor provides the most GPS data points. The balanced and fused sensors provide fewer but comparable amounts of data points. Choosing the parameters too large leads to an underestimation of both target variables. It seems that different sensors have different sets of optimal parameters.

5 Respondent interaction

Motivations for going ‘smart’ in surveys are threefold: reduce burden, improve data accuracy and provide better proxies of the topics of interest. In passenger mobility surveys that would be less burden to report all travels and more accurate estimates of traveled distances and number of stops. Furthermore, what constitutes a stop is in non-smart surveys left to respondents.

Respondents still need to be involved, however. This has three reasons, namely making sure they remain in control from a GDPR perspective, engaging them in the research, and involving them in quality control. Location sensor data may contain gaps, outliers and measurement errors and respondents may be asked to check and adjust. Potential actions by respondents may range from small adjustments on timing and locations of stops to deleting spurious stops and travels to adding complete travels. Obviously, the more advanced actions are harder but also more burdensome to perform and counteract the original motivations of going smart.

Based on field test data and in-app paradata with varying amounts of possible respondent actions, we show how willing and competent respondents are.

6 Conclusions

Using smartphone-based surveys to complement or replace traditional diary-based mobility surveys is highly important in official statistics. In this paper, we presented recent research on smart mobility surveys and related app development, methodological perspectives, and reflected on the app-interaction experiences of respondents. The questions addressed in the introduction can be answered as follows at the time of writing. Automated mobility data collection, where respondents seldom need to adjust classified states, substantially lowers the response burden and contributes to the right balance in active-passive data collection. Though automatized data collection also has boundaries, a strict parameter set may systematically lead to incorrect algorithm-based classifications in some instances. However, the experimental studies showed that comparable results between a traditional and a smart survey could be achieved. The differences are generally determined by the choice of algorithm parameters but also by different sensor technologies. These remaining differences, as well as studying the willingness and competence of the respondents to adjust errors and/or missing data, are part of current research.

References

[1] Ricciato F, Wirthmann A, and Hahn M. Trusted smart statistics: How new data will change official statistics. *Data Policy*, 2:e7, 2020. DOI: 10.1017/dap.2020.7.

[2] Smeets L, Lugtig P, and Schouten B. Automatic travel mode prediction in a national travel survey. *CBS Discussion Paper*. https://www.cbs.nl/en-gb/ background/2019/51/automatic-travel-mode-prediction, 2019.

Challenges in the Smart Household Budget Survey: how to involve & motivate respondents

Keywords: Smart surveys, CAPI, ESSNet, Household Budget Survey, sensor data

1. INTRODUCTION

Household budget/expenditure surveys contain many elements that make them very well fit for the introduction of smart survey features. These surveys tend to be burdensome, both in time and in cognitive effort, and the information requested may not always be readily available for respondents. Smart surveys introduce features of smart devices such as internal storage and computing, internal sensors, linkage to external sensor systems, access to public online data and various forms of data donation. Some of these features, such as receipt scanning or uploading, advanced product search algorithms and data donation, are very promising. They are also challenging in terms of user interface design and processing through text/image extraction methods. Besides these technical aspects, methodological uncertainties are faced; what approach strategy fits best when implementing smart features into a survey?

In ESSnet Smart Surveys a large-scale field test has been conducted testing various recruitment and motivation strategies. Randomized conditions such as mode of invitation, in-app feedback on OCR results and personalized insights were added. These conditions allow for an analysis of data quality trade-offs. Since the aim is to reduce perceived respondent burden, it is crucial to know what the boundaries are in respondent involvement and motivation, and how these depend on design features.

In the paper, we will discuss respondents' motivation and involvement for a smart household budget survey as a function of approach strategy design choices. The main question for this pilot was: how can we involve respondents into the fieldwork and how can we keep these respondents motivated throughout their writing period?

In the field test, interviewer-assistance was randomized; one sample had assistance and the other not. The non-interviewer sample thus is a control group. To quantify differences, two main questions were asked:

1) what is the impact of interviewer-assistance on respondents' participation, motivation and involvement?
2) what is the recommended role of interviewers?

Another way to keep respondents motivated is by giving them insights in their own expenses. This was varied by giving half of the sample insights into their expenses directly, whereas the other half had access to their insights page after the writing period. The main question here was:

1) What is the impact on direct versus delayed insights on registration and completion rates?

Another smart feature that was implemented on the back-end side was the logging of paradata. By this paradata, in-app behaviour of the respondents was analysed to further quantify the above mentio.

2. METHODS

For the Household Budget Survey an in-house built app within the EuroStat funded project ESSNet and @HBS, and @HBS2. The pilot described here within the ESSNet was conducted in three different countries; Luxembourg, Spain and the Netherlands. There has been a variation on three different topics:

1) Instant versus delayed insights

2) Interviewer versus non-interviewer approach

3) In-app editing of OCR

The latter variation will not be further discussed here. The different conditions are divided among the sample as follows:

Table 1: Field test outline

	ES	LU	NL
Insights instant	Planned: 400 Realized: 433	Planned: 800 Realized: 882	Planned: 800 Realized: 748
Insights delayed	Planned: 400 Realized: 433	Planned: 800 Realized: 884	Planned: 800 Realized: 737
Interviewer	Planned: 400 Realized: 433	Planned: 800 Realized: 881	Planned: 800 Realized: 685
No interviewer	Planned: 400 Realized: 433	Planned: 800 Realized: 884	Planned: 800 Realized: 800

To give indications for respondents' motivation and involvement, we looked at the response rates and how this differed between the conditions that was varied in for the table listed above. Next to this, paradata was used to further look into the respondents' behaviour in the app to reveal more about their motivation throughout time. The indicators analysed were:

1) Respons rates and completion rates
2) Time spent in the app per day
3) Number of pages visited in the app per day
4) Type of expense: manually or scanned entries

3. RESULTS

The main findings on response rates among the 3 countries for the interviewer condition were as follows:

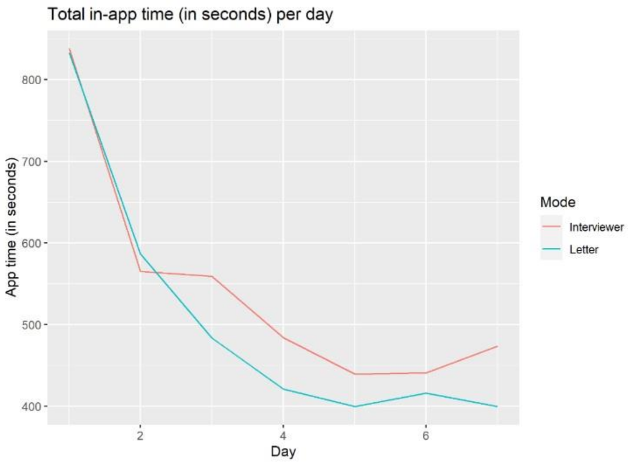
- Registration – activity – completion rates ES ○ F2F only (N = 290): 29.4%, 26.5%, 22.4% ○ F2F/tel (N = 143): 25.4%, 21.5%, 18.8% ○ Non-interviewer (N = 433): 11.3%, 10.9%, 8.3%
- Registration – activity – completion rates LU ○ Interviewer (N = 881): 30.1%, 22.9%, 18.3% ○ Helpdesk (N = 884): 28.2%, 21.4%, 17.6%
- Registration – activity – completion rates NL ○ Interviewer (N = 685): 25.6%, 23.8%, 20.3% ○ Non-interviewer (N = 800): 15.9%, 11.6%, 9.6% For the variation on the insights, the results were the following:

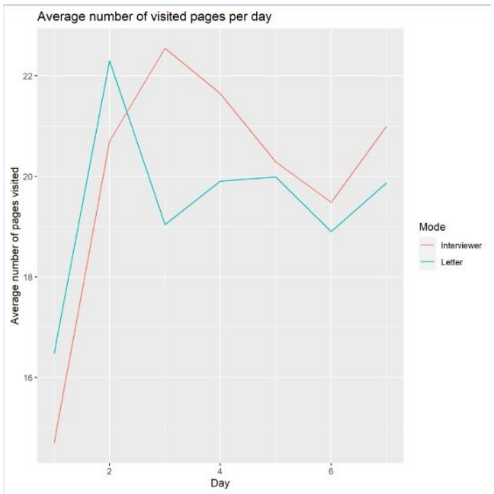
- Registration – activity – completion rates ES ○ Instant: 15.8%, 14.6%, 11.8% ○ Delayed: 20.9%, 19.1%, 15.7%
- Registration – activity – completion rates LU ○ Instant: 30.9%, 23.1%, 18.8% ○ Delayed: 27.3%, 21.1%, 17.1%
- Registration – activity – completion rates NL ○ Instant: 19.2%, 16.3%, 13.7% ○ Delayed: 21.3%, 17.8%, 15.1%

Next to these results, the paradata results showed that we did not see many differences in the interviewer versus the non-interviewer group, as can be seen in the following figures:

Figure 1: total in-app time spent per day

Figure 2: average number of visited pages per day





It is shown that people in the interviewer group keep spending more time in the app throughout the whole fieldwork period compared to the people that were invited by letter only. Respondents recruited by an interviewer also visited more pages per day.

As is looked into the number and types of purchases that respondents reported, the following can be shown:

Table 2: The number of entries (manual/scanned) and amounts, per household & per day

		Mean	1st Q	Median	3rd Q	Max	SD
Entries per day per household	Interviewer	1.05 [1.0408 - 1.0532]	0.9990	1.0117	1.1041	1.4440	0,0092
	Letter	1.05 [1.0449 - 1.0566]	1.0123	1.0588	1.0907	1.1299	0,0051
Manual entries per household per day	Interviewer	0.7932 [0.7877 - 0.7987]	0.7578	0.776	0.8181		
	Letter	0.8438 [0.8380 - 0.8497]	0.7941	0.8456	0.886		
Scanned entries per household per day	Interviewer	0.2538 [0.2517 - 0.2559]	0.2374	0.247			
	Letter	0.2069 [0.2055 - 0.2084]	0.1949	0.2083	0.2169	0.2279	0,0014

As the results show in table 7, it is seen that the number of entries is more or less the same for both groups. When the purchases are split into manual purchases and scanned ones, it is shown that the people recruited by an interviewer reported more scanned purchases than the letter group. On the other hand, the people that were invited by letter only, show more manual entries. The higher number of scanned entries versus manual entries could be caused by the interviewer itself, that at the door was able to give more information about the receipt scanner part more than logically can be done by only a letter. Not only could the interviewer make the respondent aware of the fat that receipts could be scanned, also the interviewer could show how the scanner works to make it more convenient for the respondent to work with it. The letter-only group had to find this out for themselves.

4. cONCLUSIONS

The impact of interviewer-assistance on registration and completion rates was someway what it was expected to be. The interviewer assisted approach showed higher response rates compared to the letter group. This was already expected because this is often seen in other surveys. The interviewer at the door can be of great value in convincing people to participate in the survey and to take away their worries and talk them through the app itself. This was clearly seen in the registration rates as well as in the completion rates. Both rates were significant higher in the interviewer assisted group. However, there is no strong evidence that interviewers also recruit different types of households, e.g. older households or lower income households. The number of shared auxiliary variables across the three participating countries was too small for a detailed look.

The motivation of the respondents was looked at through indicators that are predictors of motivation and involvement in the app, such as the number of purchases that were reported, as well as the average amount of these purchases. The number of purchases was the same for both groups, but the amount of the purchases was higher for the letter-only group. Besides that, an interesting effect was found, where people used the scan-function of the app more in the interviewer group when compared to the letter group. So if the in-app data entry option is further developed and perfected, then it is recommendable to have interviewers notice people that this function can be used and that it lowers respondent burden. As a final notable finding, using indicators as time spent in the app and number of pages visited, higher motivation is found within the interviewer group compared to the letter-only group.

The interviewer mode is a very expensive one and does not really affect the composition of the responding groups. Both groups spent a comparable amount of time in the app and also report the same amount of purchases into the app. For respondent burden it could be of great help that the interviewer tells the respondent about the OCR function in the app, so the respondent does not have to put all the expenses in manually.

Next to the quantity or quality of the response, interviewers also can have a helping effect for respondents. They can be the ‘face’ of the NSI, instead of just an somewhat ‘anonymous’ sent letter. Also, when a lot of materials are needed or sent to respondent, it is of great value that there is an interviewer at the door that can give some explanation about what the respondent needs to do as well as the purpose of the research. Even though the interviewer could not really help respondents with technical issues, but for app users it is of great service that they have their personal helpdesk. In this experiment, the role of the interviewer was very small by giving only a letter at the door. Interviewers might have an even greater effect if they also would conduct a start questionnaire.

REFERENCES

[1] Akkermans, J., Rodenburg, E., De Groot, J., Schouten, B., Martin Bernia, E., Balsa Criado, V., Horcajo Garcia, T., Gauche, C., Osier, G. (2022), The role of interviewers, deliverable 2.4, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

[2] De Groot, J., Oerlemans, T., Rodenburg, e., Schouten, B. Clara Lope Mariscal, A., Martin Bernia, E., Poch, J., Horcajo Garcia, T., Balsa Criado, V., Gauche, C., Osier, G. (2022), Smart Survey pilots. The Consumption study, deliverable 2.1, ESTAT project ESSnet Smart Surveys, Statistics Netherlands

[3] Schouten, B., De Wolf, N. and Oerlemans, T. (2022), Household Budget Survey app documentation, deliverable 2.1, ESTAT project @HBS2, Statistics Netherlands

Advances in Eye-Tracking and Cognitive Interviewing Methodology: Dos, Don'ts, and Decisions

Keywords: cognitive interviewing, eye-tracking, questionnaire testing.

INTRODUCTION

Cognitive interviewing is a questionnaire pretesting method by which researchers gain insights into the cognitive processes of how people interpret and respond survey questions. It is mainly used to reveal and fix problems with survey questions and questionnaire structure, with the ultimate purpose of optimising the data collection instrument. The cognitive interview (CI) is a semi-structured qualitative interview, consequently, CI studies are usually done on small, nonprobability samples. The CI simulates a standard field interview in that draft survey questions are administered to a test subject (TS), who responds them as they would normally do. The goal of a CI is to collect data on TSs' response problems and verbal narratives on their question–response process. Cognitive interviewing is more effective in pretesting interviewer-administered questionnaires. The rise of internet-based self-completion (CAWI) prompted enhancing CIs by eye-tracking, which compensates the weaknesses of CI studies on self-completed questionnaires and sheds light on hidden mental processes. However, initial efforts to integrate CIs with eyetracking (e.g. [1][2][3][4]) show teething troubles reminiscent of outdated CI methods. Therefore, a robust, scientifically reliable framework is needed for eye-tracking enhanced CI studies (ETCIs). The paper outlines how to incorporate eye-tracking into state-of-theart CI methodology, providing a list of methodology-driven recommendations on essential decisions to aid experts maintain scientific quality and validity.

METHODS

Benefits of eye-tracking in cognitive testing of CAWI questionnaires are indisputable. Its teething troubles appear to stem from failing to give heed to a pivotal debate. Kirsten Miller's [5] sharp critique of unscientific CI practices yielding impressionistic, anecdotal results brought about a methodological revolution in cognitive interviewing. It culminated in Miller and colleagues' [6] interpretive or Theme Coding [7] method: a Grounded Theory-based [8][9], rigorous methodology. Setting the standards in many aspects, this rich, descriptive approach opposed simple problem-seeking reparative aims and emphasised the interpretations and the social–cultural context. It goes without saying that introduction of a new data collection instrument should not disregard methodological standards already in place, as the basic principles remain the same. Therefore, the present paper discusses how to address critical decisions of an ETCI study. In the revision of extant ETCI practices, a 'methodology adaptation' method was used, inspired by the theory adaptation framework [10]. Problems, shortcomings, and inconsistencies were identified and resolved by applying the principles of state-of-the-art CI methodology.

RESULTS

3.1. How to design the sample and recruit participants?

Sample composition is the most salient aspect decision-making. Even a low-budget, smallsample study may yield rich and robust results by careful sample design – an oft-neglected step. Convenience sampling and sample corruption by enrolling colleagues, experts, etc. is common in ETCI studies [1][4][11][12]. Contrarily, scientific standards demand a purposive sample considering multiple factors [13][14][15][16]. Interlocked or parallel quota design is preferred [15][16]. If the questionnaire covers a wide range of topics, we recommend a complex structure: key criteria for interlocked, auxiliary criteria for parallel quotas. To avoid recruitment bias, at least two recruitment channels should be used [16]. Creating a reserve sample with the same quota structure is also needed to handle drop-outs. Administering a screening questionnaire with questions on sample and other eligibility criteria is recommended [15][16].

Concurrent vs. retrospective protocol and (not) using think-aloud

ETCI structure is determined by the relationship between its two major phases. The eyetracking phase (when the test questionnaire is completed) and the CI phase can be either sequential or simultaneous, not strictly separated in time. This decision also limits the corresponding applicable verbal data collection methods: retrospective or concurrent protocol, respectively. A concurrent probing is much more efficient in recalling cognitive processes; therefore, they are the most often used in cognitive questionnaire tests [7] and usability tests [17]. However, some of its attributes raise questions about its applicability in eye-tracking experiments, as the interviewer should be present and active during the completion of the questionnaire, which implies researcher interference, and may significantly impact the eye-movement data [1][18]. The think-aloud technique might reduce interviewer-induced bias, as well as provide concurrent data on the TSs' thought processes. Therefore, it may be tempting to ask the TS to voice their thoughts aloud as they fill out the questionnaire (see [19]). Still, think-aloud is not recommended during the completion phase, and most ETCI studies do not use it. First, many TSs are uncomfortable with thinking aloud, which poses significant risk to data quality. Second, thinking aloud might interfere with the task [13][20]. Third, most TSs need to be 'trained' for it [14][20], which might further affect their natural thought processes. Last but not least, thinking aloud alone in a room is rather unnatural in itself, which compromises the simulation. It is no wonder think-aloud faded from the standard methods of regular CI practice [13] as well.

Also related to the considerations above, another question on how to plan a CI study resurfaced with the introduction of eye-tracking. Qualitative research standards [21][22] disavow the presence of extra observers: it might bias responses or otherwise compromise the interview. This principle is reflected in recent CI standards [23], countering past practice [7][15]. As for recent ETCI studies, various observation practices have been applied. In some cases, the researcher was physically present during completion [24][19] or observed through a one-way mirror [25], while other studies took advantage of the technical capabilities of eye-tracking devices and used a separate observation room [2]. The latter two practices are in line with CI standards, but the former cannot be recommended. TSs should be alone during completion, for quality data can only be collected if the real-life situation is simulated as accurately as possible. Since the physical presence of an observer may affect the TSs' behaviour, it risks inducing bias, and thus compromising the data. Therefore, a two-phase, semi-sequential data collection protocol is recommended: the TS completes the interview alone, real-time observation happening in/from a separate room, which provides the ensuing one-on-one retrospective CI with clues for probing. This way, eye-tracking can compensate for the lack of nonverbal clue detection in a regular concurrent CI setting. Monitoring the TSs' facial and vocal expressions through a camera is also recommended (see e.g., [2]).

How and what to observe while test subjects complete the questionnaire

Observation should be done employing two additional computer screens: one for the eyemovements and another for the webcam stream [2]. However, not all gazes or fixations bear importance for pretesting purposes. To aid capturing and separating notable events, the observers should have a list of conspicuous eye movements: long fixations, skipping or disregarding a survey component, regressive saccades, etc. Eye movement patterns may differ across TSs due to individual reading skills and habits [24]. It is thus recommended that TSs start the session with reading a simple text, which can serve as a benchmark of their individual reading patterns that the observers can keep in mind during the observation.

If the observation is not assisted by computer reporting, contrarily to the interviewer-only protocol [2], we suggest that it should be done by at least a team of three: the interviewer and two observers (the team can be smaller if a computer report can be generated on peculiar eye-movements on the spot). A lot can happen, hence, be missed in the seconds of note-taking, so the interviewer's sole task is to watch and narrate the eye-movements without taking their eyes off the screen. The narration of the eye movements should focus on peculiarities that one of the observers can record with predefined signs on a printed questionnaire. Narration allows the interviewer to be fully informed for the ensuing CI and helps etching events in the interviewer's short-term memory, which makes it easier to use the notes and formulate effective probes in the CI. Accurate and prompt recording of peculiar eye-movements can be a rather challenging task in itself, therefore, the task of the other observer is to record the TS' responses. Watching the second screen and note the TS' nonverbal expressions is also the second observer's task. Although seamlessly performing this protocol requires practice, it prevents missing potentially important eye movements.

Probing: what to ask in the interview

After the TS has completed the questionnaire, the notes should be merged and shortly discussed by the team. Then the interviewer conducts the retrospective, one-on-one CI. In the course of the interview, besides the most common retrospective techniques (scripted and spontaneous probes [14][20]), two additional type of probes appear to be useful. First, it is recommended to ask questions on general (user) experiences at the beginning of the interview, for ignoring a negative experience or disturbing stimulus that may have affected the TS can mislead the analysis. Second, 'semi-spontaneous' (~observational [20]) probes based on the eye-tracking phase observations are essential for the interview [2]. Theme Coding regards interviewing as the first analytic step [26]. Considering that similar preliminary eye movement analysis is necessary for efficient probing, the observation phase can be added as a 'zeroth' step to the five-step Theme Coding analysis model. This joint, 'on-the-spot analysis' of eye movements by several observers is a clear advantage of our protocol. The same rules apply to ETCI probes as those of regular CIs: never interpreting the observations, they should be neutral, nondirective, etc. [7][13][14][20].

In case the study involves multiple interviewers, interviewing consistency across CIs has to be upheld by adequate interviewer training. Inclusion of eye-tracking in the data collection requires additional preparation and practice regarding the narrated observations and the probing derived thereof. As a joint exercise, in order to maintain a consistent level in the interviewers' preparation, interviewers (narrators) should interpret at least one eyetracking footage together. Furthermore, a previously compiled list of phrases, commonly used general probes, and expressions might well be useful during the interview as an aid in formulating semi-spontaneous and spontaneous probes.

Memory-joggers can improve the TS' recollection, mitigating a disadvantage of the retrospective CI protocol. The questions can be shown to the TS in a blank questionnaire (either printed [1][2] or on a display). Proper application of such tools in a CI as well as directing the TS' attention in the desired way also requires practice. Due to the complex and demanding task of ETCIs, it is essential to do full practice runs prior to the start of the data collection. In theory, showing the TSs cues from their gaze video could also aid probing. However, not only it may distract the TSs but they might also start fabricating explanations [11], that is, take an 'evaluator' instead of the required 'story-teller' role [26]. **3.4. How to analyse and interpret the data**

Extant ETCI data analysis practices have two major issues. First, they are inclined to conduct quantitative analysis across the whole sample (e.g. [1][2]), which is problematic on multiple levels. Quantitative results cannot be generalised as they only apply to the sample due to the nonprobability design [27], rendering them effectively useless. CI findings are factual, not statistical [26][28]; therefore, instead of prevalence-based prioritisation, equal importance should be attributed to each observation. Quantitative analysis should be restricted to within-case analyses of deviations from TSs' characteristic reading patterns as it might help spotting less obvious problems and prevent false alarms. Optimally, such individual reports are generated on the spot to assist CI probing. However, this is not always feasible, for the entire questionnaire has to be prepared for the eyetracking software with its components thoroughly (pre)defined. Second, some ETCI studies use eye movement data to validate and visually underpin CI analysis results, and identify 'main' difficulties [2]. Conversely, the two datasets should be used in a complementary manner: findings, regardless of the method that captured them, should be treated as facts in themselves. Their appearance in both datasets helps to better understand, not validate or prioritise among them.

Theme coding assumed a bottom-up approach to CI data analysis, moving away from the predefined four-stage question–response model [29][30][14]. The five-step analysis of Theme Coding enables an in-depth data reduction process, moving back and forth between raw textual data, summaries, themes emerging thereof, and conceptual claims [26]. ETCI data is mainly used in the first (and 'zeroth', observation & interview) and the second (summaries for each question) step. The latter should be divided in 2 phases. The first phase is a 'raw' analysis of eye-tracking data: thick description of observations and preliminary inferences drawn thereof. These are integrated with the CI data in the second phase. Verbatim interview excerpts are summarised so they reflect on and make sense of the findings of the first phase. Phenomena that eye-tracking cannot or did not capture (unique interpretations, problems, etc.) are also important to extract from the textual data. As for problems or other phenomena that are present in only one of the datasets, analysis is usually more informed regarding those identified in the interview excerpts. They have the potential for drawing scientifically sound inferences from them. In contrast, peculiar eye movements alone do not provide insight into their reasons [12], only allow hypotheses. However, they may shed light on noncognitive mental processes that TSs are unable to reflect on [2]. For instance, in the 2020 Hungarian Census Test, reverse saccades and long fixations on the reference period of a question was an overarching pattern, though problem was not reported in the CIs despite targeted probing. Our hypothesis was that the numeric format did not align well with the way people recall events of a fixed time period. This observation alone lead to changing the numeric format to a textual reference (‘in the last 7 days of April’), which is closer to everyday language use.

cONCLUSIONS

Capturing notable eye-movements do not necessarily indicate problems [2]. Interpretation of eye movements may be arbitrary or speculative without a CI and an adequate protocol. The paper presented a ‘how to’ guide for designing and conducting ETCI studies by a methodology-driven overview of essential decisions. Their discussion also revealed the true potential and limitations of eye-tracking enhanced cognitive interviewing and the way the new instrument impacts data collection and analysis.

rEFERENCES

[1] C. E. Neuert and T. Lenzner, Incorporating eye tracking into cognitive interviewing to pretest survey questions, International Journal of Social Research Methodology 19(5) (2015), 501–519.

[2] C. E. Neuert and T. Lenzner, Use of Eye Tracking in Cognitive Pretests, Mannheim: GESIS – Leibniz Institute for the Social Sciences (GESIS – Survey Guidelines) (2019). Accessed 10. October 2022. https://www.gesis.org/fileadmin/upload/SDMwiki/Neuert_Lenzner_Use_of_eye_tracking_13022019cc.pdf

[3] D. F. Gravem, N. Berg, F. Berglund, K. Lund and K. Roßbach, Test report from Statistics Norway’s cognitive and usability testing of questions and questionnaires. Appendix B of WP4 Deliverable 3 of the MIMOD project. Deliverable for Work Package IV of the Cooperation on Multi-Mode Data Collection (MMDC) – MixedMode Designs for Social Surveys – MIMOD. Eurostat (2018). Accessed 26. March 2021. https://www.istat.it/en/research-activity/international-research-activity/essnetand-grants

[4] D. F. Gravem, V. Meertens, A. Luiten, D. Giesen, N. Berg, J. Bakker and B. Schouten, Final methodological report presenting results of usability tests on selected ESS surveys and Census. Smartphone fitness of ESS surveys – case studies on the ICT survey and the LFS. Deliverable for Work Package V of the Cooperation on MultiMode Data Collection (MMDC) – Mixed-Mode Designs for Social Surveys – MIMOD. Eurostat (2019). Accessed 26. March 2021. https://www.istat.it/en/researchactivity/international-research-activity/essnet-and-grants

[5] K. Miller, Cognitive Interviewing, in J. Madans, K. Miller, A. Maitland and G. Willis (eds.), Question Evaluation Methods: Contributing to the Science of Data Quality. Hoboken, NJ: John Wiley and Sons (2011), 51–75.

[6] K. Miller, S. Willson, V. Chepp and J-L. Padilla (eds.), Cognitive Interviewing Methodology. Hoboken, NJ: John Wiley & Sons (2014).

[7] G. B. Willis, Analysis of the Cognitive Interview in Questionnaire Design. New York, NY: Oxford University Press (2015).

[8] B. G. Glaser and A. L. Strauss, The Discovery of Grounded Theory: Strategies for Qualitative Research. New Brunswick & London: AldineTransaction (1967).

[9] K. Charmaz, Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Thousand Oaks, CA: Sage (2006).

[10] E. Jaakkola, Designing conceptual articles: four approaches, AMS Review 10 (2020), 18–26.

[11] C. E. Neuert and T. Lenzner, A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking, Social Science Computer Review 34(5) (2015), 1–15.

[12] M. Chauliac, L. Catrysse, D. Gijbels and V. Donche, It is All in the Surv-Eye: Can Eye Tracking Data Shed Light on the Internal Consistency in Self-Report Questionnaires on Cognitive Processing Strategies? Frontline Learning Research 8(3) (2020), 26–39.

[13] S. Willson and K. Miller, Data Collection, in K. Miller, S. Willson, V. Chepp and J-L. Padilla (eds.), Cognitive Interviewing Methodology. Hoboken, NJ: John Wiley & Sons (2014), 15–33.

[14] G. B. Willis, Cognitive Interviewing. A Tool for Improving Questionnaire Design. Thousand Oaks, CA, London & New Delhi: SAGE Publications (2005).

[15] D. Collins & M. Gray, Sampling and Recruitment, in D. Collins (ed.), Cognitive Interviewing Practice. London: Sage (2015), 80–100.

[16] F. Mújdricza, A kognitív kérdőívtesztelés módszertana: mintaválasztás és toborzás, Szociológiai Szemle 28(2) (2018), 4–27.

[17] E. Geisen & J. R. Bergstrom, Usability Testing For Survey Research, Cambridge, MA: Morgan Kaufmann (2017).

[18] K. Pernice & J. Nielsen, How to Conduct Eyetracking Studies, Fremont CA: Nielsen Norman Group (2009). Accessed 10. October 2022. https://media.nngroup.com/media/reports/free/How_to_Conduct_Eyetracking_Studies.pdf

[19] E. Nichols, E. Olmsted-Havala, T. Holland & A. A. Riemer, Usability Testing Online Questionnaires: Experiences at the U.S. Census Bureau, in P. C. Beatty, D. Collins, L. Kaye, J-L Padilla, G. B. Willis & A. Wilmot (eds.), Advances in Questionnaire Design, Development, Evaluation and Testing, Hoboken NJ: John Wiley & Sons (2020), 315–348.

[20] J. d'Ardenne, Developing interview protocols, in D. Collins (ed.), Cognitive Interviewing Practice. London: Sage (2015), 101–125.

[21] Y. S. Lincoln. and E. G. Guba, Naturalistic Inquiry. Beverly Hills, CA: SAGE (1985).

[22] S. Kvale, InterViews: An Introduction to Qualitative Research Interviewing. Thousand Oaks, CA, London & New Delhi: SAGE Publications (1996).

[23] Office of Management and Budget (OMB), Statistical Policy Directive No. 2 Addendum: Standards and Guidelines for Cognitive Interviews. Office of Management and Budget (2016). Accessed 11. August 2021.

https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/directive2/final_addendum_to_stat_policy_dir_2.pdf

[24] J. K. Höhne, Eye-Tracking Methodology: Exploring the Processing of Question Formats in Web Surveys, International Journal of Social Research Methodology 22(2) (2019), 199–206.

[25] J. C. R. Bergstrom, S. Lakhe & C. Erdman, Navigation Buttons in Web-Based Surveys: Respondents' Preferences Revisited in the Laboratory, Survey Practice 9(1) (2016), 1–12.

[26] K. Miller, S. Willson, V. Chepp & J. M. Ryan, Analysis, in K. Miller, S. Willson, V. Chepp and J-L. Padilla (eds.), Cognitive Interviewing Methodology. Hoboken, NJ: John Wiley & Sons (2014), 35–50.

[27] J. Lewis, J. Ritchie, R. Ormston & G. Morrell, Generalising from Qualitative Research, in J. Ritchie, J. Lewis, C. McNaughton Nicholls & R. Ormston (eds.), Qualitative Research Practice. A Guide for Social Science Students and Researchers, London: SAGE (2014) 347–366.

[28] F. Mújdricza & M. Földvári, A kognitív kérdőívtesztelés módszertana: a kognitív interjúk elemzése, Statisztikai Szemle 96(6) (2018), 545–574.

[29] R. Tourangeau, Cognitive science and survey methods: a cognitive perspective, in T. B. Jabine, M. L. Straf, J. M. Tanur and R. Tourangeau (eds.), Cognitive Aspects of Survey Design: Building a Bridge Between the Disciplines. Washington: National Academic Press (1984), 73–100.

[30] G. J. M. E. Snijkers, Cognitive laboratory experiences. On Pre-Testing Computerised Questionnaires and Data Quality. Ph.D. Dissertation. Utrecht:

Universiteit Utrecht (2002).

Demography (JENK 1A.2)

Session Chair: **Lewis Dijkstra** *(European Commission)*

Avoiding the avoidable: Simulating long-term population changes in Germany without avoidable mortality
Jannek Mühlhan *(German Federal Statistical Office-DESTATIS)*

Estimation of mode effects in the French population census
Loreline COURT (INSEE)*; Simon QUANTIN (INSEE)

Geographic Infrastructure for the 2021 Population and Housing Census
Rossano Figueiredo (Statistics Portugal)*; Ana Santos (Statistics Portugal); Bartholomeus Schoenmakers (Statistics Portugal)

Avoiding the avoidable: Simulating long-term population changes in Germany without avoidable mortality

Keywords: Public health, avoidable mortality, dynamic microsimulation, caregiving

INTRODUCTION

The population in Germany is undergoing a rapid ageing process. The number of older people and their share of the population are increasing dramatically. This development has a direct impact on the number of people in need of care. Between 1999 and 2019, the number of people in need of care increased from around 2 million to over 4 million. Already now, employees in the care sector are reporting a care crisis. For the next few years, when the so-called "baby boomers" of the high-birth-rate cohort born between 1955 and 1970 reach the age of care, a further steep increase of people in need of care is imminent. Forecasts predict over another million people in need of care by 2040 [1]. In the same time, the workforce will decrease significantly.

A decisive factor for the development of the number of old-age people in need of care is the quality and effectiveness of health care and prevention policy. One indicator that maps both aspects, health care and prevention is avoidable mortality [2]. Avoidable mortality is widely used as indicator in health monitoring. The indicator covers specific causes of death among people under 75 that would not have occurred if more effective public health and medical interventions had been in place. In 2017, two out of three deaths of people under 75 could have been avoided in the EU. In total one million death were avoidable [3].

We use the dynamic microsimulation model MikroSim to simulate the long-term changes in population dynamics if all avoidable deaths were prevented in Germany.[9] Dynamic microsimulation allows simulating the development of a population under specific scenarios considering complex interactions and consequences at the micro level. The results show that 15 to 35 percent of deaths between the ages of 15 and 65 are avoidable. However, in absolute numbers, most avoidable deaths happen above the age of 50. Preliminary simulation results indicate that compared to the expected increase of people in need of care due to demographic changes, the additional increase due to abolished avoidable mortality is negligible.

METHODS

Different lists of diagnoses serve as a foundation for the calculation of avoidable mortality. Thus, depending on the data source different lists may apply, which can lead to fundamental differences. We use the list of diagnoses of the Permanent Working Group of the Highest State Health Authorities (AOLG) of Germany to measure avoidable mortality [4]. This list considers the high focus on data for small areas of the health monitoring of the federal states. Table 1

[9] This is a far-reaching and unrealistic scenario. Further gradations for more detailed analyses are feasible in the ongoing work.

presents the list of the AOLG as well as the respective age and sex for which the definition considers the diseases as avoidable deaths. We use age and sex specific shares of avoidable deaths for each of the 16 German federal states.

Table 1: Avoidable mortality (ICD 10)

Disease	ICD 10	Age/Sex	Sex
Malignant neoplasm of trachea, bronchus and lung	C33 - C34	15 – 64	total
Malignant neoplasm of breast	C50	25 – 64	female
Malignant neoplasm of cervix uteri	C53	15 – 64	female
Ischaemic heart diseases	I20 - I25	35 – 64	total
Hypertensive and cerebrovascular diseases	I10 - I15 I60 - I69	35 – 64	total
Diseases of liver	K70 - K77	15 – 74	total
Transport accidents	V01 - V99	total	total
Perinatal mortality	A00 - T98	Stillbirths and deaths in the first 7 days	total

Source: List according to the permanent working group of the Highest State Health Authorities (AOLG) [4]

In the simulation analysis, we utilize the multi-sectoral and regional dynamic microsimulation model MikroSim to simulate long-term changes of the German population if avoidable deaths could be completely avoided. In absence of an appropriate dataset, the MikroSim research group created a synthetic population of Germany for the purpose of microsimulation. Based on this dataset, in each simulation year, individuals and households run through a set of simulation modules such as mortality, fertility, regional mobility, education and employment [5]. In order to obtain a world without avoidable mortality from 2019 onwards, we manipulate the individual probabilities of death according to regional sex and age specific shares of avoidable deaths in 2019 and simulate population changes until the year 2060. We will present the simulation results on a small local scale using the interactive R-Shiny-Application “MikrosimulatoR”.

RESULTS

In 2019, about 53 thousand of 940 thousand deaths are classified as avoidable according the list of the AOLG. This corresponds to a share of about 5.6 percent. Figure 1 presents the number of avoidable deaths per age group and sex. It shows that most avoidable deaths occur between the ages of 50 and 65 and men are more likely to be affected than women are. In all EU member states for men the avoidable death rates are higher than for women [3].

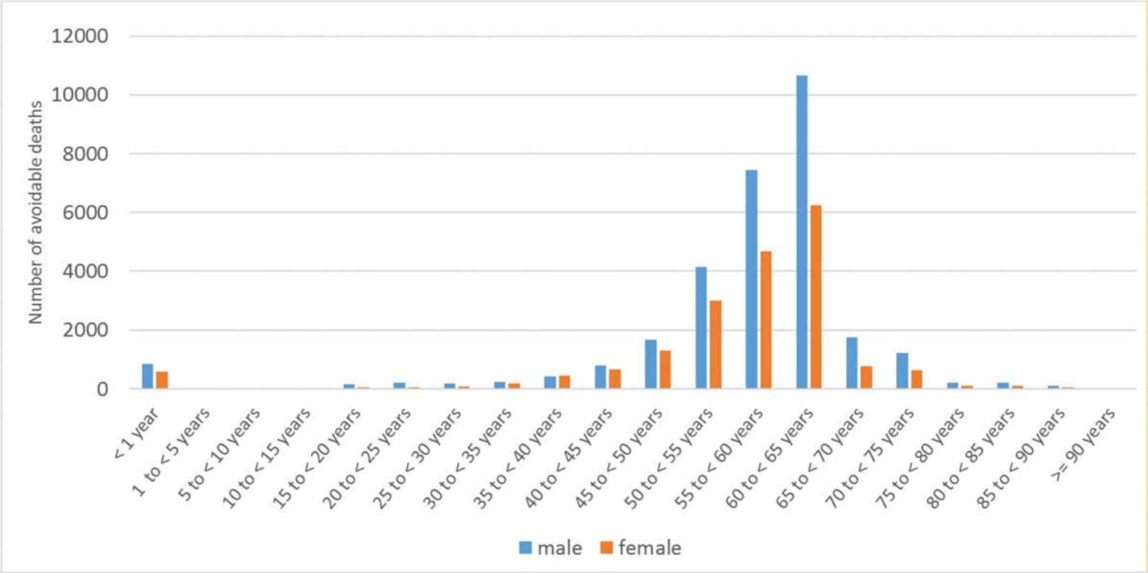


Figure 1. Number of avoidable deaths per age group and sex

However, when looking at the proportion of avoidable deaths to all deaths presented by Figure 2 we see that a significant share of deaths are avoidable for all age groups until the age of 75. For the age groups between 15 and 64 years at least 15 percent of deaths are avoidable. The highest proportions are reached for those aged between 50 and 64 years with almost 35 percent of deaths avoidable. For children aged between 1 and 15 years the only cause of death classified as avoidable are transport accidents, which cause about 5 percent of deaths in this age group. Transportation accidents are also for persons aged 75 or older the only cause of death counted as avoidable, they account for less than 1 percent of all deaths in this age group. In addition, for the age between 65 and 74, the AOLG counts persons dying from diseases of liver as avoidable deaths. Approximately 3 percent of all deaths are avoidable in this age group. One in four deaths of under 1-year-olds is avoidable. Newborns who die within the first 7 days of life and infants dying in transportation accidents account for this proportion. Stillbirths are also classified as avoidable, but not yet considered in the simulation. Although the total number is higher among men, the proportion of avoidable deaths is higher among women over the age of 30.[10] There are no major differences between the German federal states; only men from the eastern Germany have a slightly higher risk to die from an avoidable cause of death.

[10] No proportions per gender are available yet for under-1-year-olds.

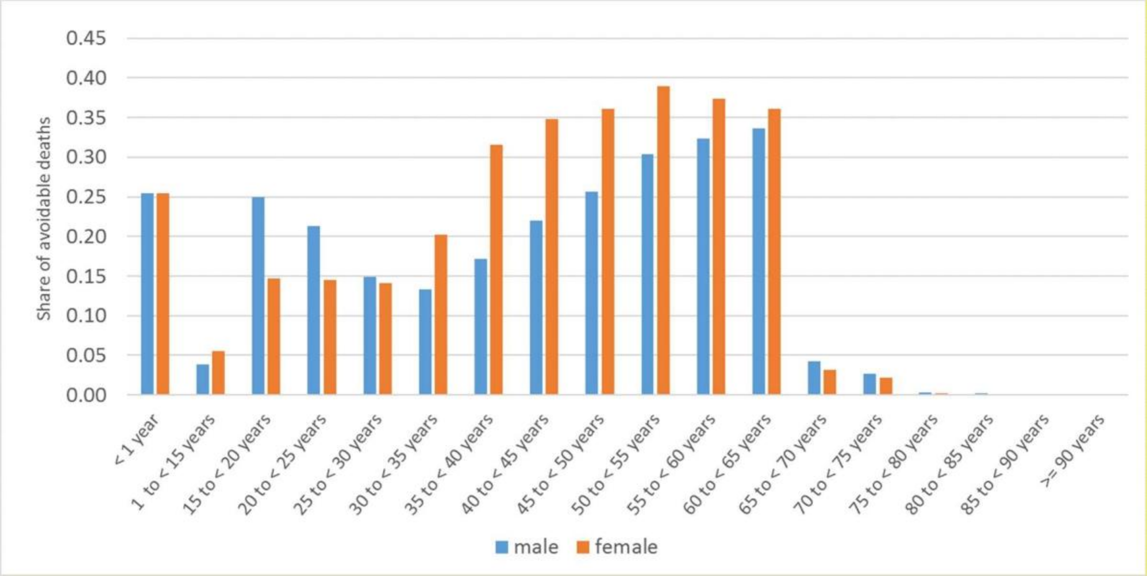


Figure 2. Share of avoidable deaths per age group and sex

Very preliminary and incomplete simulation results show that, as expected, compared to the baseline, until 2040 the population increases if avoidable deaths could be avoided. Avoiding avoidable deaths also increases the number of persons in need of care. However, compared to the simulated strong increase of people in need of care in the baseline scenario, the additional increase in the scenario without avoidable deaths appears comparatively low. However, avoidable deaths seem not to noteworthy affect the workforce in the first 20 twenty years of simulation. These preliminary results indicate that an increase in quality and effectiveness of health care and prevention policy resulting in an decrease of avoidable death would not lead to an fundamental change in the development of persons in need of care. However, the simulated required capacities in old age care for the baseline scenario illustrate the impending shortcomings in nursing care.

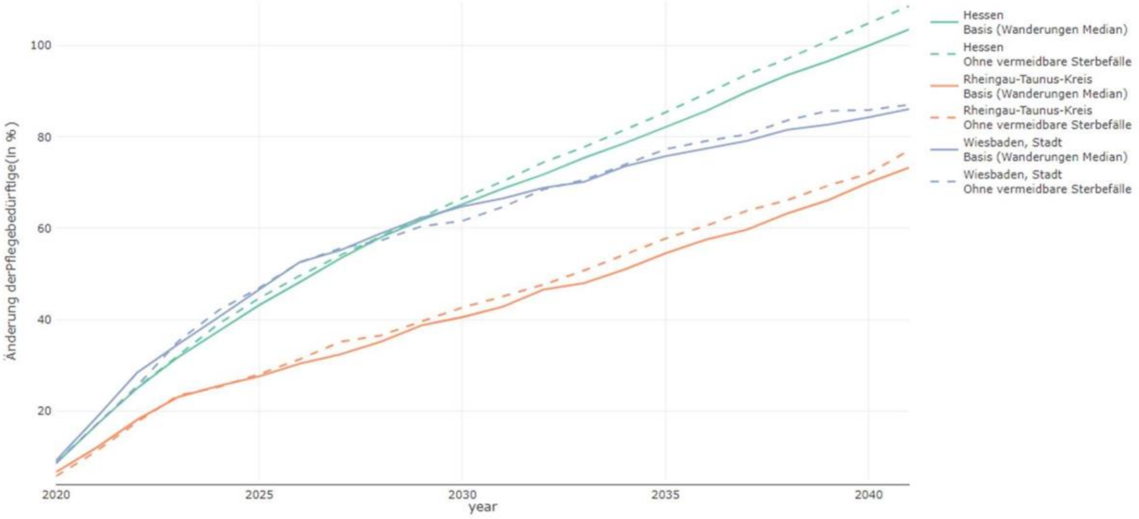


Figure 3. Simulated change of persons in need of care for exemplary regions in Germany in baseline and scenario

Figure 3 serves as exemplary presentation of the simulation results using the R-ShinyApplication MikrosimulatoR. It shows the simulation results for the federal state of Hessia and the two districts Wiesbaden and Rheingau-Taunus. The MikrosimulatoR allows to interactively showing the simulation results for different indicators, scenarios and regions. It will be published as part of "Exdat", the experimental data on the website of the Federal Statistical Office of Germany.

cONCLUSIONS

Germany is undergoing a demographic transformation and an aging of the population, which will accelerate and continue for some years to come. A strong increase in the number of persons in need of care will be a major challenge in the next decades. We simulate which effect an improved public health system resulting in decreasing numbers of avoidable deaths, we simulate here the extreme situation of total prevention of avoidable deaths, would have on the population and number of persons in need of care until 2040. Although a remarkable share of deaths are avoidable for all age groups until the age of 65, the most avoidable deaths happen above the age of 50. Preliminary simulation results show that, abolishing avoidable mortality increases the number of people in need for care slightly, while the effect on the workforce is negligible. However, compared to the already expected increase in the number of persons in need of care the additional increase is small. Preventing avoidable deaths would therefore neither the solution nor a major additional issue for the imminent caregiving crisis.

rEFERENCES

[1] BIB, https://www.demografie-portal.de/DE/Fakten/pflegebeduerftige.html (access on 14th October 2022).

[2] A. Weber, V. Reisig, A. Buschner and J. Kuhn, Vermeidbare Sterblichkeit – Neufassung eines Indikators für die Präventionsberichterstattung, Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 1 (2022), 116-125.

[3] Eurostat, Preventable and treatable mortality statistics - Statistics Explained (2022).

[4] AOLG, Indikatorensatz für die Gesundheitsberichterstattung der Länder (2003).

[5] R. Münnich, R. Schnell, H. Brenzel, H. Dieckmann, S. Dräger, J. Emmenegger, P. Höcker, J. Kopp, H. Merkle, C. Neufang, M. Obersneider, J. Reinhold, J. Schaller, S. Schmaus, and P. Stein, A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model, methods, data, analyses, 15(2) (2021), 241-264.

Estimation of mode effects in the French population census

Keywords: Mixed-mode and web data collection, Non-response, response propensity, respondent behaviour and response burden

INTRODUCTION

1.1. Survey design for the census in France and measure of employment

In France, since 2004, the population census is carried out by sampling over a five-year cycle, using the size of the municipality to define the sample design. Thus, each year, an annual census survey (ACS) is conducted among a sample of dwellings, in which each inhabitants must respond[11]. Indeed, the questionnaire provides information about the sampled dwelling (type of dwelling, type of building, number of rooms, sex, age and number of inhabitants, etc.) based on a stand-alone questionnaire, but also statistics on *each* inhabitants characteristics (occupation, citizenship, mode of transport…) based on an individual questionnaire. Since 2015, a concurrent mixed-mode design has been introduced. Even if sampled dwellings are offered the choice between paper and internet mode, within each dwelling all individuals must respond with the same mode.

In this article, we focus on the employment status which is only asked to people over the age of 14. Indeed, in France, the Annual Employment Estimates (AEE), based on several administrative sources, are the reference source for monitoring the employment situation. Until 2014, employment estimates provided by the census showed a constant difference of 700,000 jobs with the reference source (AEE), but they evolved in parallel. Since 2015 and the introduction of the mixedmode survey, the gap widened to 1,500,000 in 2019.

Two mode effects may explain this phenomenon. Selection effects, caused by the selection mechanism of the mixed-mode survey design, might have increase non-response in ACS. Recorded responses to the survey questions might also differ due to specifics of the modes employed, as respondent are potentially giving different answers to the same questions in different modes ([1]). The aim of this study is to determine whether or not the growing gap between the two series (AEE and ACS) is due to the introduction of the mixed-mode survey design and to try to disentangle measurement and selection effects, especially on non-response.

1.2. Mixed-mode ACS and non-response

Total non-response in ACS occurs when all inhabitants of a sampled dwelling fail to response. In that case, no information, including characteristics of the dwelling, is available. This only happens if the dwelling has changed status (e.g., destroyed or vacant) or if there is no sign of life during the collection time, thus total non-response will be considered missing at random and ignored for this study. Partial non-response on employment, however, encompasses two main

[11] Over five years, residents of municipalities with less than 10,000 inhabitants are all surveyed, whereas only approximately 40% of the population of municipalities with 10,000 inhabitants or more are surveyed.

reasons for non-observations. First, an inhabitant may choose not to answer the whole individual questionnaire (hereinafter referred to as *individual nonresponse*): in that case, at least, (some) characteristics of the dwelling are available, but also age and sex of every inhabitants. Second, an inhabitant responds to the individual questionnaire but incompletely (hereinafter referred to as *item non-response*). Choosing one mode or another might affect every type of non-responses.

Moreover, census individual questionnaire contains several questions on employment. Occupation (employed, apprenticeship, student, unemployed, retired, home-maker or other) is mainly measured, for each inhabitants over 14 years, with two questions : "What is your main situation regarding employment?" and "Do you currently work?". Afterwards, if the person declares to be employed or currently working, several questions about his/her job are asked (place of work, type of job, type of contract...). All these questions can of course be answered with the paper form, but not with the online questionnaire if one has first declared not to be in employment. Item and individual non-response are treated in post-collection processing. A mode-effect might then occur as item non-response is imputed by using available responses (if any) to other employment questions with the paper form but not with the online questionnaire. Moreover, individual non-response is imputed by hot-deck method disregarding the mode used.

METHODS

Estimation of mode-effects is usually obtained with matching methods between internet and paper respondents. However, identification with such an approach relies on the respect of the strong hypothesis of conditional independence. This assumption might no longer be valid when there is an unobserved characteristic affecting both the selection and the response, a common situation when all inhabitants of a dwelling don't participate, as it will relies mainly only on the characteristics of the dwelling.

The sensitivity analysis model proposed by Rosenbaum ([2], [3], [4]) that we will implement consists precisely in evaluating the impact of a relaxation of this conditional independence hypothesis by considering, for example, that after matching, one of the two respondents/dwelling, not necessarily the one using internet, is still twice as likely to use this mode. More precisely, the approach we will implement here tests the hypothesis of the existence of an effect on participation and employment response, and quantifies the extent of the unobserved selection bias that would lead to the disqualification of any causality in the correlation revealed under the conditional independence hypothesis.

FIRST DESCRIPTIVE RESULTS

This section only presents descriptive results as the estimations are yet to be finalized but will be of course available for the conference.

Since 2015, non-response behaviours are sensitive to the mode. With the paper form, individuals have a high level of item non-response (Figure 1). This item non-response of paper respondents is partly corrected afterwards, as seen above. With online questionnaire, however, item non-response is almost nonexistent thanks to visual incentives. On the other hand, individual non-response is increasing as online form becomes more frequently used by sampled dwellings.

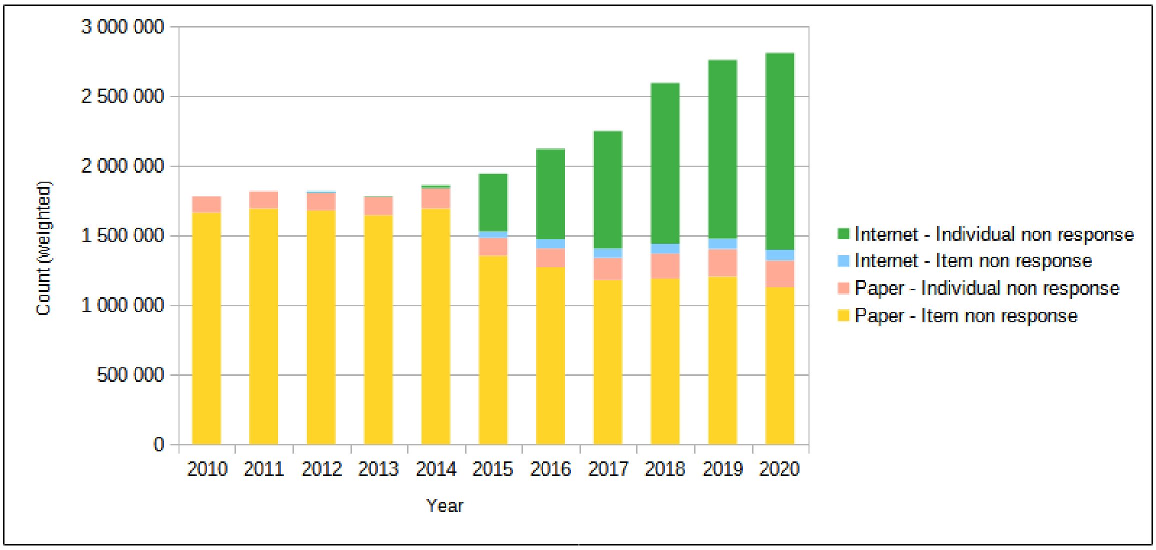


Figure 1: Individual and item non-response by mode Source : ACS (2010-2020)

Individual non-responses are imputed by hot-deck methods disregarding the chosen mode. As a result, it doesn't take into account the selection bias induced by mixed-mode and relies more and more on paper form respondents who are less likely to be employed. Thus, the proportion of online surveyed individuals imputed in employment by post-collection processing is decreasing (Figure 2).

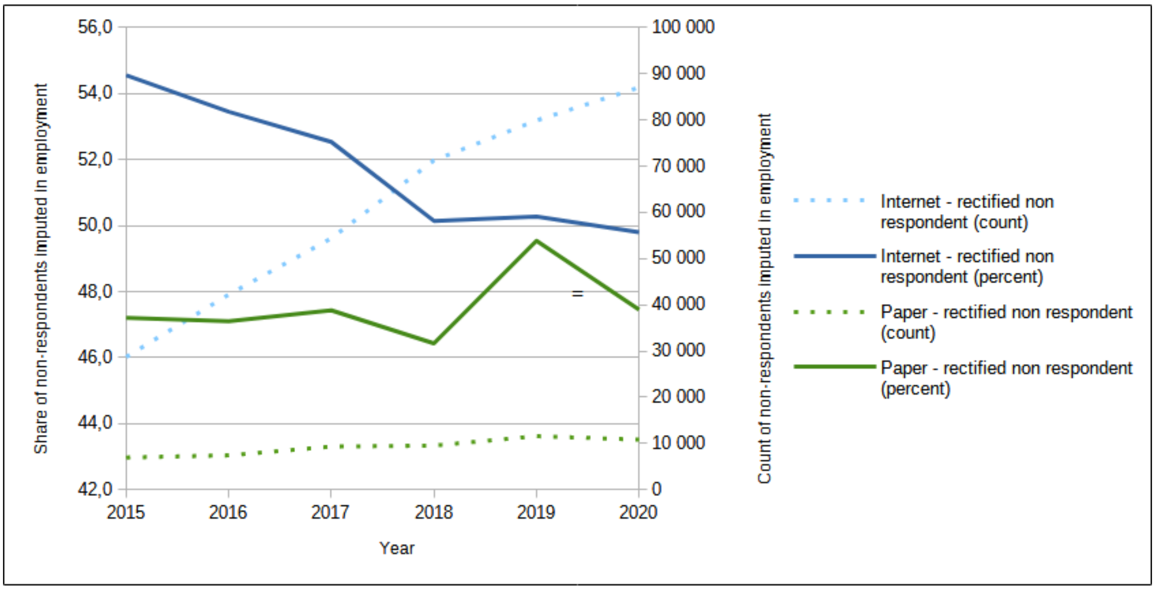


Figure 2: Proportion of individual non-respondents imputed in employment Source: ACS (2015-2020)

cONCLUSIONS

The aim of this article is to investigate the selection bias induced by the mixedmode design of the census survey, on the measure of employment. As individual non-response implies few observables characteristics, it is important to implement a method which can quantify the impact of the unobserved selection bias on our conclusions. We focus on participation bias as reweighting methods might be implement to correct the estimation.

rEFERENCES

[1] *Position paper on mixed-mode surveys*, Statistical working paper, Eurostat, 2022.

[2] Paul R. Rosenbaum: *Design of observational studies*, Springer Series in Statistics, 2010

[3] Paul R. Rosenbaum: *Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies*, Biometrics, 63:456-464, 2007

[4] Paul R. Rosenbaum: *Observational Studies*, Springer Series in Statistics, 2002

Geographic Infrastructure for the 2021 Population and Housing Census

Keywords: *census; small statistical areas; geographic information; buildings.*

Introduction

Like in many countries, the population and housing census is the cornerstone of official statistics in Portugal and a key statistical operation within all statistical activities, providing exhaustive, representative and geographically detailed statistics at the country level [1]. Portugal has been conducting census operations since the mid-19h century and following international and regional recommendations, particularly related to population and housing censuses which had their first simultaneous operation in 1970.

Maps have been supporting census data collection at Statistics Portugal since 1981 as a working tool to define small areas, firstly only for major cities and villages. In the 1991 Census, Statistics Portugal expanded the use of paper-based maps for all national territory recognising the growing need for statistical data at smaller area level, namely broken down to a level below administrative division. The 1991 Census was the first time that enumeration areas and blocks were defined for statistical purposes based on the administrative division at the parish level (4 208 parishes in 1991). The non-electronic cartography was designated “Spatial Referencing Geographical Base (BGRE 1991)” and it was used for the organisation and monitoring of data collection and for representing the spatial distribution of statistical data at a small area level.

Later in 1995, Statistics Portugal started the preparation of the digital geographical infrastructure within a Geographic Information System (GIS) environment to support the 2001 Census. This new geographical infrastructure designated as “Geographic Information Referencing Base” (BGRI 2001) was firstly digitalised and georeferenced from the BGRE 1991 and further edited according to administrative delimitation and housing criteria. Also, this digital geographic reference system containing administrative units (parishes) broken down into small statistical areas – statistical sections and subsections – has been supporting census processes so far and, at the same time, used as a tool for census data dissemination. For the years 1991 and 2001, census data was only published for the census enumeration units (statistical sections).

During the 2011 Census, a point-based spatial dataset of georeferenced residential buildings (Geographical Building Base - BGE2011) was created by enumerators with a web application called “GeoEdif”. This was the first national spatial dataset with the location of residential buildings enabling data aggregation processes and flexible and comparable geospatial statistics across space and time [2].

After the 2011 Census, several methodologies were developed to assure that this new geography could be used for the coming censuses. A National Buildings and Dwellings Register file (FNA) used for social surveys was implemented based on BGE locations and Census microdata. Through the Urban Operations Indicators System (SIOU), which registers building permits and completed construction works based on administrative data from municipalities [3] it was possible to continuously update the spatial dataset of residential buildings between census series (usually every 10 years). This procedure ensures that the stock of information on buildings and dwellings is updated for the Census – and for other statistical surveys – by adding newly constructed buildings and dwellings and excluding the demolitions.

Furthermore, Statistics Portugal has been developing the Spatial Data Infrastructure (SDI) to support other statistical activities in a permanent effort to enhance geospatial aspects – e.g., data, tools and services - in the statistical production process, including sample design, data collection, data integration, and data visualisation and dissemination phases.

Methods

To support the Census 2021 fieldwork, two sets of geographic data at national scope were created and/or updated: i) Enumeration areas (Geographic Base of Information Referencing – BGRI2021); and ii) Census buildings (Geographic Base of Buildings - BGE).

The methods used for the 2021 Census concerning the geographic infrastructure were mainly based on simplifying the census geography (BGRI) throughout the edition process. This line of work aimed to reduce the complexity of the geometry and the number of statistical subsections (blocks) to tackle the excessive territorial breakdown meanwhile ensuring confidentiality issues. This editing work was made using ESRI tools.

In this regard, the subsections were firstly defined following natural and/or anthropic boundary criteria (e.g., rivers, roads, etc.) supported by ortophotomaps, in which the enumeration areas containing urban areas were mainly delimitated using transport routes. Consequently, the redefined statistical subsections were merged, particularly in rural areas due to dispersed settlement patterns as they correspond to census locality (built-up area with its designation and with 10 or more housing units) or part of a locality. In urban areas, the aggregation of statistical subsections was less significant since most of them corresponded to well-defined neighbourhoods and there was no clear requirement to redefine their boundaries.

The definition of statistical sections was based on a developed QGIS plugin "IneSections" that allowed to automatically create the statistical division considering the classification of urban areas (2014) and the delimitation of statistical subsections within the same parish. This aggregation procedure was very time-efficient due to its automation capability.

The following table presents the number of enumeration areas and blocks from the 1991 Census to the 2021 Census.

Table 4. Enumeration areas, 1991 Census to 2021 Census

	1991	2001	2011	2021
Enumeration areas (statistical subsections)	13 705	16 094	18 074	10 401
Blocks (statistical sections)	106 626	177 893	265 995	210 170

The BGE for the 2021 Census (BGE2021) culminated from the work done in the last decade based on geometric edition processes performed in-house based on municipality licensing permits, data from social surveys and administrative data sources. This database holds the geographic location attributes (coordinates and addresses) and identifying attributes of all the buildings.

The BGRI2021 and BGE2021 are visualized in the geographic module of the “eRecenseador” application and are represented on the cartographic media in PDF format, and printed on paper, making it possible for the census taker to visualize the following elements in the respective monitoring area:

- the boundaries of the statistical section assigned to him/her and its statistical sub-sections;
- the representative points of the georeferenced buildings;
- the administrative and statistical codes to be used.

The main changes of the 2021 Census compared to previous census operations were: i) the existence of a list of buildings and dwellings; ii) the distribution process in which the delivery of a letter containing the necessary information to respond via the Internet; iii) the Internet as the main response channel; iv) the pre-filling of some variables at building level; v) the simplification of the completion and codification of open questions; vi) the reinforcement of information and communication technologies in the fieldwork (mobile using location-based technology for field operations); and vii) the increased use of administrative data as auxiliary data (e.g., age distribution, foreign population and for quality of addresses, etc.). In general, these changes aimed to improve the quality of collected data, simplify procedures, and increase the global efficiency of the operation in terms of data collection and processing. They also intended to make the filling of the questionnaires easier and more convenient for the users, reduce the data collection costs through fewer interviewers in fieldwork and reduce the use of paper considering the digital transformation paradigm.

In the context of dissemination, a web page for the download of geopackage file was developed with the boundaries of the small statistical areas (BGRI2021) and the provisional data from the 2021 Census with the main census variables: buildings, dwellings, households, and individuals (resident population). In addition, a QGIS plugin designated “downloadGeoStatPortugal” was developed to ensure easier access and use to census reference data from 1991 to 2021 at national, regional (NUTS) and local (municipality) levels. This QGIS plugin is already available in the QGIS modules’ repository.

Results

During the field data collection process, enumerators georeferenced new buildings and changed the geographic position of existing buildings. This work was carried out using the “eRecenseador” application (online and offline), installed on the smartphone (IOS and Android). The boundaries of the small statical areas (BGRI2021) remained unchanged. This procedure enabled addressing data updating and timeliness issues.

The results of the above-described methods can be present throughout the statistics based on the response mode, including electronic reply, paper and telephone. Most of the replies were electronic (99.3%), in which 87.6% responded via the internet, 7.6% used the census mobile application (*eCensos*) and 4.1% used the E-desk located in the parishes’ councils. The paper answers were very low (0.4%) compared to the 2011 Census, which recorded around 50% of non-digital replies. Lastly, the helpline registered 0.3% replies via telephone handling 200 thousand calls and 40 thousand of emails concerning support requests and/or survey questions related to filling in the survey.

Conclusions

In the context of modernising official statistics, the transition from a census traditional model to a more efficient census model based on administrative data (register-based model) several critical areas require immediate action. Some of these actions include access to administrative data sources, full coverage in all mandatory census variables, information on family structure and housing and disclosure for very fine geographical levels (grid 1km^2).

The availability of administrative data is still incomplete and, in some cases, unstructured, either due to the need for access still being protocolled or due to data that does not exist directly and needs to be worked on. In this regard, there is a clear need to further analyse and process the addresses and, through the georeferencing of buildings, develop a solution for disseminating population statistics for a geographic grid of 1km^2.

Thus, location information may play a more key role in combining different administrative data sources - including the private sector - and integrating them within the statistical production process, preferably at the record unit level and expand the analytical potential with greater geographical, demographic and socio-economic detail and completeness. Since Portugal does not have a single identification number used across the various public entities and government institutions, the location may be used as a common and underpinning reference enabling data integration from multiple data sources for census purposes. This will provide added value to statistical data based on a digital connection between a place (where), people (who) and activities (why) in a timely, detailed and reliable manner to support policies and decision-making.

References

[1] M. Srdjan, The 2020 round of population and housing censuses: An overview, Statistical Journal of the IAOS 36 (2020), 35-42.

[2] European Forum for Geography and Statistics and Eurostat, A Point-based Foundation for Statistics - Final report from the GEOSTAT 2 project, Eurostat ESSnet grant project (2017).

[3] C. Neves and F. Moreira, How to integrate statistical and geospatial data using administrative data on construction works – The case of Indicators of Urban Operations (SIOU), Statistical Journal of the IAOS 33 (2017), 627-635.

From Classifications to Artificial Intelligence: innovative services of the Publications Office (GASP1A.3)

Session Chair: **Franck Noël** *(Publications Office of the European Union)*

A successful collaboration on classifications and what is in it for the statistical community
Denis Dechandon *(Publications Office of the European Union)*, Christine LAABOUDI *(Eurostat)*

Machine Learning Support for Repository Services — Reception, Cellar, and Semi-automated Metadata Automation
Nasredine CHENIKI *(Infeurope),* Marc Küster *(Publications Office of the European Union)*

A successful collaboration on classifications and what is in it for the statistical community

Keywords: statistical classification; Linked Open Data; semantic interoperability; XKOS

INTRODUCTION

The European Data Strategy describes the vision of a common European data space, in which data could be used irrespective of its physical location of storage in the European Union in compliance with applicable law.

The establishment of domain-specific common European data spaces is proposed to pool and share data. In this framework, various domains can be covered by such common European data spaces (e.g., health, mobility, manufacturing, financial services, energy or agriculture), as well as hematic areas (European Green Deal or European data spaces for public administration or skills).

This submission, which is meant to be included in the session at the NTTS conference led by the Publications Office of the European Union (OP), covers one practical contribution to the establishment of a statistical data space. It illustrates a use case and focuses on a successful collaboration between two directorate-generals of the European Commission with positive and reusable results for the benefits of the statistical community.

A statistical classification or nomenclature is an exhaustive and structured set of mutually exclusive categories used to standardise concepts for the collection, compilation and dissemination of statistical data.

Eurostat has a high level of knowledge and experience in the development of classifications and is the custodian of several sectoral and transversal European statistical classifications. Eurostat is also responsible for covering the European dimension for international statistical classifications linked to the European ones under its responsibility (NACE, CPA, PRODCOM, and Combined Nomenclature) [3]. Each statistical classification typically exists in a statistical ecosystem, where it is normally interlinked with other classifications – either structurally, or by means of correspondence tables.

Since the early 2000s, statistical classifications and their correspondence tables used to produce European Statistics had been disseminated via RAMON, the “Eurostat Reference And Management of Nomenclatures” platform. As the platform hosting RAMON had to be phased out by the end of 2022, Eurostat seized the opportunity converting its statistical classifications into RDF format and exposing them as Linked Open Data (LOD) by the means of corporate semantic tools offered by OP to make them widely reusable by the stakeholders in the European Statistical System (ESS).

On the other hand, OP implemented the Cellar, a metadata and content repository to incorporate its publications and facilitate their reuse. The Cellar is based on semantic technologies, i.e. a framework consisting of multiple standards to share and reuse data. They normalise named resources in controlled vocabularies, which enables computers to talk and link to one another. Most controlled vocabularies used at OP and beyond are maintained, stored, disseminated and visualised using semantic applications further developed under the lead of OP.

METHODS

2.1. Statistical classifications from a Linked Open Data perspective

Eurostat has wide experience in data modelling with SDMX (Statistical Data and Metadata eXchange) [14], an international standard whose aim is to make it easier to exchange and share statistical data and metadata, supported and owned by seven international organisations, including Eurostat. In a first approach, Eurostat developed a script that converted the statistical classifications from RAMON in SDMX/XML and stored them into the SDMX Euro Registry [2]. In addition to the basic structural elements (identifier, code, label and parent code), the specific components (explanatory notes, classification levels, case law) were represented as SDMX Annotations [13], a flexible extension mechanism allowing organisation-specific metadata to be added to a SDMX Structural Artefact.

Eurostat based its second approach on the SDMX terminology, reinterpreted in the context of Linked Open Data. While there is no single formal RDF ontology that provides a full one-to-one equivalent for the SDMX Information Model, the most relevant ontology that can cover the modelling of statistical classifications is XKOS (Extended Knowledge Organization System) [1], an extension of SKOS (Simple Knowledge Organization System) [16] for representing statistical classifications that meet domain-relevant community standards and best practices. In relation to the SDMX artefacts, XKOS has the added advantage of being compliant with the semantic web technologies and allowing a richer description of the resources rendering them interoperable and machine-readable. In this second stage, only classifications developed and maintained by Eurostat as well as their correspondence tables were considered, provided that the target classifications are available in RDF.

For the storage and dissemination of our classifications in RDF, a suite of corporate semantic tools offered by OP was used.

2.2. From the availability of a corporate service offer to a successful collaboration for the benefit of semantic interoperability

The support provided by OP to the transformation of existing statistical classifications and correspondence tables maintained and disseminated by Eurostat into Linked Open Data builds on three operational pillars: (1) *reference data maintenance*, (2) *visualisation* and (3) *storing for sharing and re-use*. All of them are included in a service offer provided by OP around reference data management at a corporate level (European Commission) [8].

This offer includes the use of three semantic platforms:

- *VocBench* [6], designed to meet the needs of semantic web and linked data environments, this web-based collaborative semantic application enables the creation and maintenance of generic RDF datasets and, in particular, controlled vocabularies such as thesauri, classifications (nomenclatures), code lists. In addition to SKOS, the default data model in VocBench, it supports the combination of different ontologies for representing the data such as XKOS or COOS (the Core Ontology for Official Statistics) [15] applied for maintaining statistical metadata catalogues (e.g., statistical methodologies, standards or organisations). Furthermore, it eases the maintenance and validation of correspondence tables between two classifications or concordance between two versions of the same classification.

- *ShowVoc* [5], this additional web-based collaborative semantic application, allows an easier display and browsing of controlled vocabularies maintained with VocBench,
- *Cellar* [12], a large semantic dissemination repository (Triple Store), which exposes all OP documents (EU publications, EU legislation) and their metadata as LOD, and is the main point of interaction for systems and applications consuming the data.

The transformation into RDF of classifications and correspondence tables disseminated by Eurostat benefits from the use of the above-mentioned suite and from a strong and very constructive collaboration between relevant teams in both directorate-generals of the European Commission. As a result, the NACE classification is accessible in humanreadable formats through the EU Vocabularies website [10] and ShowVoc [11], as well as in machine-readable formats through Cellar. In this framework, OP provides a technical and theoretical support for the management of data in VocBench (in particular about the use of ontologies, the import and transformation of datasets from Excel into RDF triples, and the implementation of a persistent URI scheme), and monitors all new versions and other classifications to be added.

rESULTS

Thanks to some minor evolutions of VocBench, ShowVoc and EU Vocabularies to endorse the specificity of statistical classifications (in particular, various sorting levels and the display of explanatory notes), Eurostat statistical classifications are officially disseminated on the EU Vocabularies website, aggregated by classification families (economics, products, goods, prices, regional and geospatial statistics) under the business collection ‘Eurostat statistical classifications’ [10]. They are defined in the domain ‘data.europa.eu’, with one namespace by classification family (for example, http://data.europa.eu/ux2 identifying the NACE classification and http://data.europa.eu/ux2/nace2/ the NACE Rev.2), with a persistent URI (Uniform Resource Identifier) being assigned to each instance of the resource types: Classification, Classification item forming part of a Classification, Correspondence Table or Classification Level.

LOD compliant, they are transformed by Eurostat in RDF and uploaded in VocBench for their further dissemination. To this end, data is exported from VocBench and either directly imported into ShowVoc or further transformed to be ingested into Cellar (second transformation) to be disseminated in the Eurostat Business Collection of EU Vocabularies or as a dataset in the open data catalogue “data.europa.eu” [7].

cONCLUSIONS

This collaboration, which leads to the transformation of existing data into Linked Open Data, increases the opportunities for further collaboration between Eurostat and the ESS Members for developing, reusing and linking reference and derived classifications. The main challenge for enabling this interoperability remains the availability of these statistical classifications in RDF. A successful interoperability use case is the availability of correspondence tables established between international and EU statistical classifications (NACE – ISIC, CPA –

CPC), accessible on EU Vocabularies or on the Caliper platform [4], a project run by the Food and Agriculture Organization of the UN (FAO).

Standardisation at different levels (for instance metadata schema and data representation formats) is key to integrate data broadly, and to enable data exchange and interoperability with the overall goal of fostering innovation based on data. This refers to all types of data and data from different domains, and in particular data used for the production and modernisation of statistics. To this end, the application of standard and shared formats and protocols for gathering and processing data from different sources in a coherent and interoperable manner across sectors and vertical markets must be encouraged. It indeed eases the discoverability, retrievability and re-use of public sector information data. Additionally, data cross-reference and interoperability are facilitated using machinereadable formats and commonly agreed metadata. They can further be used by machine learning processes or consumed for automated classification processes, e.g. to derive the economic activities of businesses with NACE Code.

REFERENCES

[1] DDI (2019), XKOS – Extended Knowledge Organization System. Available at: https://ddialliance.org/Specification/RDF/XKOS

[2] Euro SDMX Registry. Available at: https://webgate.ec.europa.eu/sdmxregistry/

[3] Eurostat, The international system of economic classifications. Available at https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=NACE_background#The_international_system_of_econo mic_classifications

[4] FAO, Caliper – Statistical classifications in a Linked Open Data. Available at: https://www.fao.org/statistics/caliper/web/

[5] European Commission, Joinup – ShowVoc. Available at: https://joinup.ec.europa.eu/collection/semantic-interoperability-communitysemic/solution/showvoc

[6] European Commission, Joinup – VocBench. Available at: https://joinup.ec.europa.eu/collection/semantic-interoperability-communitysemic/solution/vocbench

[7] Official portal for European data. Available at: https://data.europa.eu/en

[8] Publications Office of the EU, Corporate Reference Data Management policy in the European Commission. Available at: https://op.europa.eu/en/web/euvocabularies/corporate-reference-data-management

[9] Publications Office of the European Union (2022), EU Vocabularies. Available at: https://op.europa.eu/en/web/eu-vocabularies

[10] Publications Office of the European Union (2022), EU Vocabularies - NACE Rev. 2. Available at: https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/nace2

[11] Publications Office of the European Union (2022), EU Vocabularies - NACE Rev. 2. Available at: https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/nace2

[12] Publications Office of the European Union (2022), Cellar: the semantic repository of the Publications Office. Available at: https://data.europa.eu/doi/10.2830/46176

[13] SDMX (2021), Guidelines on using SDMX Annotations (Version 1.0). Available at: https://sdmx.org/?page_id=4345#Annotations

[14] SDMX (2021), SDMX Technical Specifications. Available at: https://sdmx.org/?page_id=5008

[15] UNECE, Core Ontology for Official Statistics (COOS). Available at: https://linked-statistics.github.io/COOS/coos.html

[16] W3C (2019), SKOS Simple Knowledge Organization System Reference. Available at: https://www.w3.org/TR/2009/REC-skos-reference-20090818/

Machine Learning Support for Repository Services — Reception, Cellar, and Semi-automated Metadata Automation

Abstract

Annotating documents on a large scale with metadata from diverse taxonomies, with hundreds or thousands of daily documents is a hard and time-consuming task. Providing metadata recommendations for document annotation greatly supports the manual annotation task, which becomes more efficient and less costly. In this paper, we show how text embedding-based similarities combined with the reuse of existing annotated data, can improve the data annotation process of documents, including legal documents. Our approach ensures that only one model for text similarity is necessary instead of separate trained classifiers for each taxonomy. The improvement of such metadata recommendations is measured using several properties that demonstrated the usefulness of such automation.

Introduction

On behalf of the Institutions and bodies of the EU, the Publications Office of the European Union (OP) is responsible for the production and dissemination of electronic publications. These publications belong to different document collections: Official Journal and the acts, their summaries and consolidated legislation, pre-legislative documents, cases of the European Court of Justice and of national Court and general publications.

Clients want to be able to find the right publication among the thousands that OP publishes every day and the millions it has in its central repository. Associating quickly the right set of metadata to documents and allowing their efficient retrievals cross domains became a key factor to facilitate transparency and openness to the citizens. This paper intends to present the OP experience to reach this objective with some Artificial Intelligent services.

Inside OP it is in particular the "Repository services" (OP.A.2) unit who is in charge of the reception, registration and automatic validation of electronic documents and data via the CERES services as well the storage and dissemination of these documents in OP’s common repository, Cellar.

OP creates specific metadata that describe its documents precisely. This metadata may express highly specific business relationships between documents, e.g., in the legal domain, classify documents according to certain classification schemes, but can express more generic relationships between documents. The common metadata linked the published documents facilitate the documents retrieval, but the process to allocate metadata to documents is currently done either by operators, or through external contractors. This process is time-consuming task and money, and do not always allow to retrieve cross-domains documents.

Building on Artificial Intelligence, in particular on the Joint Research Centre (JRC)’s SeTA – Semantic Text Analysis – tool, the Publications Office investigates the possibility to facilitate the annotation process for the metadata operators.

Cellar – OP repository

The Cellar makes available at a single place all the metadata and digital content managed by the Publications Office in a harmonised and standardised way in order to:

- to guarantee to the citizen a better access to law and publications of the European Union;
- to encourage and facilitate reuse of content and metadata by professionals and experts; – to preserve content and metadata and access to contents and metadata over time.

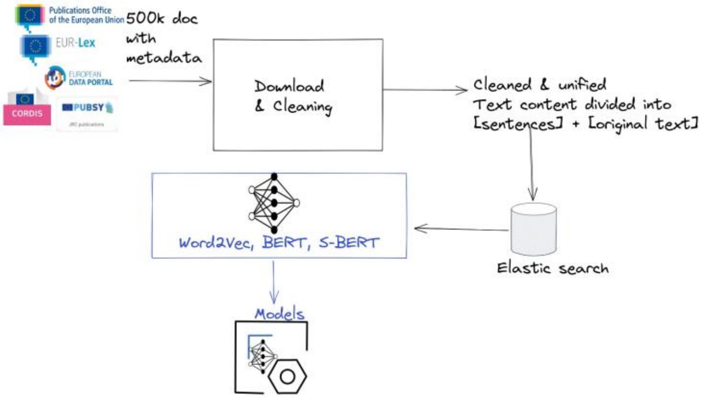
Cellar provides textual and structured data through open access to both metadata and content of documents. The metadata are openly provided and structured using W3C semantic Web standards: RDF, URIs (API with content negotiation), SPARQL, OWL ontology, etc. The metadata are formalised through an OWL ontology (CDM[12]) using RDF and linked data principles. The metadata in Cellar constitute a big knowledge graph detailing document description and interlinking them with clear semantics.

Various controlled vocabularies are used to label legal documents in Cellar. The CDM provides a variety of properties (predicates) for describing bibliographic resources (documents, agents, events, etc.). In our study we focus on the properties that are more likely to be related to the topic or the theme of documents. For this purpose, we identified a set of properties that fulfil our objectives. We selected the following metadata properties:

- *EuroVoc concepts*: EuroVoc[13], a multilingual interdisciplinary thesaurus, that allows assigning specific topics to the description of resources. With more than 8000 terms in EuroVoc thesaurus, selecting the correct values to annotate documents with an acceptable accuracy is a time consuming task, even for experts with knowledge about the content of EuroVoc and the documents to annotate.
- *rdf type*: generic document type. There are currently 505 document types to describe any document in Cellar. For instance, thematic domain, EuroVoc concept, etc.
- *Theme*: the subject of the publication – *Resource type*: the resource type of a work.
- *Subject matter*: a legal resource is about a concept expressed as a subject matter (classification tool). Very often this property is similar to EuroVoc concepts but is used for different purposes.

Semantic Text Analysis (SeTA) tool

SeTA is a tool that uses advanced text analysis techniques to help policy analysts understand the concepts expressed in large document collections, and to see



[12] Common Data Model: https://op

[13] https://op.europa.eu/en/web/eu-v

the relationships between these concepts and their development over time. SeTA uses recent machine learning techniques to compute an embedding vector for each document. Then, these vectors are

used to compute the similarity Figure 1: SeTA architecture between two documents as the

cosine of the angle between their corresponding vectors. Figure 1 shows the overall architecture of SeTA.

SeTA relies on language models (LMs) to compute documents representations which is called document embedding. LMs such as BERT [1] and its variants, e.g., DistilBERT [2], have obtained state of the art results in many monolingual and cross-lingual NLU benchmarks

[3].

BERT [1] and RoBERTa [4] had set a new state-of-the-art performance on sentence-pair regression tasks like semantic textual similarity (STS). However, it requires that both sentences are fed into the network, which causes a massive computational overhead. Sentence-BERT (SBERT) [5] is a modification of the pre-trained BERT network that is used to generate embedding for sentences or short texts up to 512 transformer tokens. SBERT is trained on a large corpus of English sentences and achieves state-of-the-art results on a number of sentencelevel tasks. The embeddings of any two texts can be compared e.g., with cosine-similarity to find sentences with a similar meaning. This can be useful for semantic textual similar, semantic search, or paraphrase mining. SeTA uses *all-distilroberta-v1*[14] sentence-transformers model which maps sentences and paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search.

Many legal documents are too long and do not fit SBERT model length limit. Based on authors findings in [6], we use the first 512 tokens of a document to compute the embedding which in most cases are enough in our document retrieval task. On the other hand, the first 512 tokens of a document are enough because the beginning of legal documents is mainly the summary and justification that is enough for document labelling task.

SeTA provides an API to access documents and their embeddings. Here is a list of the most important features for our work:

- computes the embedding of a given text or document.
- retrieve documents by EU corpus (EUR-Lex, CORDIS,etc.).
- get similar documents based on new document embeddings.

Metadata annotator@OP

In this section, we illustrate how SeTA and it’s semantic similarity calculation feature is used with Cellar to build an annotation recommender for OP publications. We also discuss the near future objectives to widely deploy, improve and adapt our prototype for different OP services.

[14] https://huggingface.co/sentence-transformers/all-distilroberta-v1

4.1 Phase-1: “Use of Artificial Intelligence in OP’s reception systems – study and prototype”

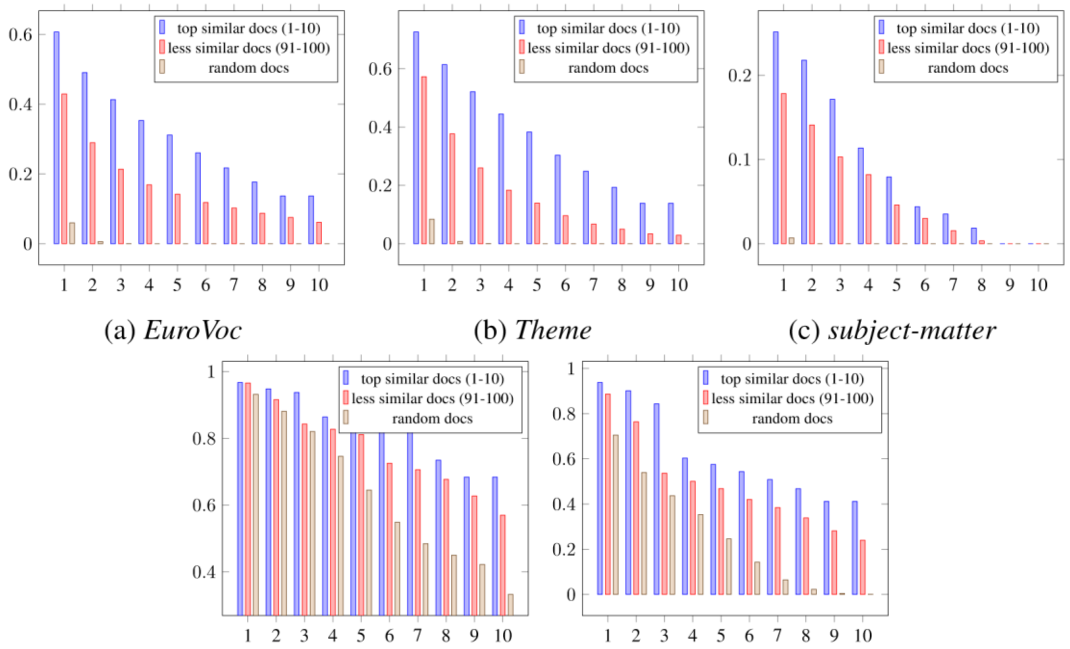
4.1.1Document similarity-based metadata enrichment

To implement our prototype, we used SeTA to generate document representations (based on SBERT language model). It takes the first 512 tokens from each document and generates an appropriate representation vector. These vectors are used to find similar documents. Metadata of these documents are then retrieved from Cellar to recommend candidates. Hence, we make the following hypothesis: ”Similar documents should have similar metadata”.

To validate our hypothesis, we conduct three times the following experiments:

– We chose randomly a set of 1000 documents from the Cellar repository (they already have their metadata).

– For each document, we get the first 10 most, less 10 (ranked 91-100) similar documents returned by SBERT, and 10 random documents for comparison.



(a) *EuroVoc*

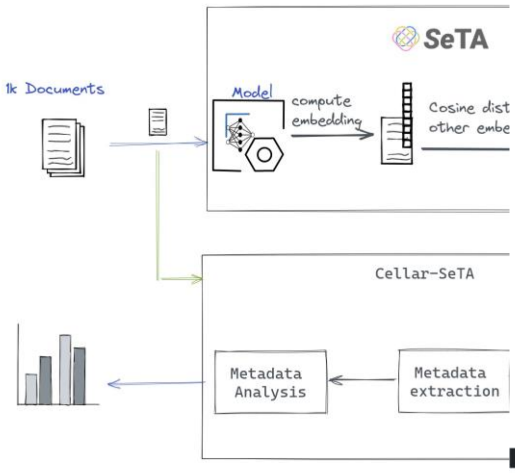
(b) *Theme*

(c) *subject-matter*

(d) *rdf:type*

(e) *resource-type*

Figure 2: SBERT metadata replication results. X-axis is minimum concept frequency, y-axis is recall



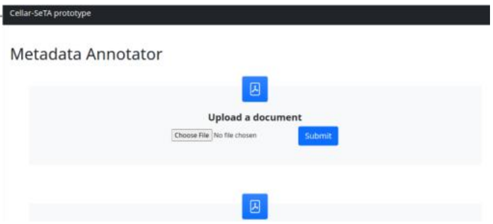


Figure 3: Metadata annotation process using SeTA

- – We check if there is any metadata in selected documents. Document itself is always discarded from the list of similar documents and never appears in the list of 10 selected documents. We take any shared metadata of similar documents. We define a hyperparameter L as an experiment parameter that filters metadata values by the frequency of their appearance in similar documents. If L=1, that means that the metadata value must appear at least once in the set of similar documents metadata. If L=10 the metadata value must appear in all 10 similar documents (i.e., all selected documents have this metadata value).

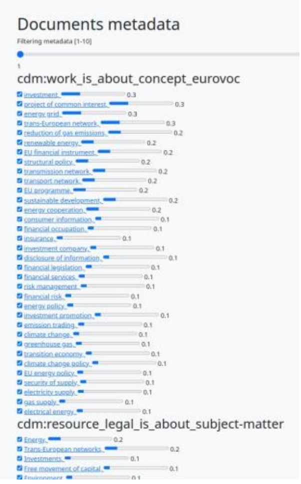
Figure 3 depicts the document metadata enrichment process.

4.1.2 Results

The results of the experiments show (see Figure 2) that we were able to retrieve significant amount of related metadata for various metadata properties such as *EuroVoc* (Fig. 2a) with 60%, *Theme* (Fig. 2b) with 70%, and *Subject-matter* (Fig. 2c) with 25% recall. However, for *rdf:type* (Fig. 2d) and *resource-type* (Fig. 2e) results are similar in all selected subsets (most/less/random). This is due to the fact that the distribution of these property values are not even. The L parameter determines how many concepts are selected. The lowers is L value the more concepts are selected. It results in higher recall which is important to have for human annotators. Nevertheless, annotator can adjust L value at any time which brings high flexibility for annotators.

4.1.3 Prototype Application

We implemented an application prototype that allow users to upload documents to get metadata annotation suggestions. Figure 4 shows two input methods; the user can upload a document or type its content into the text box. After submitting the document or its content, the application will provide annotation suggestions. Figure 5 shows the candidates metadata. User can filter results based on metadata frequency in similar documents. Afterwards, the he can export the selected metadata to generate an appropriate file to be uploaded to CERES in order to ingest and publish the document metadata.



Documents metadata

Filtering metadata [1-10]

1

cdm:work_is_about_concept_eurovoc

☑ investment, 0.3
☑ project of common interest, 0.3
☑ energy grid, 0.3
☑ trans-European network, 0.3
☑ reduction of gas emissions, 0.2
☑ renewable energy, 0.2
☑ EU financial instrument, 0.2
☑ structural policy, 0.2
☑ transmission network, 0.2
☑ transport network, 0.2
☑ EU programme, 0.2
☑ sustainable development, 0.2
☑ energy cooperation, 0.2
☑ consumer information, 0.1
☑ financial occupation, 0.1
☑ insurance, 0.1
☑ investment company, 0.1
☑ disclosure of information, 0.1
☑ financial legislation, 0.1
☑ financial services, 0.1
☑ risk management, 0.1
☑ financial risk, 0.1
☑ energy policy, 0.1
☑ investment promotion, 0.1
☑ emission trading, 0.1
☑ climate change, 0.1
☑ greenhouse gas, 0.1
☑ transition economy, 0.1
☑ climate change policy, 0.1
☑ EU energy policy, 0.1
☑ security of supply, 0.1
☑ electricity supply, 0.1
☑ gas supply, 0.1
☑ electrical energy, 0.1

cdm:resource_legal_is_about_subject-matter

☑ Energy, 0.2
☑ Trans-European networks, 0.2
☑ Investments, 0.1
☑ Free movement of capital, 0.1
☑ Environment, 0.1

Figure 5: Metadata suggestions by the prototype

4.2 Phase-2: “Perspectives - business analysis of Artificial Intelligence usage in OP”

SeTA uses only documents content to compute the semantic similarity without taking into account the graph structure of documents metadata. SeTA@OP aims to bring together both sources of knowledge. Hence, in the future, we plan more advanced graph-based experiments with metadata properties in order to evaluate, compare and combine these new models

with language models.

In future work, we will apply supervised classification models to find out whether datadriven approach provides similar results as supervised training. The other direction could be to apply Longformer models and to utilize the Dense Passage Retrievers (DPR) method.

Additionally we will collaborate with further services in OP to demonstrate the SeTA services, to analyse their needs and to write customised business specifications for their specific requirements.

Conclusion

We investigate a novel approach to automatise metadata annotation of legal documents. The approach uses document text similarity to automatically annotate documents with metadata. Semantic similarity between documents is computed based on generated vector representations called embeddings. The experiments showed that the approach is effective and efficient, and can be used to (semi-)automate metadata annotation of large collections of documents in practical applications.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for generalpurpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[5] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China, November 2019. Association for Computational Linguistics.

Confidentiality (JENK1A.3)

Session Chair: **Aleksandra Bujnowska** *(Eurostat)*

Data linkage with a search engine: identifying individuals in a population register
Heidi Koumarianos (Insee)*; Yves-Laurent Bénichou (Insee); Séverine Gilles (Insee)

Synthetic Dataset Generation for Anonymised Machine Learning Applications
Elisabeth Petersen (Deimos Space UK Ltd.)*; David Smith (Deimos Space UK Ltd.); Andrew Harvey (Deimos Space UK Ltd.); David Petit (Deimos Space UK Ltd.)

Releasing survey microdata with exact cluster locations and additional privacy safeguards
Alejandra Arias Salazar (University of Costa Rica)*; Till Koebe (Freie Universität Berlin); Timo Schmid (Otto-Friedrich-Universität Bamberg

Data linkage with a search engine: identifying individuals in a population register

Keywords: data linkage, search engine, nominative data, privacy

INTRODUCTION

Facilitate data linkage while preserving privacy

With the development of efficient information systems in many sectors, data linkage has become a very powerful way of enriching data. It allows the integration of various types of information from different universes and is less expensive to implement than field surveys. By linking information collected by different organisations, data linkage also increases considerably the possibilities of measuring and studying economic and social phenomena.

However, certain rules must still be respected, in particular the aspirations of citizens regarding the protection of their personal data. This is why a specific mechanism was put in place by a national law in 2016 with the creation of the Non-Significant Statistical Code (NSSC). The purpose of the NSSC is to facilitate the linkage of data on individuals within the Official Statistical System while guaranteeing the anonymity and protection of personal data.

In this context, all processing with the NSSC must be declared to the national Data Protection Authority. The data minimisation principle of the GDPR is also paid special attention: only the strictly necessary information for the calculation of the NSSC is exchanged between the partner organisations and the NSI. Moreover, the NSI deletes the data without delay as soon as the NSSC calculation is completed.

Use a persistent identifier to remove identifying data

The NSSC is designed to serve multiple purposes in a variety of fields and its use is not limited to the matching of a few specific files; all ministerial statistical services shall be able to use it, for all their data sources. It must therefore be based on a population reference system that covers the entire population living in the country The national population register has thus been adopted as the reference with its Identification Number, better known as the national social insurance number.

This identification number is significant: it is composed of gender – year of birth – month of birth – place of birth (municipality or country) – rank of birth in that place and a calculated key. This identifier carries individual information and must be removed from the data files.

The general principle of the NSSC is to apply a cryptographic operation to the identification number in order to obtain a matching key that guarantees the impossibility of identifying the concerned persons.

METHODS

The wealth and quality of information on people's identities varies depending on the files to be linked. Some already have the national identifier, while some others only collect identity traits such as surname, first names, sex, date and place of birth. For the same individual, the NSSC will always be the same, regardless of the source in which the individual appears and regardless of the year in which the NSSC is calculated. With this unique and permanent key, the partners can then match their files.

We will concentrate on files without identifiers, that need a linkage process based on identity features in order to determine the non-significant code.

The very large volume of the population register (it includes 130 million records) led to designing three stages of identification with a view to optimising processing times according to the difficulty of the cases to be processed.

The process was elaborated by a cross-sectional team, composed of statistical and IT staff from the demography directorate along with methodologists and data scientists from the methodology directorate.

The asymmetric problem in size and status of data (identifying people from different types of data to a large register) is a good use case for a search engine.

As one of the data files is a register, it is quite interesting to invest in good indexation of data base (needed for an efficient use of the engine), that can be updated while being quite stable for a large proportion of records.

As there were previous experiments with Elasticsearch in the NSI, a first proof of concept was set in place and concluded that good results were reachable, while allowing the processing of large volumes.

A several steps process to deal with volume and data of varying quality

Linked to the size of the register, a classical approach in several steps was chosen.

To do this, the adopted principle was to deal with simple cases using processes that require little time and to reserve complex treatments for cases that really require them.

The first step is an exact query: the identification elements (surname, first names, date of birth, geographical code of the place of birth) are searched in the register for exact matches. At this stage, identification follows a fundamental principle: it can only take place if only one echo is found.

Then comes a second stage, consisting of simple queries. This allows for some relaxation of constraints on a few variables. Here again, only echoes without competitors are retained. Five successive releases of the identification constraints are implemented (on first names, places of birth, interchanging surnames and first names).

The last step deals with the most complex cases: it allows constraints to be relaxed on several variables simultaneously. Several echoes can be found in the register and the aim is to chose the one that has the most characteristics in common with the person to be identified.

Elasticsearch engine is used to look for similar echoes, to select the most likely ones, to rank them by calculating a score.

The found register echo with the highest score is then retained and the NSSC is calculated on the identification number corresponding to this echo.

Quality assessment of the process: useful to optimise queries and to inform theuser as well

Efforts were made to monitor the quality of linkage during the optimization of the queries. It allowed us to choose between different options, such as “ngrams or fuzzy?” or “2Grams or 5grams?”. We used files containing both nominative data and identifiers, so as to evaluate true and false positives among the echoes. First, we made sure that the right echo is in the first ten results returned by the engine, and then tried to boost weights of the different criteria to have the right echo first.

We used scores to create a quality indicator, based on the score level and the ratio between scores of the two first echoes.

The estimation of true positives on test files was used to determine several levels of quality, linked to the confidence in the true status of the record, in order to provide information to the user.

RESULTS

The results are really good and encouraging to use an search engine to deal with those volumes of data (Figures 1 and 2).

The quality monitoring allows to improve the efficiency of the process and leads to better quality, in the number of identified records as well as in the true positive rate.

Figure 1 : Results of the linkage of an administrative file to the national register

Step	**Percentage of records identified by each step**	**Estimation of true positive rate**
Exact	32%	0.998
Simple release 1	56%	0.999
Simple release 2	4.5%	0.988
Simple release 3	1.5%	0.986
Simple release 4	2%	0.988
Simple release 5	0%	0.993
Approximate value query	3.5%	0.803
Records not identified	0.5%	
Total		0.994

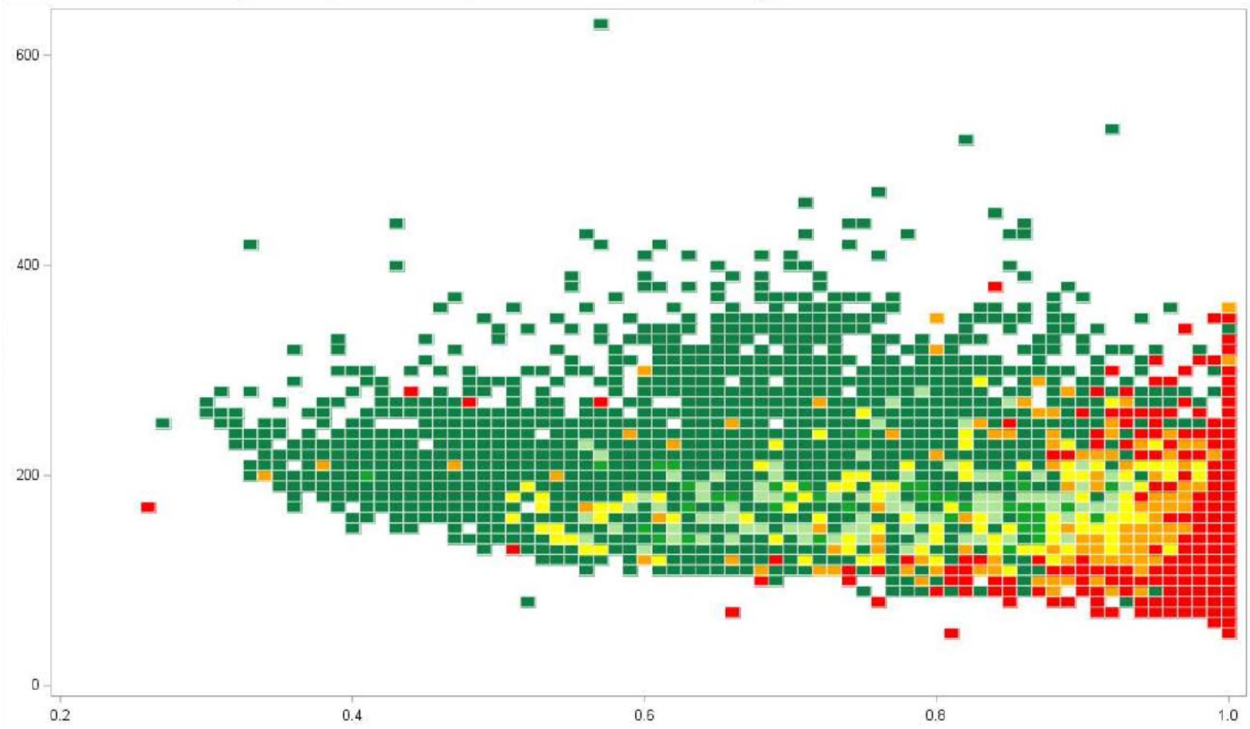


Figure 2 : Example of gradient of true positive rate for the records of the last step, depending on score and score ratio

Abscissa: ratio between the score of the second and first echo

Ordinate: score of the first echo

Gradient: true positive rate (from <0.6 to >0.98)

cONCLUSIONS

The need of a standardised process of identification and the nature of the register pushed us to look for an efficient solution.

The use of a search engine leads to technical challenges but it is worth it, as it enables to reach high quality results for different types of data files (administrative data as well as survey data).

The NSSC process has been in production for half a year.

Two projects are planning to test or to use Elasticsearch in their own identification process. One is a project of register on individuals with richer information (like place of living and household composition), and the other will be a register of addresses throughout the whole country.

Synthetic Dataset Generation for Anonymised Machine Learning Applications

Keywords: Data Anonymisation, Differential Privacy, Conditional Generative Adversarial Networks, Private Data, Survey Data, Machine Learning

INTRODUCTION

Sensitive ground survey data and geo-datasets are crucial in providing important information about different regions of the globe. Such data can be used to label satellite imagery, to delineate regional prosperity, environmental factors, or numerous other indiscreet properties useful in a variety of sectors from environmental preservation to determining the best places to provide humanitarian aid. Recently, the use of machine learning models, trained using this data, has become commonplace in predicting these same indiscreet values from satellite imagery alone.

Although, the information of individuals in our original datasets and their sensitive data would ideally be protected by using a trained neural network model, but this is not the case. A data anonymisation process must be done before, to avoid any potential privacy concerns of the survey respondent's data, which, since the introduction of data privacy regulations such as the General Data Protection Regulation in the EU, is increasingly a sensitive topic. Guaranteeing properties such as k-anonymity to the dataset, also encounters drawbacks once crossreferencing with other available datasets becomes a possibility. Although methods certainly exist to query sensitive data to reveal limited statistics, this does little to help us when confronted with the need for entire datasets to train our machine learning models.

To overcome this issue and guarantee the privacy of individuals who contribute their sensitive data to similar training datasets, the creation of synthetic data – datasets created artificially to mimic but not replicate the original – shall become an essential piece in the toolbelt of data analysts.

METHODS

Here we present two separate methods of generating synthetic datasets. First, we present the application that is based on applying a random Laplacian noise to our original dataset, resulting in differential privacy. Secondly, the use of Conditional Generative Adversarial Networks (CGAN) to generate an entirely new synthetic dataset based upon the original. Both are easily implemented in Python with the assistance of a few selected libraries. We are using the openly available Eurosat images [1] and our simulated labels to test these methods.

Differential Privacy concept

Differential Privacy is simply a mathematical property that an algorithm may have. Formally,

Let $A: X \to Y$ be a query and $x, x' \in X$ be any neighboring datasets (a dataset that differs by a single entry). A is ε-differentially private for $T \subseteq Y$ if:

$$Pr(A(x) \in T) \leq e^{\varepsilon} * Pr(A(x') \in T) \qquad (1)$$

An algorithm being differentially private means the probability of a result having come from one dataset is no greater than the probability it has come from a neighboring by a factor of e^{ε}. Hence, the effect that a single statistic can have on the result is limited, restricting how confidently someone can reverse engineer the original data from the result. The smaller the privacy parameter ε, the greater the privacy guarantee.

Privacy preserving method based on noise

The standard method of creating a differentially private result is to add random values (noise) to it, drawn from generated distribution such as Laplace or Gaussian. To generate an entirely synthetic dataset, we developed a tool that would use the basis from differential privacy and apply the noise to each individual entry not simply to the result of a single data query.
By selecting any noise mechanism to draw these random values from and applying these to our dataset always gives some level of differential privacy. The most common of these – and the one that we use here – is the Laplacian distribution. We define our distribution based upon the sensitivity value of the data in question and the desired epsilon value we would like to enforce. In python, adding Laplacian noise to a dataset can be achieved in a few lines of code. In certain cases, it can be difficult to determine the best values to use for the specified epsilon and sensitivity parameters and their potential effects to the anonymised data are unclear. Hence, we have developed a tool (see the algorithm in Figure 1) that from a single input will select an appropriate ε value for anonymisation, create a Laplacian distribution from this to draw noise from, impose bounds and then slightly adjust the output to better reflect the standard deviation of the original data.

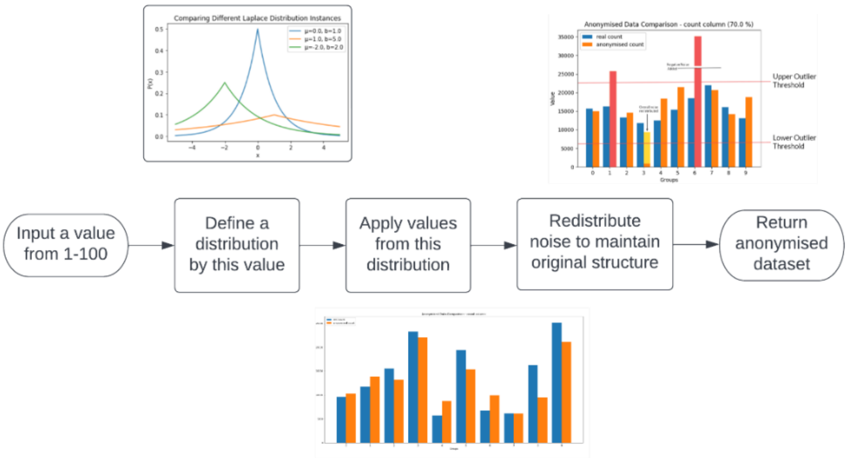

Comparing Different Laplace Distribution Instances
μ=0.0, b=1.0
μ=1.0, b=5.0
μ=-2.0, b=2.0
P(x)
x
Anonymised Data Comparison - count column (70.0 %)
real count
anonymised count
Value
Groups
Upper Outlier Threshold
Lower Outlier Threshold
Input a value from 1-100
Define a distribution by this value
Apply values from this distribution
Redistribute noise to maintain original structure
Return anonymised dataset

Figure 1. Breakdown of the Anonymisation Tool process.

Privacy preserving method based on GANs

Instead of adding noise to the original dataset, we can use a developed model CTGANSynthesizer from CTGAN class [2] to create an entirely new fake dataset made to mimic the real data.

With this method that uses CGANs, the infrastructure utilizes two separate models that compete against one another: the Generator attempts to create "fake" imitations of some data whilst the Discriminator – a classification model - attempts to classify real data and fake. Both models improve with each epoch and consequently, after training, the Generator can be used to reliably make synthetic data based on the data distribution it has learned from.

RESULTS

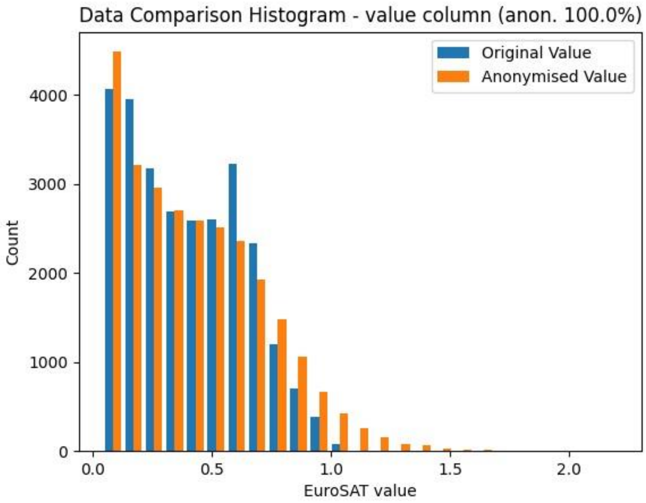


Figure 2. Anonymising Eurosat Labels with the Anonymisation Tool based on Laplacian noise.

In the above results (Figure 2) we have used an anonymisation parameter of 100% with the anonymisation tool based on Laplacian noise, to differentially privatise our dataset. At this level, the standard deviation of the noise added is approximately equal to the standard deviation of the original data, with the standard deviation of the anonymised data then brought down during the adjustment stage to the same value. Clearly, with the large values of noise added in this situation, the data has flattened a fair amount whilst individual values are altered significantly. Although varying magnitudes of noise are added to each entry, no single output is exactly the same as its original counterpart, thus ensuring that no individual's information is exposed.

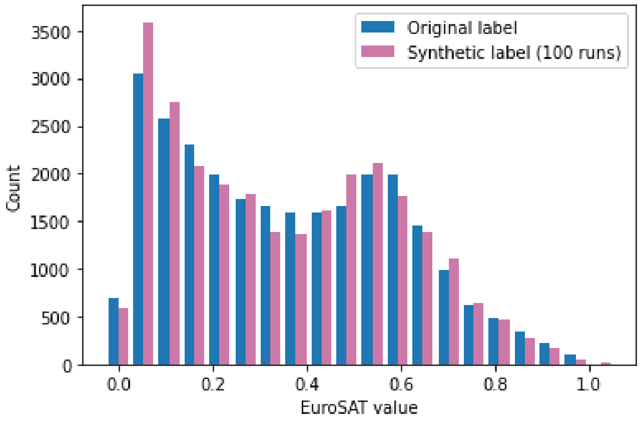


Figure 3. Generating Synthetic Labels with the CTGAN

The other option with data anonymisation can be observed from synthesising the data labels with CTGAN as seen in Figure 3. Data distribution is kept very similar to the original data and the labels are completely shuffled and new values are generated to each entry. In the results above the model was trained for 100 epochs, however this decision can be made by the user when initializing the model trainings.

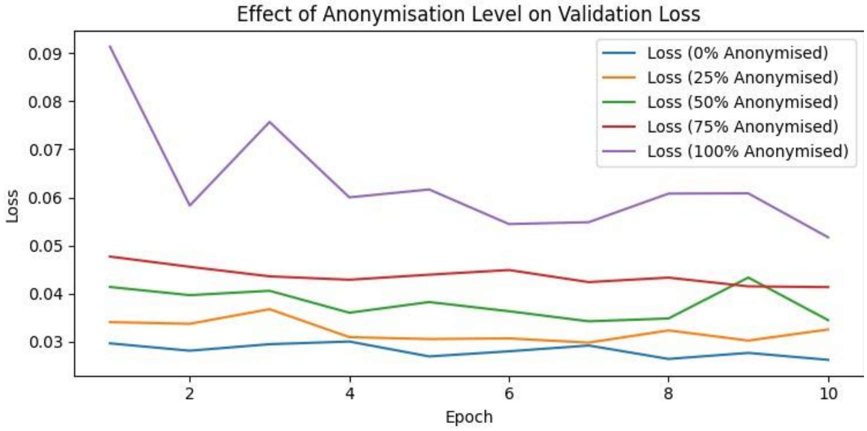


Figure 4. Comparison of validation losses from model trainings using different levels of anonymised labels.

To investigate the effects that the anonymisation of labels would bring to machine learning algorithm, we trained simple convolutional machine learning models that used the anonymised labels with different levels of anonymisation to extract features from satellite imagery. The results above (Figure 4) show how the model validation loss, which in this case is mean squared error, is affected by using more noisier labels in the model trainings. In comparing the original model training with the model using 100% anonymised labels, the accuracy has decreased by about 6% in this case. This indicates that a small level of anonymisation can be added without damaging the performance of the model.

cONCLUSIONS

To conclude, we have shown two privacy preserving techniques that are possible to implement to create synthetic datasets that anonymise the original, sensitive data. By using one or both described techniques, perhaps in conjunction with others, the risk of exposing private data can be reduced. It is also demonstrated the initial results of using synthetic labels to train a basic convolutional regression model. Work remains to observe the effect that synthetic labels have on other types of models that may be more greatly affected by discrepancies in synthetic training data.

rEFERENCES

[1] P. Helber, B. Bischke, A. Dengel and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2217-2226, July 2019, doi: 10.1109/JSTARS.2019.2918242.

[2] Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. "Modeling tabular data using conditional gan." Advances in Neural Information Processing Systems 32 (2019). Access online: https://arxiv.org/abs/1907.00503.

R/Julia/Python (MANS1A.3)

Session Chair: **Dario Buono** *(Eurostat)*

Julia as a software for Official Statistics and Social Sciences
Josep Espasa Reig *(LIS)*

New Scientist IT Infrastructure
Marius Felecan (Gopa Luxembourg)

An R package for automatically generating candidate correspondence tables between classifications
Martin Karlberg, Christine Laaboudi, Mátyás Mészáros *(Eurostat)*

Julia as a software for Official Statistics and Social Sciences

Keywords: Julia, R, Official Statistics, Computational performance, Inequality indicators.

INTRODUCTION

R and Python are widely used in Official Statistics and Social Sciences. Their open-source condition has granted them multiple advantages over proprietary software. This has resulted in a very large package ecosystem for statistics and virtually complete domination for data science tasks. However, these two programming languages bear two weaknesses.

First, because they are dynamically typed and interpreted languages, both R and Python are rather slow. Larger volumes of information and new sources of data[15] have made computational speed a rising challenge in Official Statistics and Social Sciences. So far, the traditional way of solving performance issues in R and Python has been to integrate C++ or Rust code. [16] Unfortunately, difficult languages such as C++ hinder their use for people with non-software engineering backgrounds. Hence, few Social Science packages use C++ as a 'low-level' programming language.

The second weakness of R and Python in the context of Official Statistics and Social Sciences is that developing and maintaining packages is usually done by volunteers. Few organizations in this area allocate resources to systematically develop and maintain shared statistical code. People working in both public institutions and academia rarely have incentives to devote time to these tasks. Given the amount of time required, they are even regarded as potentially detrimental for professional or academic careers. As a result, many packages related to the field end up with unsolved bugs and insufficient maintenance. [17]

A simple solution to R and Python weaknesses might come from using Julia. Julia is a high-performance language with a very similar syntax to R and Python. Experienced R or Python users will easily understand Julia's subsets, array comprehensions or anonymous functions. These users might also find Julia's syntax to be more modern and readable than older

[15] E.g. satellite data, machine learning methods, resampling …

[16] See for example the 'Rcpp' R package. Eddelbuettel, D. & Balamuta, J. (2018). Extending R with C++: A Brief Introduction to Rcpp. The American Statistician. 72(1). URL https://doi.org/10.1080/00031305.2017.1375990.

[17] See for example some of the open issues in R 'Hmisc' package, which is a widely used R package for weighted statistics: Frank, H., (2021). wtd.quantile - Issue #97 - harrelfe/Hmisc GitHub repository. Retrieved October 15, 2022, from https://github.com/harrelfe/Hmisc/issues/97

languages.[18] The design of many key Julia packages for statistics also takes into consideration that many new Julia users might already have experience in R and/or Python.[19] The main disadvantage of Julia is that it is still not a widely adopted language and

it still misses some important functions that would make it more convenient in niche areas such as Official Statistics.

Julia also has the advantage that it can be easily combined with R and Python. The three languages have packages to call functions from each other. [20] This opens the possibility of using Julia within R or Python processes and therefore avoiding some of its pitfalls. However, early evidence suggests that there might be a significant overhead when calling Julia functions from R or Python.[21]

There are ways to further speed up Julia code, such as the 'fastmath'[22] or 'turbo'[23] macros. These come at the expense of precision in estimates and/or computational checks. The adequacy and advantages of these methods for Social Sciences estimates, code and data has not been evaluated yet.

mETHODS

This paper answers the following research questions:

- How much more computationally efficient is Julia for Official Statistics? How faster is Julia code when compared to R? How much less memory do Julia processes use when compared to R ones?

[18] E.g. Julia can use unicode input which makes it more natural to read and write mathematical notation.

[19] See for example the comparison on the official documentation of the 'Data.frames' package: Kamiński,
B. et al. (2021) Comparison with Python/R/Stata - DataFrames.jl Retrieved October 14, 2022, from

[20] E.g. 'JuliaConnectoR' to call Julia from R, 'PyJulia' to call Julia from Python and 'RCall' to call R from Julia.

[21] Espasa Reig, J. (2022) [Twitter] 20 February. Available at: https://twitter.com/Josepespasa/status/1495467996034314251 Retrieved October 14 2022.

[22] JuliaLang (2022). GitHub repository. 12 March. Retrieved October 14, 2022, from https://github.com/JuliaLang/julia/blob/master/base/fastmath.jl

[23] JuliaSIMD (2022). LoopVectorization.jl GitHub repository. Retrieved October 14, 2022, from https://github.com/JuliaSIMD/LoopVectorization.jl

- What are the factors that might increase or decrease the computational advantages of using Julia for income inequality research processes? What is the 'overhead' of calling Julia from R and Python?

- What are the areas where Julia is missing key packages to become a major language in Official Statistics and Social Sciences?

To answer the first two questions, we focus on the computation of income inequality indicators. Income distributions and inequality indicators are a key part of Official Statistics and most organizations in this field produce them. We benchmark the time and memory used to compute income inequality indicators[24] and their standard errors both in R and Julia. Additionally, we benchmark modified Julia code adding the mentioned methods to speed it up. We also benchmark the Julia processes called from R in order to measure overhead.

We use the wage variable of large person-level datasets such as the Luxembourg Income Study (LIS) files for the US and Colombia. Estimates are computed both using weighted and unweighted data to simulate cases when all observations in the population are available (e.g. censuses or administrative data). Whenever possible, the code for computing R estimates come from already available packages.[25] The Julia code for computing inequality estimates uses the 'Inequality.jl' package.[26] Its functions are modified with the mentioned macros to increase computational performance.

To answer the third question, we evaluate the availability and maturity of Julia packages for producing Official Statistics. We compare these to those written in R and Python. We focus on packages for the following areas:

- importing data from datasets; [27]

https://dataframes.juliadata.org/stable/man/comparisons/ , or many of the discussions on adding new features in DataFrames.jl: Kamiński, B. (2017). Handling of duplicate column names by join · issue #1333 · Juliadata/DataFrames.jl. GitHub repository. Retrieved October 14, 2022, from https://github.com/JuliaData/DataFrames.jl/issues/1333.

[24] E.g. Atkinson index, Gini coefficient, poverty headcount, poverty gap and mean income.

[25] E.g. Zeileis, A., Kleiber, C. (2015) 'ineq' R package: Inequality, concentration, and poverty measures. Lorenz curves (empirical and theoretical). Retrieved October 14, 2022, from https://cran.rproject.org/web/packages/ineq/ineq.pdf

[26] Espasa Reig, J. (2022) 'Inequality.jl' Julia package: Julia package for computing inequality indicators Retrieved October 14, 2022, from https://github.com/JosepER/Inequality.jl

[27] Including data in proprietary formats such as '.dta' and '.sav', which might be common in Official Statistics and Social Sciences.

- interacting with SQL databases;
- manipulation of tabular datasets;
- sampling and sample survey planning (e.g. sample size calculations, optimal size allocations);
- statistical matching;
- weighting and calibration of survey samples;
- imputation and treatment of missing values;
- computation of statistical estimates and variance estimation.[28]

We classify the different areas as 'not available', 'partially available/developing' and 'mature'. The first category ('not available') refers to processes for which Julia packages are not available or the development of these do not cover the basic functions needed in Official Statistics. We labelled an area with 'partially available/developing' when the Julia packages offer limited functionalities when compared to those in R and Python, but offer all basic ones. We consider that an area is 'mature' when Julia packages are at par with those in R and Python or even offer useful additional features.

RESULTS

Our benchmarks show that a typical process is 4 to 20 times faster using Julia than R. This large range depends on the exact process implemented. For tasks where R quickly calls built-in functions, the advantage of using Julia is smaller. The difference in time is largest for those jobs involving iterations such as loops, recursive functions or tolerances. Early results also show that Julia functions are substantially more memory efficient. There are technical challenges, however, when trying to assess the memory used by Julia functions when called within R. This is because R does not track the memory used by the Julia functions.

The benchmarks also show that there is a clear overhead when calling Julia from R. This penalty tends to halve the advantage of using Julia over R. This seems to be caused by the differences in data structure between the two software. Using the 'fastmath' and 'turbo' macros can further speed up Julia code and they seem to pose little risk for processes tried (i.e. computation of inequality indicators).

The code used for the benchmarks will be made available online with an open-source license to ensure full reproducibility of the paper results. Researchers interested will be able to check the results on other machines by using their own datasets or the publicly available sample LIS datasets.

Julia has a mature package ecosystem for certain tasks. Examples of these are importing data files, interacting with SQL databases, manipulating datasets and certain types of statistical estimations. Julia has some functionalities available in the area of survey samples, variance estimation and multiple imputations, but they are currently subpar when compared to those

[28] Including multilevel analyses, standard errors from complex survey designs and clusters.

available in R. There currently seem to be no packages that cover all basic features for performing statistical matching in Julia.

cONCLUSIONS

This paper evaluates the computational advantages of using Julia and the maturity of its package ecosystem for Official Statistics and Social Sciences. Additionally, it looks at the gains of integrating Julia code into R and Python processes. We believe this might be important for researchers looking to speed up their data processes. It should offer them a rigorous comparison of different possibilities. These range from using Julia alone to writing computational demanding functions in Julia and calling them from Python or R processes.

Julia offers clear advantages in terms of speed and memory use. It also offers a more modern programming language with features that are currently not supported by R or Python. Its adoption in Official Statistics and Social Sciences might depend on the ability of its community to develop first-class packages in a few key areas.

New Scientist IT Infrastructure

Keywords: software, science, culture, open, architecture, cloud, academia.

Introduction

The IT environment for a scientist is changing rapidly.

There are cultural related reasons such as openness, transparency, reproducibility and inclusion of research.

And there are technical reasons. Mainly explored here are the challenges posed by growing research data volume and the computational intensive research.

All these reasons create the need to rethink the IT environment and create new tools that work in more advanced architectures.

Also, because of the raising research complexity, now research is done by teams. And into a team, several specializations are needed: data scientist, software engineering, data infrastructure.

Open Science

Some say that the research culture is broken [1], and one way to fix it is by open science [2].

This means openness, transparency, reproducibility and inclusion [3] in research. Of course, the architectures or tools used can't solve this problem but they can enable an environment to help obtain this better culture.

Free and Open source ecosystem

Following open science, the next stop must be open-source software. It's difficult to share algorithms when implementations are made on proprietary platforms. The same applies to data storage platforms and servers, data formats, streaming services, spreadsheets, etc.

The open source community offers also some new business models: competitions (Kaggle [4], NowCasting [11]), contributing to the community by creating open source packages, or by enabling the creation and maintenance of them (GitHub Sponsors [5]).

Architecture

All these tools can be arranged in different configurations (architectures), to be able to fulfil FAIR [6] principles: open data, tools, software, documentation and publications.

On top of FAIR principles to choose the architecture, we need to address the internal constraints of particular research. We will focus on data size and computation complexity.

Methods

This short study has two parts. One focused on the open source offer that can be used and help the scientific community and the architectures that fits the need and can be built based on open source available. The other part is a case study for a concrete implementation of such an environment here at Eurostat.

Exploring the domain

The first part is an exploration of the available open-source artefacts. Starting with the formats for managing the publication-ready research documents and papers, and data storage, continuing with the most used platforms and programming languages for processing data and finalizing with a different platform for sharing the research or that helps collaborative research.

Case study

The second part is a case study from our work in Eurostat. We took a needed tool for Eurostat data consumers, the connector to SDMX data provider service for Python, choose the tools and formats and explained our choices. We used just open source tools for all the aspects of the project: for editor, data storage, CI/DI, helps, and documentation. The problem solved, the package created, was of a small size, and test data was provided by Eurostat public databases. The process was also tailored to match the exploratory way of using a programming language that is preferred by data scientists.

Results

SEVERAL CONFIGURATIONS/ARCHITECTURES WERE EXPLORED. AS A RESULT FOR OUR CASE STUDY THE TECHNOLOGIES USED ARE AS FOLLOWS:

- PYTHON [7] WAS A REQUIREMENT, NOT A CHOICE. WE NEEDED A TOOL FOR PYTHON USERS. NONETHELESS, PYTHON IS, ONE OF THE MOST USED PROGRAMMING LANGUAGE FOR DATA SCIENCE WITH A HUGE COMMUNITY AND VAST QUANTITIES OF FREE AND OPEN-SOURCE PACKAGES.

- JUPYTER [8] IS A SOFTWARE LANGUAGE-AGNOSTIC ENVIRONMENT AND UNIQUE FILE FORMAT: WRITE PROSE, CODE, AND TESTS IN NOTEBOOKS — NO CONTEXT-SWITCHING.

- NBDEV [9] BASED ON JUPYTER FILES AND DEVELOPED FOR PYTHON IS THE TOOL THAT HELPS WRITE, TEST, DOCUMENT, AND DISTRIBUTE SOFTWARE PACKAGES AND TECHNICAL ARTICLES — ALL IN ONE PLACE, FROM YOUR JUPYTER NOTEBOOK.

- GITHUB [10] IS THE PLACE AND TOOL USED TO STORE AND SHARE OUR RESEARCH.

In addition, there is a great opportunity to extend this system, maybe in a cloud configuration, for the need of academia. In this context the environment can be used for extensive laboratory work with large number of students, and can provide the environment for knowledge testing also.

We find that this way of developing software is very well connected with the mind-set of a researcher, producing fast feedback (results) that can help a researcher more easily decide how to continue.

It helps a researcher (as a developer of tools for research) iterate faster, and produce quality code, and documentation. And because of the environment (alive code), the tests validate the code, and examples are validated immediately and give the guarantee that they will evolve at the same time as the project and will give correct results all the time.

Also, the flow of thoughts and ideas is not interrupted by switching back and forth between tools and formats which is also added a psychological dimension to the process.

This paper was produced using the same tools and development process and it used the same tools and technologies of the case study, and because of this comes out natural and effortless.

Conclusions

The process and technologies used seem promising and in our case, they were well received. Especially the flexibility of JSON Jupyter Notebook was appreciated, not just by Python users, but also by R users that now can use the same format to share research.

Still, more research is needed with different size projects and team of data engineers and data scientists. An interesting distinction can be done by having this process/technology tested for different types of developments (research, industry, education, etc.)

One-Size-Fits-All cannot work here. The tools and how they are connected must be adjusted to the internal constraints, but the Jupyter notebook format as a platform for code, results and publication-ready papers seems to be the common part of a solution and works for projects of very different size.

References

[1] Ainsworth Rachael, Research Culture is Broken; Open Science can Fix It, TEDx: https://www.youtube.com/watch?v=c-bemNZ-IqA

[2] Open science, https://en.wikipedia.org/wiki/Open_science

[3] Open-Source Science Initiative, https://science.nasa.gov/open-science-overview

[4] Kaggle, https://www.kaggle.com/

[5] GitHub Sponsors, https://github.com/sponsors

[6] FAIR Principles, https://www.go-fair.org/fair-principles/

[7] Python, https://www.python.org/

[8] Jupyter, https://jupyter.org/

[9] nbdev, https://nbdev.fast.ai/

[10] GitHub, https://github.com/

[11] NowCasting, https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220718-1

An R package for automatically generating candidate correspondence tables between classifications

Keywords: classification, correspondence table, concordance table, automation

Introduction

Interdependencies between classifications

Most statistical classifications exist in a statistical ecosystem, where they are interlinked with other classifications – either structurally, or by means of correspondence tables, with an example retrieved from the NACE Rev. 2 publication shown in Figure 9.

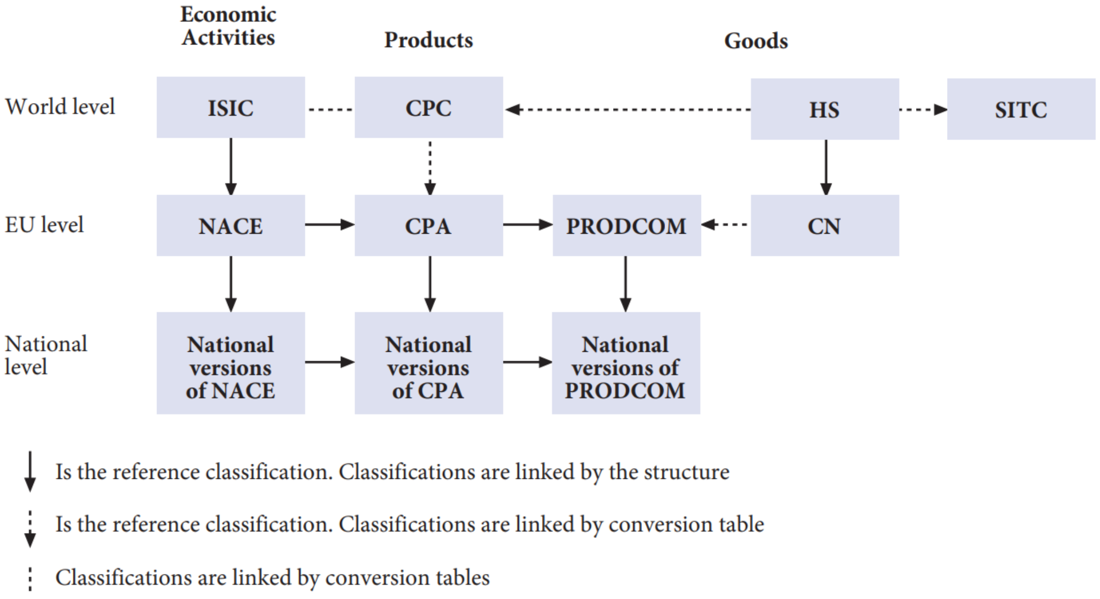

Economic Activities
Products
Goods
World level
ISIC
CPC
HS
SITC
EU level
NACE
CPA
PRODCOM
CN
National level
National versions of NACE
National versions of CPA
National versions of PRODCOM
Is the reference classification. Classifications are linked by the structure
Is the reference classification. Classifications are linked by conversion table
Classifications are linked by conversion tables

Figure 9. The international system of economic classifications (European viewpoint)

When statistics on the same topic are compiled using different classifications (for instance international production statistics based on CPC and European production statistics based on CPA; see Figure 9), either (or both) of them need to be transformed in order to become comparable – and the *correspondence table* is the tool enabling this.

Updated classifications require updated correspondence tables

When a statistical classification A that is linked to other classifications is updated to version A*, and that classification is linked to classification B by means of a correspondence table, this will trigger a request for the correspondence table between the classifications to be updated. Current best practice for creating the updated correspondence table A*:B is to use:

- the previous correspondence table A:B
- the structural (‘predecessor’) links A*:A (the *concordance table*) that would normally have been created as part of the updating process

and mechanically create a first candidate correspondence table. This first correspondence table will typically require expert review to make sure that there are no anomalies.

New applications require altogether new correspondence tables

Sometimes, we want to compare between statistics using classifications A and B that belong to a system, but are not directly adjacent to each other in that system. It might be the case that no A:B correspondence table yet exists – so we need to create one. Current best practice in that case is to use the same principles as above and 'pivot' over the 'intermediate classifications' C_1,C_2,…,C_k, using the relevant correspondence tables. (For instance, a CN:SITC correspondence table could be established via the CN:HS and HS:SITC correspondence tables.) Technically, the approach is very similar to the one described in 0 above (in particular when k=1); the main difference being that multiple correspondence tables might be used (for k≥2).

An R package to support the establishment of correspondence tables

For both of the cases described in 0 and 0 above, the 'mechanical creation' part of the process can, roughly speaking, be achieved by an 'outer join' of the (classification, concordance and correspondence) input tables involved. Thus, it could be presumed that one SQL statement could suffice to create the candidate correspondence table. However:

- the candidate correspondence table might contain literally thousands of records, and to support the expert in their review, it would be helpful to annotate it so that the cases that are most likely to need a closer look are highlighted;
- there may be input data quality issues (duplications, omissions, superfluous records) owing to the fact that one of the input tables used has not been formatted with correspondence table creation in mind.

In 2022, Eurostat decided to develop the **`correspondenceTables`** R package – mainly motivated by reason ⍰. However, during the development, plenty of issues stemming from ⍰ were encountered, and the package was rendered even more robust. We briefly describe the package in Section 0, and proceed to present a real-life application of it in Section 0, before making some concluding remarks in Section 0.

Features of the package

The package features two functions: **`updateCorrespondenceTable`** to handle the case described in 0 above, and **`newCorrespondenceTable`** to handle the case described in 0 above. The functions have purposefully been designed to be possible to include in a modular architecture, thus only focusing on their 'core tasks', leaving any possible post-processing aside (as this could easily be performed directly on the output).

As alluded to above, both functions create a candidate correspondence table by essentially carrying out an 'outer join' of all the input tables involved. The code uses, to the greatest possible extent, functions found in `base` R. Many of these functions were combined to create some basic intermediate functions (in order to develop code that is easy for users to understand and to modify if needed). This rendered code that is both maintainable, reusable, and easily extendable (allowing the addition of new features).

The remainder of this section focuses on the main added value of the package.

Flagging for review

The initially foreseen main value added of the package (in relation to a simple 'join') is the flagging of records. The flags generated cover aspects such as:

- redundant records (i.e., multiple records having identical values for A* and B because of multiple values for A). Instead of 'simplifying' by suppressing duplicate instances, the different values of A are retained to allow an expert review;
- records violating 'hierarchical expectations' (e.g., multiple records for B for a given value of A*, in cases where there ought to be a 1:M relationship between B and A*);
- code changes between A and A* (indicating whether either of the A or A* codes of a record have been involved in splits, mergers, changes or larger rearrangements).

These flags are intended to guide the expert to areas of the candidate correspondence table meriting a closer look, so that their energy is not unnecessarily spent on more straightforward, essentially unchanged, parts of the table.

Flagging for input data issues

Thanks to the inclusion of a rich body of sample data from the international system of economic classifications, we encountered a number of input data issues, and defined additional flags and error checks to handle them in a robust way. To summarise:

- Real classification and correspondence tables sometimes host multiple versions side-by-side. The functions were therefore redefined to check for the uniqueness of records with respect to the codes (for classification tables) or code pairs (in the case of correspondence and concordance tables). Thereby, the user is forced to 'clean' the input data in this regard (essentially picking the right version) before proceeding.
- For hierarchical classifications, correspondence tables are typically only established at the most granular level – although there are also examples to the contrary. By merely 'joining' the classification and correspondence/concordance tables involved, there are typically a large number of 'mismatches' (since the higher hierarchical levels will appear in the classification table but not in the correspondence table). To make sure that nothing is overlooked, incomplete records (lacking codes or data for certain classifications) are flagged and retained. The user is allowed to set their tolerance for such mismatches by means of arguments for the two functions, and does thus have the possibility to iteratively clean (and resubmit) the input data. The various types of 'mismatches' are flagged to facilitate the input data cleaning.

Experience from applying the package

Need for annual CN:CPA updates – baseline procedure applied for CN 2022

The Combined Nomenclature (CN) is a tool for classifying goods, set up to meet the requirements both of the Common Customs Tariff and of the EU's external trade statistics. The CN is also used in intra-EU trade statistics. Every year, Annex I to the basic CN Regulation is updated and published as a stand-alone Regulation.

As there is a need to maintain a correspondence table between CN and the Classification of Products of Activities (CPA), there is thus a need to update the CN:CPA correspondence table on a yearly basis. For establishing the CN2022:CPA candidate correspondence table, we used source data from RAMON, and post-processed them in MS Access (joining the classifications together; enriching the output by including contextual information; flagging unmatched classifications).

Applying the correspondenceTables package for CN:CPA

Classification data have now been migrated from RAMON and exposed in a Linked Open Data (LOD) format. In October 2022, following the yearly update of the CN (to CN 2023), we retrieved classification data (along with all their contextual data) via the EU Publications Office public SPARQL endpoint, using two SPARQL queries. We only needed to configure the column heading to to render the SPARQL output files readable by the **`updateCorrespondenceTable`** function (into which they were subsequently fed).

Applying the R package to the RDF datasets was easy (and also faster compared with the previous process in MS Access that required more system resources). The major benefit of the R package is the immediate flagging of records for review integrated into the output. This facilitates expert validation of critical cases (e.g. CN codes mapped to multiple CPA codes, changes of CN codes (new, deleted) or labels, redundant mappings) and allows navigating through the parts of the table that need a closer examination.

Conclusions

The **`correspondenceTables`** package provides classification experts with a tool that automates the first step in the process of generating candidate correspondence tables whenever an appropriate set of intermediate correspondence/concordance tables are available, and has already proven its usefulness in real-life situations. Our actual experience from applying the **`correspondenceTables`** package (see 0 above), was that it was easier and faster than the previous approach. The main advantages of using the package for creating a candidate correspondence table are the elimination of the manual work, the elimination of the risk for errors stemming from manual operations, and the tidier and cleaner look of the candidate correspondence table.

The package is available in the public domain. The stable version can be installed from CRAN and the development version can be downloaded from GitHub. The auto-generated documentation is included in the package and available at the GitHub site of the package.

In the rest of this section, some of the lessons learned during the development phase are presented, along with some ideas for future improvement.

Lessons learned

Already from the outset, the **`correspondenceTables`** package was conceived to be robust, exiting gracefully rather than crashing when there are input data issues. As described in Section 0, the issues encountered proved to be more challenging than initially thought, requiring additional error handling features to be developed. We thus learned not to underestimate data quality issues – and the utility of real-life test data.

The biggest programming challenge was the implementation in `base` R of the ‘outer join’ of all the input tables involved without excessive execution time. The second biggest challenge was the generation of flags for each record, with certain flags (e.g. the flag for violation of ‘hierarchical expectations’) requiring the joint examination of the multiple records. Finally, the development of error handling features was another programming challenge. Extensive planning and testing helped identify several types of errors that can occur and to write the code to catch them and provide informative messages to the user.

Areas for future development

The **`correspondenceTables`** package is currently only able to generate candidate correspondence tables when ‘intermediate’ correspondence tables already exist. There is a clear need to also be able to generate candidate correspondence tables completely ‘from scratch’. However, such an application would require artificial intelligence (in particular natural language processing) methods to be applied, which would add a layer of considerable complexity.

A far more modest, and hence feasible, extension of the package would be to add utility functions for retrieving classification, concordance and correspondence tables from well-structured public repositories (such as a SPARQL endpoint).

New skills in Symbolic Data Analysis for Official statistics (GASP2M.1)

Session Chair: **Rosanna Verde** *(University Vanvitelli)*

Ranking “concordance” and “discordance” between a class and a set of classes
Diday Edwin *(Paris-Dauphine University)*

An illustration of the use of the measures s-concordance and s-discordance in applications
Simona Korenjak-Cerne (*University of Ljubljana, SEB, and IMFM Ljubljana*), Jasminka Dobša (*University of Zagreb*), Diday Edwin (*Paris-Dauphine University*)

Exactly mergeable summaries
Vladimir Batagelj *(IMFM Ljubljana)*

Discovering Patterns and Trends with Distributional Data
Paula Brito *(Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC)*

An illustration of the use of the measures s-concordance and s-discordance in applications

Keywords: symbolic data, s-concordance, s-discordance, tf-idf.

Abstract

In 2019 Diday [1] introduced some concordance and discordance measures between a class and a given collection of classes. The concordance that measures the similarity between an object and a collection of objects is based on classes described by symbolic data and falls within the framework of symbolic data analysis (SDA) [2], which is why it is called s-concordance (symbolic concordance). Since then, a more precise theory has been developed and several families of these measures have been introduced.

With class, we usually mean unit obtained by aggregation of individuals (for example country with its inhabitants or school class with its students). Such data is very common in official statistics to protect the privacy of the individual. In the definitions of s-concordance and s-discordance, two variabilities are considered: variability between individuals (included with function f) and variability between classes (included with function g). These measures can be used in many situations, for example, we can measure how discordant is class c due to the collection of classes P by the value x and in this context use such a measure to identify the characteristics of a class.

In the first part, we will illustrate some possible practical applications of these measures on open datasets from education. We explore the use of s-concordance and s-discordance measures on the dataset of an international, large-scale assessment that measured student achievement in traditional reading (on paper) and reading on digital devices in several countries around the world. We examine the application of s-concordance and s-discordance measures to these data from two perspectives: a) with s-concordance, we examine how well the aggregate data for the country represent the classes/teachers within the country; b) with s-discordance, we determine in which countries the observed categorical value is characteristic.

In the second part, we will compare and study differences between s-discordance and a popular measure in text mining TD-IDF (Term Frequency - Inverse Document Frequency) that helps to identify the "relevance" of a term. In a similar way s-discordance measures relevance of a term for observed class. Data from sentiment analysis will be used to illustrate how s-discordance measure can be used for automatic acquisition of sentiment lexicons.

References

[1] E. Diday, Concordance and discordance between classes of complex data, presented at the Workshop Advances in Data Science for Big and Complex Data: From data to classes and classes as new statistical units, University Paris-Dauphine, January 10-11, 2019.

[2] E. Diday, Explanatory tools for machine learning in the symbolic data analysis framework, Advances in Data Science 2020, ISTE-Wiley, pp. 3-30.

Exactly mergeable summaries

In the analysis of large/big data sets, *aggregation* (replacing values of a variable over a group by a single value) is a standard way of reducing the size (complexity) of the data. Data analysis programs provide aggregation functions such as means (arit, geom, harm, median, modus), min, max, product, bounded sum, counting, etc. Special care has to be given to variables measured in different measurement scales.

Recently some books dealing with the theoretical and algorithmic background of the traditional aggregation functions were published (Beliakov et al., 2007; Torra and Narukawa, 2007; Grabisch et al., 2009; Bustince et al., 2013). A problem with traditional aggregation is that often too much information is discarded thus reducing the precision of the obtained results.

A much better, preserving more information, summarization of original data can be achieved by representing aggregated data using selected types of complex data such as symbolic objects (Diday, 1988), compositions (Aitchison, 1986), functional data (Ramsay and Silverman, 2005), etc. In the symbolic data analysis framework much work is devoted to the summarization process, for example, the function classic.to.sym in RSDA (Rodriguez, 2018), and SODAS or SYR software.

In complex data analysis the measured values over a selected group A are aggregated into a complex object $\Sigma(A)$ and not into a single value. Most of the aggregation functions theory does not apply directly. In our contribution, we present an attempt to start building a theoretical background of complex aggregation.

An interesting question is, which complex data types are compatible with the merging of disjoint sets of units

$$\Sigma(A \cup B) = F(\Sigma(A),\Sigma(B)), \qquad \text{for} \qquad A \cap B = \emptyset.$$

The mergeable summaries were already proposed and elaborated by Agarwal et al. (2012). In his approach, the summarization is not deterministic and allows some errors. A summary is *mergeable*, if the error and space (size of summary) do not increase after the merge. In our contribution, we will discuss *exactly mergeable* summaries “without errors”.

Citizen Science (JENK2M.1)

Session Chair: **Monica Pratesi** *(National Institute of Statistics–ISTAT)*

Data donation of Google Location History information: willingness, donation, and non-donation biases
Bella Struminskaya, Laura Boeschoten *(Utrecht University)*

To what extent do data privacy concerns and digital distrust act as barriers to smartphone sensor data collection in general population surveys and what can be done to mitigate them?
Caroline Roberts, Marc Asensio Manjon *(University of Lausanne)*; Nicolas Pekari *(FORS)*

Eurostatistics - from PDF format to interactive web visualisation using R
Rosa Ruggeri Cannata*,* Piotr Ronkowski *(Eurostat),* Johannes Buck *(Technical University of Munich)*

To what extent do data privacy concerns and digital distrust act as barriers to smartphone sensor data collection in general population surveys and what can be done to mitigate them?

Keywords: Designed data collection by (mobile) devices; Smart Surveys and Trusted Smart Surveys

Introduction

Public willingness to participate in research on smartphones remains one of most important barriers to incorporating digital data collection methods in probability-based sample surveys of the general population. While the proportion of web survey respondents completing questionnaires via a mobile browser continues to grow, when asked about hypothetical willingness to agree to smartphone sensor data collection, resistance remains high – particularly when it comes to so-called ‘passive’ data capture. Two related factors that have emerged in numerous studies as key correlates of stated *un*willingness to agree to smartphone sensor data capture are concerns about data privacy and digital trust ([1], [2], [3], [4], [5]). Less is known, however, about how closely hypothetical willingness maps on to actual participation, and hence, how important privacy concerns and digital distrust really are in specific participation decisions ([6]). Assessing this and its implications for the success of designed data collection via mobile devices is a key priority for ensuring the future success of smart surveys - as is finding suitable methods to mitigate any negative impacts on data quality.

In this paper, we present the results from a probability-based online panel survey designed to assess the impact of data privacy concerns and digital trust on both hypothetical and actual willingness to complete different types of mobile data collection. Furthermore, in embedded methodological experiments we test alternative methods aimed at reducing or offsetting the potential negative impact of privacy concerns and distrust on survey participation decisions. These include providing information (in alternative formats) designed to reassure sample members about the confidentiality of their answers and the security of their data, and offering different amounts of monetary incentive for completing data tasks. Drawing on the Leverage-Salience Theory of survey nonresponse ([7]), we assume that different methods will affect participants differently, depending on how important their prior levels of concern about data privacy and digital distrust are for their decision to complete the digital data collection tasks on their smartphone. Specifically, we address the following research questions:

RQ1 – To what extent the data privacy concerns and digital distrust relate to hypothetical willingness to participate in digital data collection tasks on a smartphone?

RQ2 - Do data privacy concerns and digital distrust influence actual participation in digital data collection tasks?

RQ3 - Does providing data protection information positively impact reported concerns about data privacy and, in turn, improve willingness to participate in digital data collection tasks and does this vary depending on the type of information or prior levels of concern?

RQ4 – Does varying the amount of monetary incentive differentially affect participation decisions for sample members with higher and lower levels of concern or distrust?

Methods

The experiments were embedded in an online panel survey (wave 1 sample of n=10,000) of a random sample of adult residents in the French-speaking region of Switzerland, conducted in February-April 2022. Subsequent experiments were conducted in August (wave 2) and in October 2022 (wave 3). The study was initially designed to measure privacy concerns and digital trust in the aftermath of the Covid-19 pandemic, as well as to test alternative ways of recruiting people to a smartphone panel (established for the purposes of question testing and methodological innovation). Named individuals in the probability-based general population sample were contacted by advance letter, which contained a URL and QR code to access the survey (together with the enclosed data privacy information (depending on the treatment group to which they were randomly assigned).

At wave 1, the experimental design compared two versions of the study's data privacy and confidentiality policy sent with the advanced letter (one a more detailed GDPR compliant text covering two sides of A4 paper. The other was less detailed, and more visually appealing, including text and illustrations, presented in the form of an A5 leaflet. A third group received no notice (besides a link to the full information on the study website). At wave 2, the experiment was extended to the no-leaflet group, which was randomly assigned to receive either the detailed version of the privacy information or the leaflet. At wave 3, experiments were conducted to assess the moderating effect of different incentive amounts on the relation between privacy concerns (measured at wave 1) and willingness to complete digital data collection tasks at wave 3: (i) provide data donations in the form of screenshots and videos of data reported in the screen time / digital wellbeing functions of their phone; and (ii) respond to ecological momentary assessments via text messages.

Results

At the present time results are only available from the wave 1 experiment. Providing privacy information with the survey invitation had a significant negative effect on willingness to participate, but not on stated willingness to share smartphone sensor data. Respondents receiving the information were significantly more likely to report that they had read the privacy information and understood it, however. Only minimal differences were observed between the two leaflet formats, although the colour leaflet appeared to have a more detrimental effect on willingness to participate. These findings will be supplemented with results from the wave 2 and wave 3 experiments to draw conclusions about the relative effectiveness of information and incentives on willingness to participate in smart surveys, as a function of prior privacy concerns and digital distrust.

References

[1] Jäckle, A., Burton, J., Couper, M. P., & Lessof, C. (2019). "Participation in a Mobile App Survey to Collect Expenditure Data as Part of a Large-Scale Probability Household Panel:

Coverage and Participation Rates and Biases." *Survey Research Methods* 13:23–44. Wenz, A., Jäckle, A., & Couper, M. P. (2019). "Willingness to Use Mobile Technologies for Data Collection in a Probability Household Panel." Survey Research Methods 13:1–22.

[2] Revilla, M., Couper, M. P., & Ochoa, C. (2019). "Willingness of Online Panelists to Perform Additional Tasks." *Methods, Data, Analyses*, 13:223–52.

[3] Keusch, F., Struminskaya, B., Kreuter, F., & Weichbold, M. (2021). "Combining Active and Passive Mobile Data Collection: A Survey of Concerns." In *Big Data Meets Survey Science*, edited by Craig A. Hill, Paul P. Biemer, Trent Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg, 657–82. Hoboken, NJ: Wiley.

[4] Struminskaya, B., Toepoel, V., Lutgtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly*, DOI:https://doi.org/10.1093/poq/nfaa044

[5] Roberts et al., (2022). Data privacy concerns as a source of resistance to complete mobile data collection tasks via a smartphone app. *Journal of Survey Statistics and Methodolog*y.

[6] Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation. *Public Opinion Quarterly*, *64*(3), 299–308. DOI: https://doi.org/10.1086/317990

Eurostatistics - from PDF format to interactive web visualisation using R

Keywords: R, data visualisations, publishing

Introduction

Making statistics accessible and meaningful for the public is an important task for statistical institutions around the globe. One of the best techniques for communicating data is to visualise the numbers in a graph. Adding interactivity to visualisations make them more interesting for the users, and can be complemented by data storytelling based on narrative. Users also like to access the freshest data.

In order to meet these expectations, we have transformed “Eurostatistics”, a Eurostat monthly publication traditionally published as a static PDF, into an interactive web visualisation. Eurostatistics presents data on the latest macroeconomic developments with a written commentary. In view of its high publication frequency, and in order to include freshest data, it was important to shorten and automatize its production process as much as possible.

Eurostatistics includes a wide set of short-term indicators with data for the European Union (EU), the euro area, Member States and some other countries. Those data come from different sources, mainly from Eurostat but also from the OECD and other data providers. Consequently, the main challenge was to set up a system that could process and visualise this vast and diverse collection of data very rapidly. Another important requirement for the system was to be agile, for both integrating easily the current month economic analysis, and being adaptable to content modifications, for example adding new indicators or showing more information on specific months. Finally yet importantly, the system should be simple enough to be managed by statisticians with some technical knowledge, but without advanced IT expertise.

Technical background

The new visualisation tool has been developed in the R programming language. It uses the R Markdown language, with the embedded R scripts primarily based on the Flexdashboard package, which allowed us to set up quickly a flexible content structure. For the interactive graphs, it uses the Plotly package. The final output of the R Markdown script is a set of HTML, JavaScript and CSS files, ready for immediate publishing.

The Flexdashboard package makes it easy to create interactive dashboards in R with R Markdown. This package:

- facilitates publishing a group of related data visualisations as a dashboard;
- provides support for a wide variety of components including html widgets; base, lattice, and grid graphics; tabular data; value boxes and text annotations;

- allows to specify row- and column-based layouts; the components are intelligently re-sized to fill the browser window and are adapted for displaying on mobile devices;
- provides storyboard layouts for presenting sequences of data visualisations and related text commentaries.

The Plotly package is an R package for creating interactive web-based graphs via plotly.js, an open source JavaScript graphing library. The Plotly package allows to create line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, and 3D (WebGL based) charts.

The output of the R visualisation tool had also to comply with a number of requirements to be in line with Eurostat's standards; this includes the following:

- be platform independent with no loss of functionality or display on resized desktop screens, smartphones and tablets,
- display well on different browser, as a minimum on Google Chrome, MS Edge and Mozilla Firefox,
- respect WCAG 2.0 level AA for accessibility (this requirement will be applied in future versions of Eurostatistics),
- respect the graphical style guide of Eurostat.

With some further minor additional adjustments, it was possible to make the layout responsive to mobile devices.

Results

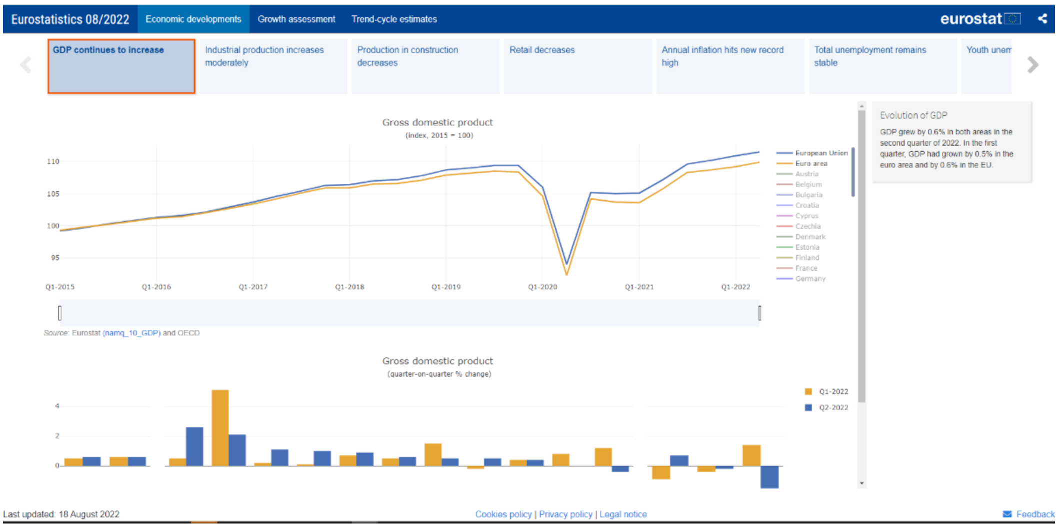
The main aim of our development has been fully reached. Thanks to the new R visualisation tool, we were able to enhance our visualisation by interactive graphs, complemented by the possibility of downloading up to date data from Eurostat dissemination environment. Users have now a number of facilities they can explore, deciding on which information is more interesting/useful for them. They can re-use proposed graphs and data in own publications, and navigate the information according to their preferences, switching among different tabs.

Concerning the production process, the R visualisation tool for Eurostatistics first loads data and then rapidly generates a set of graphs based on those data, and inserts texts from an external file we draft. The production process is much faster than traditional publishing techniques, and is therefore particularly suitable for products that are frequently updated, as in this case monthly. A set of chart functions created in Plotly makes it easy to add new graphs to the tool.

The Eurostatistics interactive web visualisation, created by the tool, includes three main themes: Economic developments (see Figure 1), Growth assessment and Trend-cycle estimates (see Figure 2).

“Economic developments” present, using a storyboard layout, a sequence of visualisations with related text commentaries for GDP (see Figure 1), industrial production, production in construction, retail trade, annual inflation, total unemployment, youth unemployment, economic sentiment and employment expectations. The visualisation for each indicator includes a line chart and a bar chart with interactive functionalities.

Figure 1: Eurostatistics - GDP



Evolution of GDP

GDP grew by 0.6% in both areas in the second quarter of 2022. In the first quarter, GDP had grown by 0.5% in the euro area and by 0.6% in the EU.

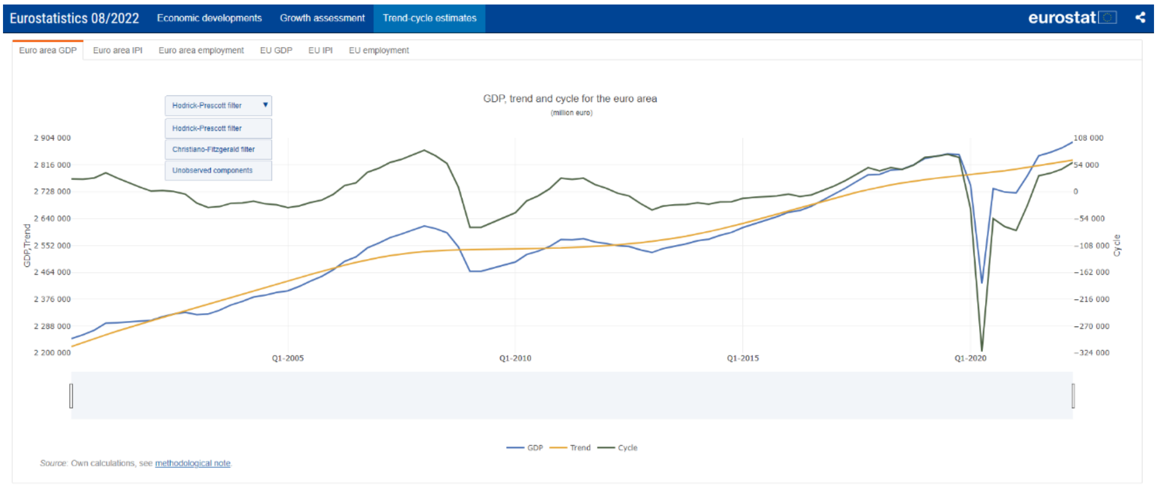
Source: Eurostat (namq_10_GDP) and OECD

Last updated: 18 August 2022

“Growth assessment” shows, also using a storyboard layout, a sequence of visualisations with related commentaries for the €-coin indicator estimated by the Bank of Italy, the Business Climate Indicator published by the European Commission’s Directorate-General for Economic and Financial Affairs, the Composite Leading Indicators (CLI) estimated by the OECD, and the Euro-Indicator published by DZ BANK. The visualisation for each indicator includes a line chart with interactive functionalities. There are also tables with forecasts for GDP and annual inflation.

“Trend-cycle estimates” present interactive line charts with trend-cycle decompositions of GDP, employment and industrial production for the EU and the euro area (see Figure 2). The results are from three methods: the Hodrick-Prescott filter, the Christiano-Fitzgerald filter and the unobserved components approach.

Figure 2: Eurostatistics - Trend-cycle estimates



Source: Own calculations, see methodological note

The storyboard layout structures the content as a narrative, so that users in addition to the data visualisations can read a commentary about them. The visuals are of high quality and easy to interpret. Interactive line charts have a number of interactive functionalities, for instance to select plot lines for countries, to change time span of time series, to read values when hovering with a mouse on plot lines. The line charts and bar charts can be downloaded.

The Eurostatistics interactive web visualisation also includes links to the data sources where the user has the possibility to explore further the data, for example by looking at more disaggregate information in a specific statistical area.

The R visualisation tool produces an output that can be browsed on mobile devices.

Conclusions

Our project has proved that R with its packages can be adopted for statistical visualisation with limited resource investment. The production process, in full hands of statisticians, is much faster compared to traditional publishing. It is particularly suitable for products that are frequently updated, as in this case monthly.

We found particular useful the storyboard layout, since it offers a user-friendly way of communicating the main messages. We also recommend the Plotly package, which allows creating interactive visuals of high quality that are easy to interpret and read.

The flexibility of the developed visualisation tool is another strength that will guarantee its maintainability. Future development could include a partial automation of the text content, although a fully automation would not probably be a feasible solution.

References

Eurostatistics interactive web visualisation (August edition):

https://ec.europa.eu/eurostat/cache/statistics_explained/eurostatistics/2022/august/visualisation.html

R documentation on Flexdashboard:
https://www.rdocumentation.org/packages/flexdashboard/versions/0.6.0

R documentation on Plotly: https://www.rdocumentation.org/packages/plotly/versions/4.10.0

Nowcasting (MANS2M.1)

Session Chair: Gian-Luigi Mazzi *(Senior Consultant)*

GDP-Flash estimates: a quality assessment
Philipp Wegmüller *(State Secretariat for Economic Affairs SECO)*

Density forecasts of inflation: a quantile regressions forest approach
Inès Moutachaker (INSEE)*; Michele Lenza, Joan Paredes (ECB)

Analyses of the connection and the interaction between retail products markets and cash-holding ratio through panics using dynamic factor
Klaudia Máténé Bella (*Hungarian Central Statistical Office*), Ildikó Ritzl-Kazimir *(Magyar Nemzeti Bank, the central bank of Hungary)*

GDP-Flash estimates: a quality assessment

Keywords: GDP, Flash estimates, Data revisions, Business cycles

Note: this abstract is based on preliminary and exploratory work

Introduction

The Corona Crisis showed the importance of quickly available data in assessing the state of the economy. This includes naturally also the most important figure: GDP. The tendency in the last few years was to shorten the time between a quarter and the first estimate of the growth rate in that quarter.

Generally, GDP is published with a delay of around 60 days with respect to the reference quarter. In order to shorten this publication lag, statistical entities have started to publish preliminary GDP estimates, so called "Flash-GDP". In the USA, such a GDP-Flash estimate is available since 1996 (first press release); the Euro Area started its publication somewhat later in 2016, while some member countries began with a publication as early as 2012. Apart from alternative monthly or higher-frequency indicators stemming from surveys, financial markets or other sources of hard data (e.g., retail sales, industrial production, employment and trade figures), the flash estimates are the most important source of information for policy makers and analysts to assess the current business cycle stance. In recent years, a broad palette of models has been developed to provide accurate nowcasts of GDP growth in the current or past quarter.[29] As those models base on indicators, they only provide an approximate signal, in which direction GDP growth will go and how large growth will be. Generally, they come along with a certain forecast error. In times of crisis, such forecast errors might be particularly large.[30] The advantage of the Flash-GDP estimate over model-based nowcasts is that GDP is calculated bottom up by means of the sum of value added from the different production sectors, it should thus provide an accurate picture of the state of the economy together with sectoral information.

In this paper, we examine the quality of the Flash-GDP estimates for 17 economies. A quality assessment seems important, as national accounts figures are subject to substantial revisions over time.[31] We raise the question, whether the Flash-estimates provide a sufficiently precise signal relative to the first official GDP release at T+60. Related to the work of Sinclair and Stekler [1], we study various characteristics of the differences between the initial estimates, available up to one month after the end of the quarter, and the first official estimates available two months after the end of the quarter: Do they have the same directional signs? Are the flash estimates unbiased estimates of the first official release? Are any observed biases related to the state of the economy? Based on a variety of statistical tests, we find strong evidence that

[29] To mention just a few: Giannone, Reichlin, and David Small [4], Bok, Brandyn, et al. [5]

[30] See for instance Siliverstovs [6]

[31] See for instance Faust, Rogers and Wright [7].

the early data is helpful to obtain a realistic, timely picture of what had happened in the economy in the previous quarter.

The reasons for revisions between release of the Flash-estimate and the first official release (or compared to the final available vintage) are manifold. On the one hand, fewer statistics are included in earlier national accounts calculations of a quarter than in later calculations, as more information becomes available on an ongoing basis. On the other hand, the basic data used in the calculation are often revised. In addition, the definitions and calculation methods used for the basic data as well as for the national accounts themselves may change. In the case of quarterly national accounts, for example, the methods used for seasonal adjustment or temporal disaggregation may change. In general, national accounts results are expected to show some inaccuracies in the short and medium term, which decrease or disappear over time, so that published values eventually converge to the "true values". In fact, the revisions of earlier calculations of a quarter are usually larger than the revisions of later calculations. Thus, there is a fundamental trade-off between earlier availability of decision-relevant information and its accuracy, which has to be weighed up by the responsible departments.

The quality and quantity of the available data basis and the reliability of the applied methodology play a major role. Using a broad international sample, this paper compares the quality of quarterly GDP flash estimates on the basis of various indicators on the basis of various key figures. This allows certain conclusions to be drawn about the quality of GDP figures and their flash estimates. The focus is placed in particular on revisions between the flash estimate and the first publication after the end of the quarter.

Overall, it can be seen that growth rates are not subject to systematic revisions in the short and medium term, i.e. they are faithful to expectations. In the long-term comparison, there are some indications of systematic revisions to a minor extent. However, these are mainly due to extraordinary (i.e., "Benchmark") revisions. Moreover, the level of GDP revisions is low by international standards. In the years of the Covid-pandemic, however, revisions were substantial.

Methods

We compiled real-time vintages for 17 countries. For the USA, the vintages start as early as 1996:Q2 (first press release), for the EuroArea the data starts in 2016:Q1, while we were able to gather data spanning back to 2012:Q1. For some member countries (Belgium, Spain, Germany, Hungary, Latvia, Portugal, and Romania), we were able to collect vintages starting in 2012:Q1.

We apply a variety of statistical ratios to assess the revisions. Here follows only a brief list of the most important quantities considered in this study: Mean Revision (MR): The (arithmetic) mean revision gives an indication of whether revisions systematically deviate from zero, i.e. whether GDP tends to be revised in a certain direction. Standard deviation of revision (SDR): The standard deviation of revision is a measure of the dispersion of revisions. Mean Absolute Revision (MAR): The (arithmetic) mean absolute revision indicates how large the revisions usually are. The root mean squared revision (WMQR): The root mean squared revision indicates how much the revisions vary in size, since the quadratic function makes larger revisions more important than small ones. Correlation (KORR): The correlation provides information about the extent of the (linear) relationship between two variables. It is normalized between -1 and 1. Is in the present case the correlation between revisions, that occur within different time intervals is significantly different from zero, it means that one revision can be used to predict another revision. News and Noise: By examining the correlation between GDP growth rates and revisions, it is possible to assess whether national accounts results are revised primarily because of so-called noise (“Noise“), for example calculation and measurement errors, or because of new information (“News“). The latter is desirable. In the case of “News“, the WMQR for longer-term revisions usually increases. For “noise“, the opposite is true.

Apart from these statistical ratios, we study directional errors and systematic errors. For the former, we base our analysis on the work of Joutz and Stekler (1998) [2], for the later we apply Mincer and Zarnowitz (1969) [3] regressions.

Results

In the following, we present some preliminary results. In Figure 1, we display the mean absolute revision between Flash-GDP and the final available data vintage. We consider here the whole time span, which is available for each country. The final vintage of all countries corresponds to 2021:Q4.

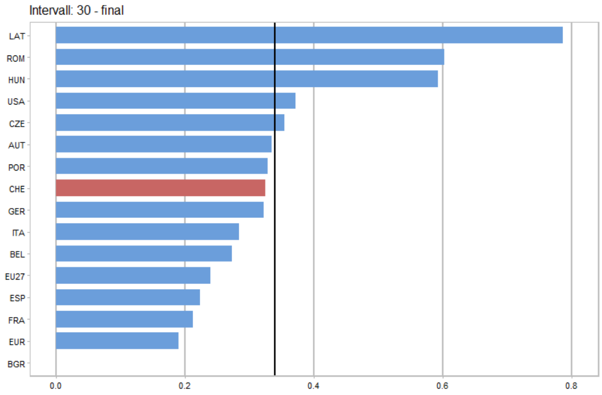


Figure 10. Mean absolute revision across countries

Evidently, the EuroArea as an aggregate features the lowest MAR with merely 0.2 percentage points. The average across countries is 0.35 percentage point. We also note that the revisions for the USA are slightly above average. Notably, most countries exhibit similar mean absolute revisions.

In Figure 2, we show the mean absolute revision of all countries across time. While the revisions have been relatively small prior to 2020, it is evident how the crisis of the Corona-pandemic has lead to substantial revisions in GDP figures between 2020 and 2022. While the reasons for these substantial revisions have to be further investigated, we hypothesize that reasons lie in difficulties with indicators to capture the pandemic-related outfalls and shortages, model uncertainty and model revisions.

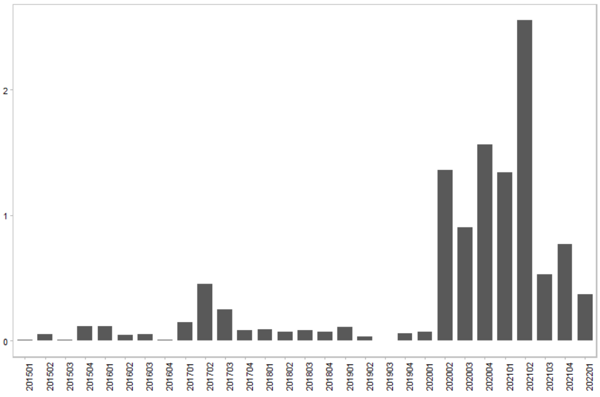


Figure 2. Mean absolute revision across time

Conclusions

In the present, preliminary work, we investigate the revision patterns of Flash-GDP estimates relative to the official GDP release. We show that across countries, the revisions are comparable in size; however, they increased substantially in the period of the Coronavirus pandemic. Given that GDP is the most important indicator to assess the current stance of the business cycle, it is important that statistical agencies put a lot of emphasis to ensure high quality standards and reliability in providing GDP-Flash estimates. The revisions of Flash-GDP estimates to their first official release are relatively small and the benefit of having an early assessment of GDP growth exceeds by far the cost incurred by the revisions.

References

[1] Tara M. Sinclair, H.O. Stekler, Examining the quality of early GDP component estimates, International Journal of Forecasting, Volume 29, Issue 4, 2013, Pages 736-750

[2] F.L. Joutz, H.O. Stekler, Data revisions and forecasting, Applied Economics, 30 (1998), pp. 1011-1016

[3] J. Mincer, V. Zarnowitz, The evaluation of economic forecasts J. Mincer (Ed.), Economic forecasts and expectations, National Bureau of Economic Research, New York (1969)

[4] Giannone, Domenico, Lucrezia Reichlin, and David Small. "Nowcasting: The real-time informational content of macroeconomic data." Journal of monetary economics 55.4 (2008): 665-676.

[5] Bok, Brandyn, et al. "Macroeconomic nowcasting and forecasting with big data." Annual Review of Economics 10 (2018): 615-643.

[6] Siliverstovs, Boriss. "Assessing nowcast accuracy of US GDP growth in real time: the role of booms and busts." Empirical Economics 58.1 (2020): 7-27.

[7] Faust, Jon, John H. Rogers, and Jonathan H. Wright. "News and noise in G-7 GDP announcements." Journal of Money, Credit and Banking (2005): 403-419.

Density forecasts of inflation: a quantile regressions forest approach

Abstract

Assessing the future implications of economic shocks for the inflation dynamics is a fundamental challenge for medium-term oriented central banks, such as the European Central Bank (ECB). We show that a random forest, capturing a general non-linear relationship between euro area inflation and a large set of determinants, performs well in an out-of-sample evaluation of point and density forecasts over the last two decades. The random forest performs similarly to state-of-the-art linear benchmarks over the full sample under analysis and, hence, it complements rather than substitutes more conventional techniques. Remarkably, however, the random forest proves particularly competitive in capturing the long period of low inflation in the pre-COVID decade and combining its forecasts with those of linear models leads to gains in forecasting accuracy.

Keywords: Inflation, Non-linearity, Random Forest, Combination.

JEL Codes: C52, C53, E31, E37

Introduction

The mandate of the European Central Bank (ECB) is to maintain price stability over the medium term. Because of such medium term orientation, in its regular economic analysis, the ECB assesses the current and prospective driving forces of inflation and interprets their nature. If such forces have only expected temporary effects on inflation, the ECB is more likely to look through them. If, instead, such forces are of a more persistent nature, they may have an impact on monetary policy decisions. Hence, inflation projections, which condense the views of the Eurosystem staff on future inflation dynamics are a crucial element of the monetary policy briefing.

However, despite a very large literature, modelling inflation dynamics in the euro area remains elusive, owing to the many potential driving factors of inflation trends and cyclical fluctuations (see, for example, Koester et al., 2021) and the difficulty in capturing their relationship with inflation dynamics. One key point pertains to whether inflation dynamics are better characterized by a linear or a non-linear relationship with its potentially many determinants.

A survey of the models used in the Eurosystem economic analysis in Darracq Pari`es et al. (2021), conducted in the context of the recent ECB strategy review, reveals that the Eurosystem modeling toolbox consists mostly of linear models. Yet, the relationship of consumer prices with their determinants may be characterized by non-linearity. For example, among others, Lind´e and Trabandt (2019); Costain et al. (2022) argue that inflation dynamics are essentially state-dependent and that some form of state-dependence could help to explain the inflation dynamics over the last fifteen years. For these reasons, the study of non-linearity in macroeconomic dynamics can be considered as a relevant gap for the Eurosystem modelling toolbox. In this paper, we help to fill this gap by proposing a new non-linear model to forecast inflation in the euro area. We measure inflation in terms of rate of change of the Harmonized Index of Consumer Prices (HICP). The determinants of inflation in our model include measures of inflation expectations, cost pressures and real activity, broadly inspired to the Phillips Curve framework. For what concerns the relationship of the determinants with inflation, we employ the random forest model (Breiman, 2001), which is able to capture very general functional forms, encompassing non-linearity.

The random forest is an ensemble technique, combining a number of non-linear predictive models, called regression trees. Regression trees split the sample of the predictors in (potentially many) sub-samples and form a prediction of the target variable according to the average value or the quantiles of the distribution of the latter in each particular sub-sample. These models can capture very general forms of non-linearity because they do not assume any specific parametric relationship between predictors and the target variables. However, regression trees typically suffer from overfitting, displaying a rather poor out-of-sample forecasting performance. The random forest tackles the issue of overfitting by combining a potentially large set of forecasts from different regression trees.

In our empirical application, we evaluate the random forest model in a recursive out-of-sample exercise and we focus both on point and density forecasts. Our first finding is that the random forest outperforms the standard autoregressive benchmarks of non-forecastability. We also compare the Random Forest to a state-of-the-art linear benchmark, i.e. a combination of bivariate vector autoregressive models estimated with bayesian techniques (BVARs), each including HICP and one of the inflation determinants. We look at a combination of BVARs so that our benchmark differs from the random forest mainly for the assumption of linearity in the inflation dynamics. We find that the performance of the random forest is comparable to that of the linear benchmark, on average over the sample. This result may suggest that euro area inflation dynamics are not characterized by pervasive non-linearity. However, the comparable forecasting accuracy across models masks a relevant heterogeneity in sub-samples(see Giacomini and Rossi, 2009, 2010; Rossi and Sekhposyan, 2016). More specifically, the linear model outperforms the random forest in the run up to, during and in the aftermath of the Great Recession, while the latter was faster to adapt in the period characterized by low inflation and the attainment of the zero lower bound for nominal interest rates in the euro area. The "diversity" brought by the random forest implies that combining its forecasts with those of the linear benchmark produces forecasts which are, on average, more accurate than those from the individual models(Timmermann, 2006).

Our paper relates to a growing literature on the virtues of (machine learning) ensemble methods, which are becoming more and more popular in the econometric literature for prediction (Athey et al., 2019; Avramov, 2002; Bai and Ng, 2009; Cremers, 2002; Faust et al., 2013; Fernandez et al., 2001; Inoue and Kilian, 2008; Jin et al., 2014; Ng, 2013; Rapach and Strauss, 2010; Sala-I-Martin et al., 2004; Varian, 2014; Wager and Athey, 2018; Wright, 2009). Giannone et al. (2021) shows that ensemble methods may be particularly successful thanks to their ability to appropriately handle model uncertainty. Medeiros et al. (2021) shows that the random forest model helps to predict US inflation. We focus on the euro area and we look also at the uncertainty surrounding the point forecasts, which is extremely important for monetary policy.

We also contribute to the literature on potential non-linearity in inflation dynamics. Specifically, a large literature studies the likelihood of changes in the shape of the Phillips Curve and the factors which may potentially explain such changes. For an extensive survey and a systematization of the debate, see Del Negro et al. (2020). Several papers in this literature point to a different relationship of inflation with its determinants in high and low inflation regimes (see, for example Akerlof et al., 1996; Costain et al., 2022; Fahr and Smets, 2010; Benigno and Ricci, 2011; Lind´e and Trabandt, 2019; Forbes et al., 2021).

The rest of the paper is organized as follows. In section 2, we discuss our data and the empirical strategy. Section 3 presents the results. Section 4 concludes.

Empirical models, data and out-of-sample evaluation

Empirical model

We adopt a "direct" forecasting scheme according to which, for a generic forecasting horizon *h*, we estimate the relationship of inflation at time *t* with its determinants at time *t*-*h*. Then, we apply the estimated model on the data at time *t* to produce an inflation forecast at time *t*+*h*. The target variable in our exercises isπ_{t+h}^{h}, i.e. the annualized growth rate of the Harmonized Index of Consumer Prices (HICP, defined P_t), at the forecast horizons of 3, 6 and 12 months ahead (h=3,6,12):

$$\pi_{t+h}^{h} = (12/h) \times [ln(P_{t+h}) - ln(P_t)]$$

Formally, we would like to estimate a non-linear relationship between our target concept of inflation π_t, its lags and a set of determinants x_{t-h} and their lags:

$$\pi_t^h = m(\pi_{t-h}^1 ... \pi_{t-h-p}^1; x_{t-h} ... x_{t-h-k}) + \varepsilon_t$$

and then obtain an inflation forecast as

$$\hat{\pi}_{t+h}^h = m(\pi_t^1 ... \pi_{t-p}^1; x_t ... x_{t-k})$$

Rather than tightly parameterizing *m*(.) by committing to a specific form of non-linearity, we capture quite general forms of non-linearity by resorting to machine learning techniques. In particular, we estimate the potentially non-linear relationship of inflation with its determinants by means of a random forest (Breiman, 2001). A random forest is an ensemble method which combines the results from a certain (potentially large) number of non-linear models, called regression trees.

A regression tree fits a specific target variable (inflation, in our case) by repeatedly splitting the sample of the potential predictors in different sub-samples. Once the final split is achieved, the predicted value of the target variable associated with a specific sub-sample is represented by the sample mean or median of the target variable in that sub-sample, for point prediction. Instead, density prediction can be carried out by computing the empirical quantiles of the target variable associated with each specific sub-sample of the predictors (see Meinshausen, 2006). In our paper, we focus both on point and density prediction. The sub-sample splits in a regression tree are obtained through a process known as binary recursive partitioning, an iterative process that splits the data into partitions. The process continues until the splits achieve an improvement in terms of a statistical criterion, such as the mean squared error in the fit for inflation or, alternatively, until the splitting process hits a stopping rule which, in our case, is that any final sub-sample on which predictions are computed should include at least ten data points for the target variable. To develop different regression trees and to maximize the advantages of combining them, the original data are bootstrapped with replacement before constructing any new tree[32] and the splits are computed, at each node, only by looking at a randomly selected set of the regressors. The default choice for the size of the latter set, which we follow in this paper, is to draw a third of the variables for each split. To complete the description of the random forest we set the number of combined regression trees to the default value of 500.[33]

We compare the predictions from the random forest to two benchmarks. First, we consider a popular benchmark of non-forecastability, i.e. the random walk (RW) model, according to which the forecast of inflation is the last recorded value of inflation similarly to Atkeson and Ohanian (2001)

$$\hat{\pi}^h_{t+h,RW} = \pi^h_t$$

Second, we compare the random forest to a equally weighted combination of 62 bivariate VAR forecasts. The bivariate VAR models feature inflation and one of the 62 determinants described in the data sub-section. The models are estimated using bayesian techniques. The prior distributions for the lag and error variances are in the Normal-Inverse Wishart class and are parameterized to shrink the model estimates toward the parameters of a random walk model for each variable, in the tradition of the Minnesota prior Litterman (1979); Doan et al. (1984); Banbura et al. (2010). The prior hyperparameters are treated as random variables and their value is drawn from their posterior, following Giannone et al. (2015).

Data

Beside headline HICP, our target variable, our database contains 62 variables. The data is obtained from the ECB Statistical Data Warehouse (SDW) and comes from a variety of original sources. The choice of the variables is similar to de Bondt et al. (2018). Broadly speaking, the dataset is inspired by the Phillips Curve framework, covering different areas of the economy.

Specifically, we include measures of external cost pressures (for example, commodity prices, exchange rates, global indicators); domestic price and cost variables (for example, wages and

[32] Notice that inflation is very likely to be auto-correlated, so we also include several lags of inflation in the inflation determinants, so that the bootstrap procedure does not impair the ability of our model to account for the autoregressive dynamics of inflation

[33] See Probst et al. (2019) for a discussion of the default specification choices for random forests and of the techniques to tune the model.

producer prices); survey and hard data on economic activity (for example, Purchasing Manager Index and European Commission surveys on prices, employment expectations, confidence measures, industrial production, euro area business cycle indicators, various productivity measures); measures of inflation expectations (for example, survey and market-based measures over different forecast horizons); and financial variables (for example, interest rates, monetary aggregates, asset prices, bank lending).

Our sample ranges from December 1991 to December 2019 and the frequency of the data is monthly. We stationarize the data, when needed. Appendix A gives more details on the data, their source and the transformations we apply before estimating the models.

[APPENDIX A - DATA APPENDIX - TO BE ADDED AT THE END]

Out-of-sample evaluation

Our out-of-sample exercise is carried out according to a recursive updating scheme. Specifically, first, we estimate our models until December 2002, and we produce forecasts for inflation at the three, six and twelve months horizon (spanning change in prices until March, June and December 2003). Then, we continue to update the estimation sample by adding one month at a time, and we repeat all the steps of the forecasting exercise until exhaustion of the sample.

For the point forecasts, we take the median of the posterior distributions of the random walk, the BVAR and the Random Forest. We assess the accuracy of the point forecasts by computing the associated Mean Squared Forecast Error (MSE). To assess the accuracy of the density forecasts, instead, we compute the Continuous Ranked Probability Score (CRPS) which, roughly, measures the distance of the cumulative distribution function predicted by a specific model from the true one (see Gneiting and Raftery, 2007).

We focus both on the average accuracy over the whole sample and on the evolution of the measures of accuracy over time, to gauge the ability of the different models to capture the dynamics of inflation in different regimes (Giacomini and Rossi, 2009, 2010; Rossi and Sekhposyan, 2016).

Results

Table 1 shows the results of our out-of-sample comparison. The results are reported in terms of ratio to the MSE of the random walk model, for point forecasts (panel a), and ratios to the CRPS of the random walk model, for density forecasts (panel b). Values smaller than one of the two ratios imply that the specific model under analysis outperforms the random walk. The last column of the table reports results for an equal weight forecast combination of the Random Forest and the combined BVAR forecasts.

Table 1: Ratios of MSE and CRPS of different models vs the RW

Horizon	Random Forest	BVAR	Combination
Panel a: MSE ratios			
h=3	0.63	0.63	0.60

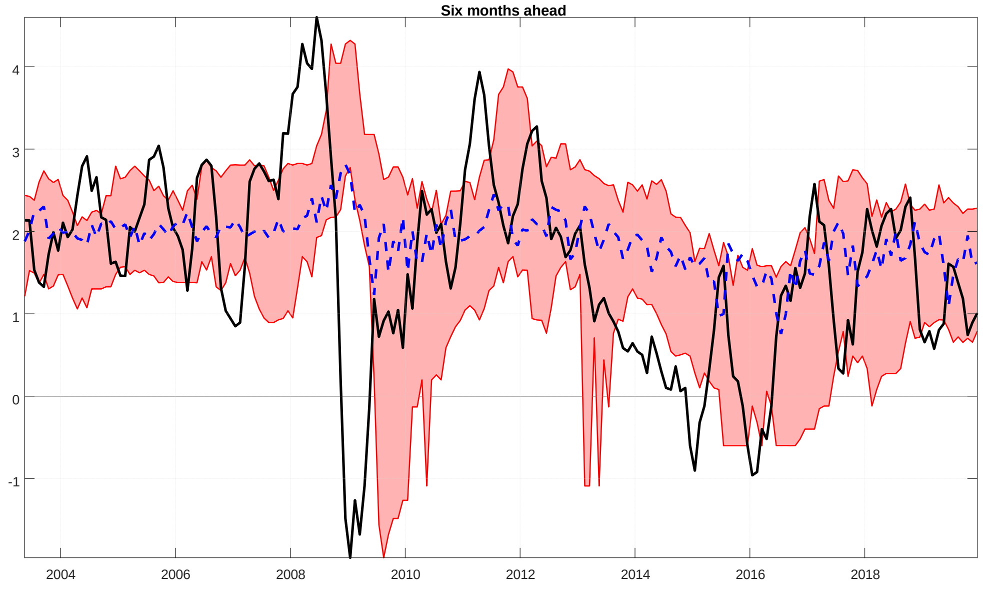
h=6	0.74	0.70	0.68
h=12	0.69	0.72	0.65
Panel b: CRPS ratios			
h=3	0.77	0.77	0.75
h=6	0.84	0.84	0.81
h=12	0.86	0.86	0.82

The random forest outperforms the random walk by a comfortable margin, at all horizons, both in terms of point and density forecasts. Instead, the performance of the random forest is similar, on average across the whole sample and both for density and point forecasts, with respect to the linear benchmark. Hence the random forest and the VAR combination are good forecasting models, but one should not lightheartedly discard the hypothesis that inflation dynamics are approximately linear.

However, interestingly, the forecast combination of the random forest and the VAR forecasts outperforms both individual forecasts. This suggests that the forecasts based on the non-linear and the linear models are sufficiently “diverse”.

In order to dig deeper on this point, figure 1 shows the forecasts of the random forest (orange area) and the median forecast from the BVAR combination (dashed blue line) together with realized inflation (black solid line), at the horizon of six months ahead.

Figure 1: Inflation forecasts, six months ahead horizon

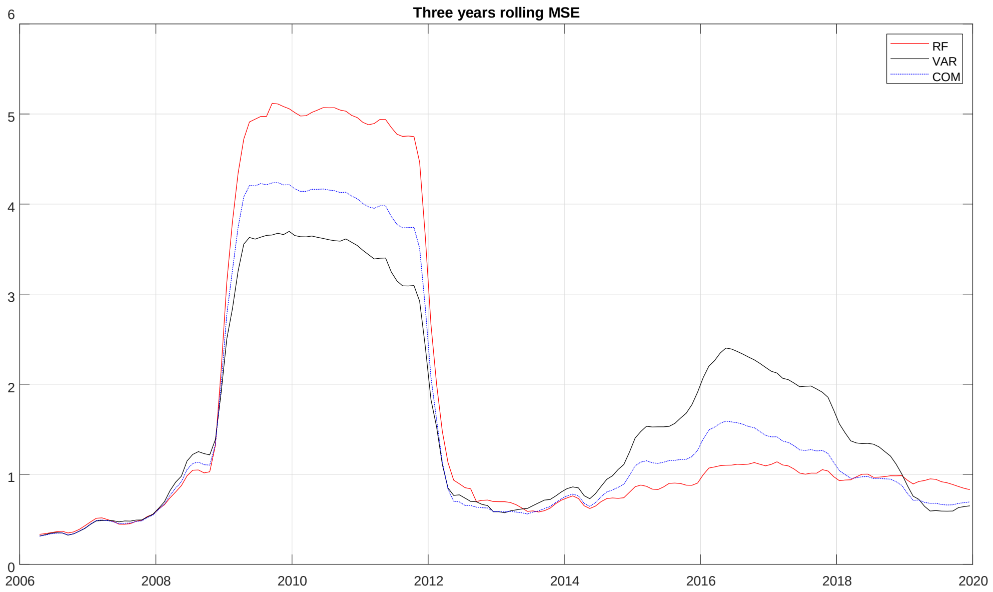


Note: ADD NOTE

Despite some periods with relevant overlaps, the two forecasts display relevant differences over specific parts of the sub-sample and it is visible how their relative accuracy changes over time. The evolution over time in the relative accuracy of the two forecasts can be appreciated in figure 2, which shows the rolling MSE of the random forest, the VAR combination and the combination of the linear and the non-linear forecasts.[34]

Figure 2 shows that the linear model is better able than the random forest to account for the quick inflation rebound post Great Recession, detecting earlier than the random forest the inflation trough and having been less reactive than the random forest forecast throughout the crisis period.

Figure 2: Three years rolling MSE



Note: Add Note

Generally, the random forest lags the inflation dynamics on the run-up to, during and in the aftermath of the Great Recession, casting some doubt on the view that to reconcile the inflation dynamics in that period with economic theory one would need to invoke strong non-linearity. This result is in line with Bobeica and Jarocin´ski (2019) which, in an ex-post conditional forecasting exercise, show that a linear model is able to accurately describe the inflation dynamics post-Great Recession. At the same time, the random forest adapts much faster than the VAR forecasts to the prolonged period of low inflation characterizing the pre-COVID decade. The accuracy of the non-linear model in this episode suggests that low inflation regimes may be characterized by different inflation dynamics than high inflation regimes as in Forbes et al. (2021). However, as a caveat, our evaluation sample is relatively short compared to sample analyzed in Forbes et al.
(2021), making the identification of high and low inflation regimes potentially challenging.

[34] For the sake of brevity, we don't report the chart with the rolling CRPS, which gives a very similar message as the chart of the rolling MSE.

Conclusion

In this paper, we show that the random forest, a non-linear model, could be a useful addition to the current modelling toolbox to forecast euro area inflation, which is heavily skewed toward linear models.

Specifically, we show that the random forest has comparable accuracy to state-of-the-art BVAR models. At the same time, its forecast is“‘diverse” enough from that of BVAR models, and combining linear and non-linear approaches leads to an improvement of the forecast accuracy over individual models(Timmermann, 2006).

The similar accuracy of linear and non-linear models over the full sample suggests that nonlinearity is not a pervasive feature of euro area inflation dynamics but, when looking at subsamples, the non-linear model turned out to be more accurate over the relatively long period of low inflation which has characterized the previous decade.

A Database

To be added

References

Akerlof, G. A., W. R. Dickens, and G. L. Perry (1996): "The Macroeconomics of Low Inflation," *Brookings Papers on Economic Activity*, 27, 1–76.

Athey, S., M. Bayati, G. Imbens, and Z. Qu (2019): "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, 109, 65–70.

Atkeson, A. and L. E. Ohanian (2001): "Are Phillips curves useful for forecasting inflation?" *Quarterly Review*, 25, 2–11.

Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64, 423–458.

Bai, J. and S. Ng (2009): "Boosting diffusion indices," *Journal of Applied Econometrics*, 24, 607–629.

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25, 71–92.

Benigno, P. and L. A. Ricci (2011): "The Inflation-Output Trade-Off with Downward Wage Rigidities," *American Economic Review*, 101, 1436–1466.

Bobeica, E. and M. Jarocinski´ (2019): "Missing Disinflation and Missing Inflation: A VAR Perspective," *International Journal of Central Banking*, 15, 199–232.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Costain, J., A. Nakov, and B. Petit (2022): "Flattening of the Phillips Curve with statedependent prices and wages," *The Economic Journal*, 132, 546–581.

Cremers, K. M. (2002): "Stock return predictability: A Bayesian model selection perspective," *Review of Financial Studies*, 15, 1223–1249.

Darracq Paries, M., A. Notarpietro, J. Kilponen, N. Papadopoulou, S. Zimic, P. Al-` dama, G. Langenus, L. J. Alvarez, M. Lemoine, and E. Angelini (2021): "Review of macroeconomic modelling in the Eurosystem: current practices and scope for improvement," Occasional Paper Series 267, European Central Bank.

de Bondt, G., E. Hahn, and Z. Zekaite (2018): "ALICE: A new inflation monitoring tool," Working Paper Series 2175, European Central Bank.

Del Negro, M., M. Lenza, G. Primiceri, and A. Tambalotti (2020): "What's up with the Phillips Curve?" *Brookings Papers on Economic Activity*, Spring.

Doan, T., R. Litterman, and C. Sims (1984): “Forecasting and conditional projection using realistic prior distributions,” *Econometric reviews*, 3, 1–100.

Fahr, S. and F. Smets (2010): “Downward Wage Rigidities and Optimal Monetary Policy in a Monetary Union,” *Scandinavian Journal of Economics*, 112, 812–840.

Faust, J., S. Gilchrist, J. H. Wright, and E. Zakrajˇssek (2013): “Credit Spreads as Predictors of Real-Time Economic Activity: A Bayesian Model-Averaging Approach,” *The Review of Economics and Statistics*, 95, 1501–1519.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): “Model uncertainty in cross-country growth regressions,” *Journal of Applied Econometrics*, 16, 563–576.

Forbes, K. J., J. E. Gagnon, and C. G. Collins (2021): “Low inflation bends the Phillips curve around the world: Extended results,” Working Paper Series WP21-15, Peterson Institute for International Economics.

Giacomini, R. and B. Rossi (2009): “Detecting and Predicting Forecast Breakdowns,” *Review of Economic Studies*, 76, 669–705.

——— (2010): “Forecast comparisons in unstable environments,” *Journal of Applied Econometrics*, 25, 595–620.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): “Prior Selection for Vector Autoregressions,” *The Review of Economics and Statistics*, 97, 436–451.

——— (2021): “Economic Predictions With Big Data: The Illusion of Sparsity,” *Econometrica*, 89, 2409–2437.

Gneiting, T. and A. E. Raftery (2007): “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102, 359–378.

Inoue, A. and L. Kilian (2008): “How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation,” *Journal of the American Statistical Association*, 103, 511–522.

Jin, S., L. Su, and A. Ullah (2014): “Robustify Financial Time Series Forecasting with Bagging,” *Econometric Reviews*, 33, 575–605.

Koester, G., E. Lis, C. Nickel, C. Osbat, and F. Smets (2021): “Understanding low inflation in the euro area from 2013 to 2019: cyclical and structural drivers,” Occasional Paper Series 280, European Central Bank.

Linde, J. and M. Trabandt´ (2019): “Resolving the Missing Deflation Puzzle,” CEPR Discussion Papers 13690, C.E.P.R. Discussion Papers.

Litterman, R. B. (1979): “Techniques of forecasting using vector autoregressions,” Tech. rep.

Medeiros, M. C., G. F. R. Vasconcelos, Alvaro Veiga, and E. Zilberman´ (2021): “Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods,” *Journal of Business & Economic Statistics*, 39, 98–119.

Meinshausen, N. (2006): “Quantile Regression Forests,” *Journal of Machine Learning Research*, 7, 983–999.

Ng, S. (2013): “Variable Selection in Predictive Regressions,” in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, Elsevier, vol. 2 of *Handbook of Economic Forecasting*, chap. 0, 752–789.

Probst, P., M. N. Wright, and A.-L. Boulesteix (2019): “Hyperparameters and tuning strategies for random forest,” *WIREs Data Mining and Knowledge Discovery*, 9, e1301.

Rapach, D. and J. Strauss (2010): “Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth,” *Econometric Reviews*, 29, 511–533.

Rossi, B. and T. Sekhposyan (2016): “Forecast Rationality Tests in the Presence of Instabilities, with Applications to Federal Reserve and Survey Forecasts,” *Journal of Applied Econometrics*, 31, 507–532.

Sala-I-Martin, X., G. Doppelhofer, and R. I. Miller (2004): “Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach,” *American Economic Review*, 94, 813–835.

Timmermann, A. (2006): “Chapter 4 Forecast Combinations,” Elsevier, vol. 1 of *Handbook of Economic Forecasting*, 135–196.

Varian, H. R. (2014): “Big Data: New Tricks for Econometrics,” *Journal of Economic Perspectives*, 28, 3–28.

Wager, S. and S. Athey (2018): “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Journal of the American Statistical Association*, 113, 1228–1242.

Wright, J. H. (2009): “Forecasting US inflation by Bayesian model averaging,” *Journal of Forecasting*, 28, 131–144.

Analyses of the connection and the interaction between retail products markets and cash-holding ratio through panics using dynamic factor

Keywords: Panic, cash-holding ratio, dynamic factor model, Google Trends data.

Introduction

In recent years, we have experienced several panics in various areas of the economy. These include stock market panics, panic buying, bank runs and cash demand shocks. Such panic was observable during the financial crisis in October of 2008 in Hungary, or after the outbreak of the Russian-Ukrainian war in Central and Eastern Europe and in the initial period of COVID-19 pandemic worldwide.

Panics can be examined from a sociological, economic and psychological point of view. The panic is a collective flight from real or imagined threat by definition of sociology. (Smelser, 1962) In economic concept the panic contradicts the assumption of rational behaviour. In terms of panics Keynes highlighted the role of "animal spirit". When someone faces with an uncertain decision-making situation avoids this or yields to the spontaneous urge to action. (Keynes, 1936) This urge to action can be affected by psychological factors, like uncertainty, perceptions of severity, scarcity and anxiety. (Omar, et al. 2021) Kiss et al. proved the effects of negative experiences on the decision makers. (Kiss et al. 2018) Harmon et al. analysed the single day panics using network analysis. They found that the panics are preceded by market mimicry process while the influence of external news are weak. (Harmon et al. 2015) The panics have effects on discretional spendings, and the media has a significant effect on it. (Loxton, 2020) The psychology underlines that the behaviour of people can be changed in case of natural disasters, healthcare crises and terrorist attacks. (Forbes, 2017)

This research focuses on the Central European countries, Austria, Czech Republik, Germany, Hungary, Poland, Romania and Slovakia. It identifies the panic buying and panic cash withdrawals between 2004-2022, and it examines the similarities between the panic buying and panic cash withdrawals.

Methods

Data

In order to identify the panics and their interactions the cash holding ratio was calculated. The retail trade data would be appropriate for the purpose of analyses, but the monthly indicator in long time series were not available. We collected the search terms indices from the Google Trends, which illustrates the developments of retail trade in case of some specific search term, in addition they draw up the formation of panics. The cash holding ratio were calculated by the Cambridge formula. The chain linked and seasonally adjusted GDP was used for proxy for real income. We disaggregated the quarterly time series by Chow-Lin method, the seasonally adjusted volume index of industry was applied as an indicator time series. For the selected countries the harmonized consumer price indices were downloaded.

Monthly indices were calculated from Google Trends data for the following search terms: firewood, fuel, passport, solar panel, yeast, AdBlue, toilette paper. The search terms were used in the official language of the selected countries.

Applied method

Using the database containing monthly frequency data we examined the cointegration of available time series by Johansen cointegration test. In the next step we checked the stationarity of time series, because the same order of integration is necessary. The stationary was tested by Augmented Dickey-Fuller test.

The time series gained from Google Trends indices illustrates only single fields of panics. Therefore, we ordered these indices into dynamic factors in order to estimate the so-called "panic indicator" in this step we used the whole set of Google Trends indices for all analysed countries.

In the third step we examined the factors and identified this factor which summarizes the effects of panics. Finally, we fitted a vector error correction model (VECM) to the cash holding ratio and factor time series in order prove the connection and the interaction of simultaneous panics in several areas.

Results

According to our results the examined time series are cointegrated and have the same level of integration. Using the Google Trends data, we calculated the factors which show the effect of panics in retail products market. The cointegration of factors and cash holding ratio time series was tested, as well. Finally, we fitted a VECM to the time series, because the cash holding time series and factor were cointegrated by each other.

Conclusions

We identified the fields which were exposed to the effects of panics and the main panics in the examined period.

We calculated the so-called panic factor on the goods market.

We proved there is a connection and an interaction between cash holding ratio and goods market through panic factor.

References

[1] N. J. Smelser, Theory of Collective Behavior Routledge, Oxon, (1962)

[2] J. M. Keynes, The General Theory of Employment, Interest, and Money. Palgrave Macmillan, (1936)

[3] N. A. Omar and M. A. Nazri and M. H. Ali and S. S. Alam, The panic buying behavior of consumerts during the COVID-19 pandemic: Examining the influences of uncertainty, perceptions of severity, perceptions of scarcity, and anxiety, Journal of Retailing and Consumer Services, 62, (2021)

[4] H. J. Kiss and I. Rodriguez-Lara and A. Rosa-Garcia, Panic bank runs, Economics Letters Volume 162, January (2018), pp. 146-149

[5] D. Harmon and M. Lagi and M. A. M. de Agular and D. D. Chinellato and D. Braha and I. R. Epstein and Y. Bar-Yam, Measures of Collective Panic, PLoS ONE 10(7), (2015)

[6] M. Loxton and R. Truskett and B. Scarf and L. Sindone and G. Baldry and Y. Zhao, Consumer Behaviour during Crises: Preliminary Research on How Coronavirus Has Manifested Consumer Panic Buying, Herd Mentality, Changing Discretionary Spending and the Role of the Media in Influencing Behaviour, Journal of Risk and Financial Management, (2020) 13(8) 166.

[7] S. L. Forbes, Post-disaster consumption: analysis from the 2011 Christchurch earthquake. The International Review of Retail, Distribution and Consumer Research. Volume 27, Issue 1. (2017)

European Master in Official Statistics (EMOS) (GASP2M.2)

Session Chair: **Mariana Kotzeva** *(Eurostat)*

Moderator: **Maja Islam** *(Eurostat)*

Estimating Spatial Distribution of the NEET Rate in Italy by Small Area Methods
Sara Sodini *(University of Pisa)*

Measuring competitiveness: a new composite indicator for Italian municipalities
Anna Scaccabarozzi *(University of Bergamo)*

Interactions within a multi-layer EU inter-bank network
Chloé Maisonnave *(Ensai/Rennes 1)*

Estimating Spatial Distribution of the NEET Rate in Italy by Small Area Methods

Introduction

Studying the distribution of the NEET phenomenon [1] [2] has become more and more relevant over the last few years. The NEET label, which signifies that a young person is not in education, employment or training, first emerged in the United Kingdom towards the end of the last century. Due to the absence of a standardised definition at a global level and given the size of the category of reference, governments have adopted the NEET concept, applying it to the specific contexts that characterise their countries. Its use became widespread after 2010, when the European Union adopted the NEET rate as a reference indicator regarding the state of the younger population.

Formally, the NEET indicator is computed by Eurostat as a ratio where the numerator is the number of young people who are neither included in a work or study path nor in training, while the denominator is all young people in the population, having the same age as the reference group. It is important to distinguish it from the Unemployment rate, and we can find two main reasons: the NEET denominator is composed of all young people, while in the unemployment rate's denominator we find only the economically active population; the NEET indicator manages to capture not only the unemployed component, and therefore the active part, but also those who are inactive, even if this last category is at the same time the most heterogeneous, complicated and difficult to analyse.

Heterogeneity is a characteristic feature of the NEET population. The profiles found in the category, considering the 15-34 age group, are multiple and become clearer especially when the reasons that push young people to leave the training system and place themselves outside the labour market are analysed. According to the Eurofound [1], different macro-categories can be identified in order to describe the NEET category, such as the **Unemployed**, the **Unavailable**, the **Discouraged workers** and **Other inactive**.

With respect to the size of the phenomenon, Eurostat data showed that Italy had higher levels than the European average even before the economic crisis (18.8% in 2007 compared to 13.2% in EU-28). The phenomenon increased most during the crisis (the rate rose to 26.2% in 2014 against 15.4% EU-28). According to the latest data available (2019, ISTAT), Italy has the highest percentage (22.2%) of young people in this condition in the entire euro area. This means that, even if in relative terms the percentage has slightly decreased each year, in absolute numbers the stock of young people is still very high. In fact, since 2013 the NEET quota has not fallen below 2 millions of young people. In addition, concerning the last year, in Southern Italy, the incidence of NEETs was more than double (33.0%) that in the North (14.5%) and much higher than that recorded in the Centre (18.1%).

Extensive research has investigated the several and interlinked factors that cause the NEET status. The first set of factors are the ones related to each individual, such as the level of education, gender and all the other factors linked to the socio-economic status of the family. A peculiarity typical of Italy emerges with young people, as well as those already into their thirties, being dependent on their parents. Another Italian peculiarity, which certainly contributes to a higher NEET rate, is represented by the high share of the underground economy, where there is a large part of undeclared work. Last, there are also several factors at a macro level linked to problems associated with the economic and social reality of the country, such as the difficult path that young people are facing in the transition from school to work.

Objective

The main studies published on this subject, such us by Eurofound [1], as well as the most authoritative statistics produced on NEETs (e.g. those published by EUROSTAT or ISTAT), give us an in-depth, high-quality and, in many cases, internationally comparable information framework. However, these estimates come from sample surveys (Labour Force Surveys),that are representative at a macro-areas territorial level, such as countries and regions. Unfortunately, this does not allow us to go into a deeper detailed level (beyond those considered in the design of the surveys), and this weakens the cognitive value of the estimate, especially when we are moving from a theoretical plan to the planning and design of policies.

One of the objectives of this work is therefore to estimate the spatial distribution of the NEET rate at the Italian level, for those areas that are not planned in the stage of the survey design, such as the DEGURBA-Region level. In this categorisation the different regions are classified through the criterion of the degree of urbanization, thus identifying three main areas: cities, towns and suburbs and rural areas.

Since the typical direct estimation techniques do not produce reliable results for this level, the small area estimation methodology has been applied [3]. The data used to obtain estimates come from the European EU-SILC survey [6]. However, this is not recognised as the official source for obtaining the NEET rate. Therefore, another aspect of this work has been the creation and subsequent estimation of the NEET rate in the Italian regions using EU-SILC data. An analysis was performed to verify whether starting from a different survey would make it possible to obtain a similar rate to the one that results from the official Labour Force Survey source, obviously applying the same definition that the National Statistical Institute (ISTAT) gives to the NEET category.

Another reason for the choice of the EU-SILC survey has been related to its possibility to estimate several important indicators, such as the at-risk-of-poverty or social exclusion

(AROPE), employed to depict the poverty situation in Italy.

Methods

In order to measure the Italian territorial distribution of the NEETs, it was first necessary to reconstruct the variable itself within the EU-SILC data sets (2016). In particular, following a preparatory phase of the datasets, some variables were selected and through them it was possible to create the NEET dichotomous variable. To define the NEET category, the definition provided by ISTAT was used. In the 2016 data collection, 21,325 households and more than 40,000 individuals were interviewed.

One important adjustment that was performed in the whole dataset, was the re-calibration of the survey weights. In the EU-SILC file, every interviewed person has a weighting component, representing his or her probability of being selected in the sample and aspects such as the non-response factor. In order to obtain the real weight of a person living in a specific region, they need to be re-proportioned to the true population size in that region in the selected year.

In the last decades, the increasing demand for statistical indicators that described that the situation of local areas has led to the development of new methodologies. In particular, there was the need to obtain timely and detailed information also for those areas in which the sample size is not sufficient to obtain a reliable direct estimate [3]. An estimator of the parameter of interest for a given domain is said to be a “direct” estimator when it is based only on sample information from that specific domain. For example, the EU-SILC survey is designed to obtain reliable estimates at NUTS2 level in the European countries. Therefore, direct estimators can be computed only at regional level, and they cannot be used to compute the same parameter for an unplanned domain such as Provinces (NUTS3) or Regions by DEGURBA. Applying a direct estimation methodology, for those unplanned domains could lead to a not acceptable large standard error and very high coefficient of variations (CV) and therefore it is not possible to consider the result as reliable.

We have followed the guidelines of Statistic Canada on data reliability, with 3 main categories in interpreting coefficient of variations (CV) [7]. In this case, the aim of small area (domain) estimation (SAE) methods is to produce reliable estimators for the variable of interest under budget and time constraints. These methods try to overcome the problem of poor information for each domain by borrowing strength from the sample information belonging to other domains, trying to decrease the coefficient of variations. The increase in efficiency of SAE is obtained using the information on units belonging to other areas considered geographically close or similar with respect to structural characteristics to the small area of interest. Small area estimation is a model-based method in which a model is usually constructed, where some dependent variables are expressed as a function of some independent variables. The models are classified into two broad types: Area Level which was used for the first time by Fay, R.E. and Herriot R.A.[4], and The Unit-level Estimator, used by Battese,G.E., Harter,R.M. and Fuller,W.A [5]. In this work the Fay-Herriot model has been applied since only area level summary data were available for the auxiliary or response variable.

In particular, the basic FH area level model rely on the following assumptions, and it is based in two subsequent stages. It is assumed that a basic unbiased direct estimator $\hat{\theta}_i$ is available, with θ_i being the true unknown value of the target parameter. Then, it is assumed a linear relationship between θ_i and a set of covariates $X_i{}^T$ whose value are known for each domain of interest *i.* Combining the two model components, we obtain the linear mixed model i.e. the Fay- Herriot model:

$$\theta_i = X_i^T \beta + u_i + e_i$$

Where u_i are the domain effects assumed to be normally distributed (as a white noise) and e_i the sampling errors, whit zero mean and constant known variance

Our final interest is to find a predictor at small area level with a lower variance that the direct one. In this case, the Best Linear Unbiased Predictor (BLUP) of the target parameter can be obtained. In the work, in order to support the Fay-Herriot models, several covariates were selected from the ISTAT website “a misura di Comune”, such as Gross income per capita, Share of resident population by gender, etc.…

Results

As we can see in the Fig.1, in 2016 the italian NEET rate was 25.35%, that corresponds to more or less 3 millions of young people, that compared to the official one provided by ISTAT, 26%, is quite close but a little bit lower.

According to our estimates, several southern regions were far above this percentage. The highest value was in Campania, followed by Sicilia, Puglia and Sardegna. More generally, all the regions in North-eastern area were those with lower values, whereas the central regions were those more in line with the Italian value. The situation depicted by the NEET rate, seems to trace the classical Italian dichotomy between the southern and northern part of the country. We should not be surprised if the regions with the highest youth unemployment rates had also the highest NEET rates.

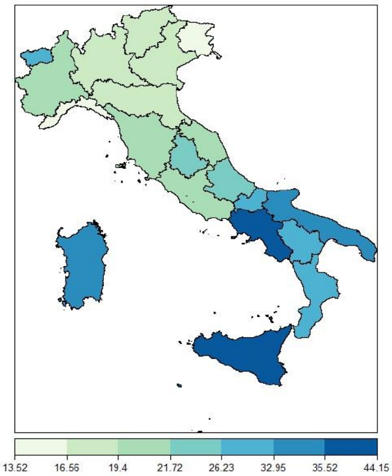


Figure 1 NEET Rate Direct Estimate

It was fundamental to check how reliable these estimates were. For this purpose, the values of the coefficient of variation were computed. In this case, the number of regions with a CV larger than the critical threshold of 16.6% was zero, that is all estimates seemed to be reliable. Confidence intervals were also computed. Since the larger is the sample size, the smaller and precise is the confidence interval, if there big differences in the sample sizes of the groups are found, these can affect the widths of the intervals and may give misleading results. On the opposite, if sample sizes are approximately the same, we can be confident that these differences are primarily due to different variation. In our case, the regions with a confidence interval larger than the others seems to be only three. Anyway, some of these intervals are overlapping to each other's. Therefore, the difference in means could be statistically significant or not.

Another element of the analysis pertained the comparison between the Labour Force Survey (LFS) NEET estimate from ISTAT and the one obtained in this framework using EU-SILC data. One important difference is that EU-SILC considers a smaller age group, from 16 to 34 years, while instead the LFS has the possibility to include in the interview people from 15 years old. At the end, the NEET national EU-SILC estimate was quite close to the LFS, 25.35% and 26.02%.

For what concerns the different regions, in some cases the value were basically the same, such as in Piemonte, Veneto, Marche, Abruzzo, Sicilia, Sardegna and Puglia. In other cases, estimates differentiated as much as 1-3% from the actual real-world distribution such as in Lombardia, Emilia-Romagna and Lazio. Unfortunately, there were few cases in which the two were not so close such as for Valle d'Aosta (with EU-SILC is overestimating the LFS) and Calabria (with EU-SILC underestimating the LFS).

Focusing on a different territorial level, the degree of urbanisation within the regions, the results were more variegated. In particular, three labels were assigned to each region, "cities", "towns and suburbs" and "rural areas". The only exception was Valle d'Aosta, whose classification had only two levels (no urban areas). As a result, 59 sub-domains were obtained for the Italian peninsula, and these domains were not planned in the survey. Therefore, we expected less reliability in terms of Coefficient of Variations (CVs). In fact, considering the 59 total domains, more than half of them presented a CV higher than 16.6%. Just the 29% of the domains presented a reliable direct estimator. For this reason, it was decided to apply a small area methodology, the Fay- Herriot model, in order to try to obtain NEET estimates with lower variability.

After applying Fay- Herriot model, the first important result was the gain in terms of variability. With the model 28 areas, representing the 48% of the analysed areas, had a CV below the 16.6%, threshold. The estimates are plotted in the figure below, with the FH estimates represented in the map on the right-hand side.

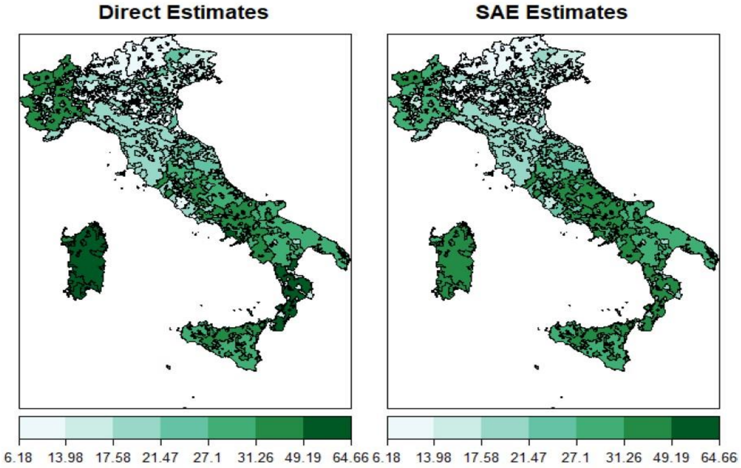
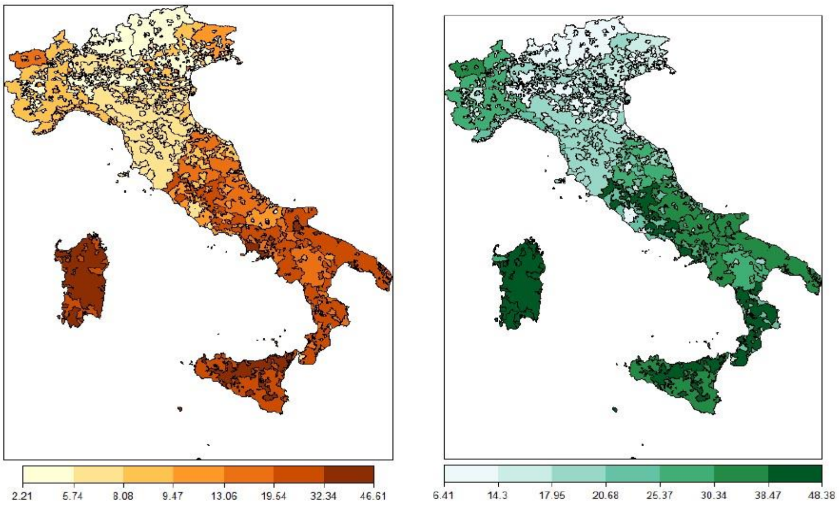


Figure 2 Direct and FH NEET total estimates

The gap between the North and South of the country is still present, but with this type of territorial classification, another aspect has emerged. Infact we can see that in sevaral cases the NEET rate was higher in those areas classified as rural, such as Sardinia or Calabria. Generally, it had been seen that less inhabited areas with fewer job opportunities for young people and less educational services had also a larger share of NEETs. [1] The same methods were applied also for a gender analysis. The NEET male rate estimated from the EU-SILC dataset was 21.64%, corresponding to about 1,338 million of young males. This average value was computed as the ratio between the NEET male populations over the young male population of the same age group. The mean value for Italian young female was about 18.72%, corresponding to 5,951 million of females all over the country. The female rate was therefore a little bit smaller with respect to the value of male population. A possible reason could be the slight gender imbalance for the young population sample, where the male observations are more frequent than the female ones.

A diagnostics analysis was also performed. Since the goodness of small area estimations relies on the quality of the specified model, it is appropriate to evaluate the implemented model, in order to verify if some assumptions are violated or if there exists a potential bias. In addition, it was very important to check the correlation between the direct and Fay-Herriot estimates, that turns out to be highly correlated especially for the gender variable.

Finally, since the analysis is based on EU SILC data, it was possible to also compute the AROPE (share of total population at risk of poverty or social exclusion) indicator at both regional and regions by DEGURBA level. This is a very important indicator in the Europe 2020 strategy to monitor the percentage of the at-risk population. It is a composite indicator, that is computed as the sum of people who are either at risk of poverty or severely materially deprived or living in a household with a very low work intensity.
It was very important to address the NEET issue also from this point of view since the more detailed information it is possible to get for this category, the more it would be the possibility to apply targeted policy measures. Therefore, by combining the AROPE rate, one of its dimensions (economic), which is the at-risk-of-poverty rate, and the total NEET rate it was possible to outline the final scenario. Unfortunately, as we can see from Fig.3 the areas where there was a higher poverty rate and a higher AROPE were also those with a higher NEET rate, such as the southern regions of Campania and Sicilia.



(a) *At risk of poverty and social exclusion* (b) *NEET rate region by degurba*

Figure 3 AROPE and NEET, region by degurba

Contribution

In this work, in order to estimate the NEET spatial distribution throughout Italy, EU-SILC data were employed. Compared to the Labour Force Survey, which is the official source for estimating the NEET rate, the EU-SILC allowed to obtain reliable national data, which can also be used for country comparisons at European level. Using EU-SILC data it was possible to study the NEET phenomenon according to the specificity of the Italian territory, using the regions by degree of urbanization classification (DEGURBA), and to verify if some spatial patterns were identifiable. In addition, as the EU-SILC survey is one of the most important European sources for the estimation of poverty indicators (such as AROPE), this made it possible to better contextualize the NEET rate. Indeed, considering the NEET category also from this point of view allowed to better understand the potential criticalities to which they are exposed, and how it is important to delineate specific issues so that policymakers can tackle the problem providing targeted interventions.

References

[1] Eurofound. (2012). NEETs – Young people not in employment, education or training: Characteristics. Luxembourg: Publications Office of the European Union.

[2] Alfieri S., Rosina A., Sironi E., Marta E., & Marzana, D. (2015). Who are Italian "Neets"? Trust in institutions, political engagement, willingness to be activated and attitudes toward the future in a group at risk for social exclusion. Rivista Internazionale di Scienze Sociali, n. 3, pp. 285-306.

[3] Molina, I., & Rao, J.N.K. (2015, (Second Editon)). Small Area Estimation. Wiley Online library.

[4] Fay, R., & Herriot, R. (1979). Estimates of Income for Small Places: An Application of JamesStein Procedures to Census Data. Journal of the American Statistical Association, 74(366), 269277. doi:10.2307/2286322.

[5] Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83(401), pp. 28–36.

[6] Eurostat. (May 2017). Methodological Guidelines and Description of EU-SILC Target Variables.

[7] Statistics Canada. (2016). Guide to the Labour Force Survey. Technical report, Statistics Canada, Catalogue 71-543-G, available at http://www.statcan.gc.ca/pub/71-543-g/ 71-543-g2016001-eng.pdf.

Contact information:

University: University of Pisa, Department of Economics and Management

EMOS programme: Master of Science in Economics

Applying student:

Mr/Ms	Ms
Surname	Sodini
First name	Sara
E-mail address 1	sarasodini@gmail.com
E-mail address 2	
Telephone number	
Mobile number	+ 39 3316119533

EMOS Programme responsible supporting the application:

Surname	Corsini
First name	Lorenzo
E-mail address 1	lorenzo.corsini@unipi.it
E-mail address 2	mse@ec.unipi.it
Telephone number	
Mobile number	+39 333 2783342

Qualitative assessment for the selection Max. 2000 words	The master thesis by Sara Sodini is an original study on the estimation of the NEET rate at the local level in Italy using EU-SILC data. The first innovation in the thesis is the use of EU-SILC data to estimate the NEET rate, officially computed by Eurostat using Labour Force Survey data. In the thesis an original algorithm to compute the NEET rate using EU-SILC variables is proposed and tested. Then, the NEET rate is estimated for the local areas represented by the DEGURBA by Region classification. In this categorisation the Italian regions are classified through the criterion of the degree of urbanization, identifying three main areas: cities, towns and suburbs and rural areas. Additionally, also the gender classification is considered. This allows to explore the NEET phenomenon under a new perspective that can be highly valuable for policy decisions. As the DEGURBA by Region areas are unplanned domains for the EU-SILC survey, from a methodological point in the thesis small area models are estimated and carefully tested. Finally, being based on EU-SILC data, in the thesis a comparison of the NEET and the AROPE rates at the local level is also performed and commented. Summarizing, the thesis by Sara Sodini makes original use of official statistics data, carefully exploring the definition of the indicator of interest and estimating it by using appropriate methodologies. The work done in the thesis has also an European perspective, in the preliminary descriptive analysis performed to introduce the NEET phenomenon, and in the use of the EU-SILC data set itself, that could allow the extension of the current work to compare the NEET phenomenon at European level.

The European Commission is committed to personal data protection. Any personal data is processed in line with the Regulation (EU) 2018/1725 All personal information processed by Eurostat is treated accordingly.

Measuring competitiveness: a new composite indicator for Italian municipalities

1. Introduction

Competitiveness is a phenomenon of particular interest for territorial bodies, for both economic and social reasons. For instance, it strengthens nations' economy (Bhawsar and Chattopadhyay, 2015) and it is a tool which can improve living conditions and welfare (Competitiveness Advisory Group, 1995).

The concept of competitiveness is elusive and complex. Competitiveness is a multidimensional phenomenon which is studied at different levels, and several definitions are present in the literature based on the specific context of application.

The main objective of this thesis is measuring territorial competitiveness at local level in Italy, particularly competitiveness at municipal level, through the proposal of a composite indicator to be applied in the Italian context. The proposed indicator allows to get precise insights into territorial competitiveness in Italy, in terms of competitiveness scores and relative position of the municipalities in the rankings.

2. Objective

The main objective of this research is measuring the competitiveness at local level in Italy, particularly the competitiveness of Italian municipalities. In Italy, municipalities are basic public territorial entities having own autonomy in governing their territory (Istat, 2021). They belong to the system of Local Administrative Units (LAUs) (Eurostat, 2021), which identifies the municipalities and communes of EU member states. The focus on this detailed geographical level may allow to develop more targeted actions and policies to improve growth performance and well-being.

The main objective implies that a proper strategy to measure municipal competitiveness needs to be delineated. This is a challenging task, because competitiveness is a multidimensional phenomenon. Different methods to measure multidimensional phenomena are present in the literature. Among them, building a composite indicator is a valid strategy, because it allows to take into account the dimensions and to aggregate them into a unique number, reducing dimensionality (Mazziotta and Pareto, 2017; 2013). While an individual indicator is a quantitative or a qualitative measure of a phenomenon, a composite indicator combines several individual indicators, following a framework established by the composite indicator builder (OECD and JRC, 2008; Freudenberg, 2003).

This research aims to introduce a composite indicator which is useful to measure competitiveness at municipal level in Italy. Such indicator should be able to effectively distinguish between different levels of competitiveness, and to compare competitiveness over time, since the objective is to allow a dynamic analysis on the competitiveness of Italian municipalities.

3. Methods

The objectives imply the definition of the methodology to build the proposed composite indicator and then to analyze it. The building process of a composite indicator is stepwise: it entails a sequence of connected phases (Mazziotta and Pareto, 2013; 2017; OECD and JRC, 2008; Freudenberg, 2003). Jointly considering these phases and the main objective of this thesis, the methodology used in this research may be summarized into the following main steps: 1) Definition of municipal competitiveness and of its dimensions, 2) Selection of the data sources and the individual indicators, 3) Normalization, weighting and aggregation of the indicators, 4) Analysis of the scores and the rankings, and influence analysis, 5) Study of two specific groups of municipalities. These phases are described in detail in the following.

1) Definition of municipal competitiveness and of its dimensions

The main objective of this thesis implies, first of all, that the phenomenon of interest needs to be clearly identified. The economic literature does not provide neither a specific definition nor a theoretical framework related to competitiveness at municipal level, which could be used in the proposed indicator. However, regions and municipalities share some features: for instance, the adjustment mechanisms which operate at national level do not similarly apply neither at the regional nor at the municipal level. The underlying theoretical framework and definition of municipal competitiveness were based on the literature on regional competitiveness.

The proposed composite indicator is based on a definition of regional competitiveness which belongs to the context of official statistics: the definition adopted by the Regional Competitiveness Index (RCI) (Dijkstra et al., 2011; Annoni and Kozovska, 2010), the first composite index which measures regional competitiveness in the EU at NUTS 2 level. Hence, based on the literature review, in this thesis municipal competitiveness was defined as the ability of a municipality to offer an attractive and sustainable environment for firms and residents to live and work.

The RCI was also used as the main reference to identify the dimensions of competitiveness and to select the individual indicators. The proposed composite indicator entails seven dimensions: *Education*, *Job*, *Economic wellbeing*, *Territory and environment*, *Entrepreneurship*, *Innovation*, and *Infrastructures and mobility*. They were identified based on the theoretical framework of the RCI and the extensive literature review, including for instance Bhawsar and Chattopadhyay (2015), and OECD (2016).

2) Selection of the data sources and the individual indicators

The individual indicators were selected according to two criteria: first, their assumed relevance for competitiveness based on the theoretical framework; second, their time availability for the same periods. Given the dynamic analysis objective, the aim is to analyze at least one comparison over time for the composite indicator.

Data were mainly retrieved from *A misura di Comune* (Istat, 2016) for the periods 2014 and 2015. *A misura di Comune* is an Italian statistical multi-source system gathering data from different sources, mainly administrative ones, for Italy. An univariate analysis was performed to understand the main features of these data.

The final dataset includes data for 7159 municipalities and 17 individual indicators, for 2014 and 2015. A multivariate analysis was carried out to explore the data structure and the suitability of the individual indicators to measure competitiveness.

3) Normalization, weighting and aggregation of the indicators

The individual indicators were normalized to achieve comparability. The choice of the normalization method considered that the objective is carrying out a dynamic analysis of the competitiveness of Italian municipalities, hence the chosen method should allow comparisons of absolute changes over time. The chosen normalization method is the rescaling (or min-max) (Mazziotta and Pareto, 2017).

As each dimension includes at least two individual indicators, for each year, the individual indicators were aggregated using a two-step procedure: first, a sub-index for each dimension was computed, then these were aggregated in the final composite indicator. In both steps, a system of equal weights was adopted, because it is not possible to find a clear indication in the literature about the most appropriate differential weights for the individual indicators and for the dimensions. The proposed municipal-level indicator is positive: higher scores indicate improvements of municipal competitiveness.

The sub-indexes were created using the COMposite Indices Creator (COMIC) software (Massoli and Pareto, 2017). Three versions of the sub-indexes were obtained applying three aggregation techniques: the mean of 0-1 indices, the geometric mean, and the Adjusted Mazziotta-Pareto Index (AMPI) (Massoli and Pareto, 2017; Mazziotta and Pareto, 2017). In the formulation adopted in COMIC, these methods include a rescaling step to normalize the data.

Adopting a consistent approach, the sub-indexes were combined into the final composite indicator using the same aggregation methods used for the individual indicators. A code written using the software R was implemented for the final aggregation. COMIC was not used at this step, to avoid normalizing the data twice.

The two-step aggregation resulted in six composite indicators: three versions, namely the mean, the geometric mean and the AMPI, for each year.

4) Analysis of the scores and the rankings, and influence analysis

To study the ability of the proposed indicator to measure municipal competitiveness, a detailed analysis was carried out, comparing the scores and the related rankings, and identifying the most influential dimensions for competitiveness. The results were compared between the three aggregation methods, and over time.

The scores were analyzed using maps to study the geographical distribution, scatter plots to examine heterogeneity of the scores, and velocity-acceleration score plots. The rankings were analyzed computing Spearman’s rank correlation and summary statistics of the absolute differences of rank.

Besides, an influence analysis was carried out to identify the most relevant dimension for municipal competitiveness. In each of seven simulations, one dimension was removed and the absolute differences of rank were computed between the original ranking and the new ranking. This process was repeated for the three versions of the composite indicator and for the two periods. The most influent dimension is the removed one which is associated to the highest coefficient of variation of the absolute differences of rank.

5) Study of two specific groups of municipalities.

Finally, the study of two specific groups of municipalities further explored the properties of the composite indicator. It also showed the usefulness of disaggregating the composite indicator into its dimensions, which can highlight specific strengths and weaknesses of the municipalities. This final analysis provided additional interesting insights which completed the detailed picture of territorial competitiveness in Italy provided by the proposed composite indicator.

4. Results

In order to reach the desired objectives, a composite indicator of municipal competitiveness was built following the methodology described above in Section 3. The aim was creating a composite indicator which is able to properly measure municipal competitiveness by distinguishing between different levels of competitiveness, and to compare competitiveness over time. To evaluate such ability, the composite indicator was applied to a dataset of 7159 municipalities and 17 individual indicators for two periods, 2014 and 2015. As described in Section 3, three aggregation methods, namely the mean, the geometric mean and the AMPI, were applied and their results compared.

The 2016 edition of the RCI (Annoni et al., 2017) shows that NUTS 2 Italian regions have either a low or medium level of competitiveness. The proposed composite indicator gives a geographically detailed picture at municipal level, providing further insights about territorial competitiveness in Italy.

Figure 1 shows the geographical distribution of municipal competitiveness scores in 2014: it was similar for the three aggregation methods. Municipal competitiveness scores were heterogeneous within each NUTS 2 region, with particularly high coefficients of variation in the regions of the South and the Islands. Every NUTS 2 region hold some municipalities with high competitiveness scores, but their concentration was more variable: on a general basis, it was higher in Northern Italy than in Southern Italy. The geographical distribution of the scores showed few differences between 2014 and 2015. In terms of competitiveness scores, most of the municipalities did not notably change between 2014 and 2015, independently of their level of competitiveness in 2014, their geographical position in terms of NUTS 1 regions, and the aggregation technique which was applied. However, the interpretation of these results should take into account that the two periods considered are very close in time.

The competitiveness scores were used to create rankings of the municipalities. In each period, the Spearman's rank correlation between each pair of methods was higher than 0.98, while the mean of the absolute differences of rank was high: on average, a municipality shifted by at least 130 positions when two methods are compared. Table 1 shows the summary statistics of the absolute differences of rank for 2014 data, the results for 2015 are similar. To sum up, the rankings for different methods were not perfectly equal, suggesting that it is important to be cautious in the interpretation of the relative position in the rankings.

The most influential dimension for municipal competitiveness did not coincide in the two time periods considered. The influence analysis showed that the most influential dimension for municipal competitiveness is Innovation in 2014, and Entrepreneurship in 2015, regardless of the aggregation method applied. Though, Innovation and Entrepreneurship are the two most influential dimensions in both years. Table 2 shows the summary statistics of the absolute differences of rank for the 2014 influence analysis, for the arithmetic mean composite indicator.

Furthermore, two specific groups of municipalities were studied: the ten least competitive municipalities in Northern Italy and the ten most competitive municipalities in Southern Italy. The analysis of these specific cases, also involving the disaggregation of the composite indicator into its dimensions, gave further interesting insights into municipal competitiveness. Particularly, the disaggregation can be used to identify the dimensions which represent specific strengths and weaknesses of the municipalities. This concluded the geographical analysis and contributed to create a detailed picture of territorial competitiveness in Italy.

In addition, the analysis of these groups of municipalities highlighted specific issues arising from using administrative data sources. It was not possible to properly measure some features of the residents, for instance of those who live in municipalities in Northern Italy but work in Switzerland. This issue may have partially impacted the scores and the rankings. However, the advantage of administrative data sources is that they allow to carry out analyses even at a very disaggregated level of analysis. The building of a composite indicator at municipal level could not be achieved without using administrative data sources.

The initially set objectives were essentially met. The results showed that building a composite indicator to measure territorial competitiveness at municipal level, which is a very disaggregated level of analysis, is actually feasible. The research carried out in this thesis resulted in the proposal of a composite indicator which is useful for providing detailed insights into competitiveness in Italy, by measuring municipal competitiveness over all Italian municipalities.

5. Contribution

The contribution of this thesis can be summarized into two main points.

First, it proposes a measure of competitiveness at a very disaggregated geographical level: a composite indicator which is useful to analyze municipal competitiveness in Italy. A similar measure is not available in the Italian literature on competitiveness. Besides, an original feature of the proposed indicator is that it uses individual indicators resulting from an experimental program. *A misura di Comune* is a multi-source system which integrates data from both traditional and experimental sources, the latter being particularly valorised. This corresponds to an integration between traditional and more innovative data production methods. The experimental program provides data at highly disaggregated territorial level.

Second, the proposed indicator allows the comparison of competitiveness of all Italian municipalities: it is not built for a subgroup of municipalities or for one specific municipality, but it is created to provide a comprehensive picture of municipal competitiveness in Italy. The proposed indicator has a detailed geographical focus which lead to detailed insights into territorial competitiveness in Italy, in terms of scores, rankings, and strengths and weakness of the municipalities. Its insights add to the one provided by the RCI, helping in capturing the heterogeneity which characterizes the competitiveness in the Italian NUTS 2 regions.

The proposed composite indicator of municipal competitiveness is based on data retrieved from experimental data sources, mainly administrative ones, which allow to gather data at the desired geographical level. This composite indicator represents a valid example of use of such data sources, which contribute to analyze phenomena that it should be not otherwise possible to measure. Due to the data availability issues which typically characterize the municipal context, data were retrieved only for 2014 and 2015, and the composite indicator was applied for those periods. The proposed composite indicator may be seen as a first prototype, the eventual update of new data would allow its computation for other periods.

In this research, the proposed composite indicator is applied in the Italian context. Actually, its applicability may be extended beyond Italian borders. It is recalled that, according to the NUTS classification, Italian municipalities are LAUs, a group which include all municipalities and communes of EU member states. The proposed indicator may be applied to other LAUs in the EU. Though, this potential extension would require further research, to understand the degree of data availability at LAU level in other EU countries and, for the available data, the extent to which they are comparable.

Figure 1: Maps showing the level of competitiveness of Italian municipalities in 2014. They refer to the composite indicators computed with the arithmetic mean, the geometric mean and the AMPI, respectively. Darker colours represent lower scores of the competitiveness indexes while warm colours refer to higher scores of competitiveness.

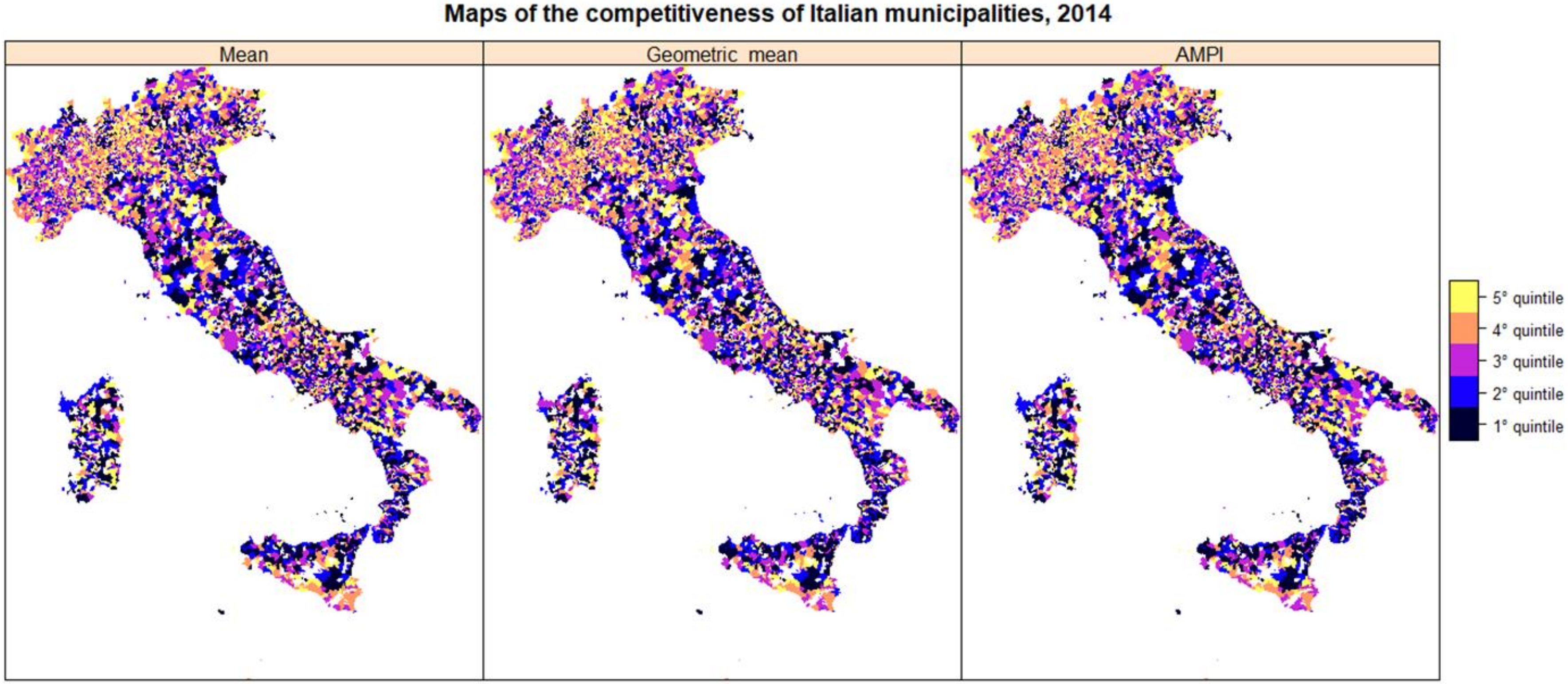


Table 1: Summary statistics of the absolute differences of rank, 2014

Absolute differences of rank	Mean	Variance	Standard deviation	Coefficient of variation
Mean-geometric mean	251.675	74074.98	272.167	1.081
Mean-AMPI	137.128	19445.92	139.449	1.017
Geometric mean-AMPI	150.739	33553.25	183.175	1.215

Values are rounded to three digits.

Table 2: 2014 influence analysis, summary statistics of the absolute differences of rank. Composite indicator: arithmetic mean.

Removed pillar	Mean	Variance	Standard deviation	Coefficient of variation
6.Innovation	439.442	326884.347	571.738	1.301
5.Entrepreneurship	127.791	24133.749	155.35	1.216
3.Economic wellbeing	383.987	165125.933	406.357	1.058
2.Job	206.319	44192.517	210.22	1.019
7.Infrastructures and mobility	302.814	82465.545	287.168	0.948
1.Education	282.699	67379.351	259.575	0.918
4.Territory and environment	698.681	386441.13	621.644	0.89

The removed pillars are arranged according to the size of the coefficient of variation, from the highest to the lowest.

References

Annoni, P., & Kozovska, K. (2010). *EU Regional Competitiveness Index 2010.* Luxembourg: Publications Office of the European Union.

Annoni, P., Dijkstra, L., & Gargano, N. (2017). The EU Regional Competitiveness Index 2016. European Union Regional Policy Working Papers, no. 02/2017.

Bhawsar, P., & Chattopadhyay, U. (2015). Competitiveness: Review, Reflections and Directions. *Global Business Review* , 665-679.

Competitiveness Advisory Group. (1995). *Enhancing European Competitiveness: First report to the President of the Commission, the Prime Ministers and the Heads of State.* Luxembourg: OPOCE.

Dijkstra, L., Annoni, P., & Kozovska, K. (2011). A New Regional Competitiveness Index: Theory, Methods and Findings. In *Working Papers: A series of short papers on regional research and indicators.* Eric VON BRESKA.

Eurostat. (2021). *Local Administrative Units (LAU)*. Retrieved from Eurostat: https://ec.europa.eu/eurostat/web/nuts/local-administrative-units

Freudenberg, M. (2003, November 12). Composite Indicators of Country Performance: A Critical Assessment. *OECD Science, Technology and Industry Working papers* . OECD Publishing.

Istat. (2016). Retrieved from http://amisuradicomune.istat.it/aMisuraDiComune/

Istat. (2021). *Glossario statistico*. Retrieved from Istat: https://www.istat.it/it/metodi-e-strumenti/glossario

Massoli, P., & Pareto, A. (2017). *COMIC-Guida all'uso.* Retrieved from Istat:

https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/analisi/strumenti-di-analisi/comic

Mazziotta, M., & Pareto, A. (2013). Methods for constructing composite indices: one for all or all for one? *Rivista Italiana di Economia Demografia e Statistica , LXVII* (2), 67-80.

Mazziotta, M., & Pareto, A. (2017). Synthesis of Indicators: The Composite Indicators Approach. In F. Maggino (Ed.), *Complexity in Society: From Indicators construction to their Synthesis* (pp. 159-191). Springer International Publishing AG.

OECD. (2016). *OECD Regions at a Glance 2016.* Paris: OECD Publishing. doi:http://dx.doi.org/10.1787/reg_glance-2016-en

OECD, & JRC. (2008). *Handbook on constructing composite indicators. Methodology and user guide.* Paris: OECD.

Interactions within a multi-layer EU inter-bank network

Introduction

In the last two decades, financial and banking research papers granted a lot of interest to assess systemic risk in a banking system. Those papers investigated the mechanisms at stake in an inter-bank market in the wake of a financial crisis and suggested that markets tend to contract before a crisis (Minoiu and Reyes 2013); it means that the number of interconnections between banks increased. Papers on network theory in finance demonstrate that conventional banking measures such as banks balance sheets are not sufficient to evaluate banking stability. This lack led to a change of perspective from "Too big to fail" to "Too interconnected to fail" (Hüser 2015) in order to take into account bank interconnections in the way systemic risk is assessed. Looking at an inter-bank market through a network perspective gives more insights on its mechanisms as it incorporates market externalities. Allen and Gale 2000 assessed that a great amount of connections between banking groups increase the resilience of a banking system. This statement has been qualified recently, arguing that "An intermediate level of connectivity" seems to enable banks to absorb shock instead of propagating them (Battiston et al. 2012b). A new way of applying network theory to financial networks is to analyse multi-layer networks in order to incorporate different types of interrelations between banks. It is motivated by the fact that the connections between banks do not follow the same logic depending on the market considered.

Most of the studies on financial networks focus on the interactions within a market, studying the mechanisms over time and during a financial crisis (Martinez-Jaramillo et al. 2014). A few papers assess systemic risk using global network metrics[35](Brandi, Di Clemente, and Cimini 2018). In this study, we aim at taking the analysis to the next level by evaluating systemic risk of individual banking groups using local network features[36]. We will study three types of markets: two lending markets (short-term and long-term loans) and a crossholding market. We will particularly focus on the effect of local network metrics, e.g indegree centrality (Nieminen 1974), closeness centrality (Freeman 1978), betweenness centrality (Freeman 1977) and clustering coefficient (Fagiolo 2007) on a proxy of systemic risk, DebtRank (Battiston et al. 2012a).

This study is mainly driven by the recent access to financial granular datasets collected from European banks thanks to a common European framework. These new data sources fulfill the need to measure financial stability ; they contain bilateral exposures between banks for several

[35] In a network, a global metric is a measure that describes the whole network, e.g the number of nodes or edges.

[36] In a network, a local metric is a measure that describes a single node in the network, e.g the number of neighbors of a node.

inter-bank markets. To the best of my knowledge, this is the very first big data analytics using real world data on the impact of network topology on systemic risk at the banking level. Before, the effect of connections between banks was obtained using artificial data based on simulations (Nier et al. 2007; Montagna and Kok 2016). To develop our work, section 2 will introduce the objective of the study, section 3 will describe the method used to solve the problem, section 4 will summarize the main results and section 5 will present the key contributions and conclusions of the work.

Objective

The goal of the paper is to study the evolution of a multi-layer European inter-bank network. A multi-layer network considers different types of interactions and interrelations of institutions, e.g. long or short-term loans, cross-holdings, etc. The underlying idea is to define the role of banking interconnections in network stability on several inter-bank markets. We aim to underline if some network configurations threatens banking stability more than others by increasing individual systemic risk.

Studying a multi-layer network with time dimension will bring novel insights on interbank market stability for two reasons. Firstly, taking into account several layers in a financial network enable to compare the topology and the structure of different layers, i.e, of different inter-bank markets. The relative position of banking groups differs according to the layer, because banks behave differently on distinct markets. A bank can be a central player in the loan market and be more backward in the cross-holdings securities market. Secondly, by including time dimension in the study, we aim to capture the dynamics of the multi-layer network through the time, especially during the COVID-19 health crisis. Comparing local network metrics is an efficient and relevant way to compare different networks, either between distinct markets or over time.

Beyond this, we investigate whether the evolution of local network metrics have an impact on systemic risk at the banking level. Special attention will be given to assessing the impact of COVID-19 on systemic risk. Indeed, the health crisis likely forced banks to modify their behavior to adapt to this unusual context. Thus, we can assume that bank interconnections have changed during the crisis, both in quantity and type of interconnection.

Method

3.1 *Integration of financial granular datasets*

To create the suitable datasets for our analytical purposes, we started from the tools

and methods delivered by the Data Committee on Advanced Analytics Project of the ECB (Aarab et al. 2022). This project aimed to integrate several structured and unstructured granular data to provide users with clean and easy to use financial datasets. For the thesis, we integrated highly confidential financial granular datasets on a quarterly basis from September 2018 to December 2021. We extended the DCAAP tools to create multiple aligned multi-layer networks, which capture the evolution of the interactions of banks over time. From this, we built a panel data containing the identifier of each banking group and their associated local measures by following these steps.

First, we retrieved the list of significant institutions[37] and their group structure. To do so, we used the Repository of the SSM[38] Supervised Institutions (ROSSI) which contains the list of significant banking groups head. Also, we worked with the Register of Institutions and Affiliates Database (RIAD) to complement the banking groups with their group structure, i.e. the list of entities belonging to the group.

Second, we enriched the list of significant banking groups with their balance sheets items such as cash, deposits, capital and total assets. Individual Balance Sheets data (IBSI), Common Reporting framework data (CoRep) and Financial Reporting data (FinRep) provided us with these conventional key banking features.

Third, we added the bilateral exposures of banking groups on the three studied markets: the long-term loans market, the short-term loans market and the cross-holdings market. For the loan network data, AnaCredit data provides the basis for the required information. Analytical credit datasets report information on individual bank loans above 25,000 euros to legal entities in the euro area; it started to be reported in September 2018. To create the cross-holdings layer, we combined Centralised Securities Database (CSDB) that contains information on the issuer of a security, with Securities Holding Statistics (SHSG) that provides information on the securities held within the Euro Area.

Finally, we created a directed multi-layer network per period in order to retrieve the local metrics of each banking group in the network. For the loans layers, the creditor (resp. debtor) represents the source (resp. target) of the edge. For the cross-holdings layer, the holder (resp. issuer) represents the source (resp. target) of the edge. From this, we built an unbalanced panel dataset of 14 periods and 85 to 105 banking groups. The number of banking groups in the data is varying because not all banks interact systematically with other banks. In the case a bank does not establish an edge with at least one other bank, it is excluded from the network for this layer and this period only.

3.2 *Statistical model*

Regression with panel data

We aim to assess the effect of local network metrics (in-degree centrality, closeness centrality, betweenness centrality and clustering coefficient) on a proxy for systemic risk, DebtRank (Battiston et al. 2012a). This metric allows to assess for systemically important banks, that is, banks that hold a great part of financial exposures of the network. We will follow the approach of Dong and Yang 2016 who assess the impact of local centrality measures on innovation in pharmaceutical industry with panel data, using NPD (New Product Development) as an indicator, but applying it to a European multi-layer financial network.

To do so, we implemented an Ordinary Least Square regression with individual fixed effects on each layer of the network; thus, we dealt with three different models. We controlled with conventional balance sheets items and global

[37] Significant institutions are banking groups verifying a list of criteria on assets size, finance public assistance etc.

[38] SSM: Single Supervisory Mechanism

network metrics: the number of nodes and edges, the total amount of assets, and the level of leverage[39]. All the variables are scaled, except from DebtRank values to which we applied a cube root transformation[40]. We introduced a dummy variable with a value of 1 starting from March 2020 to control for the possible effects of the Covid crisis. We used clustered standard errors to remove the remaining heteroskedasticity of the residuals (Zeileis, Köll, and Graham 2020).

Hypotheses

Our proxy for systemic risk refers to the impact of a bank going into default on the overall system. Thus, the higher the exposures of a bank are relatively to the total exposures of the network, the more this bank is systemically important. Then, we conjecture that DebtRank is positively linked to the number of incoming edges. This leads us to the first hypotheses:

H1.a A higher banking group's in-degree centrality makes the value of DebtRank increase. H1.b In-degree centrality is the biggest determinant of DebtRank values.

Empirically, it is reasonable to believe that the more central a bank is, the more impact it will have on the network if it goes into default. Thus, we want to verify the following hypothesis:

H2. Higher banking group's centrality measures make the value of DebtRank increase or have no effect.

In the same logic, the number of connections between the neighbors of a bank is likely to increase contagion in case of default. Then, we introduce the last hypothesis:

H3. A higher banking group's clustering coefficient makes the value of DebtRank increase.

Results

Table 1 presents the results of our regression models for the three layers: (1) refers to the long-term loans layer, (2) is assigned to the short-term loans layer and (3) corresponds to the cross-holdings layer.

H1.a and H1.b are supported for the long-term loans layer and the cross-holdings layer, as the parameters associated with the indegree centrality in (1) and (3) are positive and significant; in-degree appears as the biggest determinant of DebtRank values. The value of the parameter in the long-term loans layer is higher than the one for the cross-holdings layer. We postulate that it appears riskier to take an additional loan than to issue an additional security. Although none of the parameters associated with centrality measures are significantly negative, H2 is not supported for all the layers. It appears that closeness centrality is the local metric that has the biggest impact on DebtRank for the short-term loans layer. Also, the closer the banks are in the cross-holdings layer, the more systemic risk increases. H3 is only supported for the longterm loans layer. Consequently, cross-holdings securities between the neighbors of a given bank do not increase this bank's systemic risk.

Not all assumptions made are supported by every layer; this demonstrates, as already showed in the literature (Bargigli et al. 2015) that layers in a multi-layer network are very different. Beyond

[39] Leverage Tier 1 is the ratio between capital Tier 1 and total assets.

[40] DebtRank density was highly skewed for all the layers. It also contains a great amount of zero values that prevented us to apply a box-cox transformation.

this, we can assess that network topology has a different impact on systemic risk depending on the layer considered. Moreover, the parameter associated to the Covid crisis is negative and significant for all the layers. Running permutation tests of equality of densities (Bowman and Azzalini 2021) on all the variables, we can state that the density of some local metrics significantly changed during the Covid crisis period.

For instance, for the cross-holdings layer, the in-degree centrality and the closeness centrality are statistically different before and during the crisis. For the short-term loans, it is the closeness centrality and the clustering coefficient that are significantly different between the two periods. We also noticed that the total number of edges in the cross-holdings layer significantly increased during the Covid crisis. It demonstrates that network topology for European banking groups significantly changed during the Covid crisis. So, the Covid-crisis seems to have an impact on DebtRank values through the shift it generated on local network metrics; the structure observed during the crisis seems to make the system less vulnerable to individual bankruptcy as it decreases systemic risk of banking groups.

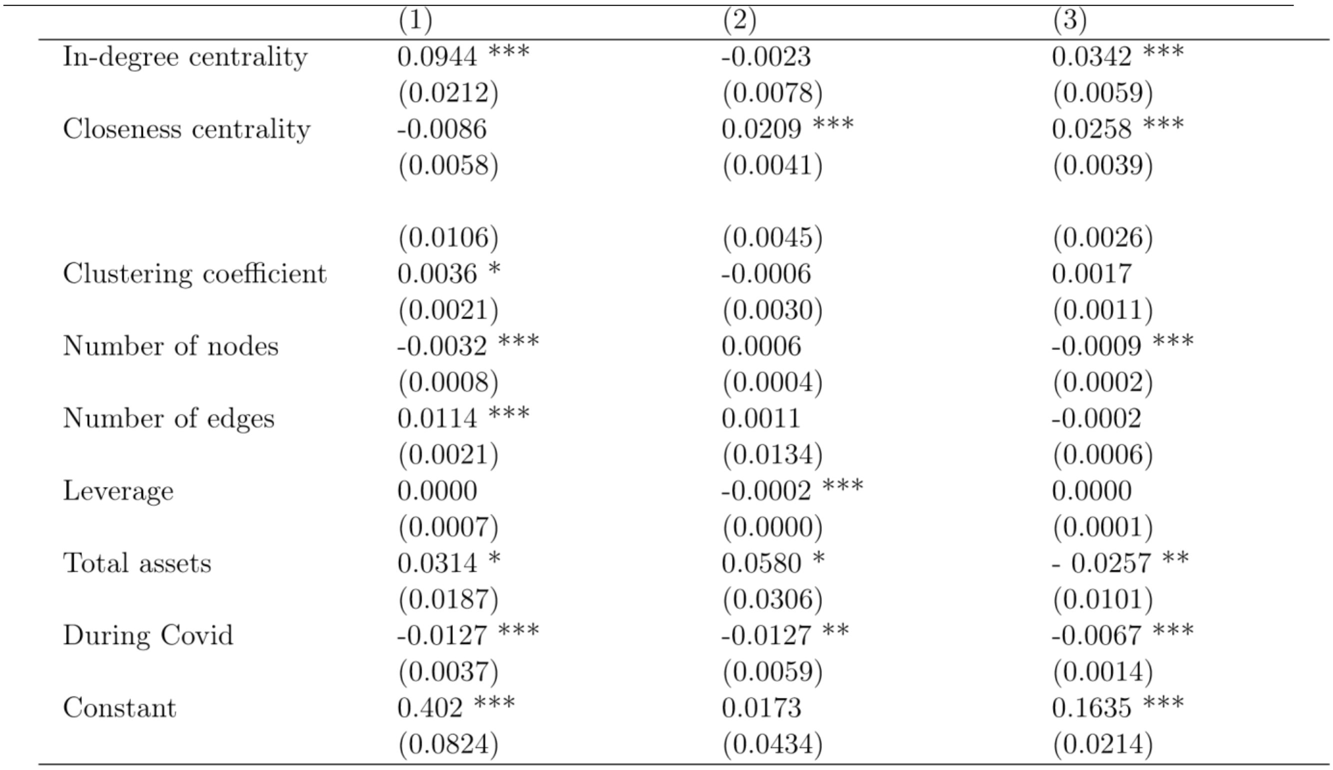
Contributions

This paper revealed the importance of taking into account local network metrics to assess systemic risk using real world data. The recent access to financial granular datasets given to various business areas in the European Central Bank enabled to develop a novel approach to assess systemic risk ; and thus to extend the previous work done with simulated data (Nier et al. 2007, Montagna and Kok 2016). To the best of my knowledge, it is the first time that a multi-layer network analysis focusing on banking groups is ran with a scope as broad as the European level. Among this, the disrupted context of the health crisis meant that we had to ensure not to neglect the impact of the crisis on banking group’s behavior, whether real or anticipated. Then, we adapted classical statistical tools to fully incorporate the exogenous factor in the study.

This paper contributed to demonstrate that local network metrics provide a detailed understanding of the different mechanisms between the layers, and whether these mechanisms ensure the stability of the network. The paper helped to complement the work that described inter-bank markets focusing only on banks financial information and on global network characteristics (e.g number of nodes, number of edges etc.). More than depicting the macro market dynamics, we ensured to prove the importance and to define the role of market mechanisms at the micro level. Thus, we can state that being central in a European network increases systemic risk at the banking level. Indeed, the more the banks establish connections with other banks, the greater the impact on the system will be in case of default.

Our approach is all the more novel in that it succeeds to combine an assessment of individual systemic risk and time dimension with several inter-bank markets. The multi-layer approach enabled to draw comparisons between layer structures and their evolution over time, particularly during the COVID-19 crisis. It finally contributed to produce comparisons between the impact of different layer topologies on a proxy for systemic risk.

Long-term loans layer Short-term loans layer Cross-holdings layer



	(1)	(2)	(3)
In-degree centrality	0.0944 ***	-0.0023	0.0342 ***
	(0.0212)	(0.0078)	(0.0059)
Closeness centrality	-0.0086	0.0209 ***	0.0258 ***
	(0.0058)	(0.0041)	(0.0039)
	(0.0106)	(0.0045)	(0.0026)
Clustering coefficient	0.0036 *	-0.0006	0.0017
	(0.0021)	(0.0030)	(0.0011)
Number of nodes	-0.0032 ***	0.0006	-0.0009 ***
	(0.0008)	(0.0004)	(0.0002)
Number of edges	0.0114 ***	0.0011	-0.0002
	(0.0021)	(0.0134)	(0.0006)
Leverage	0.0000	-0.0002 ***	0.0000
	(0.0007)	(0.0000)	(0.0001)
Total assets	0.0314 *	0.0580 *	- 0.0257 **
	(0.0187)	(0.0306)	(0.0101)
During Covid	-0.0127 ***	-0.0127 **	-0.0067 ***
	(0.0037)	(0.0059)	(0.0014)
Constant	0.402 ***	0.0173	0.1635 ***
	(0.0824)	(0.0434)	(0.0214)
Betweenness centrality	-0.0078	0.0084 *	0.00508 **
Adjusted R^2	0.7729	0.7521	0.9779
Individual fixed effects	Yes	Yes	Yes
n	1409	1214	1420

Table 1: Regression results

Reading note: *p < 0.1; **p < 0.05; ***p < 0.01. The value in parentheses are clustered standard errors.

Dependant variable is DebtRank value computed on each layer.

References

1. Aarab, Ilias et al. (2022). “Analysis of a Multi-Layer Network of Euro Area Banks: unsing granular data to facilitate advanced analytics”. In: Poster presented at the Dutch National Bank conference, Amsterdam, Netherlands.
2. Allen, Franklin and Gale, Douglas (2000). “Financial contagion”. In: *Journal of political economy* 108.1, pp. 1–33.
3. Bargigli, Leonardo et al. (2015). “The multiplex structure of interbank networks”. In: *Quantitative Finance* 15.4, pp. 673–691.
4. Battiston, Stefano et al. (2012a). “Debtrank: Too central to fail? financial networks, the fed and systemic risk”. In: *Scientific reports* 2.1, pp. 1–6.
5. Battiston, Stefano et al. (2012b). “Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk”. In: *Journal of economic dynamics and control* 36.8, pp. 1121–1141.
6. Bowman, A. W. and Azzalini, A. (2021). *R package sm: nonparametric smoothing methods (version 2.2-5.7)*. University of Glasgow, UK and Università di Padova, Italia. url: http://www.stats.gla.ac. uk/~adrian/sm/.

7. Brandi, Giuseppe, Di Clemente, Riccardo, and Cimini, Giulio (2018). "Epidemics of liquidity shortages in interbank markets". In: *Physica A: Statistical Mechanics and its Applications* 507, pp. 255–267.

8. Dong, John Qi and Yang, Chia-Han (2016). "Being central is a double-edged sword: Knowledge network centrality and new product development in US pharmaceutical industry". In: *Technological Forecasting and Social Change* 113, pp. 379– 385.

9. Fagiolo, Giorgio (2007). "Clustering in complex directed networks". In: *Physical Review E* 76.2, p. 026107.

10. Freeman, Linton C (1977). "A set of measures of centrality based on betweenness". In: *Sociometry*, pp. 35–41.

11. — (1978). "Centrality in social networks conceptual clarification". In: *Social networks* 1.3, pp. 215–239.

12. Hüser, Anne-Caroline (2015). "Too interconnected to fail: A survey of the interbank networks literature". In.

13. Martinez-Jaramillo, Serafin et al. (2014). "An empirical study of the Mexican banking system's network and its implications for systemic risk". In: *Journal of Economic Dynamics and Control* 40, pp. 242–265.

14. Minoiu, Camelia and Reyes, Javier A (2013). "A network analysis of global banking: 1978–2010". In: *Journal of Financial Stability* 9.2, pp. 168–184.

15. Montagna, Mattia and Kok, Christoffer (2016). "Multi-layered interbank model for assessing systemic risk". In.

16. Nieminen, Juhani (1974). "On the centrality in a graph". In: *Scandinavian journal of psychology* 15.1, pp. 332–336.

17. Nier, Erlend et al. (2007). "Network models and financial stability". In: *Journal of Economic Dynamics and Control* 31.6, pp. 2033–2060.

18. Zeileis, Achim, Köll, Susanne, and Graham, Nathaniel (2020). "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R". In: *Journal of Statistical Software* 95.1, pp. 1–36. doi: 10.18637/jss.v095.i01.

Online Data (JENK2M.2)

Session Chair: **Matyas Meszaros** *(Eurostat)*

Development of an OJA gold standard to classify occupation
Anca Maria Kiss *(Sogeti)*, Gabriele Marconi *(Sogeti)*, Alexandros Bitoulas *(Sogeti)*, *Fernando Reis (Eurostat)*

URL finding methodology
Heidi Kühnemann *(Statistik Hessen)*

National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption
Gergely Buda *(Barcelona School of Economics)*,Vasco Carvalho *(Faculty of Economics)*, Stephen Hansen *(Imperial College London)*

Development of an OJA gold standard to classify occupation

Keywords: online job ads, human labelling, quality of classifier algorithms, data classification, Doccano.

Introduction

Aim of the work

Online job advertisements (OJA) are a relatively *new non-traditional* source to produce data of high relevance. It provides great opportunities to produce official statistics, but also pose challenges, in particular on what concerns data quality assessment. Therefore, the ability of this new web data source to generate accurate, consistent, and comparable data needs to be addressed, assessed, tested, and improved.

This paper addresses the quality monitoring procedure that is being put in place to assess the algorithms used to harvest this new data source. It sets up a procedure to evaluate the classifiers used for the occupation variable in OJAs. The procedure seeks to combine human and machine intelligence to maximize accuracy, to assist human tasks with machine learning to increase classifiers efficiency. The creation of a gold standard for OJA represents one way to address this need for evaluation and quality improvement of algorithms.

OJA use case

Online job advertisements (OJA) refer to advertisements published on the World Wide Web revealing an employer's interest in recruiting workers with certain characteristics for performing certain work. OJA include data on the characteristics of the job (e.g., occupation and location), characteristics of the employer (e.g., economic activity) and job requirements (e.g., education and skills). The data used in this paper is the Web Intelligence Hub's Online Job Advertisement (OJA) database, developed by Cedefop and Eurostat [1]. This dataset covers over 100 million ads posted in EU countries and the UK since July 2018 and collected from more than three hundred web sources including job search engines and public employment services' websites. We focus our analysis on the period Oct 2021 – Jan 2022, excluding the last active quarter.

Methods to improve quality of OJA data: Gold Standard

Natural Language Processing dataflow

To assess the quality of OJA data, we use as much relevant information as available, i.e., the *full job description* –together with additional text extracted from structured fields (e.g., raw text on job title, salary, etc.). The main goal of this work is to conduct quality analysis and improvement of OJA data quality, by improving classification algorithms.

In view of these quality control purposes, the OJA data collection includes a new dataflow with additional information on OJAs. This dataflow will allow the assessment of the precision of the automatic classifiers that extract skills, occupations, and the other statistical variables from the text in natural language present in the job descriptions and other structured fields present in the job portals. By comparing the text in the job advertisements with the results of the classifiers, statisticians and users will be able to assess the correctness of those results.

Data annotation and OJA gold standard

Data annotation. It is the process of labelling data and refers to the human classification of raw data. Different methodologies exist for enriching linguistic data collection with annotations for quality purposes. The annotated corpus -single set of data annotated with the same specification- is a crucial input in the OJA data classification process. The annotated corpus consists in series of keywords list called ontologies Each OJA variable has an ontology associated, i.e., list of keywords for each specific classification category.

Quality of annotated data. Verifying OJA data is a way to improve the annotated corpus, used to train, validate and test machine-learning algorithms or to estimate the precision of the classifiers, combining human and machine intelligence in applications using AI (Artificial Intelligence). Some measurements to evaluate the quality of annotated data are accuracy (i.e., how close a label is to the truth), and consistency (i.e., the degree to which multiple annotations on various training items agree with one another). Standard methods exist to assure annotated data of high quality, such as the use of gold standards [2], consensus (use of multiple annotators to label the same raw OJA), and auditing (expert review and spot-checks the labels).

Gold standard (or "ground truth"). It represents the best available benchmark to evaluate classification outcomes. It is obtained through human labelling work on a predefined randomly selected sample with some desirable requirements (e.g., uncertainty, diversity, and random sampling). Human errors in training data can be more or less important depending on the use case. Supervised learning models get more accurate with more labelled data of quality. Active learning will help deciding which data to sample for human annotation. [3].

Doccano as tool to gather annotations. Different platforms exist for data annotation. Doccano is an open-source tool for text human annotation that provides features for text classification, sequence labelling and sequence-to-sequence tasks. It allows collecting labelled data by creating a project, uploading data, and starting the annotation. We used Doccano in the context of the human-in-the-loop evaluation procedure developed and proposed for the OJA data.

Human annotation of OJA

Gold standard sample

The sampling frame for the OJA gold standard is the NLP dataflow. In this study, we analyze the occupation classifier. We create a stratified evaluation dataset containing the raw data and the classification outcomes. This dataset consists of a set of raw data extracted for a sample of ads, including the full text of the job description as listed by the entity posting the ad, stratified by the classification outcomes (up to ISCO-08 level 4) and combinations of country and language. It also includes the job description tokens and dictionary terms that were matched by the algorithm (the classification is performed by an ontology-based algorithm). The stratification ensures that some ads will be included in the evaluation dataset for each classification outcome. The sample contains OJAs not available anymore online, to prevent any potential negative effect on the websites owners from where the content is retrieved. Other classification algorithms could also be evaluated in the future (type of contract, salary, working time, education, economic activity and required work experience).

We will put in place a cyclical evaluation procedure that leads to iterative cycles of feedback and improvement. Annotators will look at selected ads in the evaluation dataset, and report if they agree or disagree with the classification outcomes, using a data annotation tool. They can also propose changes to the algorithms to fix the problems identified. The evaluation dataset released yearly will lead to:

- Evaluation metrics for the classification algorithms (e.g., the accuracy rate),
- Suggestion for improvement of the sets of keywords (i.e., "ontologies"),
- A set of human-labeled data growing over time for training ML (Machine Learning) models.

Definition of specific metadata labels for OJA annotation

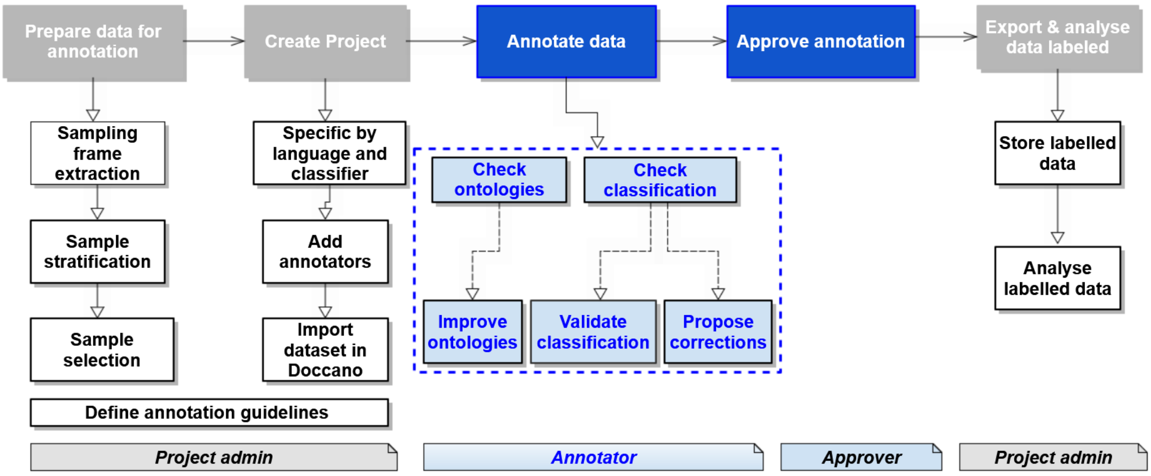
The annotation sought is the classification of the occupation into ISCO classes at 4 digits level (or lower levels if not possible at 4 digits). The labels include several ISCO classes and "metadata" labels designed to capture other situations. The metadata labels defined for this human annotation are as follows: "*Correct*" – 0, "*Incorrect*" – 1, "*Comment*" – 2, "*No reference to occupation in the description*" – 3, "*Impossible to classify at 4th level*" – 4, "*Wrong language*" – 5, "*Not a job ad*" – 6, "*Job description missing*" – 7, "*Multiple ISCO (International Standard Classification of Occupations) labels*" – 8, "*Misspelling*" – 9.

Annotators

Members of the Web Intelligence Network from several countries performed annotation. The annotators included a mix of labour market statistics experts and web intelligence specialists.

Human annotation workflow

We defined in Doccano one project per pair of country and language spoken in the country. For each project, we created the specific labels that were used for the human annotation exercise. The selected sample is uploaded into the project, the annotators are assigned to each project and the annotation work can start.



3. Results

A first analysis of the labeled data for four countries (AT, BG, IT, SI) shows that there are ways to improve the classifier used for the occupation variable in the OJA data.

Table 5. Analysis of labelled OJA datasets *

Country	**AT**			**BG**			**IT**			**SI**		
OJA label	**n**	**A (%)**	**A_w (%)**	**n**	**A (%)**	**A_w (%)**	**n**	**A (%)**	**A_w (%)**	**n**	**A (%)**	**A_w (%)**
Correct - 0	161	41.93%	49.87%	147	45.37%	47.43%	185	47.68%	51.87%	143	45.05%	58.54%
Incorrect - 1	223	58.07%	50.13%	177	49.86%	45.58%	203	50.37%	46.59%	177	53.62%	39.96%
Comment - 2	40	10.10%	11.07%	0	0	0	8	1.99%	0.73%	0	0	0
No reference to occupation in description - 3	20	5.05%	3.67%	1	0.28%	0.01%	1	0.25%	0.01%	1	0.31%	0.01%
Impossible to classify at 4th level	16	4.04%	4.79%	0	0	0	1	0.25%	0.00%	2	0.62%	0.37%
Wrong language - 5	2	0.51%	0.04%	28	7.89%	1.75%	0	0	0	0	0	0
Not a job ad - 6	3	0.76%	0.88%	7	1.97%	0.10%	11	2.73%	2.46%	6	1.86%	0.84%
Job description missing - 7	82	20.71%	16.06%	5	1.41%	0.64%	0	0.00%	0.00%	1	0.31%	2.43%
Multiple ISCO labels - 8	45	11.36%	9.97%	0	0	0	5	1.24%	0.74%	0	0	0
Total ads labeled	**403**			**355**			**397**			**329**		

* A – Accuracy rate (%); A_w – weighted accuracy rate ((%)) (weighted on the proportion of job ads in each occupation compared to the total job ads); n – number of times a label was used by the annotator to annotate the OJA sample selected for a specific country. The percentage sum of A (and of Aw) is not 100% because of multiple labeling of the same ad (for example, the labels 1, 2 and 8 are often used together). In addition, the sum of "Correct" and "Incorrect" <100% suggests that some ads were not labeled at all.

Conclusions

The results of the first round of labelling exercise for the gold standard samples show a need for improvement of the classifier for the occupation variable. Future rounds of human labelling will help achieve a gold standard for OJA data. The extension of this methodology for more countries and for additional variables (e.g., classifiers) and considering suggestions for improving ontologies will contribute to a higher quality of OJA data.

References

[18] Cedefop (2020). Cedefop and Eurostat formalise joint approach to online job advertisement data.

[19] Wissler, L., Almashraee, M., Monett, D., Paschke, A. (2014). The Gold Standard in Corpus Annotation. 10.13140/2.1.4316.3523.

[20] R. (Munro) Monarch, Human-in-the-Loop Machine Learning, Active learning and annotation for human-centered AI, Simon and Schuster, 2021.

URL finding methodology

Keywords: web scraping, statistical business register, machine learning, text mining

INTRODUCTION

Enterprise websites can be a data valuable source to improve the statistical business register, e.g. to update economic activity codes or to fill data gaps in the available contact information. Knowing enterprise URLs is a necessary precondition of using website data to enhance the statistical business register. Even if enterprise URLs are available from administrative or survey data sources, those sources often do not cover the whole business register or are not of sufficient quality. In order to fill data gaps, statistical offices can either purchase URL data or identify URLs with an automated procedure.

This contribution presents results from the ESSnet Web Intelligence Network, Work Package 3, Use Case 5 on Business Register Enhancement. It focuses on URL finding approaches that send automated requests to a search engine and use the results to identify the correct enterprise URLs. It aims to provide an overview of URL finding approaches across the ESS. URL finding approaches from the NSIs of the Netherlands, Italy, Austria, Bulgaria and Finland as well as from Statistics Hesse (Germany) were taken into account [1, 2, 3]. The following chapters are structured as follows. In chapter 2, I present the general methodology, variations due to country-specific preconditions and performance evaluations. In addition, I describe the process of scraping, data extraction and model creation in more detail with the example of URL finding in Statistics Hesse. I conclude with a discussion on current challenges and potential improvements.

METHODS

Overview of URL finding methodology

Figure 1 shows the usual steps involved in URL finding. Starting from a list of enterprises, each enterprise is searched in a search engine – often using the name and the location of the enterprise as search term. Search results are retrieved either by using an API or by scraping search engine result pages. This results in a number of candidate URLs – URLs that are potentially the correct enterprise URLs. These are subsequently filtered by applying a blocklist that contains unwanted URLs, e.g. business information websites, which are expected to appear frequently in the results. The remaining URLs are then scraped. To create features, the scraped data are compared with the business register data (such as enterprise name, address, register numbers). Features indicate either the presence of the enterprise's data on the website or are similarity scores of the enterprise data and parts of the obtained web data. A machine learning model is then trained and used to determine which of the candidate URLs are correct. Finally, evaluation scores are computed to evaluate the model performance.

One variation of this general approach is to skip the scraping part: only search results are used as data source and no additional scraping is done. This approach has been implemented first by Statistics Netherlands [1], who sent six different search queries for each enterprise to the Google API. The search terms gave in part different results and therefore supplemented each other. Instead of scraping the resulting URLs only the search result data (URL, title, snippet, etc.)

were used as data source to create features. Statistics Netherlands quantified the agreement between search results and enterprise data with string similarity measures. Compared to scraping the resulting URLs, this approach is computationally inexpensive and fast.

Another variation is to use domain registration data as additional or only data source for URL finding. This data can be used in the process at different stages: for example, it can replace or supplement the results of the search engine or it could be added as a feature to the machine learning algorithm. Statistics Finland is currently in the process of developing a URL finding procedure using data from their national domain registrar Traficom.

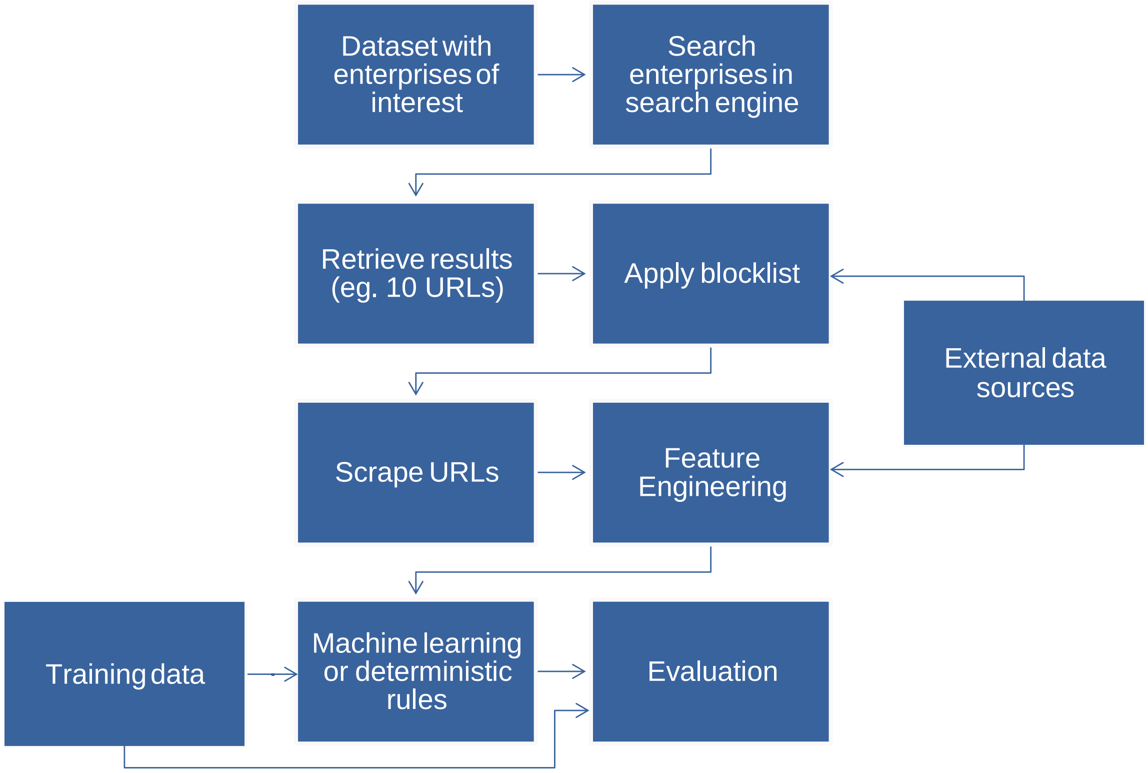

Dataset with enterprises of interest
Search enterprises in search engine
Retrieve results (eg. 10 URLs)
Apply blocklist
External data sources
Scrape URLs
Feature Engineering
Training data
Machine learning or deterministic rules
Evaluation

Figure 1. General overview of URL finding process

URL finding at Statistics Hesse

Statistics Hesse has their own URL finder written in R. It searches and scrapes enterprise websites and subsequently searches for contact information and other identifying information of enterprises with regular expressions. Machine learning is applied to identify the correct URLs (an enterprise can have more than one URL).

The Google search API is used with the name and municipality of the enterprise as search terms and supplies up to 10 results per search query. Scraping was done with the headless browser PhantomJS. German enterprises are legally required to have an imprint (“Impressum”) page on their website, which lists basic contact details and identifying information about the enterprise. Using regular expressions, Statistics Hesse searches for links within the website that contain strings like “Impressum” and scrapes these website subpages since it is assumed that they will give the best information for linkage. Statistics Hesse searches for the name, address, and different register numbers (Chamber of Commerce, European VAT ID, German tax ID) in the website texts to identify correct websites. Since there are no URL data available from Official Statistics data sources in Germany, Statistics Hesse manually created a training set of 2000 enterprises in retail trade. This training data has been used to predict the URLs of all retail trade enterprises in Hesse (29490 legal units) in 2021.

RESULTS

Evaluation scores for URL finders within the ESS

Determining how well URL finding works in general as well as for different countries and statistical offices is challenging. On the one hand, most statistical offices have developed their own URL finding software using slightly different methodologies. On the other hand, high quality training and evaluation data is hard to come by. Most frequently, statistical offices use already existing URL data in the statistical business register as training data. It is difficult to assess if already existing URL data is up to date and relatively error free. For an attempt to compare URL finding performances, see Table 1. For the URL finders of Bulgaria (BNSI), Italy (Istat), Statistics Hesse and the Netherlands (CBS), the table reports the F1 score as well as the evaluation score that the author(s) preferred [3]. Overall, evaluation scores indicate that URL finding performs sufficiently well but could be further improved.

Table 1. Evaluation scores of different URL finders

URL finder	F1 score (level: websites)	Evaluation score „of choice“
BNSI	0.59	Precision: 0.72
Istat	0.81	Accuracy: 0.79; F1: 0.81
Statistics Hesse	0.82	82.3% of enterprises correct
CBS	0.77	F1 score on enterprise level: 0.84

Detailed results for URL finding in Hesse

This subsection presents the results of URL finding at Statistics Hesse in more detail. Statistics Hesse first tested a deterministic approach for identifying correct URLs. If one of two tax IDs (the German tax ID and the European VAT ID) were found on the website, the URL was considered to be correct. Table 2, section “Deterministic approach”, presents the resulting evaluation scores. The results show that, by far, not every enterprise has a tax ID on its website. The VAT ID is the more common and valuable criteria to determine if the URL is correct. However, only 64% of the correct websites contain this information. To conclude, while deterministic linkage can achieve high precision, it will miss many correct URLs due to the low recall.

The best performing ML algorithm, Gradient Boosting, achieved a substantially higher recall (76%) than the best deterministic approach with just a slight decrease in precision. The correct URL was found for 82.5% of the enterprises. According to the manual search, the most frequent error was that no website was found for enterprises with URLs (9.6%). 5.7% of the enterprises were assigned a URL even though they did not have a website according to the manual search, and 2.4% of the enterprises were assigned a wrong URL.

Table 2. Results of URL finding at Statistics Hesse

Classifier	Precision	Recall	F1
Deterministic approach			
VAT ID	0.91	0.64	0.76
German Tax ID	0.98	0.06	0.12
VAT ID or German Tax ID	0.92	0.66	0.77
Machine Learning			
Gradient Boosting	0.89	0.76	0.82

cONCLUSIONS

Different statistical institutes within the ESS have adopted URL finding procedures with search engines results. Most institutes have written their own software for that purpose with different methodologies and used different data sources for training data. This makes evaluating and comparing the performance of URL finding difficult. In general, URL finding is a promising way for official statistics to fill data gaps on enterprise URLs in their business register. However, more work should be done to evaluate the quality of URL finding and to determine at what point more standardization would be useful.

Statistics Hesse is currently in the process of repeating the case study from 2021 on retail trade enterprises with improved scraping procedures and better feature engineering steps. Subsequently, it aims to bring URL finding closer towards production by applying it to the whole statistical business register.

rEFERENCES

[1] A. van Delden, D. Windmeijer and O. ten Bosch, Finding enterprise websites. European Establishment Statistics Workshop (2019). https://www.researchgate.net/publication/336995371_Finding_enterprise_websites

[2] G. Barcaroli, M. Scannapieco and D. Summa, On the Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web, Italian Review of Economics, Demography and Statistics 70 (2016), p. 25–41. http://www.sieds.it/listing/RePEc/journl/2016LXX_N4_RIEDS_2541_Scannapieco.pdf

[3] H. Kühnemann, A. van Delden, D. Summa et al., URL finding methodology, Joint report for Work Package 2 and Work Package 3, Use Case 5, Web Intelligence Network (2022). https://ec.europa.eu/eurostat/cros/system/files/20220131_url_finding_methodology.p df

National Accounts in a World of Naturally Occurring Data: A Proof of Concept for Consumption

Keywords: National Accounts, Transaction Data, Consumption, Big Data, Household Survey

7. Introduction

The workings of modern payment systems and financial institutions generate a complete ledger of everyday transactions. This large, naturally occurring and unstructured transaction-level data is increasingly available to researchers and holds the promise of reshaping economic measurement. And yet, despite recent advances, national statistical agencies still rely on more traditionally structured survey data and slow-moving censuses. These are commonly perceived to be facing both increased funding and, on occasion, political pressures [1][2][3].

Against this background, this paper provides a first proof of concept that naturally occurring transaction data, arising through the decentralized activity of millions of economic agents, can be organized via national accounting rules and then harnessed to produce a large-scale, high-quality and highly-detailed consumption survey. This, in turn, can then be deployed to produce national account objects via simple aggregation. In particular, we show how comprehensive transaction-level data from one of the largest banks in the world can be organized to (i) reproduce current official statistics on aggregate consumption in the national accounts with a high degree of precision and, (ii) as a result of the richness of the underlying transaction data produce novel, highly detailed distributional accounts for consumption. Additionally, we show that (iii) the panel nature of our data can offer new insights on the nature of individual-level consumption risk and consumption dynamics, overcoming challenges associated with purely cross-sectional or short-panel consumption surveys.

8. Methods

Our data covers the universe of retail accounts of BBVA, a large Spanish bank, and yields an unprecedented granular ledger, allowing us to track expenditure as it flows out of these accounts, transaction by transaction, for a total of 3 billion individual transactions by 1.8 million BBVA customers, from 2015Q2 to 2021Q4.

8.1. SAMPLE FRAME AND DEMOGRAPHIC WEIGHTING

We define a balanced panel of active customers who make at least ten consumption-related transactions in each quarter. There are 1,827,866 such customers, much more than the sample frame of the Spanish Household Budget Survey (HBS), allowing us to make much finer cuts of the data.

We form an estimate of active customers' household size using auxiliary information on co-signed financial contracts. When we come to form aggregate consumption measures, we address the demographic imbalances of our data appropriately. Demographic cells used for proportional upweighting are defined at the gender, age group and neighbourhood income quintile levels. Here, we also account for non-active customers and family members who share the consumption of active customers, but also potentially generate consumption spending outside the BBVA universe.

2.2 From spending to consumption

For non-housing consumption, our overall strategy is to use transaction metadata to classify individual purchases as either consumption- or non-consumption-related and, if the former, to assign a COICOP classification. In doing so, we closely follow national accounting principles from the European System of Accounts (ESA 2010). There are three main types of transaction data in the sample: card payments, direct debits, and irregular transfers. Each payment class has different associated metadata, which we use to classify transactions using, among others, Merchant Client Codes (MCCs), NACE codes linked to the beneficiary's tax ID and text mining in different data fields. Purchases from multiproduct retailers are allocated into COICOP categories according to official sales-by-product statistics. Cash withdrawals are considered consumption and are allocated over COICOPs according to the proportions observed in offline card transactions.

We use a regression model to estimate the imputed housing consumption for each household. Here, we use utility payments and income as explanatory variables and month as the fixed effect to estimate rental payment; the latter is extracted for 16,977 households with continuous rental payments identified through the free-text field of manual transfers. The resulting model is used to impute housing consumption according to the principles of national accounts.

There is indeed a large distance between raw spending and consumption. This suggests that using the former as a proxy for the latter in the absence of appropriate metadata is likely to produce a poor approximation.

3. Results

3.1 Measuring Aggregate Consumption

Figure 1 plots the level of consumption according to the "Final Household Consumption" in the Spanish national accounts against our approach in the left panel, and quarter-on-quarter growth rates in the right panel. The striking result is that naturally occurring and official data align exceedingly well in both levels and growth rates at quarterly frequency. This is despite their quite different constructions.

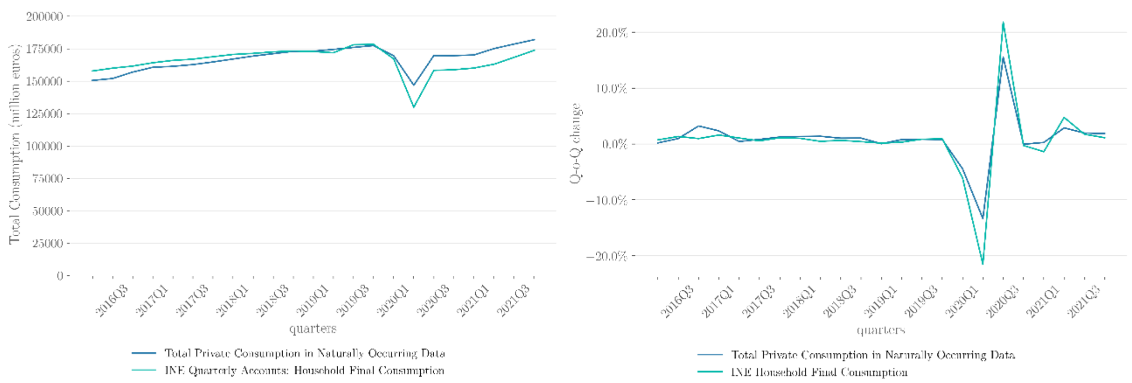


Figure 1: Aggregate Naturally Occurring Consumption vs. National Accounts

The National Accounts is a complex object which relies on dozens of underlying data series that include household and firm surveys with different samples and for different goods categories, often using synthetic data that is not publicly available. Indeed, one virtue of our approach is its simplicity: our philosophy is to design an accurate individual consumption survey which can then aggregate directly to the national level. The average coverage ratio across quarters of our level series with respect to national accounts is 1.01. Our series is thus a direct measure of national consumption, not simply a coincident indicator. The correlation in quarter-on-quarter growth rates is 0.974.

In **Figure 2**, we compare the distribution of aggregate consumption across COICOP categories according to national accounts, the HBS, and our naturally occurring data in 2019. In general, there is a strong relationship between national accounts COICOP-specific consumption levels and those of the HBS and naturally occurring data. Overall, though, naturally occurring data achieves better coverage of national accounts: the average absolute error with respect to national accounts of naturally occurring data (HBS) is 0.266 (0.333) log points across all COICOP categories.

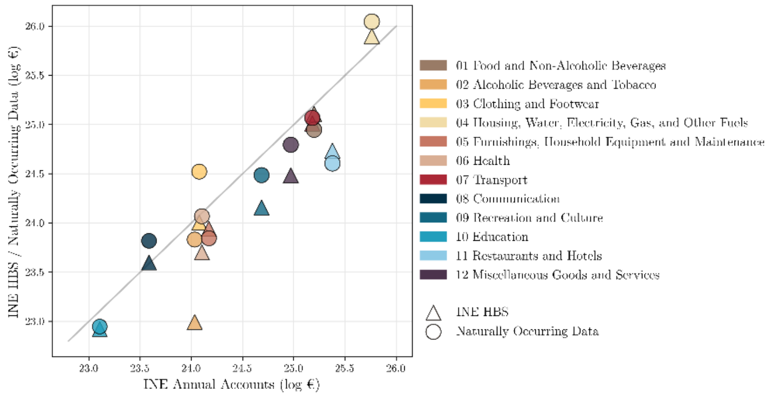


Figure 2: Distribution of Spending across COICOP Categories

Card spending is one of the most widely available forms of financial transaction data, and has recently been used to track the effects of COVID-19 in several papers [6][7]. We find that the card series has poor aggregate coverage of national accounts consumption which underlines the need for tracking all payment methods. Moreover, its growth rate has a notable upward bias not present in the full naturally occurring consumption measure.

Further in this section of our paper, we create novel aggregate objects that go beyond what is already available from statistics agencies, such as consumption nowcasting by using a daily frequency, evolution of consumption in different COICOP categories and evolution of means of payment over time.

3.2 Distributional National Accounts

Following the work of Piketty et al. [8], distributional national accounts for income already exist for a large number of countries. Yet, to the best of our knowledge, distributional national accounts for consumption are virtually non-existent. The left panel of **Figure 3** overlays the consumption distribution gained from our data with the post-tax national income distribution for Spain (by income percentiles in 2017), made available by the World Inequality Database. The right panel plots the implied Lorenz curves for these two distributions. Clearly, inequality in consumption is substantially smaller than income inequality.

Further in this section of our paper, we present distributional national accounts disaggregated by demographic characteristics, in particular by age and gender. For instance, we see that adult consumption grows throughout the 20s and 30s, peaks in middle age, and declines smoothly thereafter.

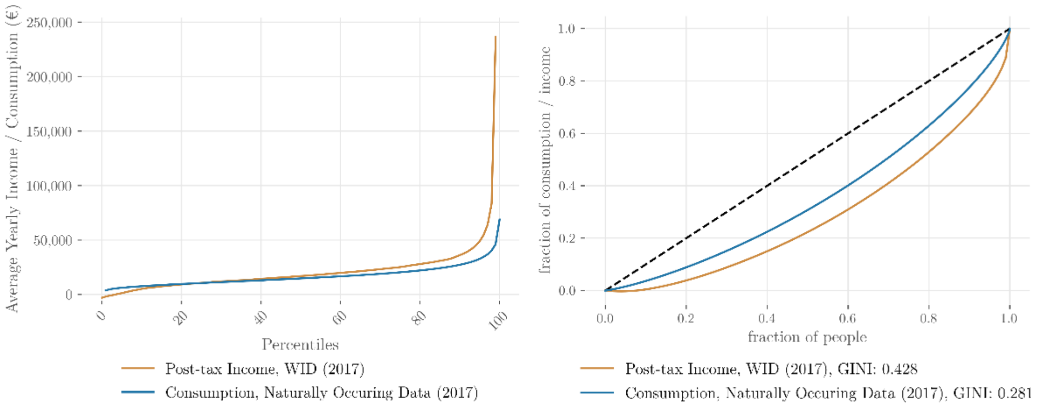


Figure 3: BBVA Consumption Inequality vs Income Inequality

One additional advantage of high-resolution transaction data is that it allows us to conduct distributional analysis at varying time frequencies, which is likely important to policy-makers when considering the real-time implications of major shocks. The Lorenz curves over the entire distribution of aggregate consumption implied by different sampling frequencies (daily, monthly, yearly etc.) indicate that the inequality in the distribution of total consumption declines strongly with the sampling frequency.

3.3 Individual Consumption Dynamics Across the Consumption Distribution

Observing the data at an individual level, it becomes clear that - just as Guvenen et al. [9] report for individual income growth in the US – consumption growth does not seem to be well approximated by a Gaussian distribution. Rather, the linear log-log relationship suggests a form of Pareto distribution for consumption growth in both the left and the right tail.

Similarly, we investigate different moments of the consumption growth by age groups. We observe strong and fast mean reversion. Agents in the lowest percentiles of the consumption distribution have large average increases in consumption over the following year (of about 10%), while agents that are in the highest percentiles sustain a severe decline.

4. Conclusions

Our paper advocates the use of this unstructured, but readily available data for both the construction of national aggregate and distributional accounts as well as the study of the microstructure of economic activity. Our proof of concept results imply that simple and transparent procedures followed by bottom-up aggregation tracks with remarkable accuracy not only the growth rate of consumption in national accounts, but also its level. Further, due to its granularity it allows immediate decomposition across goods, demographics, space or time frequencies. In particular, the good aggregation properties of the data allow for a distributional analysis of aggregate consumption, providing a rich, macro-consistent description of consumption inequality and its time-evolution. Finally, we have seen that this same data that aggregates into national accounts can be used to analyse the microstructure of the economy.

References

[1] European Commission (2010). Report on Greek Government Deficit and Debt Statistics. Technical report, European Commission, Brussels.

[2] Vinik, D. (2017). Is Washington bungling the Census? Politico.

[3] AEA Committees on Economic Statistics and Government Relations (2020). Statement on the Need for Accurate and Reliable Data from the 2020 Decennial Census. Technical report, American Economic Institution, Nashville.

[4] Attanasio, O., Hurst, E., and Pistaferri, L. (2014). The Evolution of Income, Consumption, and Leisure Inequality in the United States, 1980–2010. In Improving the Measurement of Consumer Expenditures, pages 100–140. University of Chicago Press.

[5] Barrett, G., Levell, P., and Milligan, K. (2014). A Comparison of Micro and Macro Expenditure Measures across Countries Using Differing Survey Methods. In Improving the Measurement of Consumer Expenditures, pages 263–286. University of Chicago Press.

[6] Andersen, A. L., Hansen, E. T., Johannesen, N., and Sheridan, A. (2020). Consumer Responses to the COVID-19 Crisis: Evidence from Bank Account Transaction Data. SSRN Scholarly Paper ID 3609814, Social Science Research Network, Rochester, NY.

[7] Vavra, J. (2021). Tracking the Pandemic in Real Time: Administrative Micro Data in Business Cycles Enters the Spotlight. Journal of Economic Perspectives, 35(3):47–66.

[8] Piketty, T., Saez, E., and Zucman, G. (2018). Distributional National Accounts: Methods and Estimates for the United States. The Quarterly Journal of Economics, 133(2):553–609.

[9] Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2021). What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics? Econometrica, 89(5):2303–2339.

Data Literacy (MANS2M.2)

Session Chair**: Remco Paulussen** *(Statistics Netherlands-CBS)*

Creating, testing and maintaining a functional library for statistical methods
Susie Jentoft *(Statistics Norway)*

Eurostat Education corner
Romina Brondino and Nina Jere *(Eurostat)*

Creating, testing and maintaining a functional library for statistical methods

Keywords: standardisation, statistical methods, programming, modernisation

Introduction

There is an increasing interest within statistical agencies to move towards open source software solutions. These are often attractive not only for the financial reasons, but for their ease in integrating into new modern platform solutions. There is also a broad range of tasks they solve. Both python and R programming languages provide a wealth of opportunities for programming statistical production processes; both having a modular nature. There are over 18 000 R packages publicly available on CRAN [1] while python has over 400 000 modules on its central repository PyPI [2]. However, with all these possibilities comes the question of which to use to solve the problem at hand and do they really do what they say do?

Statistics Norway has had several developments in the last 5 years leading to an increased need for competencies in open source software, specifically R and python. The establishment of its common data platform called DAPLA has directly increased the need for programming skills for employees. R and python are the main programming languages available on the new platform. Additionally, a National Programme for Official Statistics has been established allowing other organisations to produce official statistics for the country [3]. This has increased the demand for the Division of Methods to provide open source solutions to methodological challenges.

Here, we describe how the methods library, *Metodebibliotek*, was created at Statistics Norway. Building a library of statistical methods' functions provides both a framework for methods architecture and a practical solution for the implementation of those methods. It is one way to strengthen methodological skills both within Statistics Norway and with external statistical producers; an important recommendation in Statistics Norway's recent peer review rapport [4]. The *Metodebibliotek* is essentially an organised list of functions written in either R or python, which meet standardised criteria.

Methods

We have classified the challenges faced when building the *Metodebibliotek* into the groups: contents, organization and technical. The contents and organisation aspects are described in this section, and guidelines were developed by a project working group. The implementation of these relates to the technical aspects of the *Metodebibliotek* and are described in the results section.

One important area to establish has been the criteria for including a function in the library, i.e. the contents. Three specific criteria were established for both internally written and externally published functions. Firstly, we determined that functions must be currently used within a statistical process at Statistics Norway or used within internal statistical methods courses to be included in the library. Secondly, as the quality of the content is essential for the success of the project, functions would always be checked by the Division for Methods prior to inclusion. This was particularly important for functions developed outside of Statistics Norway. A general check of the functionality is undertaken, and insurance on adequate theoretical documentation of the method used. It must be considered a reasonable statistical method for official statistics production. Thirdly, unittests should be written for all functions included in the library to provide automated tests for a base level of functionality. These can be used to check when there are updates in either the packages/modules, program or platform.

For the *Metodebibliotek* to be used broadly it needs to be organised and presented in an understandable way. It should be easy to search and find functions. It must also be possible to implement the functions smoothly into a production process. Most employees within the organisation are familiar with the GSBPM [5] or at least can identify with the process steps. Therefore, the project working group decided to use this to organise the functions. Additionally, we identified a need for categories that span across the GSBPM, for example the statistical domain.

Results

In 2021, we created a public Github repository for the *Metodebibliotek* [6] which has been developing since then. This solution allows access to the content for both internal and external users. The functionality available in Github repositories has been developed greatly over the years and provides an ideal technical solution to the requirements specified in the project.

We built a website using *Rmarkdown* and *distill* in R to organise and communicate the functions. While functions are generally not directly located within the repository, it acts as a list with details on where and how the functions can be used. Documentation can be created automatically based on the help (Rmd) files for R functions or manually. Categories are used to provide better searchability and several theme pages have been created. This site is hosted on *Github* pages allowing an easily integrated system.

Unittests have been written using *testthat* for R and *pytest* in Python to allow automatic testing. We utilised *GitHub Actions* to automate the testing process and to communicate the status of these using badges. A matrix configuration is used to run the tests on multiple system setups to mimic the various programming environments within Statistics Norway.

Conclusions

The *Metodebibliotek* is established and uses a defined organisational structure. It provides a toolbox of functions for methods which can be implemented into a production processes at both Statistics Norway and other official statistics organizations. We have seen the need for good methods to be available and easy to implement so they used in production. We believe providing a set of tested and assured functions will aid in improving the quality of the statistics we produce and save resource for those setting up new production processes.

Further work includes building the contents of the library to cover more methods. More automation will help the Division for Methods to extend the library and minimize the ongoing resources needed to maintain the site. For example, assessing whether the use of *pkgdown* for R and/or *sphinx* for python can contribute to the automation processes of the library.

Alternately, *Quarto* [7] has shown interesting developments for building cross-language websites and will also be considered in future.

References

[21] The comprehensive R network, accessed 15.10.2022. https://cran.r-project.org/

[22] Python package index, accessed 15.10.2022. https://pypi.org/

[23] National Programme for Official Statistics, accessed 16.10.2022. https://www.ssb.no/en/omssb/nasjonalt-program-for-offisiell-statistikk

[24] M. Bruun, J. Delbas, H. Ottosson, K. Zeelenberg (2021) *Peer review report on the compliance with the European Statics code of practice and further improvement and development of the National Statistical System: Norway.*

[25] GSBPM: Generic Statistical Business Process Model. European Comission, accessd 15.10.2022. https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en

[26] Metodebibliotek, accessed 15.10.2022. https://github.com/statisticsnorway/metodebibliotek

[27] Quarto, accessed 15.10.2022. https://quarto.org/

POST02: Poster session

Using Digital Trace Data to Generate Representative Estimates of Disease Prevalence [COVID 19 Infections] in Belgian Municipalities
Sen Dishani *(KU Leuven)*

Integration of land use vector data from administrative sources for agri-environmental analysis
Pietro Macedoni *(University of Bologna)*

Development of methodology for automated crop mapping in Greece using Neural Networks and Sentinel-2 satellite imagery
Eleni Papadopoulou *(Aristotle University of Thessaloniki)*

Evaluation of some methods of detecting outliers
Adaku Obikee *(University of Agriculture and Environmental Sciences Umuagwo)*

Pre and post COVID-19 related statistics analysis for identifying future opportunities based on historical dynamics
Sergio Gallego García *(UNED)*

Unsupervised ranking and categorisation of companies using web scraping and machine learning
Michael Reusens *(Statistics Flanders)*, Cedric De Boom *(Statistics Flanders)*

Integrating multimode survey data with VTL
Benoit Werquin *(National Institute of Statistics–INSEE)*

Modelling local level housing demand based on the Multi-sectoral Regional Microsimulation Model (MikroSim)
Sarah Bohnensteffen (*German Federal Statistical Office-DESTATIS)*

Non-traditional data and methods for identifying patterns of inequality to digital access in the EU
Patrizia Sulis *(European Commission)*

Natural Language Processing to automate data coding
Costas Diamantides *(Statistical Service of Cyprus-CYSTAT)*

Maritime mobility statistics using open data
Danila Filipponi *(National Institute of Statistics–ISTAT)*

The use of synthetic data for data sharing : An application on survey data
Isabella Corazziari *(National Institute of Statistics–ISTAT)*, Loredana Di Consiglio *(National Institute of Statistics–ISTAT)*

Statistical Data Triplification : The case of semi-automatic generation of RDF triples from relational databases
Stamatios Theocharis *(Ministry of the Interior)*

Local population projections with Bayesian hierarchical models
Violeta Calian *(Statistics Iceland)*

Data circularity for the circular economy
Firuza Nahmadova *(Datastake)*

Evaluation of Some Methods of Detecting Outliers

Abstract

This research work employed both Real life Data Application and a Simulated Data to evaluate some outlier detection techniques such as t-statistic, Modified Z-Statistic, Cancer Outlier Profile Analysis (COPA), Outlier Sum-Statistic (OS), Outlier Robust T-Statistic (ORT), and the Truncated Outlier Robust T-Statistic (TORT) to verify which technique has the highest power of detecting and handling outliers on the bases of their rank values, P-values, true positives, false positives, False Discovery Rate (FDR) and their corresponding Area Under the Receiver Operating Characteristic (ROC) curves (AUC) respectively. Using the real life data, we observed that among the first three outlier methods Z, COPA and OS, the performance of OS is outstanding with a smaller FDR, better FP and TP followed by COPA with a better FDR, better FP and TP while Z has no FP and a poor FDR. Among the last three outlier methods, T, ORT and TORT, the performance of T is better with a better FDR and FP followed by ORT with a better FP while TORT has no FP and poor FDR. Then for the simulated data, we observe that among the first three outlier methods Z, COPA and OS, OS performed better with the highest rank value followed by Z and COPA. We observed using the p-values that OS has the least number of True Positive and the highest number of False Positive followed by COPA and Modified Z. We also observed that OS has the smallest FDR followed by COPA and Z. From the ROC curves, we observed that COPA has the highest AUC which indicate better sensitivity and specificity followed by Z with AUC that is on the reference line while OS has AUC that is under the reference line. Among the last three outlier methods, T-statistic, ORT and TORT, the T-statistic performed better with a better rank value followed by ORT and TORT. We observed using the p-values that T has a better number of True Positive and False Positive followed by ORT and TORT. T has a smaller FDR followed by TORT and ORT in that order. From the ROC curves, we observed that ORT and TORT has better significant AUC which indicate better sensitivity and specificity while T has AUC that is under the reference line.

Keywords: ***Area under the ROC Curve (AUC), Sensitivity, Specificity, P-Value, Rank Value, False Discovery Rate (FDR), Real life Data Application, Simulation.***

Pre and post COVID-19 related statistics analysis for identifying future opportunities based on historical dynamics

Keywords: COVID-19, Big Data, Time series forecasting, Hypothesis Testing, Correlation Analysis, Regression Analysis, Vaccines, Mortality.

INTRODUCTION

COVID-19 has changed our world. From the initial impact until the long-term consequences, it has reshaped many areas of our society. Thus, the data collected [1, 2, 3, 4, 5, 6] is expected to provide the effects of this impact on different indicators. However, in many cases data alone does not provide information about the influence on one factor on the indicator under study. To address this gap this paper aims to provide an approach to analyse the impact of an event on society based on statistical analysis and on a selection of key indicators. In this case the event is the COVID-19 pandemic with its related policies such as vaccine policy, service management, etc. The analysis is performed for different regions showing the impact, explainability, and significance of the event on the changes on the selected indicators.

METHODS

2.1. Time series forecasting

The suitability of a forecasting method depends on the pattern [9]. Therefore, several models were selected. The forecasts of the different methods in the conceptual model were compared by means of the forecast error. There are different methods to measure and provide conclusions regarding the accuracy of the used forecast method [10]. The mean absolute deviation (MAD), the mean square error (MSE), and the mean absolute percentage error (MAPE) as methods that produce consistent results when comparing different forecasting methods [8]. The methods used are [7, 8]:

1. Moving average and cumulative moving average.
2. Linear regression.
3. Exponential smoothing of first, second, and third order.
4. Croston method.

In this paper, the forecasting methods are applied to the forecast of the excess mortality (%) as well as for the total mortality (# number), among others.

2.2. Hypothesis testing

Hypothesis testing is a set of formal procedures used to either accept or reject statistical hypotheses. In this study, it is applied to test the hypothesis regarding the mortality level before and after the COVID-19 pandemic as well as during the vaccination process and after a high level of vaccination rate of the population. Among others, one of the hypothesis tests is performed for Spain comparing two samples of mortality rates before 2022 (since the beginning of COVID-19 cases in Spain) and after 2022 to identify if the average excess mortality rates of the two samples are equal (H_0) or are different (H_1). Thus, one can determine if the vaccination process has had an impact on the average mortality rate or not. Furthermore, other of the hypothesis testing has been performed to compare the mortality rates between 2016-2019 and during the COVID-19, and after the vaccination process to obtain conclusions regarding the explainability of the excess mortality rates.

2.3. Correlation and regression analysis

Correlation and regression analysis are essential tools in statistical analysis. While correlation captures the interrelationship between two variables, regression measures one variable's effect on another indicator. In this publication, these statistical analyses are applied to study the correlation and regression analysis between vaccination rate level of population and excess mortality for a given region, among others.

2.4. Society impacts and Key Performance Indicators Selection The main

indicators and factors analysed are:

1. COVID-19 cases (# number).
2. Monthly total deaths (# number).
3. Monthly excess mortality (% of additional deaths compared with average monthly deaths in 2016-2019).
4. Vaccinated population evolution (% of total population).
5. Registrations of new businesses (% and total).
6. Declarations of bankruptcies (% and total).
7. Other factors: seasonal effects, policies, etc.

RESULTS

An extract of results is shown:

1. Correlation and regression analysis for Spain: excess mortality with and without vaccine in percentage (vaccinated level impact on excess mortality):

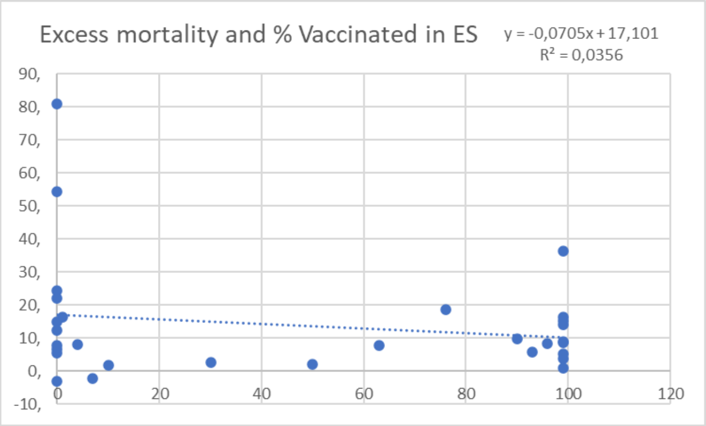


Figure 1. Excess mortality in Spain (%) versus vaccinated population (%)

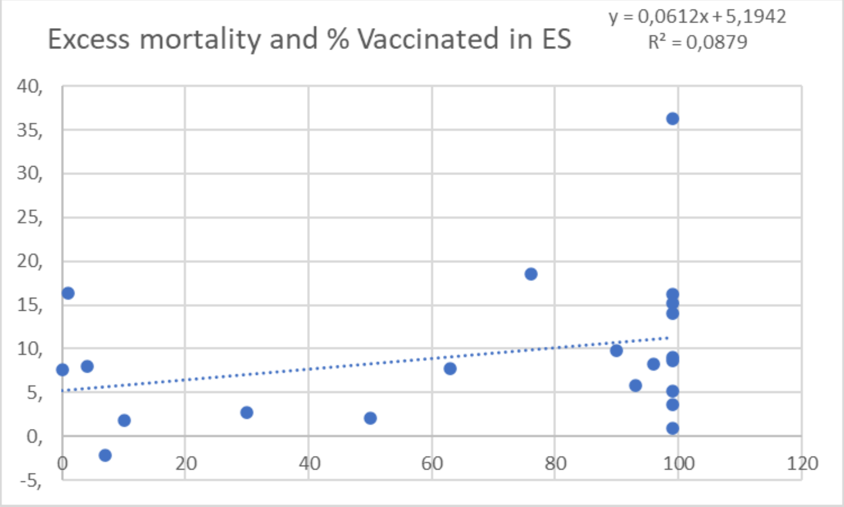


Figure 2. Excess mortality in Spain (%) versus vaccinated population (%) since vaccination started

2. Peak deaths associated with mismanagement among other causes: For instance, Spain presents one main peak in March and April 2020 with over 50% mortality rates higher than in period 2016-2019.

3. Service quality/level and impact on other diseases (processing time).

4. Mortality before versus mortality after covid with forecasting of mortality impact on the short, medium, and long-term:

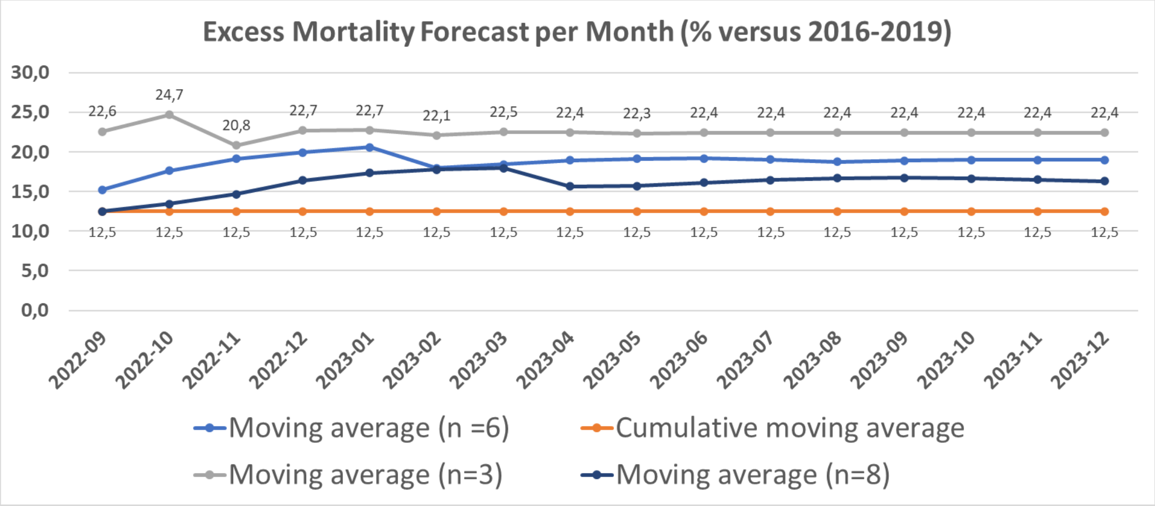


Figure 3. Extract of results: forecast of excess mortality in Spain from 2022-09 to 2023-12.

5. Comparison for ES - IT - FR – DE, as well as for regions within Spain: the hypothesis tests performed show that the average of excess mortality before and after the high level of vaccination could be considered the same for different samples and p-values for the different countries. Therefore, it does not show a reduction of the excess mortality rates thanks to vaccination. Moreover, the hypothesis test to check if the excess mortality after COVID-19 and vaccination of most part of the population could be random in comparison with period 2016-2019 is rejected.

cONCLUSIONS

First the model supports the forecast of impacts as for instance in excess mortality in the short, medium, and long-term showing for instance levels from 12,5% to over 20% using the moving average with different alternatives. Then the hypothesis testing allows describing if the excess mortality is due to COVID-19 and other effects, as well as if the vaccination level had an impact on cases and excess mortality. The analysis also provides a syndication of potentials for management during the crisis and for service management in the process. Finally, as shown correlation and regression analysis allow to see how higher vaccination levels does not imply less excess mortality rates.

rEFERENCES

[1] https://ine.es/

[2] https://www.destatis.de/DE/Home/_inhalt.html

[3] https://www.istat.it/en/

[4] https://www.insee.fr/en/information/2107711

[5] https://ec.europa.eu/eurostat/databrowser/view/DEMO_MEXRT__custom_309801/bookmark/line?lang=en&bookmarkId=26981184-4241-4855-b18e-8647fc8c0dd2

[6] https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/pbiVacunacion.htm

[7] Schuh, G.; Stich, V.; Wienholdt, H. Logistikmanagement.; Springer: Berlin/Heidelberg, Germany, 2013.

[8] Meyer, J.C.; Sander, U.; Wetzchewald, P. Bestände Senken, Lieferservice SteigernAnsatzpunkt Bestandsmanagement; FIR: Aachen, Germany, 2019.

[9] Gallego-García, S.; Reschke, J.; García-García, M. Design and simulation of a capacity management model using a digital twin approach based on the viable system model: Case study of an automotive plant. Appl. Sci. 2019, 9, 5567.

[10] Schönsleben, P. Integrales Logistikmanagement: Operations und Supply Chain Management Innerhalb des Unternehmens und Unternehmensübergreifend; Springer: Berlin/Heidelberg, Germany, 2011.

Unsupervised ranking and categorisation of companies using web scraping and machine learning

Keywords: web scraping, natural language processing, official statistics, machine learning

INTRODUCTION

This paper presents a method to automatically categorise companies based on the text scraped from their website. The method is demonstrated by applying it to categorising companies as being active in the domain of artificial intelligence (AI) or not.

This study has been set up to evaluate if the text scraped from company websites can be used as a complement to existing data sources to produce official statistics. Today, most official statistics are produced using survey data (such as the Community Innovation Survey [1]) and administrative data sources. The web scraping methodology proposed in this paper has the following advantages compared to these traditional data sources. First, scraping instead of surveying alleviates the response burden of companies. Second, using the web scraping method, companies can be surveyed at any desired frequency. This leads to more up-to-date data, resulting in increased quality and timeliness of the company statistics. Finally, our approach can be generalised to any categorisation of interest, resulting in the ability of statistical organisations to create new company statistics relatively quickly.

In recent years, there have been other studies that show the opportunities of web-scraped company information for use in statistics production [2,3]. Our study contributes to these existing studies in the following ways. First, we apply and evaluate a method that to the best of our knowledge has not yet been applied for use in official statistics. Next, our approach allows for the ranking of companies without any labelled data, and for the categorisation of companies with only a small amount of labelled data. Existing methods of categorising companies based on their website texts require a relatively large set of labelled data. Finally, the method evaluated in this paper is generically applicable to a large amount of different company categorisations.

METHODS

As a specific use case to demonstrate the method, we discuss our experiments categorising companies as being active in AI or not. Keep in mind that the same method can be applied to other company categorisations (e.g. active in bioeconomy or not, being a transportation company, etc.).

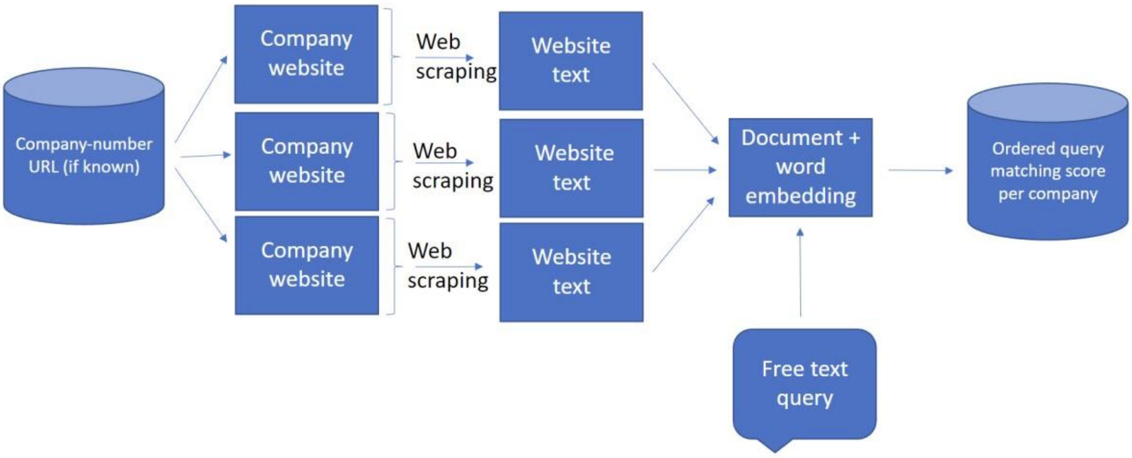


Figure 1. Overview of the method

Figure 1. shows an overview of the method. In the following subsections, we will discuss each step in the figure.

Input data

The input data for the method is a list of companies described by their company number and URL if known. For our experimental dataset we used the list of all companies with a legal entity in Belgium, excluding one-person businesses. The one-person businesses are excluded for privacy reasons. This results in a dataset of 914,000 Belgian companies, of which 320,000 had a known URL. The URLs in this dataset were purchased from a business partner. The challenge of automatically finding and validating company URLs is out of the scope of this study.

Web scraping and data cleaning

For each of the URLs, the visible text from the homepages was scraped using Python in combination with the requests [4] and beautifulsoup [5] libraries. Scraping text from deeper webpages (such as the 'about us' page) is an improvement we will tackle in future work. The following cleaning steps are performed on the resulting texts. Only texts in English and Dutch are retained. Language detection are done using the langdetect [6] library in Python. Only texts with more than 50 characters are retained and stop words and a custom list of web-technology words are removed from the texts. The goal of the cleaning is to obtain texts that are dense in information on company activities. After scraping and cleaning, the dataset is reduced to 200,000 clean texts.

Document and word embedding

The cleaned texts and individual words are jointly embedded using a fine-tuned version of a pretrained multilingual transformer model [7]. For the implementation of this joint embedding the Top2Vec [8] library is used. The resulting embeddings of company texts and words lie close together if they are semantically similar and far apart if they are dissimilar. The embedding model used in our demonstration allows for the combination of 16 different languages for which the model is pretrained. This makes it trivial to deal with different companies using a different language on their website (as long as the languages are included in the pretrained set of languages).

Query selection and embedding

Next, a free text query is defined that describes the categorisation of interest. For our example use case, AI, we concatenated the Wikipedia introductions of ‘artificial intelligence’ and of 'machine learning' with a description of ‘data science’ found on an IBM webpage. There are infinitely many options to define a query. Designing a method to find the optimal query for a given categorisation will be tackled in future work. Once it is defined, the query is embedded using the same model as the company texts.

Company ordering and categorisation

The distance between the embedded query and each embedded company text is calculated. This allows for a ranking of companies with the first company having a website text that has the shortest distance to the query and the last company having the website text with the highest distance to the query. If a sorted list of companies given a specific activity is the desired output, the method can stop here and is completely unsupervised. For businessfacing government agencies this is already a valuable outcome. If a binary categorisation is needed, such as for statistics production, some labelled data is necessary, making the approach semi-supervised. To go from ordering to categorisation, a cut-off score must be decided. Companies with a shorter distance to the query than the cut-off are considered part of the category, the others as not part of the category. The choice of cut-off can be made by optimising recall@N and precision@N, with N the number of companies being included by the cut-off.

RESULTS

In order to validate the approach, a dataset of 50 known AI companies was created.

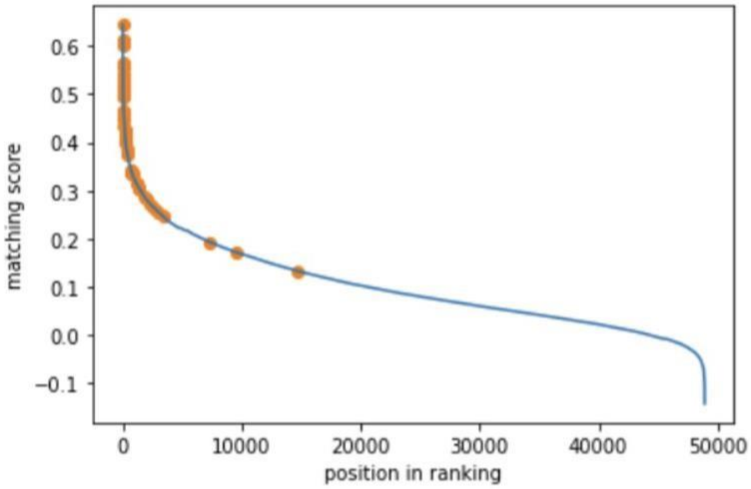


Figure 2. Matching score per company (blue) and placement of known AI companies (orange)

Figure 2. shows the sorted matching scores per company given its position in the ordering. This figure also shows that the known AI companies are placed highly in the ordering, which is desirable. The quantitative performance of the method can be seen in Table 1. The median ranking of the known AI companies is 199, with median score of 0.41.

Table 1. Quality of ranking known AI companies

	Mean rank	Median rank	Mean score	Median score
Known AI companies	1290	199	0.40	0.41

Inspection of 50 randomly selected unknown companies that were ranked higher than the median rank showed only 1 company not active in AI. An inspection of 50 random companies with matching score lower than the lowest-scoring known AI company showed no companies active in AI. This indicates desirable false-positive - and false-negative rates.

cONCLUSIONS

The method presented in this paper demonstrates ta new way of complementing traditional company information with website texts. Our first experiments show that the method is successful in ordering companies active in AI.

Following these promising results, we identify the following gaps in this paper for future research. First, a more elaborate validation approach needs to be set-up to assess the quality of the ordering and give guidance to the choice of cut-off point for categorisation. In order to do so, the ordered list of AI companies is currently being used by business-facing government consultants, who provide further feedback on the quality of the results of this method. Next, the performance of the method for other categorisations should be verified. Besides AI, we are investigating categorising companies as being active in bioeconomy and circular economy. Finally, further research should be done on the general properties of website texts for the production of company statistics. For example, bias in the type of companies having a website could be of concern.

References

[1] https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey, (2022)

[2] P. Daas and S. van der Doef, "Using Website texts to detect Innovative Companies.", (2021).

[3] https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en, (2022)

[4] Python Software Foundation, https://requests.readthedocs.io/en/latest/, (2022)

[5] L. Richardson, https://www.crummy.com/software/BeautifulSoup/, (2022)

[6] https://github.com/Mimino666/langdetect , (2022)

[7] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3 , (2022)

[8] D. Angelov, "Top2vec: Distributed representations of topics.", arXiv preprint arXiv:2008.09470, (2020).

Integrating multimode survey data with VTL

Keywords: Data integration, multimode, VTL

1. Introduction

2. Context: diversification of survey modes

For several years, Insee has been working on renovating its information system for the collection of business and individual surveys [1]. If the collection modes and protocols were previously simple at Insee, the survey collection information system must now respond to an increased move towards more multimode and more evolved and diversified protocols.

The development of multimode surveys requires a delicate and rigorous orchestration. This involves, among other things, constructing Internet questionnaires (CAWI) that are consistent with survey questionnaires (CAPI/CATI) or with paper questionnaires (PAPI). Emblematic Insee surveys (LFS, Housing, Resources and Living Conditions, etc.) have been or are being constructed to move towards this type of complex protocols.

3. Issues/challenges on protocols and process

- Data collection

Beyond the delicate statistical or methodological questions raised by the advent of such protocols, it is appropriate to consider, in light of the experience acquired, the operational complexity induced by multimode, imposing real technical challenges on the collection platforms for operating the questionnaires (GSBPM: from "Design/ Design Collection" to "Collect/Run collection").

Insee has recently renovated its collection information system (with the "Metallica" program) for surveys of individuals based on the concept of active metadata: a single questionnaire specification (expressed in DDI) generates several collection instruments (multimode) within the platforms. Several surveys by Internet and paper were thus operated in 2021 (for example the "Daily life and health" survey: VQS) and by Internet and telephone in 2022 (for example the "Housing" survey) as the program works on these collection platforms.

- Data integration

The data, once collected in different modes, needs to be processed to be integrated later in the statistical operation process. The Metallica program has therefore pursued technical investments to reconcile the data from the different modes (GSBPM: "Process/Integrate data") and start the first processing (GSBPM: "Process/Classify and code").

4. Methods

In order to implement the integration of multimode survey data, Insee uses the Validation and Transformation Language (VTL) proposed by the SDMX initiative.

5. Before VTL at Insee

Before the redesign of the tools in charge of metadata-driven survey collection and the use of formal specifications feeding the collection process (questionnaires, variables, processing), each survey had its own specific tool. Specifications were written by the survey designer (questionnaire model, dictionary of variables, initial processing) and then implemented (in Blaise) by teams of developers to build the collection instruments and a set of statistical processes (using SAS or R) to extract the data and ensure initial processing (data tabulation, multimode reconciliation, recoding, etc.). This work organization, which is still in place for some surveys before their migration to the new system, is quite costly because it requires developing and testing the chain from start to finish, including for a change in the questionnaire from one survey edition to the next. Complex survey protocols (panel, sequencing...) have processing chains developed in SAS that are quite complex and very difficult to maintain over time.

6. VTL - Validation and transformation language

VTL is a standard language [2] for defining validation and transformation rules (set of operators, their syntax and semantics) for various kinds of statistical data.

VTL is intended to be used by statisticians and is at the "business" rather than the technical level.

VTL processing rules are used and interpreted thanks to Java and JavaScript implementations (Trevas [3]) in the tools of the new Metallica collection system. The designer can directly write these rules and they can be directly integrated into the system when it comes to specifying the expected treatment on the data, specific to the survey without additional developments.

VTL is already used in the questionnaire design tool Pogues [4] to specify logical expressions within the questionnaire (conditional expressions, checks and filters).

7. Results

For the VQS survey, a questionnaire in paper format and a questionnaire in Internet format were proposed to almost 240,000 respondents using the same specification. These data had to be reconciled because the response formats were not exactly the same. For example, a single-choice question can be implemented in web format by a set of checkboxes and a control of the uniqueness of the answer (usually a radio button) but will be implemented in paper format by a set of checkboxes where it will not be possible to control the uniqueness of the answer. The reconciliation processing of data from several modes will then consist in specifying, for the paper response, what to do in cases where several boxes are ticked: retain none, the first, the one consistent with other responses, etc.

Beyond the specification of the questionnaire in DDI, a new type of metadata must be added to specify these processing phases, specific to the survey and the question.

8. Conclusions

9. Assessment of the solution

While the redesign of the collection information system has led to a great deal of work on standardizing processing, there are still a number of specificities to be taken into account for each survey. The use of the VTL processing language, dedicated to the designer and interoperable with the rest of the highly standardized system, has already made it possible to optimize the implementation and renovation of certain household surveys (all Insee household surveys will be migrated to this system within the next 3-4 years) while guaranteeing the specificities of each one. The VTL grammar makes it possible to cover the vast majority of needs in terms of post-collection processing specific to each survey, even in the case of complex protocols.

10. Prospect

The next step will be to further develop the concept within complex panel and mutlimode processes (e.g., the use of VTL rules for the post-collection processing necessary for re-collection or change of mode, including through the use of paradata) and to develop a tool dedicated to the designer's work: the simplified specification of VTL rules for post-collection processing in a working environment, integrated with the one that already exists for the specification of questionnaires (Pogues).

References

[1] E. Sigaud and B. Werquin, "La mise en musique d’enquêtes multimodes", Courrier des statistiques n°7 (2022) - (english version to come) https://www.insee.fr/fr/information/6035936?sommaire=6035950

[2] "Validation and Transformation Language (VTL)" on the official site for the SDMX community. A global initiative to improve Statistical Data and Metadata eXchange https://sdmx.org/?page_id=5096

[3] "Transformation engine and validator for statistics (Trevas)" on github.com https://github.com/InseeFr/Trevas

[4] F. Cotton and T. Dubois, "Pogues, a questionnaire design tool", Courrier des statistiques n°3 (2019) https://www.insee.fr/en/information/5014167?sommaire=5014796

Modelling local level housing demand based on the Multi-sectoral Regional Microsimulation Model (‘MikroSim’)

Keywords: Dynamic microsimulation, housing, policy evaluation

INTRODUCTION

Adequate and affordable housing as a basic need of society is an ongoing political and societal challenge. Due to its high influence on the quality of life and therefore on the living standard of the population, it affects large parts of the society. In Germany, an increasing number of people live in overcrowded dwellings; a considerable proportion of the population spends more than 40% of their disposable household income on housing. Housing accounts for around 37% of the consumption expenditure of average German household [1, 2, 3].

While housing scarcity and rising rents are observed in densely populated areas, the circumstances differ strongly at the local level: structurally weak regions are often characterised by high vacancy rates. Quantitative and qualitative housing demand strongly depend on demographic development, household composition, migration and regional mobility patterns [4, 5].

Housing is also the focus of political efforts, with the German government is currently aiming to build 400,000 new homes annually, announcing measures such as making construction faster and cheaper through serial and modular construction, building a quarter of these flats as social housing, as well as taking ecological aspects into account [6]. For the evaluation of such housing policy measures, detailed and high-quality data basis are required.

(Dynamic) microsimulation is a powerful tool for creating such a data basis, which has already proven its strength in a variety of applications. The microsimulation model ‘MikroSim’ is a dynamic discrete-time microsimulation model for Germany at the level of individuals and households starting in 2011 [7]. To model long-term housing demand, we build on the existing ‘MikroSim’ modules in which fundamental demographic events such as household formations and dissolutions, employment, income and regional mobility are simulated annually for the entire German population on a local scale. Basis of ‘MikroSim’ is a partially synthetically representation of Germany's population on a local scale [7]. The first phase of the implementation of the housing module focusses on the modelling of housing demand, while housing supply will be added to the simulation in a second step in order to subsequently model housing prices and rents. To this end we explore suitable data source to take the demand of different household types for different dwellings into account and present a modelling strategy. Apart from first results we discuss existing challenges as well as open questions.

METHODS

To extend the ‘MikroSim’ model by housing module, the synthetic population and the exiting households formed by it need to be allocated housing information. For this, a basic stock basic dataset must be created by given data about the distribution of different types of household to different types of dwellings on a local level. Information on this is obtained mainly from the German census and microcensus, which has a dedicated housing programme every four years with broad information on the housing characteristics of German households.

Given this initial distribution, the simulation of transitions can be be tackled in order to model the development of the housing situation in the future. The focus is on modelling transition probabilities for the event of a relocation. Relocations can either affect the entire household or only single individuals, such as the separation of (previously) cohabiting partners or the departure of a child from the parental home or the move into or out of a shared flat. In the cases mentioned, the relocation processes of single individuals are already modelled in other modules of ‘MikroSim’. Other types of relocation of parts of the household are conceivable, such as when family members who are not part of the nuclear family move into the household, children come back into the household or non-family persons such as au pairs are taken into the household. The modelling of the movement of persons and households between dwellings is therefore divided into 1. the probability of moving out or in at the individual level and 2. the probability of a relocation of the entire household.

We include two basic dynamics into the simulation; namely the demographic development and secondly behavioural developments on the housing market. These trends basically refers to all behavioural dynamics that take place independently of or in addition to demographic developments, such as the trend towards more single households. Although this work is focused on the demand side of housing, the spatially and qualitatively segmented structure of the housing market will be already taken into account. By only considering the demand side, we temporarily assume a completely elastic supply. However, the moving behaviour in our data will be based on past relation between demand and supply, which is why should be no relocations that are completely out of touch with the market. Data on the moving behaviour is currently not available in official statistics and is therefore obtained from longitudinal survey data.

Our model strategy uses a two-step approach at the level of houselholds, in which we first estimate the overall probability of relocation and then move on to modelling the concrete choice of residence. For how to model this choice, we use the work of Hansen et al. [8] as a reference, who have also used microsimulation to predict housing demand in Denmark. The model strategy follows a hierarchical approach, where variables are estimated sequentially but conditionally on the previous decisions, using the following parameters in the following order:

Region ($\mathrm{reg}_{\mathrm{D}}$) → Dwelling type ($\mathrm{type}_{\mathrm{D}}$) → Dwelling category ($\mathrm{cat}_{\mathrm{D}}$) → Dwelling size ($\mathrm{size}_{\mathrm{D}}$) → City size($\mathrm{city}_{\mathrm{D}}$) → Dwelling age .

To estimate combined probabilities would not be viable due to insufficient data coverage. Hence, we use a binary stochastic variable to model the relocation event of a household, depending on the characteristics of the household, with the result set a relocation (1) or no relocation (0). Given that a household is going to move, we estimate the (conditional) probability of a change of region. Given the change of region we estimate the specific region where the household is going to move. From there, we estimate the dwelling type (rent or owner-occupied), the household is going to live in, and so on:

$$P(X_H|\boldsymbol{Z}_{H,t}); X_H(\omega) = \{\begin{matrix} 0, if\ \omega = no\ relocation \\ 1, if\ \omega = relocation \end{matrix}\quad ; \omega \in \{0,1\}$$

$$P(\Delta reg_{DH,t}|\boldsymbol{Z}_{H,t}, X(\omega) = 1)$$

$$P(reg_{DH,t+1}|\boldsymbol{Z}_{H,t}, X(\omega) = 1, \Delta reg_{DH,t} = 1)$$

…

$$P(age_{DH,t+1}|\boldsymbol{Z}_{H,t}, X(\omega) = 1, reg_{DH,t+1}, type_{DH,t+1}, cat_{DH,t+1}, size_{DH,t+1}, city_{DH,t+1}),$$

with D describing the individual dwelling, t the time in years, H the individual Household, X_H the relocation event and $\boldsymbol{Z}_{H,t}$ being a matrix of household specific variables (like age of members, household size, family type, children, education, origin, and employment).[41]

RESULTS

To take the heterogeneity of dynamics of housing demand and supply in different areas into account when modelling the housing situation, both long term trends and currently most important impact variables on housing have been analysed. Figure 1 resumes some of the most important factors that have been identified as influential on the housing demand. Different regional patterns of housing demand occur largely due to internal and external migration and demographic developments. Moreover, due to the decreasing average household size, the number of households has increased significantly more than the size of the population in the last decades. The age effect and cohort effect show an increase of living space consumption over the course of life and between cohorts. First results of the approach outlined above show that the implementation of a housing module in the dynamic microsimulation model ‘MikroSim’ can be a valuable tool to predict these developments in detail and therefore create the basis for quantifying effects of policy measures.

[41] Due to lack of space, only the last one (probability of a certain age of the dwelling) is shown here, but the hierarchically organized variables are estimated in the same way. The expected value of this variable depends, as seen, on all previous decisions (like the size of the dwelling, its category, etc.).

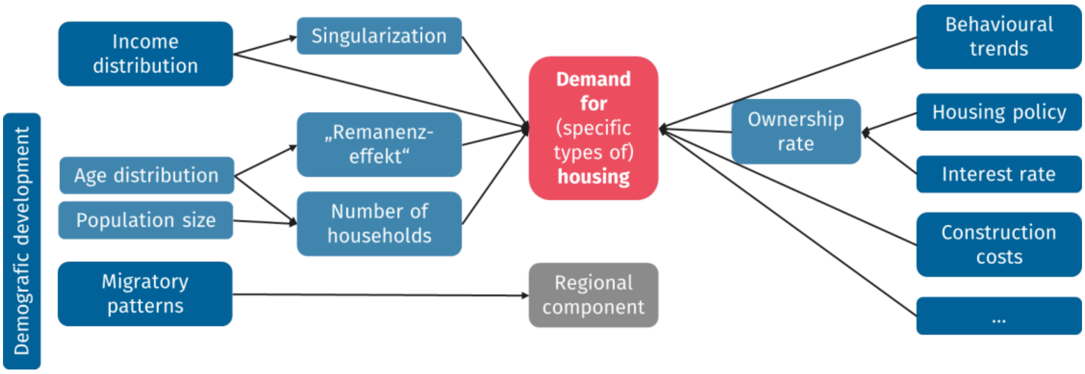


Figure 1. A causal diagram for determinants of housing demand

At this early stage of the project, numerical results of the estimated trends and dynamics as well as their visual representation still have to be waited for and be available in the coming weeks and months.

cONCLUSIONS

In this work we have made use of the georeferenced German Microcensus and its extended module on housing in 2018 as well as the German census 2011 for modelling the current distribution of types of households to different types of dwellings locally. In addition, we have explored suitable data source to model the future demand for different dwellings not only based on demographic developments but by the movement behaviour of households. The modelling strategy will be incorporated into a spatial model, as soon as computational performance will allow for it.

During the construction of the module, we are encountering numerous challenges, some of the being:

- Numerically rare or compositionally complex household forms are not differentiated in most data sets.[42]
- Missing official data on movements within regions at a spatially granular level lead to high errors at the local level.
- The use of panel data causes new questions such as how to deal with the circumstance that panel mortality overly affects moving households, when the data is used to model moving behaviour.
- Without control mechanisms, the estimations based on historical relocation behaviour lead to very strong population growth in urban agglomerations, which are not realistic with regard to the availability of building land and densification possibilities.

In future work, we therefore plan to add a simulation of housing supply to the module in order to tackle these difficulties and subsequently model housing prices and rents. On the basis of this

[42] From the German microcensus, shared flats can be defined on the basis of the information that there are several households in the flat, combined with some demographic restrictions. However, a definition like this is not compatible with other data sources like the census.

work other innovative questions may be explored supported by the use of new digital data, such as the feasibility of web scraping the lifespan of housing ads to proxy housing demand and housing scarcity.

rEFERENCES

[1] Statistisches Bundesamt, Destatis (2020): Press release. 14% of the population affected by excessive housing costs in 2019. Available at: https://www.destatis.de/EN/Press/2020/10/PE20_428_639.html (accessed: July 2022).

[2] Statistisches Bundesamt, Destatis (2021): Press release. 8.5 million people in Germany lived in overcrowded dwellings in 2020. Available at: https://www.destatis.de/EN/Press/2021/11/PE21_506_63.html (accessed: July 2022).

[3] Statistisches Bundesamt, Destatis (2022): Consumption expenditure. Housing accounts for more than a third (37%) of monthly household final consumption expenditure. Available at: https://www.destatis.de/EN/Themes/Society-Environment/IncomeConsumption-Living-Conditions/Consumption-Expenditure/current.html (accessed: July 2022).

[4] P. Deschermeier and R. Henger, Die Bedeutung des zukünftigen Kohorteneffekts auf den Wohnflächenkonsum. In: *IW-Trends-Vierteljahresschrift zur empirischen Wirtschaftsforschung* (2015), 42 (3), 23–39.

[5] Wolff and D. Rink, Strukturen von Wohnungsleerstand in Deutschland. Eine Analyse der Gebäude-und Wohnungszählung 2011 in deutschen Gemeinden. In: Raumforschung und Raumordnung| Spatial Research and Planning (2019), 77 (3), 273–290.

[6] Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (Ed.) Bündnis bezahlbarer Wohnraum. Maßnahmen für eine Bau-, Investitions- und Innovationsoffensive (2022) Berlin.

[7] R. Münnich, R. Schnell, H. Brenzel, H. Diekmann, S. Dräger, J. Emmenegger and others A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model. In: *methods, data, analyses* (2021), 15 (2), p. 24. Available at https://mda.gesis.org/index.php/mda/article/view/2021.03.

[8] J. Z. Hansen, P Stephensen and J.B. Kristensen, Modeling Household Formation and Housing Demand in Denmark using the Dynamic Microsimulation Model SMILE. Danish Rational Economic Agents Model, DREAM (2013).

Non-traditional data and methods for identifying patterns of inequality to digital access in the EU

Keywords: non-traditional data, spatial patterns, inequality, machine learning

INTRODUCTION

High-speed broadband represents a fundamental infrastructure for social and economic development. The lack of accessibility to a broadband network can reinforce existing disparities, acting as a barrier to social, technological, and market development. Such disparities among places and citizens can leave some areas behind in terms of access to services and opportunities, with direct consequences on the quality of life of residents (Proietti et al., 2022).

Access to broadband is particularly important for areas that are marginalised and disconnected from a territorial perspective, characterised by poor access to physical services (schools, hospitals), as the broadband infrastructure represents a potential alternative of accessing fundamental services and have a tangible impact on businesses and residents. However, the deployment of high- speed broadband across the EU appears to proceed slower than expected.

The EU sets specific targets for the 2020: 'access to 30 Mbps or above by all citizens and at least 50% of household with a connection over 100 Mbps' (Feijóo et al., 2018). Unfortunately, such targets have not been reached yet. Furthermore, the deployment of broadband infrastructure has not been even across regions. This connectivity gap, recognised as an urbanrural digital divide, represents an important challenge (DESI project, 2020). The focus of this analysis is to investigate the spatial accessibility of broadband and the quality of the connection in relation to the degree of urbanisation of the European territory.

METHODS

This work employs a dataset provided by Ookla® containing information about the performance of broadband network for the last quarter of 2020 in the 27 EU Member States. Each record includes attributes for fixed and mobile broadband networks. The analysis uses population figures for 2018 to assess the share of population, and the population density aggregated at municipal level by degree of urbanisation.

The analysis is performed at municipality (LAU) level by applying descriptive statistics and geospatial computational tools to analyse spatial patterns of broadband network across the EU. The network speed is classified into three categories: below 30 Mbps, between 30 and 100 Mbps, higher than 100 Mbps. To identify patterns of spatial disparities, a Machine Learning unsupervised technique is applied, using the average speed and latency to establish the presence of a reliable connection, population density as proxy for density of urbanisation; remoteness classification as proxy of distance from major urban centres.

RESULTS

Results show that for fixed broadband most the population in all countries has access to speeds above 30 Mbps and often above 100 Mbps (except for Greece). For mobile broadband, most residents in each country have access to speed up to 100 Mbps, with a higher percentage of citizens only accessing speeds under 30 Mbps.

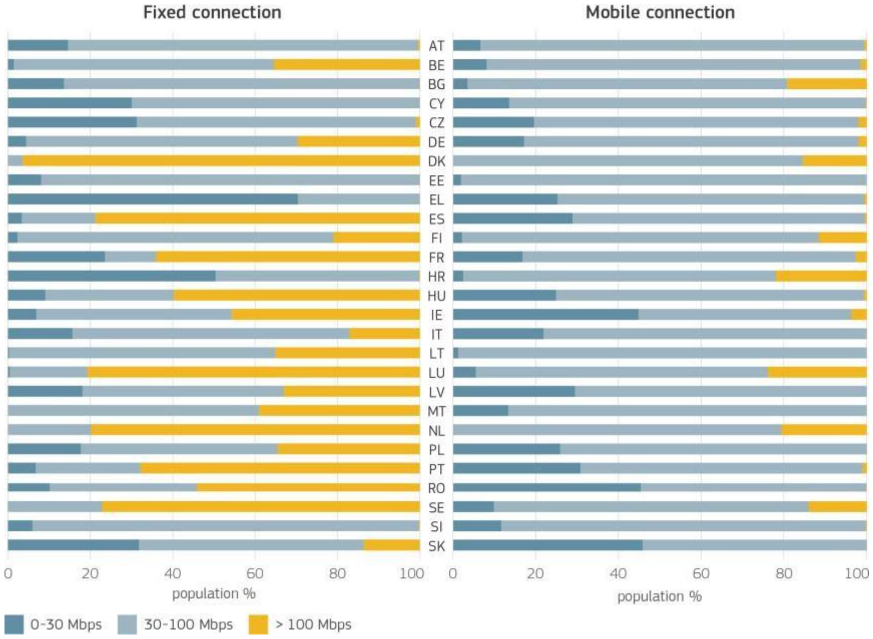


Figure 1. Share of population with access to broadband at country level.

The highest values in fixed broadband are in northern Europe, with Denmark and the Netherlands presenting a homogeneous access to high speed connection. Croatia and Greece show the opposite trend, with a low speed connection across the whole country. In France and Spain, the situation is heterogeneous, with significant differences across the territory and a fast average speed in the major cities. Regarding mobile broadband, the average speed is lower than fixed broadband, with a few areas having more than 100 Mbps. Ireland and Romania show poor access to broadband connection, whereas the Benelux region presents a homogeneous picture.

A neat digital divide is noticeable between urban and non-urban areas. The urban population has good access to speed over 30 Mbps, and a relevant percentage also to 100 Mbps. In rural areas, many residents have access only to speed below 30 Mbps. These results are confirmed overlapping the spatial patterns of average speed and the classification of municipalities according to degree of urbanisation (see Figure 2 below). Urban areas present the highest speed in broadband connection, revealing how areas already connected in terms of physical networks (i.e., with roads and railways) are also the most connected from the digital point of view.

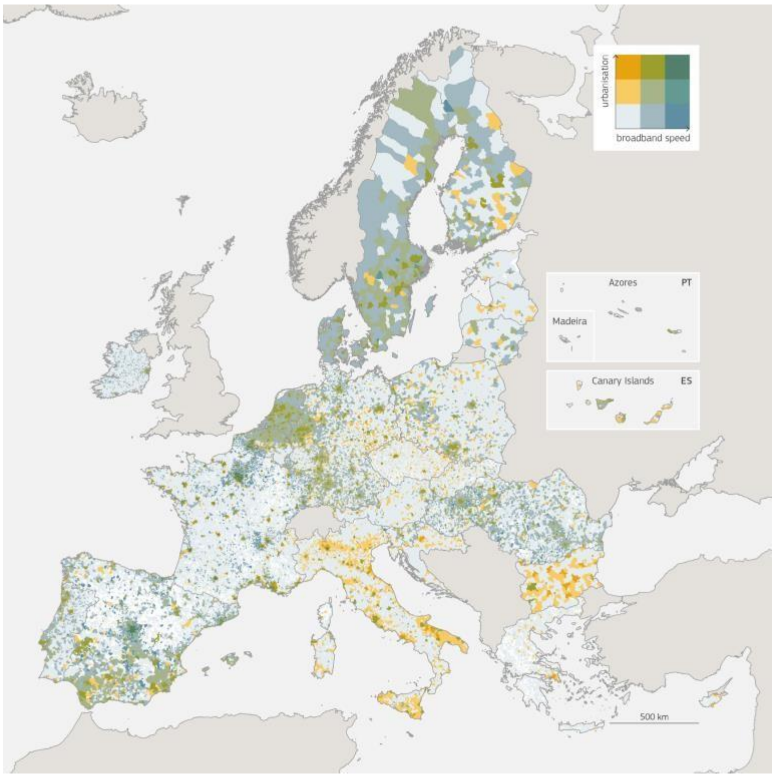
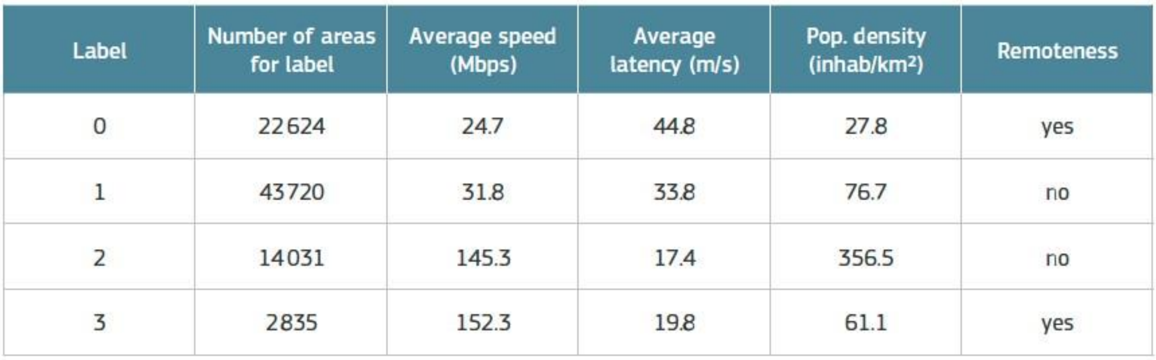


Figure 2. Urban-rural digital divide: comparison urbanisation – broadband quality

To identify the most vulnerable and marginalised areas in terms of scarce access to digital connection, more factors are added to perform a cluster analysis on all municipalities, regardless of the degree of urbanisation. The cluster analysis assigned data to different labels according to: average speed, average latency, population density, remoteness classification (see Table 1).

Table 1. Results of clustering analysis



Label	Number of areas for label	Average speed (Mbps)	Average latency (m/s)	Pop. density (inhab/km²)	Remoteness
0	22624	24.7	44.8	27.8	yes
1	43720	31.8	33.8	76.7	no
2	14031	145.3	17.4	356.5	no
3	2835	152.3	19.8	61.1	yes

Results show that the most disconnected areas (cluster 0) present the following characteristics:

- very low speed connection (average speed below 30 Mbps);

- high latency (slow responsive connection);

- low population density (scarcely populated rural areas);

- classified as remote (places far from major urban centres).

cONCLUSIONS

The poor access to high-speed broadband might leave some areas behind. Unveiling spatial patterns of access to broadband network is critical to inform policy with quantitative evidence. Overall, results confirmed existing disparities highlighting the spatial patterns of European areas that do not have access to high-speed connection. These areas experience the worst conditions in terms of physical and digital accessibility, directly affecting the quality of life of citizens residing in such disconnected places.

rEFERENCES

[1] Proietti, P., Sulis, P., Perpiña Castillo, C., Lavalle, C., Aurambout, JP., Batista e Silva, F., Bosco, C., Fioretti, C., Guzzo, F., Jacobs-Crisioni, C., Kompil, M., Kučas, A., Pertoldi, M., Rainoldi, A., Scipioni, M., Siragusa, A., Tintori, G., Woolford, J., *New perspectives on territorial disparities. From lonely places to places of opportunities* (2022), Proietti, P., Sulis, P., Perpiña Castillo, C., Lavalle, C. (eds), EUR 31025 EN, Publications Office of the European Union, Luxembourg, doi:10.2760/847996, JRC126033.

[2] Feijóo, C., Ramos, S., Armuña, C., Arenal, A., & Gómez-Barroso, J.-L, A study on the deployment of high-speed broadband networks in NUTS3 regions within the framework of digital agenda for Europe, *Telecommunications Policy* 42(9) (2018), 682-699.

[3] DESI Project, Digitisation: Economic and Social Impacts in Rural Areas (2020), accessible from: *https://desira2020.eu/the-project*

Natural Language Processing to automate data coding

Keywords: Machine Learning, Natural Language Processing, Coding, Occupation, Economic Activity, Multinomial Naïve Bayes Classifier, Random Forest Classifier, Support Vector Machine, Neural Networks

Introduction

It is estimated that in the Census of Population of 2021 around 370,000 people have been recorded for whom their occupation and economic activity has been declared. Traditionally, the coding of descriptions is done by trained coders. Due to the large number of descriptions that needs to be handled, the process of coding is time-consuming and requires a significant number of human resources and thus, leading to high costs. As a result, the coding process of occupations and economic activities in each census is determined based on the available budget and human resources.

For the Census of Population 2021, it was decided to utilize modern technology, specifically Machine Learning. In this context, a system was developed using NLP algorithms, which help accomplishing a high success rate of automatic correct coding. The coding process is very fast as it takes only 10 seconds to code 50,000 descriptions. From all the initial checks carried out so far, the success rates are very satisfactory.

Methods

In order to be able to implement this project, the use of Natural Language Processing (NLP) techniques took place. NLP is a branch of Data Science which deals with text data. Apart from numerical data, text data is available which is used to analyse and solve business problems. But before using the data for analysis or prediction, processing the data is important [1].

To prepare the text data for the model building we perform text pre-processing. It is the very first step of NLP projects. Some of the pre-processing steps carried out are the following: Remove punctuations like . , ! $ ( ) * % @; Remove URLs; Remove Stop words; Lower casing; Tokenization; Stemming; Lemmatization [2],[3].

The implementation considered the evaluation of four algorithms as follows: (a) The Multinomial Naïve Bayes Classifier, (b) Random Forest Classifier, (c) Support Vector Machine and (d) Neural Networks.

Multinomial Naïve Bayes Classifier

Naive Bayes methods in the field of Machine Learning are a set of supervised learning algorithms based on applying Bayes' theorem with the strong assumption of conditional independence between the features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_i through x_n.

Naive Bayes learners and classifiers are usually extremely fast compared to more advanced algorithms since they require very few calculations to predict the target variable, but this does not make this algorithm less powerful. This algorithm is also less susceptible to problems stemming from the curse of dimensionality since each distribution can be independently estimated as a one-dimensional distribution.

Random Forest Classifier

A Random Forest algorithm is a meta estimator (an estimator which takes another estimator as a parameter) that trains a number of decision tree classifiers on different sub-samples of the dataset and uses averaging to select the final predictive outcome. This combination of multiple decision trees improves the predictive accuracy of the model and handles overfitting much better than a single decision tree. The number of decision trees, and their parameters, and the number of sub samples that are used in the training is controlled by the parameters of the random forest classifier.

Decision Trees are a non-parametric supervised learning method used for classification and regression. When a testing row of input data goes into the tree, based on the values of that row, it will lead into the corresponding branch of the tree and the response will be the predicted value for that row.

Support Vector Machine

In Machine Learning, Support Vector machines are used as supervised learning models for both classification and regression problems. The way this algorithm works is by constructing a hyperplane which can be used for the type of problem selected, and even outlier detection. This model separates the data by the hyperplane that has the largest distance to the nearest training-data point of any class, this separation is called functional margin since the greatest the margin of separation is the lower the generalization error of the prediction. The problem that is usually encountered in real life problems, is that the data is not easily separable or not even possible to be separated in a linear space, so this algorithm solves this problem by mapping the data into a higher dimensional space where the data could be separable. This mapping of the data to the higher dimensional space happens with the help of various kernels.

In premade SVM classifier algorithms, the kernel selection is very limited and for this reason we implemented a new SVM algorithm from scratch, using a premade SVM classifier that takes data after the kernel process is performed and combining it with custom made kernels, creating a new classifier that has multiple kernels to be tested with. In our case, we tested all of these kernels and based on the results the best performing one was the log kernel, so we continue using the SVM algorithm with the logarithmic kernel.

Neural Networks

Neural Networks is a collection of connected units or nodes called Neurons, each of them transmits information to all of the neurons in the next layer after processing the information received from all the neurons in the previous layers. A Neural Network consists of an Input Layer and an Output Layer and in between those two layers are the so-called Hidden Layers which can be as many as the constructor of the architecture of the Network decides that fits the purpose of the problem the best. Further to those connections between the neurons where the information is transmitted in the form of “Weights” between the neurons, there are also activation functions selected based on the type of input and output of the problem. Finally, the Neural Network makes use of a loss function and backpropagation to update its weights and ultimately improve its predictions by minimizing the loss function.

In our case the input of the neural network was the vector of the description of the occupation/ economic activity created by the TFIDF vectorizer, and the output of the network was the probability of the description belonging to each class in a one-hot-encoding form. The hidden layers of the network were two dense layers with the first having a Rectified Linear Unit (ReLu) activation function and the second dense layer a Sigmoid activation function. The loss function used was a categorical cross-entropy function provided by TensorFlow. The Neural Network was trained for 50 epochs with validation splits that can inform us about the performance of the learning for each epoch, using batch sizes of 32.

Results

After implementing the models, we tested a variation of different combination of parameters for our models and selected the best fitting parameters where the performance was higher without the risk of overfitting to the training data. Running the example above for the Random Forest Algorithm, SVM model and Neural Network model, we get the results shown in Figure 1.

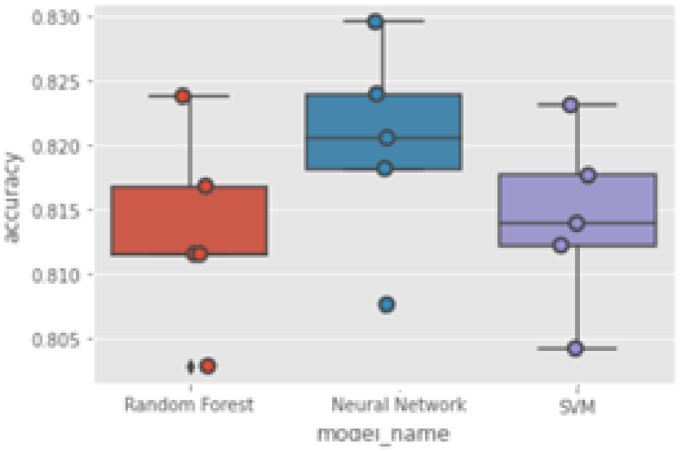


Figure 11: Comparison of the three algorithms

As shown in the figure depicted above, it is clear that the Neural Network (LSTM) algorithm outperforms the other two algorithms since its accuracy is better in each split and thus, its average accuracy is also the highest.

In addition to all the previous tests carried out during the development phase it was decided to test the system with 100 occupation descriptions and 100 economic activity descriptions extracted from the Census database. At first stage the descriptions were coded by the NLP system and then, at second stage, the same descriptions with the corresponding codes from first stage were loaded into another system developed with Blaise in which two experienced coders confirmed or corrected the codes. The coding process followed by the two coders was "double blind", i.e., each coder was not aware of the code given by the other coder. The coding of occupation and economic activity are based on the ISCO-08 (4-digit) [4] and NACE Rev2 (5-digit) [5] classification systems, respectively.

As it turned out, the success rates of automatic coding were very satisfactory at all coding levels (e.g., 5-digit, 4-digit etc). In particular, as regards NACE Rev. 2 at the 5th digit level the average success rate was 88% and at the letter level 92%. For ISCO-08 the average success rate at the 4th digit level was 81% and at the 1st digit level 90%.

Conclusions

The results of all the tests carried out so far are very encouraging, however, in order to decide whether to use the NLP system for the needs of the census and for all the statistics production processes in general, more testing needs to be carried out. Moreover, one major conclusion drawn up to now is that the system works very well when the descriptions are good. In the cases where the descriptions provided are meaningless then the system although it provides suggestions with lower prediction rates it cannot, as expected, provide reliable suggestions. The interface developed for the NLP system is user friendly and very easy to interact with and it offers two options for coding. The first option is for bulk coding which is implemented by uploading a csv file in a standardized format whereas the second is for single cases where the coding is carried out by entering manually the description. A major advantage of the NLP system concerns its capability to continuously improve the predictions by feeding into the system additional cases, i.e., descriptions with correct codes given by coders.

As regards to the census data, based on the preliminary test results it has been decided, for the time being, not to proceed with recruitment of coders for manual coding but to produce the codes with the NLP system. Then a representative sample of descriptions will be drawn and automatically coded by the NLP system and the results will be compared with manual coding from experienced coders. Moreover, certain occupation and economic categories that are known beforehand to be problematic in coding will be examined in more detail. In addition, the distribution of codes in the various occupation and economic activity groups will be compared with other sources such as the business register, the labour force survey and the structural business statistics surveys.

Finally, the prediction rates provided by the NLP system will be analysed in order to assess whether a threshold can be defined for the accuracy of the results, i.e., all codes provided with a certain prediction rate (the threshold) and up will be automatically considered as correct.

References

[1] Manning, C. D. & Schütze H., (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press

[2] Brill, E.: Transformation-based error driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21(4), 543–566 (1995)

[3] 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, Cardiff, UK, June 23-25, 2010, Proceedings

[4] International Labour Organization, International Standard Classification of Occupations 2008 (2012)

[5] Eurostat, NACE Rev. 2: Statistical classification of economic activities in the European Community (2008)

Maritime mobility statistics using open data

Keywords: Big-data, AIS, statistics, vessels, maritime

1. INTRODUCTION

In recent years we are witnessing an increased number of experimental studies on the use of alternative data sources by statistical offices to assess their use in the production of official statistics. In this context, Eurostat has launched several initiatives for the development of IT, methodological and quality frameworks, tools and infrastructure suitable for Big Data processing. In this work we present some results of the research undertaken within Eurostat project on Transport and Mobility Innovation in the area of maritime transports. The aim of the project is to detect new/non-traditional data sources for the maritime transport, together with the identification and description of methodologies and tools suitable for the implementation of the new data sources in official statistics. In the search for Big Data, priority was given to data sources open or free, available on a European scale, frequently updated, stable over time, relevant and coherent with official statistics.

In the area of maritime transports, the source that seems to be most comprehensive for this purpose is provided by the AIS, Automatic Identification System. Nowadays, AIS data are used to produce a big amount of information covering different domains. For official statistics, AIS data could be consider a pillar source into the statistical system related to maritime statistics. The research on the use of AIS for official statistics has been the target of different projects and publications introducing significant improvements for the production of statistics on maritime transport, but not only. Among all it is important to mention the results obtained within the ESSnet Big Data ([1], [2]) The results obtained by ESSnet on Big data regarding the statistics on port visits seems to be really promising, however the size and the complexity of the data may be difficult to handle by the NSI. Using port calls close to the port area reduces the size and complexity of the data while still keeping track on incoming and outgoing vessels with a better AIS-receiver coverage. Then this data can be easily treated to produce different indicators that range from indicators for official statistics and experimental one like now-casting economic indicators ([3], [4]).

Currently, there is a growing number of private web-services offering data access to ship tracking using the automatic identification system to display the real-time location of the ships and to obtain historical information. The main Vessel Tracking Websites are MarineTraffic, VesselFinder, FleetMon, Shipfinder, Vesseltracker, Cruisemapper. Unfortunately, the access to raw AIS signals is limited; many services are free of cost, while more advanced functions come with a fee, which is a limitation for the project. As far as we know, AISHub is the only private web-services which provides totally free access to real time ship positions, through custom APIs, only to active users sharing their own AIS feed.

2. METHODS

2.1. AISHub data

The International Convention for the Safety of Life at Sea (SOLAS) of the International Maritime Organization (IMO) requires all ships over 300 gross tonnages (GT), and all passengers' ships, to have on board a VHF radio transponder that constantly broadcasts the ship's position, speed and other dynamic, static and voyage related information. Radio AIS signals are sent directly to land-based radio stations located along coastlines within a distance of about 40 nautical miles, or through satellite AIS transceivers when the ship is in open sea. Port authorities or other public bodies dealing with maritime safety at sea can receive AIS information and view the vessel traffic condition.

As mentioned, AISHub is a web service that provides active users with access to real-time AIS data globally. Active users are those who install an AIS receiver with a VHF antenna at their location and send the received signals to the AISHub server through a dedicated TCP port. They offer the possibility to use a simple app, AIS Dispatcher, running on Linux or Raspberry PI devices, to assist in the data forwarding. Users providing active data sources receive credentials to send data access requests via the system's API. It is possible to download the entire real-time dataset in json, xml or csv format. In requests, it is possible to filter the data by ship identifier (IMO or MMSI) and by geographic area, defining minimum and maximum latitude and longitude. Unfortunately, AISHub does not publish historical data that are sold by other providers. However, it is possible to automate the download of the dataset for the area of interest with a time frequency that depends on the purpose.

For the goals of the project, an AIS receiver was installed in a coastal area of central Italy and starting from the end of August we are collecting AISHub messages every 15 minutes. In order to reduce the data dimension, we consider only the areas related to the Port of Lisbon. Within the identified port boundaries, we consider all the AIS messages regardless the reported speed and the navigational status. A Java programs have been developed for the acquisition of AIS data through the APIs offered by the site https://data.aishub.net/ and using a PRIVATE_KEY. The downloaded data, after a parsing activity, are then stored into an Open Source (MariaDB) database.

The information collected using the AISHub API are not sufficient to identify the vessels in the scope and to produce the proposed indicators. To integrate AISHub data we carried out a web scraping activities from the Vessel Tracking Websites VesselFinder using a Java program that allows to link the information provided by the AIS signal with the specific information of the ship through the MMSI identification code. The data scraped are: IMO identification code, name of the vessel, type of vessels, Gross Tonnage, Deadweight, the length and the width.

2.2. Indicators from ASI data

Maritime activities are important measures of real economic and have a relevant impact on the overall macroeconomic indicators during the same reference periods [5]. Moreover, it is always more important to have measures of economic activity available at higher frequencies to support policy and to inform business communities. AIS information together with other source of data can be used to produce different indicators related to the Maritime activities. Possible indicators are:

- Vessel traffic: number of vessels in port. Data refer to the activity of the port during a defined time interval that can be week, month or quarter. Quarterly data correspond to the official tables F1 and F2 described in the Directive 2009/42/EC of the European Parliament and of the Council of 6 May 2009.

- Time-in-port indicator: Total time vessels spend in port for the load and unload activities by port, broken down by type of vessel. Time in port is an indicator related to the loaded and unloaded of goods at each port call. It can be argued that the greater the number of goods moved the longer is the time spent at the port.

- Cargo-load indicator: Total load of vessels measured on the deadweight tonnage and the draught reported in AIS messages. Cargo load indicators is also an indicator related to the loaded and unloaded weight of goods at each port call in time period T in the port.

To derive the inwards and outwards port calls we consider the variables Navigation status according to AIS specification declared into the message and the difference in speed, time and the distance of position occurring between two AIS message of the same vessels (identified by MMSI). A port call is counted as new arrival when the vessel enters the selected area, reduce the speed and move toward a terminal. The Arrival time is the one corresponding to the first message with the Navigation status equal to “moored” or “anchor”. A port call is counted as departure when the navigation status changes from moored to “under way” and the vessel speed increases. The Departure time is one corresponding to the first message with the Navigation status equal “under way”. A new arrival of the same vessel is counted when the vessel approaches once more the port and it exists a reasonable difference in time with respect to the previous AIS message. The port calls in scope are then identified according to the type of vessel and the anchored zone. The time in port variable is defined as difference between the arrival and departure times. If the ship is at anchor and waiting for a berth is not counted in these time interval since we recorded only the anchored or moored vessels outside the anchorage area. Finally, we derive a measure of a load of the ship as the product between the dead weight of the vessel and the relative variation of the draught in the port call.

We evaluate the coverage of port calls derived from the AISHub using as benchmark the data from the port authority of Lisbon. We link the two dataset using as key variables the IMO code date of arrival/departure. The coverage analysis shows that the two data sources are coherent both in terms of number of vessels (table 1), type, gross tonnage and deadweight.

Table 1 Cross classification of number of port call detected by AISHub data and Port authority by type of vessel (inward declarations). September 2022

	AISHub data					
PA data	Liquid bulk	Dry bulk	Container	General cargo	Missing	Total
Liquid bulk	16	0	0	0	0	16
Dry bulk	0	11	0	0	0	11
Container	0	0	44	0	5	49
General cargo	0	0	0	33	1	34
Missing	5	4	2	1	0	12
Total	21	15	46	34	6	122

3. RESULTS.

The Time-in-port indicator and Cargo-load indicator have been evaluated based on AIS data from the port of Lisbon. However, in order to have longer data series than those available with AISHub data, the data has been integrated with the one from the port authorities, available since 2014. Both indicators are derived on a monthly basis and the series are transformed to represent month percentage changes. For all series we apply a seasonal adjustment.

The relevance of the proposed indicators can be tested by associating them both with the tonnage of goods carried and with the international trade volume during the same period. Figure 1 shows the comparison of monthly percentage changes of vessels total time in port obtained from the port authority data with the monthly results of imports of Portugal published by National Statistical Institute of Portugal from 2018 until 2022.

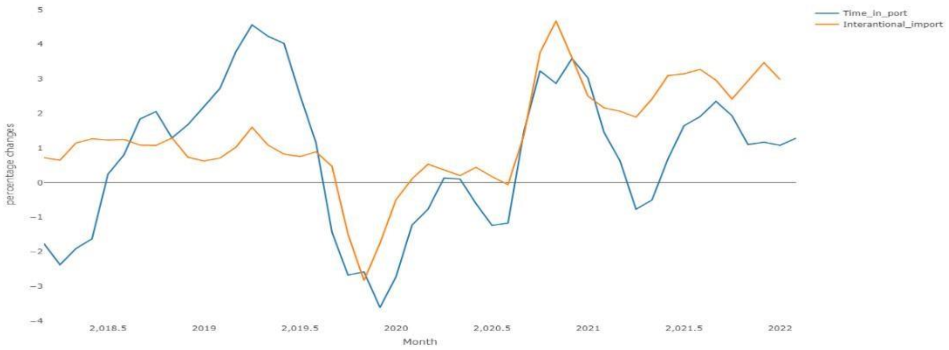


Figure 1 - Monthly percentage changes of Vessels total time in port of Lisbon (PA data) vs International Import Portugal (Ine data) - Seasonally Adjusted Data

The proposed indicator appears to follow closely Portugal’s import volume, even if we are comparing Lisbon results with the national trade volume. In particular, the negative peak at the end of 2019 and the positive peak at the beginning of 2021 in trade activity are quite well detected by the time in port indicator.

4. cONCLUSIONS

This exercise shows that AIS signals can be relevant for the production of maritime statistics. The data acquisition method via API is fast and it is possible to link AIS data to other datasets by foreign key or by spatial relationship, such as proximity, inclusion, and connection. Statistics analysis shows an excellent match with data produced by the port authority or by Eurostat. Finally, real time indicators, can potentially be used as a nowcast predictor of real imports or Gross weight of goods.

rEFERENCES

[1] ESSnet Big Data II, Deliverables Work Package E Tracking ships, (2020).

[2] ESSnet Big Data I, Deliverables WP4 AIS data, (2018).

[3] S.Arslanalp, M.Marini, P. Tumbarello, Big Data on Vessel Traffic: Nowcasting Trade Flows in Real Time, (2019), IMF Working Paper 19/275.

[4] Cerdeiro, Komaromi, Liu and Saeed, World Seaborne Trade in Real Time: A Proof of Concept for Building AIS-based Nowcasts from Scratch, (2020), IMF Working Paper 20/57

[5] International Maritime Organization. Introduction to International Maritime Organization, (2019).

Statistical Data Triplification : The case of semi-automatic generation of RDF triples from relational databases

Keywords: LOSD, triplification, RDBMS.

Introduction

Decision-making on critical issues by the executives of organizations and businesses is based on data and information analysis in order to estimate specific dimensions and examine statistical hypotheses. Usually the necessary raw data are either stored in relational databases with limited access or in the form of open data in special repositories. In any case, stakeholders must search the Internet for the necessary data and combine information from different sources and systems that usually do not interoperate, in order to produce the necessary information and knowledge. At the same time, the interest of the scientific community is focused on the further development of the Web and its evolution into the so-called semantic web. The semantic web is expected to contribute to the more intelligent access and management of the information traded on the internet through new technologies and the development of corresponding new applications. The overall vision concerns the transition from the existing Internet of static pages to a network of dynamic service providers as web services that automatically discover the information sought, taking into account the semantics of the concepts mentioned by the user. Thus the interest turns to the development of tools and methods based on the development and management of common vocabularies and new standards for the representation of information and generated knowledge.

Subject of this work

In this paper we examine the possibility of exploiting the raw data stored in existing relational databases with the help of semantic web technologies for the benefit of the production of Linked Open Statistical Data. Our goal was to semi-automatically generate RDF triples and develop ontologies based on a common vocabulary to describe the domain of official statistics and then feed the knowledge base from existing data retrieved from relational databases.

The use of existing databases is particularly important, since considerable resources have been invested so far and a huge amount of digitized data stored in relational databases has been developed. Thus, their eventual replacement by other technologies may prove uneconomical in terms of time and money.

Motivation

This work was motivated by our research in the field of generating personalized statistical products from existing information systems using semantic web technologies. Our interest is mainly focused on the utilization of the data held in the existing relational databases for the extraction of information based on dynamic queries. Existing database systems that support the management of public information for the production of statistical products use relational databases. Searches in such databases are usually based on standardized and inflexible and static SQL queries which are usually submitted to the database by the user and are built in advance. Customizing and adapting them to new needs usually requires the assistance of database administrators/developers. The whole process is characterized as time-consuming and uneconomical.

The situation we described above can be addressed with semantic web technologies where information is encoded in the form of RDF triples and can be accessed by software based on SparqL queries [1]. Thus, intelligent and dynamic queries can be implemented that search for information based on specific properties and rules, rather than simple data checks performed by traditional relational databases.

Methods & Tools

Field of Interest

In this paper, we have examined the case of the statistics produced by the Citizen Register - National Municipal Registry maintained by the Ministry of the Interior and accessed by the country's municipalities through the registries. According to the architecture of this system, a central database is maintained at the Ministry of the Interior as well as individual databases in the municipalities, which, under their responsibility, synchronize the local databases with the central database maintained at the Ministry of the Interior. The central hub of the Ministry of the Interior provides all Web Services to the public bodies that call its services through a web site installed on the portal of the Ministry of the Interior. Under the responsibility of the Ministry of the Interior, specific statistics are produced from the raw data stored in this system.

The experience, so far, from the databases, points out problems regarding the observance of a common vocabulary and rules, when managing the records in the local databases. Thus, for example, problems are reported during the search in cases of double surnames or double names of citizens, which is a very common phenomenon. The use of sometimes Latin and sometimes Greek characters or the possible abbreviations used in the names of citizens also creates a problem. Another issue is related to the search for possible relationships (such as uncle-nephew or grandfather-grandson) between members belonging to different family registers. The relevant search is based on the registrations and deletions from the family registers in which, however, no family relationship is recorded except that of spouses and children. Thus, parallel tables must be maintained to track changes in the municipal registry resulting in the storage of redundant or duplicate information. These problems constitute a risk of malfunctioning of the existing system and therefore a risk for the quality of the statistics produced. On the contrary, as are going to see below, the process of finding direct relatives of any degree, with the help of the ontology, is handled with simple descriptive questions on the knowledge base maintained by the ontology [1].

Tools

The number of Semantic Web tools is constantly expanding, while there is already a particular interest in the development of technologies for converting data from traditional formats to formats that can be handled by Semantic Web technologies. These technologies are mainly based on specific programming languages such as Java, Python, OWL and SparqL, while a number of standards such as RDF, RDF Schema, XML, HTTP URI's and tools such as CKAN, D2RQ

[2] etc. they support the management and linking of open data and knowledge modeling through ontologies. In the context of this work, we reviewed published works on the conversion of data from relational schemas to RDF triples format, which are mentioned in the relevant section of the work, while the extended version of this work provides an extensive presentation of the tools we used.

Methods – Basic steps

The most systematic and efficient processing of heterogeneous data can be achieved when these data can be transformed into the same format, but without losing information. In addition, it is desirable to assign semantics to the various data and then link them together. In this way, the interoperability of the system data is increased. An interesting and particularly easy-to-use data format that serves the above is formatting in RDF triples. Of particular interest is the automated generation of RDF files from various formats such as .XLS, .CSV and .JSON formats but also from tabular data or data stored in relational databases. The architecture of the system - methodology that we followed in this work can be seen in the image below [fig. 1a] while the basic steps of the process are analyzed below.

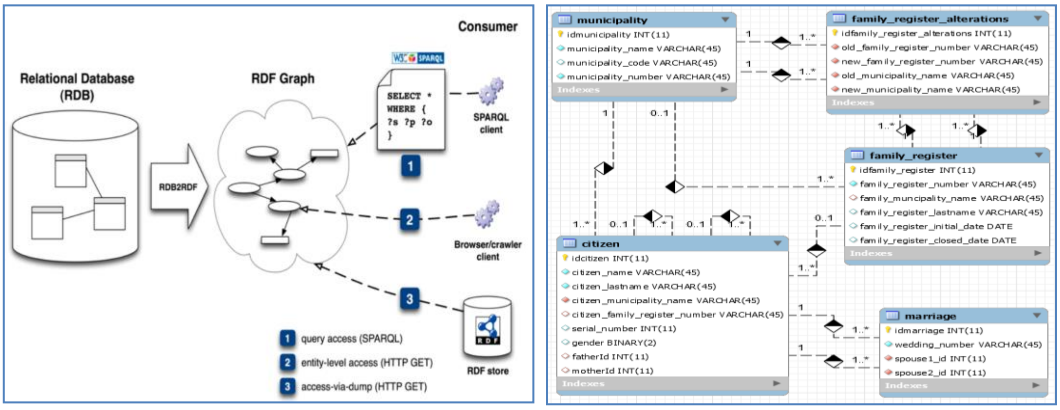
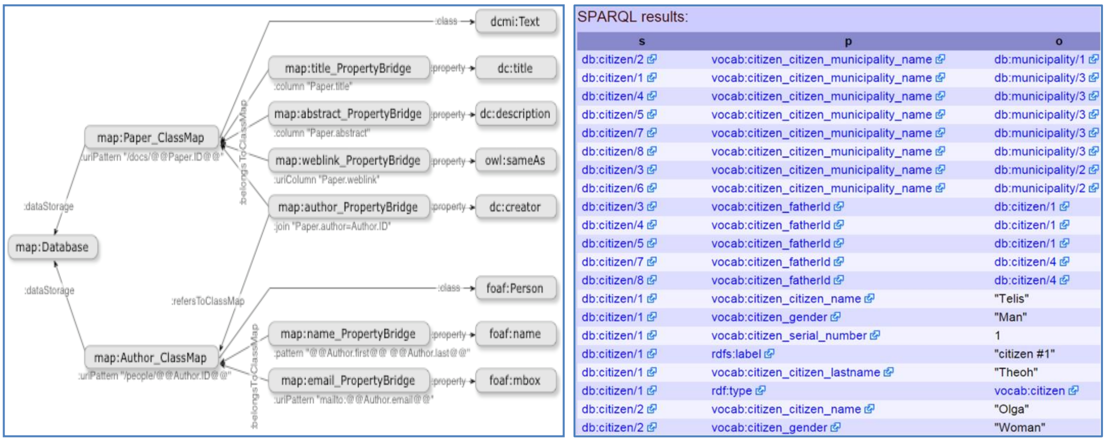

Consumer
Relational Database
(RDB)
RDF Graph
RDB2RDF
W3C SPARQL
SELECT *
WHERE {
?s ?p ?o
}
SPARQL
client
1
2
3
Browser/crawler
client
RDF
RDF store
1 query access (SPARQL)
2 entity-level access (HTTP GET)
3 access-via-dump (HTTP GET)
municipality
idmunicipality INT(11)
municipality_name VARCHAR(45)
municipality_code VARCHAR(45)
municipality_number VARCHAR(45)
Indexes
family_register_alterations
idfamily_register_alterations INT(11)
old_family_register_number VARCHAR(45)
new_family_register_number VARCHAR(45)
old_municipality_name VARCHAR(45)
new_municipality_name VARCHAR(45)
Indexes
family_register
idfamily_register INT(11)
family_register_number VARCHAR(45)
family_muncipality_name VARCHAR(45)
family_register_lastname VARCHAR(45)
family_register_initial_date DATE
family_register_closed_date DATE
Indexes
citizen
idcitizen INT(11)
citizen_name VARCHAR(45)
citizen_lastname VARCHAR(45)
citizen_municipality_name VARCHAR(45)
citizen_family_register_number VARCHAR(45)
serial_number INT(11)
gender BINARY(2)
fatherId INT(11)
motherId INT(11)
Indexes
marriage
idmarriage INT(11)
wedding_number VARCHAR(45)
spouse1_id INT(11)
spouse2_id INT(11)
Indexes
1
1..*
0..1

Figure 1a & 1b:The architecture of our system & the E-R diagram of the our Dbase

Below we outline the steps to first "build" a simulation of the Home Office database, the corresponding ontology, and then transform the relational data into semantically linked data[1].
1. Creating the relational database. In the MySQL environment we created a simulation of the national census database with a model containing the census core elements, so that we could evaluate the process of conversion to RDF triples. The basic elements were grouped into corresponding tables which are connected with the appropriate relationships, as shown in the image above [fig. 1b]
2. Triplification using the D2R server[3]. To retrieve our relational database data and convert it into RDF triples, we used the D2RQ-0.8.1 package, which works directly with MySQL data and provides, among other tools, generate-mapping and d2r - server. The generate-mapping tool creates a D2RQ mapping file [fig. 2a] by analyzing the schema of an existing database. This mapping file can be used as-is or can be customized. The results of the process are obtained after a related SparqL query and can be seen in the image below [fig. 2b].



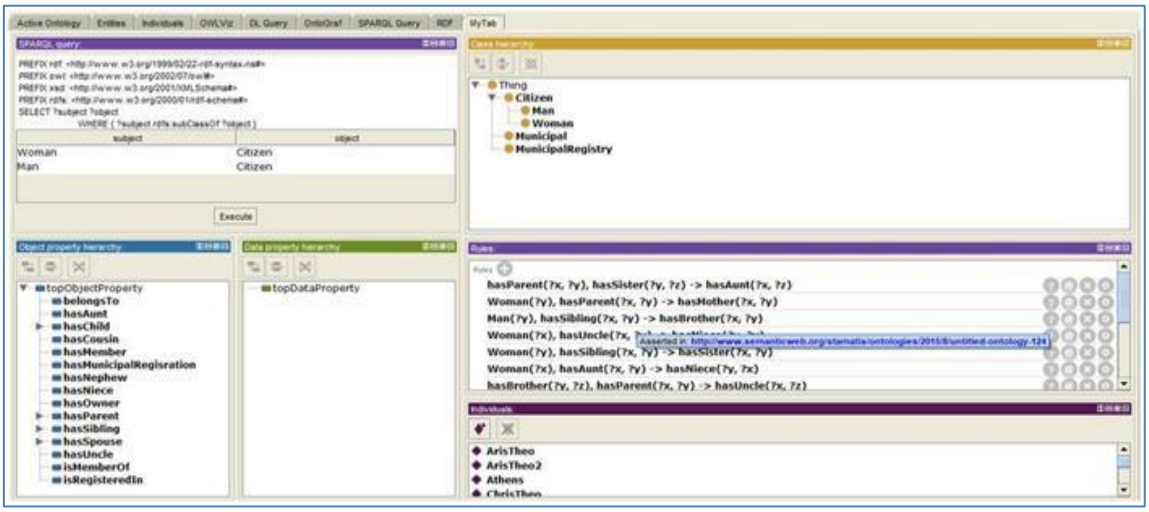
SPARQL results:

s	p	o
db:citizen/2	vocab:citizen_citizen_municipality_name	db:municipality/1
db:citizen/1	vocab:citizen_citizen_municipality_name	db:municipality/3
db:citizen/4	vocab:citizen_citizen_municipality_name	db:municipality/3
db:citizen/5	vocab:citizen_citizen_municipality_name	db:municipality/3
db:citizen/7	vocab:citizen_citizen_municipality_name	db:municipality/3
db:citizen/8	vocab:citizen_citizen_municipality_name	db:municipality/3
db:citizen/3	vocab:citizen_citizen_municipality_name	db:municipality/2
db:citizen/6	vocab:citizen_citizen_municipality_name	db:municipality/2
db:citizen/3	vocab:citizen_fatherId	db:citizen/1
db:citizen/4	vocab:citizen_fatherId	db:citizen/1
db:citizen/5	vocab:citizen_fatherId	db:citizen/1
db:citizen/7	vocab:citizen_fatherId	db:citizen/4
db:citizen/8	vocab:citizen_fatherId	db:citizen/4
db:citizen/1	vocab:citizen_citizen_name	"Telis"
db:citizen/1	vocab:citizen_gender	"Man"
db:citizen/1	vocab:citizen_serial_number	1
db:citizen/1	rdfs:label	"citizen #1"
db:citizen/1	vocab:citizen_citizen_lastname	"Theoh"
db:citizen/1	rdf:type	vocab:citizen
db:citizen/2	vocab:citizen_citizen_name	"Olga"
db:citizen/2	vocab:citizen_gender	"Woman"

Figures 2a & 2b : the D2RQ map & the relevant SparqL results

Ontology Design And Development

In the context of this work and in order to prove the retrieval of information concerning the kinship relationships between citizens through appropriate DL queries, we created the ontology of the poll in Protégé 4.3. We have supplied the ontology with the necessary SWRL classes, properties and rules, as shown in Figure [5]. In order to check the correctness and consistency of our ontology, we manually fed it with virtual data and then after submitting the ontology to the evaluation by activating the reasoners it has, we also submitted specific queries to the DLQuery Tab which returned the expected results.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?subject ?object
        WHERE { ?subject rdfs:subClassOf ?object }
```

subject	object
Woman	Citizen
Man	Citizen

- Thing
  - Citizen
    - Man
    - Woman
  - Municipal
  - MunicipalRegistry

- topObjectProperty
  - belongsTo
  - hasAunt
  - hasChild
  - hasCousin
  - hasMember
  - hasMunicipalRegisration
  - hasNephew
  - hasNiece
  - hasOwner
  - hasParent
  - hasSibling
  - hasSpouse
  - hasUncle
  - isMemberOf
  - isRegisteredIn

- topDataProperty

Rules

```
hasParent(?x, ?y), hasSister(?y, ?z) -> hasAunt(?x, ?z)
Woman(?y), hasParent(?x, ?y) -> hasMother(?x, ?y)
Man(?y), hasSibling(?x, ?y) -> hasBrother(?x, ?y)
Woman(?x), hasUncle(?x,
Woman(?y), hasSibling(?x, ?y) -> hasSister(?x, ?y)
Woman(?x), hasAunt(?x, ?y) -> hasNiece(?y, ?x)
hasBrother(?y, ?z), hasParent(?x, ?y) -> hasUncle(?x, ?z)
```

Asserted in: http://www.semanticweb.org/stamatis/ontologies/2015/8/untitled-ontology-124

- ◆ ArisTheo
- ◆ ArisTheo2
- ◆ Athens
- ◆ ChrisTheo

Figure 5. The municipal registry Ontology general view from Protege 4.3

Conclusions

It is a fact that there is a large amount of data on the web stored in relational databases, while at the same time any obsolescence with the development of new technologies cut from existing standards and tools will become uneconomical. So it is very important to "produce" statistical semantic data using the already stored data in whatever form it is. In order to use database data in the Semantic Web to produce linked statistical data, it is necessary to use a mechanism that will map the elements of the relational schema to the elements of an ontology and, based on these mappings, enrich the ontology with data from the relational database.

References

[28] Theocharis, Stamatis & Tsihrintzis, George. (2016). RDB Data Triplification vs. Ontologies: The Case of Municipal Registry Data. International Journal of Computer and Electrical Engineering. 8. 44-56. 10.17706/IJCEE.2016.8.1.44-56.

[29] Richard Cyganiak et al., The D2RQ Mapping Language, retrieved 8/2015, available: http://d2rq.org/d2rq-language#examples

[30] Chris Bizer, Richard Cyganiak, D2R Server: Accessing databases with SPARQL and as Linked Data, retrieved 8/2015, available:http://d2rq.org/d2r-server

Local population projections with Bayesian hierarchical models

Keywords: small area estimation, population forecast, R open source code

Abstract

In this paper we describe a new approach to local and national population forecasting. It is based on Bayesian hierarchical models which are able to efficiently solve small area/population issues as well as to account for complex (auto-) correlation structures and to incorporate qualitative and quantitative prior information and even expert assumptions.

We provide two classes of models, implemented in open source R code, depending on training data: individual response models, if input microdata is available and aggregated/rates response models if counts data is available. Time and age (auto-) correlations are incorporated via Gaussian process priors and/or flexible non-linear smooth terms while spatial (municipality identifiers) and social-demographic characteristics are included in a natural multilevel model setting.

As a result, fertility and mortality models show non-significant variation of rates by municipality but depend on family related and education characteristics. Stable forecasting models by age, time and gender and/or citizenship are straightforwardly built for mortality and fertility rates. Migration counts and rates, by contrast, have strong fluctuations and are sensitive to a richer set of factors, therefore Gaussian process priors, in addition to hierarchical dependence on these factors, performed particularly well for this type of non-stationary and auto-correlated data.

Introduction

The goal of population projections is to predict the future values, and their uncertainty measures, of regional/total population by age, gender, time and other demographic or spatial characteristics, based on past observed values over a reasonably long time. The projection method should satisfy the following conditions: to be based on statistical modeling and be able to efficiently solve small area/small population and rare events issues as well as to account for complex (auto-) correlation structures and to incorporate qualitative and quantitative prior information and/or expert assumptions.

Currently there are three main classes of methods for producing population projections: based on assumptions/scenarios, on functional models [1] and based on Bayesian models [2]. Statistics Iceland employed in the past a mixture of such methods, by forecasting fertility and mortality with functional models while modeling short term migration with econometric/ARDL models. In addition, modeling the time correlation between emigration and lagged immigration has been employed in order to further improve the predictions. Only national projections were produced until now.

In this paper we test a new approach to population forecast which can successfully fulfil the requirements mentioned above. It is based on Bayesian hierarchical models and it is implemented in open source R code (shared at https://github.com/violetacln/SIPP). We *provide* two classes of solutions, depending on data: individual response models, if input microdata is available and aggregated/rates response models if only count data is available. Time and age (auto-) correlations are incorporated via Gaussian process priors or flexible non-linear smooth terms while spatial and social-demographic characteristics are included in a natural multilevel model setting. For instance, we show that fertility and mortality rates have a non-significant variation by municipality but depend on family related and education characteristics. When based on count data, they provide very stable forecast models by age, time and gender (in the case of mortality) or citizenship (as in the case of fertility). Migration rates, by contrast, are highly fluctuating and sensitive to a richer set of factors as described in what follows. Gaussian process priors are thus most useful and they are mainly characterized by the mean function, which dominates the long-term behaviour, and by the covariance function, which describe the correlation between ant two response values.

Methods

We formulate the general statistical problem to be solved as estimating and predicting with a model for response data which consists either of: (i) Poisson rates, when modelling aggregate counts (of death, giving birth and migration in or out of the country events) per exposed population and time interval, when the input is *count* data, or (ii) binomial response data, in the case of modelling same type of events, when the input is *microdata*.

Our solution is to build generalised additive models with hierarchical structures to account for any clustering effects (by location or by other characteristics such as citizenship or education) and differences between/within groups of observations. In addition, we required the models to have correlation structures such that to account for temporal effects. This type of models is estimated in Bayesian framework, with priors chosen according to data exploratory analysis and expert knowledge. The Gaussian process priors (over time and age) are particularly suitable due to their ability to describe complex dynamics and phenomena specific to time series, i.e. non-stationarity or periodic and trend components.

The expected value of the response is therefore written, via a link function when needed, as a sum of functions over a set of predictors: (i) time, age, location, gender, citizenship in the case of aggregated models and (ii) education, family size, municipality size, in addition to time, age, ***location,*** gender, citizenship, for event modelling. These functions may be smooth one, such as splines and their tensor products or unknown ones, defined by a prior stochastic process and updated by the observed data points.

The implementation was made straightforward by exploiting reliable R-packages such as *mgcv* [3](which accommodates a wide selection of smoothers and Gaussian process kernels), *lme4* [4] (for frequentist fast estimates and testing multilevel models), and *brms* [5] (Bayesian, based on a *stan* engine [6]). We also share our R-code on the *github* page of the *SIPP* (Statistics Iceland's Population Projections) repository linked in the previous section.

Results

In this section we illustrate the age effect on fertility, mortality and migration rates (see Figures 1-3) as fitted by hierarchical models. We continue by showing forecasting results for some of these rates (Figures 4 only, due to limited space), generated by the models described in the previous section.

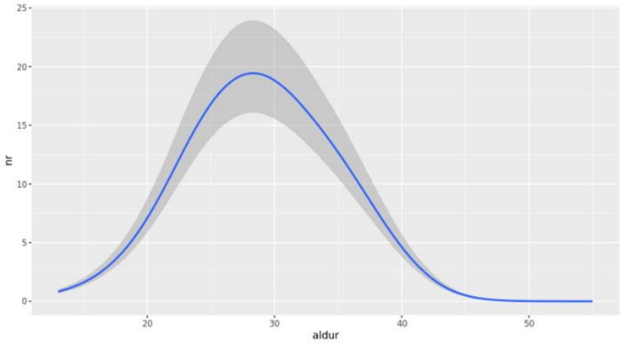


Figure 12. Age dependence of fertility rates

Figure 1 shows a well-known pattern of the last almost 20 years, with the most likely age of mothers close to 30. Testing for significance of location-dependence of fertility rates showed that most municipality effects (very few exceptions) are non-significantly different from the national level. A significant effect was however found for citizenship, the Icelandic women having higher fertility rates than the ones with foreign citizenship.

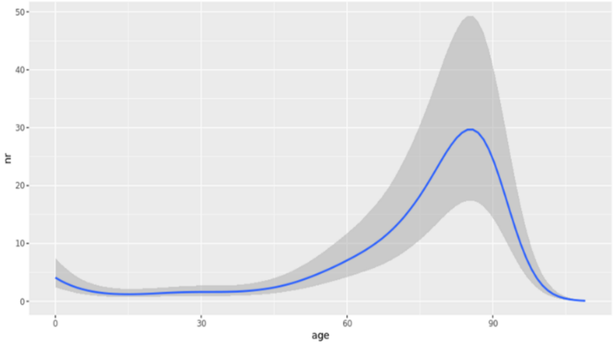


Figure 2. Age dependence of mortality rates

This confirms both the most likely ages of death and the growing estimation uncertainty with age. The models also proved that location (municipality) does not have a significant effect on mortality rates.

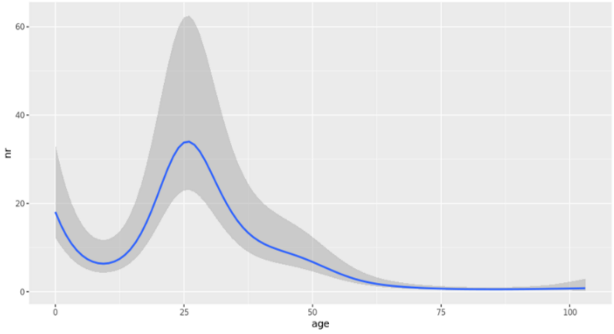


Figure 3. Age dependence of migration rates

The most likely age of migrants is around 25, followed closely by the very young ages of their children. Migration in and out of the country show similar patterns, but here we showed the emigration only. The rates however depend on municipality, in addition to age, citizenship and gender. The models allowed us to estimate and forecast this more involved structure.

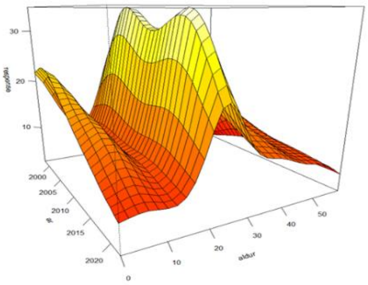


Figure 4. Estimated and predicted age-time surface of migration rates

The most complex pattern of age-time variation is displayed by the migration rates, which require more involved tests in order to choose the covariance kernels with best performance. This variation proved to be dependent on location as well in a statistically significant way. By contrast, fertility rates regional variation is non-significant for most municipalities and always non-significant for mortality rates.

Conclusions

We have analysed, modelled and forecast fertility, mortality and migration time series data at local and national levels. We share our R-code and use open source R-packages for implementing the hierarchical models with best performance.

References

[1] Hyndman, R. J., Booth, H., (2008) Stochastic population forecasts using functional data models for mortality, fertility and migration, International Journal of Forecasting 24 (2008) 323–342.

[2] Jakub Bijak & John Bryant (2016) Bayesian demography 250 years after Bayes, Population Studies, 70:1, 1-19, DOI: 10.1080/00324728.2015.1122826

[3] Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

[4] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[5] Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

[6] Stan Development Team (2022). “RStan: the R interface to Stan.” R package version 2.21.7, https://mc-stan.org/.

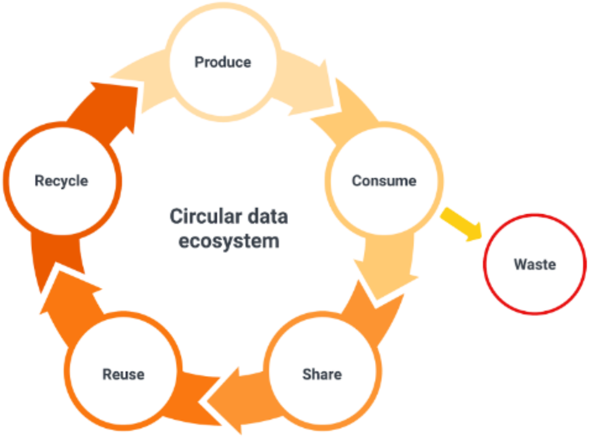
Data circularity for the circular economy

Keywords: data revolution, big data, circular economy, datafication

1. INTRODUCTION

The circular economy has become the centre of attention of many initiatives at national, regional, and international levels. Be it private or public sector, organisations are focusing on achieving a more circular economy, as in one that does not simply produce and waste, but also reuse and recycle resources. Data is essential for achieving a circular economy and yet, the global data ecosystem also lacks a circular mindset. How can we measure our impact on global challenges, such as climate change, poverty, or sustainability, if the data to support these initiatives is not reliable or readily available?

While 2.5 quintillion bytes of data were produced every day in 2021, this number is expected to grow exponentially in the coming years. Some researchers have named this datafication [1]. With the constantly increasing flow of data, organisations around the world are struggling to process, structure, store. and use the data produced leading to missed opportunities and inability to track progress and monitor successes and failures. Often data is used once for a specific purpose, in isolation from other sources, and lost or forgotten about after this initial use. Far from being useless, this data can be used and reused for other purposes by other organisations and people.



In this paper, I argue that a circular mindset to data is essential in preventing poor-quality data, as well as low data valorisation, reusability, and accessibility. The lack of it prevents efficient data-driven decision-making at a global scale. The main tool in achieving data circularity is the maximisation of interconnectedness between the data generated. The more interconnected the data, the easier it is to find, share, use, reuse, and recycle it.

Data initiatives such as open-data are often not sufficient to break data silos as they are often reliant on expert knowledge that is difficult to collect. Furthermore, each open-data initiative has its own standard and the lack of a common data architecture leads to difficulties to reuse the data in different contexts. The lack of high-quality and accessible data prevents impact measurement, hinders global decision making [2], dilutes accountability, and ultimately leads to less effective global action. New data solutions need to stimulate the generation of structured, consolidated, and eventually tradeable data. The objective is to turn data into a valuable, recyclable asset.

2. METHODS

The research methodology will draw upon both primary and secondary methods with an interdisciplinary approach. The first step in the analysis is to review all previously available research on the subjects of datafication and the circular economy.

A second step is to explore the results of a survey conducted among the users of a data marketplace, Datastake, that uses a data circularity approach to its solutions. Datastake commoditizes information with a standardised data taxonomy that identifies information connections across organisations, use cases and sectors. Holders of information, from corporations to development programs, may use the service to identify additional sources to consolidate from, or new clients to extend their data value chain. It lets users decide which information to acquire, and on what terms to share it.

The survey will identify the main advantages of such an approach in terms of data availability, quality, and useability.

3. CONCLUSIONS

Data circularity is an approach that can effectively achieve the generation of high-quality, accessible, and usable data. This approach stimulates coordination within and between organisations leading to better decision-making through standardisation, reconciliation, consolidation, and trade of information.

REFERENCES

[1] K. Cukier and V. Mayer-Schoenberger, The rise of big data: How it’s changing the way we think, Foreign Affairs#92 (2013), 28–36.

[2] Joseph T. Bonivel Jr. Ph.D and Solomon Wise, THE DATA DIVIDE How Emerging Technology and its Stakeholders can Influence the Fourth Industrial Revolution, Atlantic Council (2022).

Statistics during pandemics (GASP2A.1)

Session Chair: **Eniel Ninka** *(Eurostat)*

SORS decision making support system in COVID-19 emergency situations
Branko Josipovic, Nebojsa Tolic *(Statistical Office of the Republic of Serbia)*

The Effect of Non-pharmaceutical Policy Interventions on Population Mobility During the COVID-19 Pandemic
Jonas Klingwort, Joep Burger*, Jan van den Brakel (Statistics Netherlands-CBS)*

The behaviour of alternative trend-cycle decomposition methods during the pandemic
Ferdinando Biscosi *(GOPA Luxembourg)*, Gian-Luigi Mazzi *(Senior Consultant)*, Piotr Ronkowski *(Eurostat)*, Rosa Ruggeri Cannata *(Eurostat)*

SORS decision making support system in COVID-19 emergency situations

Keywords: Data collection, analytics, Covid-19 surveys, decision-making, statistics

Introduction

Statistical Office of the Republic of Serbia (SORS) has responded to The Covid-19 outbreak and increasing demands for statistical data, especially in domain assessment of the economic impact of the crisis. SORS introduced new changes in the organizational aspect to preserve business continuity. Fast, responsive, predictive statistics were also introduced through the concept of a decision-making system as a new role of official statistics, showing that the position of the statistical institute is very important in these unpredictive situations.

Methods

In order to obtain a high level of statistical production it was necessary to introduce different organizational and operational changes as follows:

- Rapid growth of teleworking - all SORS employees have been provided with the technical possibility to work from home, including interviewers who used cloud technology for the CATI data collection method.
- To obtain qualitative information on the current economic situation and to predict short-term trends in the business of economic entities it was crucial to keep conducting Business climate and population consumption surveys (BCS) using the CATI method.
- In addition to standard CATI surveys, seven new ad hoc CATI and MIX mode surveys (CATI+CAWI) were introduced, related to businesses and the COVID-19 crisis. Data collected and processed in these surveys, monitoring, and auditing was visualized in a way that’s easy to use and helps to gain deeper data insight.
- Increasing use of Cloud services and use of artificial intelligence services.

A set of fast responsive surveys were also introduced in order to help the government to mitigate the pandemic effect:

- Estimate of current industrial production
- Monthly assessment of industrial capacity engagement in the industry in the Republic of Serbia
- Monthly assessment of the implementation of the construction activity plan in the Republic of Serbia
- Monthly assessment of wholesale and retail trade turnover index in the Republic of Serbia
- Estimation of expected capacity utilization in accommodation facilities in the Republic of Serbia for July and August 2020
- Analysis of the package of a financial support program to economic entities, ad hoc
- Realization of contracted projects with the Republic of Serbia

- Expected trends in the current month in the areas of professional, scientific, technical, administrative, and other ancillary services in the Republic of Serbia
- Analysis of measures taken by local governments to facilitate the functioning of the economy during the pandemic Covid19
- The impact of COVID-19 on the realization of teaching in primary schools
- The impact of Covid-19 on the operation of mini grocery stores
- The impact of Covid 19 on the business of micro and small enterprises and entrepreneur
- Survey on foreign tourists
- Reserved number of overnights in the next six months in catering facilities in the Republic of Serbia
- Burden of reporting units survey
- Register of industrial zones

Thanks to territorial organization, in the situation when special measures caused by the COVID-19 virus were implemented, SORS managed to successfully monitor surveys through the regional statistics data centres.

Results

Since the beginning of the COVID-19 outbreaks, interest in statistical data and information has drastically increased. Serbian government was the main stakeholder in statistical data, who needed information on the different statistics at a faster rate, to be fully prepared for making a proper decision in that domain. Regarding this, all resources from our office have been regrouped.

The benefits are various. SORS, from its side, has proven its concept of including data-driven Decision-Making Support, as very useful. The need for data was shown. The resources have been identified. The organization was moved in this direction. The requested data have been collected, analyzed, and sent to the government. The government, public, researchers, and academia were informed of the statistics data in time.

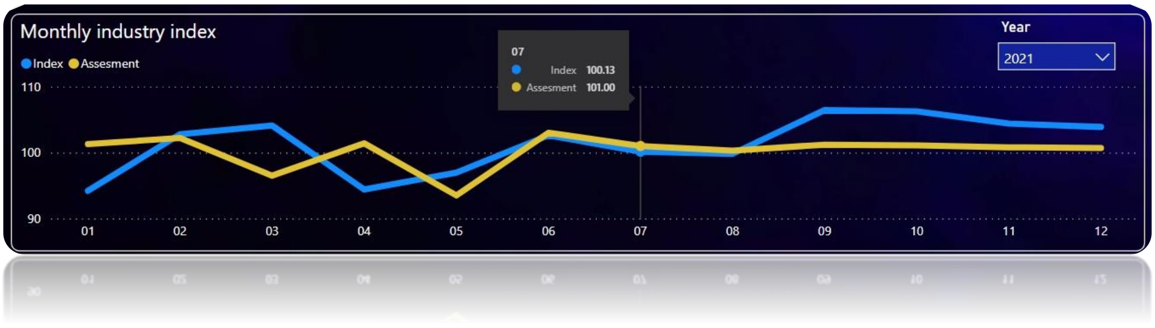


Figure 1: Monthly prediction of industry index (new rapid responsive survey)

Statistics on some statistical domains, based on data administrative and statistical register, have been processed, since the introduction of the state of emergency, on weekly basis instead of monthly.

Conclusions

SORS developed close and lasting cooperation with other state institutions and entities (ministries, chambers of commerce, working groups, etc.) The integration of all available statistics and all levels of statistical expertise together, as a response to the increasing needs, of local communities strengthens the new and significant role of official statistics and at the same time creates completely new analytical perceptions at the state level. When it comes to making any analysis and creating any strategic plans statistics need to be the focal point whether in the case of state level or local level domain.

The COVID-19 outbreak showed us that we need to maintain a different communication channel with reporting units as our valuable source of data. At the same time, official statistics must be able to quickly respond to unforeseen circumstances, and the best way for that is to add new and fast statistics surveys which can help us to better anticipate and react in those situations.

Regarding IT infrastructure, the decision to develop and invest its resources, in highly automated IT was proven, again, as the right decision.

SORS was pointed to as the institution of trust, by having the ability to adapt to new circumstances and to provide a fast response to the arose requests from the stakeholders from all levels. At the same time, SORS has practically shown the integration and the leadership role in the national statistical system, by gathering all relevant public and private companies and associations and giving direction in providing reliable and accurate data.

The Effect of Non-pharmaceutical Policy Interventions on Population Mobility During the COVID-19 Pandemic

Keywords: sensor data, structural time series model, Kalman filter, coronavirus, SARS-CoV-2, population willingness to comply

Introduction

The COVID-19 pandemic has shown that policymakers require timely and accurate information on the effectiveness of policy interventions, such as vaccination programs and non-pharmaceutical interventions (NPIs) that reduce virus transmission. An important aspect is people's willingness to comply with such regulations. In this paper, we evaluate the effectiveness of NPIs in reducing population mobility. Using structural time series modeling, we estimate the effect of policy stringency on pedestrian frequency. The regression coefficient that describes the relation between NPIs and pedestrian frequency is modeled dynamically to investigate how this relation evolves over time during the COVID-19 pandemic. Policy stringency is obtained from the Oxford Covid-19-Government Response Tracker (OxCGRT) project. Pedestrian frequency is obtained from location-based sensors in metropolitan areas during and before the pandemic. The results show a gradual decline in the effect of stringency on population mobility, suggesting a reduced willingness to comply.

Research Background

Limiting social contacts within a population to reduce the transmission of the SARSCoV-2 coronavirus has been highly effective in combating the COVID-19 pandemic. Here, NPIs such as stay-at-home orders or closing (non-essential) businesses lead to decreasing population mobility causing reduced social contacts [1, 2]. The targeted effects were evaluated using mobile phone data or were surveyed retrospectively. The drawbacks of these approaches and data sources are discussed by [3]. In this study, we propose using passive sensor measurements of pedestrian flows. An advantage of this data is the availability of measurements from years before the pandemic. Thus the data include a control of the natural experiment induced by the pandemic.

Data

The sensor data on pedestrian flows in metropolitan areas is provided by the company hystreet.com GmbH [4]. The data collection started on 01.05.2018 at 27 locations in Germany. The sensor network has been continuously expanded and works in 92 cities at 220 locations in 6 different European countries. The sensors are attached to facades and measure the pedestrian flows on a minute level. Hence, pedestrians are unaware of being recorded and cannot consciously avoid the recording. No personal data are collected, only aggregate counts. For the analysis, a selection of sensors was made that were installed in Germany and started recording before 2019. This selection leads to 20 selected cities and 42 locations. The period under study is 01.05.2018 to 31.03.2022. All locations' daily counts were averaged per week to obtain a weekly average national count.

The OxCGRT database provides quantifications of the NPIs. From this database, the stringency index is used. It provides daily measurements since 21st January 2020 and consists of 9 indicators informing about the severity of NPIs. These indicators are school closures, workplace closing, cancel public events, restrictions on gatherings, public transportation, stay-at-home order, restrictions on internal movement, international travel controls, and public information campaigns. They primarily aim to reduce people's mobility behavior [5]. The stringency index is calculated using all ordinal containment and closure policy indicators plus an indicator recording public information campaigns. It ranges from 0-100 (100=strictest). For the analysis, the daily measurements were averaged per week. A binary indicator is used in the model to capture the peaking Advent period at the end of each year, which is 1 for the last weeks of the year and zero otherwise. The data used for analysis is shown in Figure
1.

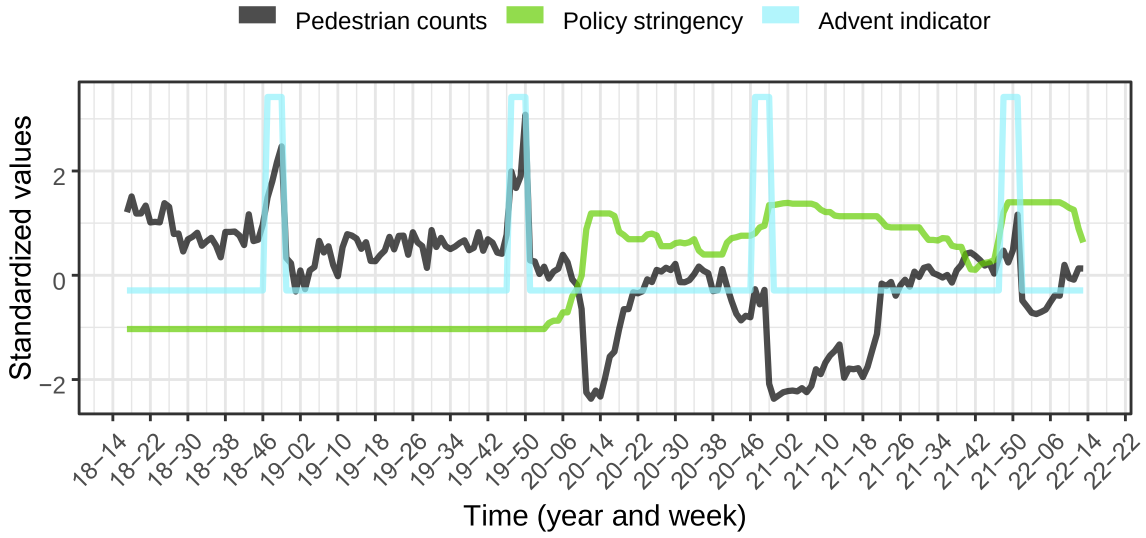


Figure 1: Time series of average weekly pedestrian counts, policy interventions, and indicator for Advent period. For visual comparability, standardized values are shown.

Methods

The weekly pedestrian counts are described with a structural time series model that contains a trend and two regression components, namely policy measures, and a binary Advent indicator. To fit the structural time series model with the Kalman filter, it is expressed as a state space model that consists of an observation equation and a state equation. The observation equation states how the observed series depends on the trend and the regression components and is defined as

$$y_t = Z_t \alpha_t + \epsilon_t \tag{1}$$

$$Z_t = \begin{bmatrix} 1 & 0 & x_t & d_t \end{bmatrix}$$

$$\alpha_t = \begin{bmatrix} L_t & R_t & \beta_t & \gamma_t \end{bmatrix}^\top$$

$$\epsilon_t \sim N(0,\sigma_\epsilon^2),$$

where y_t is the observed average number of pedestrians in week t, Z_t the design vector, α_t the vector of unobserved state variables, σ_ϵ^2 the variance of the observation disturbance ϵ_t, L_t the level of the trend, R_t the slope of the trend, β_t the regression coefficient of the policy index x_t, which is modeled dynamically using a random walk, and γ_t the regression coefficient of the Advent indicator d_t, which is also modeled dynamically using a random walk. The state equation, which defines how the states evolve over time, is given by

$$\alpha_t = T\alpha_{t-1} + \eta_t \qquad (2)$$

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \eta_t \sim N\left(\mathbf{0}_{[4]}, \mathrm{diag}\left(\begin{bmatrix} 0 & \sigma_R^2 & \sigma_\beta^2 & \sigma_\gamma^2 \end{bmatrix}\right)\right),$$

where T is the design matrix and σ^2 the variances of the state disturbance vector η_t. Note that the trend only contains a disturbance term for the slope R_t but not for the level L_t since $\sigma_L^2 = 0$. This trend model is known in the literature as the smooth trend model. The regression coefficients β and γ are time varying since they are modeled dynamically with a random walk. The smooth trend model appeared to be flexible enough to capture the seasonal fluctuation in the weekly observations. Therefore a seasonal component is not required. Kalman filtering and smoothing are applied to fit the model. The variance components σ_R^2, σ_β^2 and σ_γ^2 are estimated with maximum likelihood (ML) using a numerical optimization procedure. The state variables in the Kalman filter are initialized using a diffuse initialization. For details on state space models and the Kalman filter we refer to [6].

Results

Figure 2 shows the observed average weekly pedestrian counts, and the red line shows the smoothed signal with its 95% confidence intervals. The signal is defined as the sum of the trend and the two regression coefficients according to observation equation (1). In Figure 3, the regression coefficient of the policy index is shown. Initially, a negative correlation between the policy index and the observed sensor count is observed. Thus, the NPIs initially reduced population mobility. However, the negative correlation gradually disappears over time. Accordingly, the weekly number of pedestrians increased while NPIs were still in place. This might be a potential effect of pandemic fatigue. Note that the constant negative value of the regression coefficient before the start of the pandemic is the result of applying a smoothing algorithm to the filtered estimates.

The smoothed signal does not precisely follow the weekly changes, indicating no overfitting. Moreover, the smoothed signal is almost always (only with a few exceptions) within the error band of two standard deviations. The standardized innovations (onestep forecast error) are primarily within the interval of two standard deviations (not shown). Data visualization showed that the standardized innovations were normally distributed, indicating a good model fit.

Conclusions

In this paper, we present an analysis of the effectiveness of NPIs during the COVID19 pandemic. Almost the entire period of the pandemic is considered, and a period of two years before the start of the pandemic, so the data includes a control of the natural experiment introduced by the pandemic. Location-based sensors passively measuring pedestrian frequencies and structural time series modeling were used for this purpose. Thus, we could successfully demonstrate how the NPIs worked and how their effect evolved throughout the pandemic. The weakening of the correlation between NPIs and mobility suggests a reduced willingness to comply with coronavirus measures. Currently, we are studying the individual effects of the separate indicators from which the overall policy index is composed. Moreover, we are implementing a multivariate state space model on separate time series of German federal states. These research initiatives highlight the importance of integrating new data sources, routinely collected administrative data, and sound methodology. Such results can be used for (real-time) policy evaluation. As some European countries have installed such sensor technologies, we see much potential for practical implementation to develop real-time indicators for governmental bodies, policymakers, and official statistics.

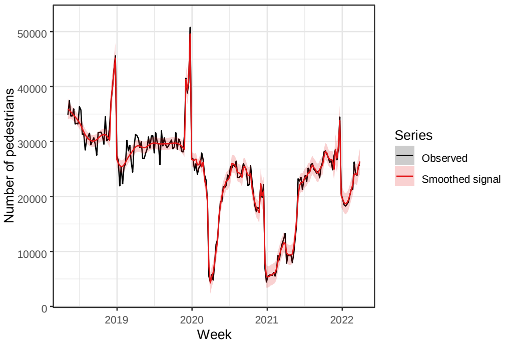


Figure 2: Observed average weekly pedestrian counts and smoothed signal.

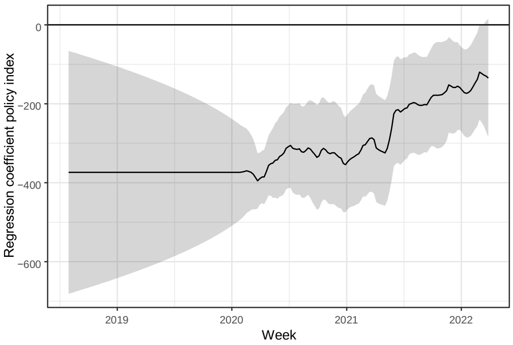


Figure 3: Regression coefficient of policy index.

References

[1] Brauner JM, Mindermann S, Sharma M, Johnston D, and Salvatier J et al. Inferring the of Government Interventions Against COVID-19. *Science*, 371(6531), 2021.

[2] Flaxman S, Mishra S, Gandy A, Unwin HJT, and Mellan et al. Estimating the effects non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584 (7820), 2020.

[3] Klingwort J., De Broe S, and Brocker SA. Sensing pedestrian flows for realtime assessment of non-pharmaceutical policy interventions during the COVID-19 pandemic. *International Journal of Population Data Science*, 5(4), 2020.

[4] hystreet GmbH. Pedestrian frequencies: Timely. precise. transparent, 2022. URL https://hystreet.com.

[5] Hale T, Angrist N, Goldszmidt R, Kira B, and Petherick A et al. A Global Panel Database of Pandemic Policies (Oxford Covid-19 Government Response Tracker). *Nature Human Behaviour*, 5(4):529–538, 2021.

[6] J Durbin and SJ Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford, 2 edition, 2012.

The behaviour of alternative trend-cycle decomposition methods during the pandemic

Keywords: business cycle, filtering techniques, unobserved component models, data analysis

Introduction

Researchers and business cycle analysts are particularly interested in obtaining timely, reliable and objective statistical information to be used as input in their analysis and statistical institutions appear the ideal partner for researchers and analysts. Providing input for the classical cycle analysis does not require any additional effort for statistical institutions since this information is part of the seasonal adjustment output, the situation is different when looking at the growth cycle because further statistical elaboration is required. For this reason, statistical agencies have been reluctant to be involved in this additional activity, even if in the more recent years the situation appear to change.

In this respect, since several years Eurostat is publishing at monthly frequency alternative trend-cycle decompositions for the Gross Domestic Product (GDP), Industrial Production Index (IPI) and Employment based on different decomposition methods. In 2020, the pandemic was really risking to affect dramatically the trend-cycle decomposition provided by Eurostat. As a consequence, in the last two year and half an intense methodological and empirical activity has been conducted aiming at preserving the reliability and credibility of the trend-cycle decomposition provided to the public.

This paper compares three popular methods used by Eurostat namely, the Hodrick-Prescott (HP) method, the Christiano-Fitzgerald (CF) approximation the ideal band pass filter and the Harvey and Trimbur (HT) unobserved components model to compute trend and cycle estimates. In particular, these three methods were applied to Euro Area (EA) monthly IPI as well as quarterly GDP and Employment figures that cover the last three decades, including the COVID pandemic. The latter brought to the introduction of an additive outlier correction to handle the most volatile observations. The results section shows only the quarterly GDP series estimates.

Methods

Two-side Hodrick and Prescott filter

The Hodrick and Prescott (HP) filter decomposes the time series y = $(y_1, \ldots, y_T)$ into a cyclical component ψ and a trend component τ:

$$y_t = \psi_t + \tau_t$$

The two-sided HP (HP2s) filer estimates the trend component by solving the following minimization problem:

$$(\hat{\tau}_{1|T,\lambda},\ldots,\hat{\tau}_{T|T,\lambda}) = \text{arg } min_{\tau_1,\ldots,\tau_t} \sum_{s=1,\ldots T} (y_t - \tau_s)^2 + \lambda \sum_{s=2,\ldots T-2} (\tau_{s+1} - 2\tau_s + \tau_{s-1})^2$$

where λ controls the smoothness of the trend estimates. The higher its value, the smoother the extracted trend component will be. The HP model can be rewritten with the following state space, where

$$\tau_t = 2\,\tau_{t-1} - \tau_{t-2} + \epsilon_t$$

for the unobservable trend (see [1], [2] and [3] for further details).

Christiano-Fitzgerald (CF) approximation

The Christiano-Fitzgerald (CF) filter is the approximation to the ideal band pass filter for time series [4]. The 'ideal' band pass filter can be used to isolate the component of a time series that lies within a particular band of frequencies. CF identifies one approximation, which, though it is only optimal for one particular time series representation, nevertheless works well for standard macroeconomic time series.

Harvey and Trimbur model

The model of Harvey and Trimbur [5] for the trend-cycle decomposition is given by:

$$y_t = \mu_t + \psi_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

where μ_t represents a series-specific trend component for series and ψ_t represents the stochastic cycles (BC for business cycle) which are common to all series in y_t. The BC captures the medium-term business cycle that is the main object of this simulation. The disturbance term ϵ_t, also referred to as the irregular component, is assumed to be normally distributed, and serially and cross-sectionally uncorrelated.

The trend component is specified as a local linear trend given:

$$\mu_t = \mu_{t-1} + v_{t-1} \quad and \quad v_t = v_{t-1} + \xi_t \qquad where \;\; \xi_t \sim N(0, \sigma_\chi^2)$$

where v_t represents the gradient of the trend component μ_t and is often referred to as the growth or slope term (see [6] and [7]).

One of the main characteristic of Harvey and Trimbur model is the smoothed trend component, that resemble a I(2) model. This assumption guarantees a very smooth long-term trend, necessary requirement to have a reasonable cycle. The disturbance term ξ_t is assumed to be serially and cross-sectionally uncorrelated and independent of ϵ_t, contemporaneously and for all leads and lags.

The stochastic cycle component ψ_t is modelled as nested stationary dynamic processes and are formulated by the trigonometric specification:

$$\begin{pmatrix} \psi_{1,t} \\ \psi^*{}_{1,t} \end{pmatrix} = \rho \begin{bmatrix} cos\,\lambda & sin\lambda \\ -sin\lambda & cos\lambda \end{bmatrix} \begin{pmatrix} \psi_{1,t-1} \\ \psi^*{}_{1,t-1} \end{pmatrix} + \begin{pmatrix} \omega_t \\ 0 \end{pmatrix}, \quad (\omega_t) \sim N(0, \sigma_\omega^2)$$

$$\begin{pmatrix} \psi_{i,t} \\ \psi^*{}_{i,t} \end{pmatrix} = \rho \begin{bmatrix} cos\,\lambda & sin\lambda \\ -sin\lambda & cos\lambda \end{bmatrix} \begin{pmatrix} \psi_{i,t-1} \\ \psi^*{}_{i,t-1} \end{pmatrix} + \begin{pmatrix} \psi_{i-1,t} \\ 0 \end{pmatrix}, \;\; i = 2, \dots, n$$

where the frequency of the cycle λ is measured in radians with $0 \leq \lambda \leq \pi$ leading to a period of the stochastic cycle of $2\,\pi/\lambda$. The index i refers to the number of harmonic components and adding more components smooths the cycle. The persistence parameter

ρ is restricted within the interval $0 \leq \rho \leq 1$ to ensure a stationary process for the cycle. The disturbances ω_t is mutually, serially, uncorrelated and independent of all other disturbances in the model.

Outlier correction

Large shocks associated to extreme events such as the financial crisis in 2008-2009 and even more importantly the COVID pandemic introduce instability and parameter breaks in the model. To deal with it, we apply an automatic procedure for detection of outliers in time series. However, after analysing the series, we selected as outliers 2020M4 and 2020M8 for the IPI monthly series and 2020Q2 for GDP and Employment quarterly series.

Results

This section presents the results for the Euro Area quarterly GDP data.

Figure 13 shows the HP, CF and HT trend estimates. The HT estimate is the smoother and less variable. The HT diverges in the last four years of the sample (from 2018), when the HT trend is more positive and keeps a higher value during the COVID pandemic.

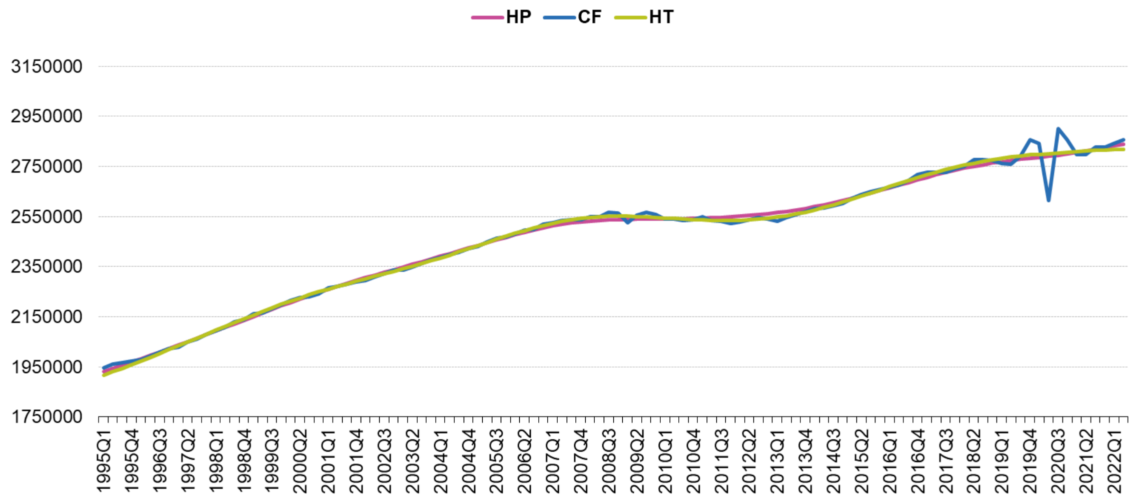


Figure 13: EA quarterly GDP – trend estimates

Figure 14 shows the cycle estimates. The three lines have a very similar pattern and only minor differences in the peak before the financial crisis and at the throw of the COVID-19 pandemic in 2021Q3-2021Q4 when HP and HT cycles are more negative and the following recovering when HT increases less than the other three estimates. However, the general conclusion is that the three trends and cycles of EA GDP are very similar.

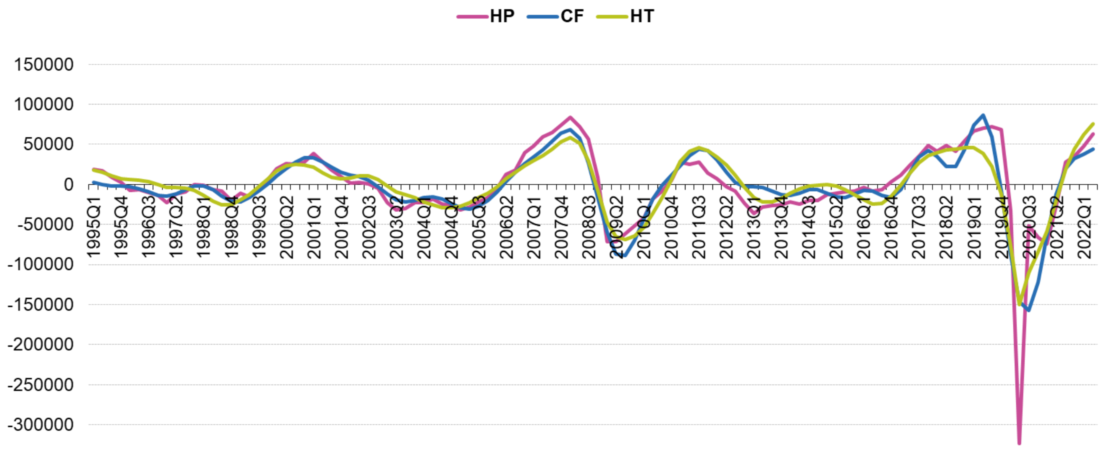


Figure 14: EA quarterly GDP – cycle estimates

Conclusions

Results indicate that estimates are for EA GDP and Employment are very similar across the three filters; whereas the estimates for EA IPI given by the HT indicate that the trend is higher in period of IPI expansion and lower in period of contraction such as the financial crisis and the COVID pandemic.

References

[31]R. J. Hodrick and E. C. Prescott, Post war U.S. Business Cycles: An Empirical Investigation, Journal of Money, Credit and Banking, 29(1), (1997).

[32] K. Kuttner. Estimating potential output as a latent variable, Journal of Business and Economic Statistics, 12(3) (1994).

[33] L. Ljungqvist and T. J. Sargent, Recursive macroeconomic theory, Cambridge, Massachusetts, MIT Press (2004).

[34] L. J. Christiano and T. J. Fitzgerald, The Band Pass Filter, International Economic Review, 44(2) (2003), 435–65.

[35]A. C. Harvey and T. M. Trimbur, General Model-based Filters for Extracting Cycles and Trends in Economics Time Series, Review of Economics and Statistics, 85(2) (2003), 244-245.

[36]A. C. Harvey, Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge University Press (1989).

[37]J. Durbin and S. J. Koopman, Time Series Analysis by State Space Methods (2nd), Oxford University Press (2012).

Equality (JENK2A.1)

Session Chair: **Sorina Vâju** *(Eurostat)*

Promoting the collection and use of equality data
Rossalina Latcheva (European Union Agency for Fundamental Rights)

Beyond the gender pay gap
Marina Perez Julian *(Eurostat)*, Denis Leythienne *(Eurostat)*

New statistics on discrimination and gender based violence at EU level
Aura Leulescu, Merle Paats *(Eurostat)*

Promoting the collection and use of equality data: measuring ‘racial or ethnic origin’[43]

Keywords: equality data, racial or ethnic origin, EU Subgroup on equality data, discrimination, victimisation surveys

Introduction

Equality and non-discrimination are founding values of the European Union, enshrined in its Treaties, in the Charter of Fundamental Rights of the European Union and an integral part of the European Pillar of Social Rights. The European Union has in place an advanced legal framework with which to promote equality and non-discrimination. All 27 EU Member States have transposed this legal framework into national laws, often going beyond the minimum standards included in the Racial Equality Directive and the Employment Equality Directive.

Despite this, data collected by the European Union Agency of Fundamental Rights (FRA) show that significant proportions of people in the European Union experience discrimination, inequality and social exclusion on a regular basis. This can be based on disability, sex, age, racial or ethnic origin, skin colour, religion or belief, sexual orientation, and gender identity, or a combination of these, as evidence consistently shows. Recent research shows that the Covid-19 pandemic might fuel discrimination and inequality for minority groups as well as for women and specific age groups and impact negatively on equal opportunities.

Equality data are essential for assessing the situation of ethnic minorities and other racialised groups and so effectively tackling racism and structural inequalities.[44] Data makes the nature and extent of discrimination and inequality visible and provides the substance for evidenced based policy making. When collected regularly and systematically, equality statistics enable Member States to assess the proper application of anti-discrimination legislation, monitor compliance with human rights obligations, and track progress in achieving goals towards equality – as set by EU economic governance instruments such as the European Semester or by global agendas such as the UN 2030 Agenda for Sustainable Development.

Nonetheless, there is still a lack of comparable and regular data collection on equality and non-discrimination, which limits effective monitoring of the application of the core legal EU frameworks in this area. The absence of robust and systematically collected equality data, combined with the very small number of discrimination cases brought to the attention of

[43] “The European Union rejects theories which attempt to determine the existence of separate human races. the use of the term ‘racial origin’ does not imply an acceptance of such theories”. See Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (6). See as well EU Anti-Racism Action Plan 2020-2025, p.1.

[44] The Council Recommendation on Roma equality, inclusion and participation defines systemic or structural discrimination “as being evident in the inequalities that result from legislation, policy and practice, not by intent but resulting from a range of institutional factors in the elaboration, implementation and review of legislation, policy and practice” (p.20).

relevant authorities, competent bodies and courts, paints an incomplete picture of the reality of discrimination in the EU.[45]

To date, few countries operate comprehensive systems or have a coordinated approach to collecting and using equality data that would uncover inequalities based on racial or ethnic origin. If available, such data are often not up to date or lack comparability across time and geographical regions; frequently they are of limited scope as well as are not transparently available in the public domain.

Acknowledging this, in 2018 the EU High Level Group on Non-Discrimination, Equality and Diversity (HLG) set up a Subgroup on Equality Data (Subgroup) to support Member States in their efforts to improve the collection and use of equality data. It tasked the EU Agency for Fundamental Rights (FRA) to facilitate the work of the Subgroup, in line with the Agency's mandate to develop methods and standards to improve the comparability, objectivity and reliability of equality data at European level.

EU Subgroup on Equality Data: Guidance note on the collection and use of equality data based on racial or ethnic origin

According to the European handbook on equality data,[46] 'equality data' means any piece of information that is useful for the purposes of describing and analysing the state of equality. Such data are indispensable to informing an evidence-based assessment of the application of non-discrimination policies at EU and Member State levels, and to empowering population groups at risk of discrimination. When collected regularly and systematically, such information is essential to helping Member States assess their compliance with human rights obligations.

Acknowledging the importance of equality data and the fact that to date, few Member States operate comprehensive systems or a coordinated approach to the collection and use of equality data, in 2018 the EU High Level Group on non-discrimination, equality and diversity set up the Subgroup on Equality Data, with a view to support Member States in their efforts to improve the collection and use of equality data. It tasked the FRA to facilitate the work of the Subgroup, in line with the Agency's mandate to develop methods and standards to improve the comparability, objectivity and reliability of equality data at European level.[47]

Methods

Equality statistics can be compiled from multiple data sources, such as population censuses, administrative registers, household and individual surveys, victimisation surveys, and attitudinal surveys. Other sources could encompass complaints data (including aggregate profiles of complainants and offenders, for example), criminal justice data (including court statistics and data on outcomes of court cases, as well as compensation offered/sanctions applied, for example), as well as other avenues of data collection, encompassing discrimination testing,

[45] Ibid., p.19.

[46] European Commission (2016), European handbook on equality data, Luxembourg, Publications Office, p. 15.

[47] Council Regulation (EC) No 168/2007 of 15 February 2007 establishing a European Union Agency for Fundamental Rights, OJ L 53, 22.2.2007.

diversity monitoring by employers, CSO and service providers, and data used to train algorithms for artificial intelligence (AI) and machine learning.

Data disaggregated by certain personal characteristics – including age, sex, racial or ethnic origin, religion or belief, disability, sexual orientation, and gender identity – can be used for producing equality data, at an aggregated level for statistical purposes, if this is done voluntarily, in full compliance with legal provisions and the corresponding exceptions.

In the EU, the collection of personal data disaggregated by sensitive personal characteristics, such as racial or ethnic origin, is protected by constitutional norms, EU data protection law and the Charter of Fundamental Rights. Data protection rules do not stand in the way of collecting equality data for statistical purposes. On the contrary, the rules allow for the correct processing of data while ensuring the respect of fundamental rights.[48]

Individuals should have the option to disclose or withhold information about their personal characteristics. This means that providing personal information about one's ethnic affiliation in a survey or administrative source should be optional for the potential respondents and this has to be clearly communicated to them in the instructions preceding the question on how they self-identify in racial or ethnic terms.

The concept of 'racial or ethnic origin'

To make informed policy choices for countering discrimination and fostering equal treatment, legislators and policymakers need data on people's social positioning and experiences of racism and discrimination based on racial or ethnic origin. However, introducing categories such as 'racial or ethnic origin' in official statistics bears the risk of such categorisations being socially reproduced and used to incorrectly label people.[49] This can have negative consequences for members of certain social groups, stemming from the biased (potentially stereotypical) belief systems that such social categorisation can support. To address this, the use of statistical (analytical) categories for any data collection or for the purpose of data disaggregation should always be led by the overriding human-rights based principle of **doing no harm**, as set out by the UN High Commissioner for Human Rights (OHCHR) in the Human Rights-based Approach to Data. Doing no harm means that no data collection activity should create or reinforce existing discrimination, bias, or stereotypes and that the data collected should be used for the benefit of the groups they describe and society as a whole.

According to the Guidance note on the collection and use of equality data based on racial or ethnic origin, race/racial origin and/or ethnic origin are social constructs and as such they are weak proxies for the genetic diversity of humankind. While some individuals may self-identify as 'white' or 'black', racism and racial or ethnic discrimination are often shaped by how society categorises individuals in racialised terms. Ideas about race/racial origin are often ascribed to or

[48] See FRA (2021), Equality in the EU 20 years on from the initial implementation of the equality directives, Publications Office, Brussels; European Commission (2021), Round table on Equality Data in September 2021.

[49] Two points need to be acknowledged in this regard: (a) it is not possible to limit the use of language to informed communicators in a society, and (b) the conscious intentions of communicators are not the only factors shaping the social meaning of a concept or a category. See Guidance note on the collection and use of equality data based on racial or ethnic origin.

imposed on people, and individuals or groups can be racialised by others in ways that negatively affect their experiences and how they are treated. The social construct of race/racial origin is distinct from but may overlap with how people identify themselves, which can be much more varied and complex.

In line with the applicable EU legislation[50], the Guidance note on equality data based on racial or ethnic origin, and the EU Agency for Fundamental Rights (FRA) when conducting research in this field, refer to 'racial or ethnic origin' with respect to its being **a cause of discrimination.** Aligned with practice that has been established in some countries and EU Member States – which use the category 'racial or ethnic origin' for statistical purposes, including to highlight discrimination and inequality – we further refer to 'racial or ethnic origin' as:

- **a generic statistical (analytical) category** that allows for disaggregation of any data, to assess the state of equality in society,
- **an aspect of a person's self-identification and ethnic attachment**, that is, as a personal characteristic.

As mentioned in the UNECE Guide to Data Disaggregation for Poverty measurement[51], "ethnic identity can be measured using a variety of concepts, including ethnic ancestry or origin, ethnic group, cultural origins, nationality, race, [skin] colour, minority status, tribe, language, religion – [and in numerous cases through proxy variables such as country of birth, country of birth of parents, citizenship] - or various combinations of these concepts".

When it comes to experience of discrimination, questions about how others (those who discriminate) perceive someone – based on perceived external attributions – become important. Attributions made by others may not necessarily relate to a person's self-identification. In addition, one's identity encompasses multiple, intersecting characteristics that must be recognised; not just racial or ethnic origin, but also sex, age, sexual orientation, (dis)ability and other personal traits.[52]

Results: FRA approach

According to the UN Principles and Recommendations for Population and Housing Censuses related to Ethnicity (para 4.183.), data on ethnicity provide information on the diversity of a population and can serve to identify subgroups in a population.

For data collection, the populations of interest, surveyed by FRA are **self-defining**, which means that the parameters of the population cannot be imposed by an external party or assigned through imputation or proxy. This principle is essential, especially when 'racial or ethnic origin' refers to an aspect of a person's attachment to or identification with an ethnic or any other minority group. The ascription/identification of a respondent's 'racial or ethnic origin' attributed

[50] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin ('Racial Equality Directive') as well as Article 10 and Article 19 of the Treaty on the Functioning of the European Union (TFEU).

[51] United Nations Economic Commission for Europe (UNECE) (2020), Poverty Measurement – Guide to Data Disaggregation, p. 33.

[52] Farkas, L. (2017), The meaning of racial or ethnic origin in EU law: between stereotypes and identities, Luxembourg, Publication Office of the European Union, 2017, p. 37.

by a third party (e.g., interviewer or a service provider) does not necessarily correspond with a person's actual identification.

Following the UN Principles and Recommendations for Population and Housing Censuses related to Ethnicity (para 4.185.), classifying ethnic/minority groups also requires the inclusion of the finest levels of ethnic/group affiliations, self-perceived groups, regional and local groups, as well as groups that are not usually considered to be ethnic groups, such as religious groups and those based on nationality. Precoding or pre-classification of ethnic groups at the time of data capture may have a tendency to lose detailed information on the diversity of a population unless space to record 'other' unspecified ethnic identities/groups and free-text responses are possible.

For example, having consulted with all relevant Roma organisations in Bulgaria during the implementation of the Fundamental Rights Agency's EU-MIDIS II survey, the survey used a showcard with 25 categories for self-identification during the screening for eligible Roma respondents at the doorstep. The list included for example categories such as: Roma, Gypsy, Ierlii, Demirdjii, Bakardjii, Dzambasi, Rudari/Ludari, Kopanari, Vlasi etc. Therefore, data collectors have to decide which categories are the most important and meaningful for the respondents, which should be ideally done in cooperation with the communities/groups from which data is going to be collected.

Examples from surveys by the Fundamental Rights Agency (FRA)

The Agency's EU-MIDIS surveys targeting immigrants and descendants of immigrants use a multiple approach to 'racial or ethnic origin':

1. To sample and screen eligible respondents, proxy information about respondents' country of birth and their parents' country of birth is used, as there are almost no available sampling frames in the Member States (such as address or individual registers) that would include self-identification information based on racial or ethnic origin.
2. To measure self-identification, the Agency uses a combination of separate survey questions asking respondents for information related to different aspects of group affiliations/attachments based on racial or ethnic origin, as shown below.
3. The Agency uses a set of questions measuring respondents' experiences of discrimination based on racial or ethnic origin in key areas of life.

Source: EU Survey on immigrants and descendants

IN09 Self-identification as a person of African descent or a black person

Question wording: **Would you describe yourself as a person of African descent/ a black person?**

SINGLE RESPONSE

1 Yes

2 No

-96 Prefer not to say

-97 Don't understand the question

-98 Not applicable

-99 Don't know

RA02 ***Self-identification as national/European/country national of respondent's country of birth /country of birth of parents***
ASK ALL

Question wording: **People might see themselves in different ways. The following question is about how you see yourself. On a scale from 1 to 5, where 1 equals "not at all" and 5 "very strongly", to what extent do you feel.. ?**
SHOW CARD ; RA02 READ OUT RA02_1 TO RA01_5

RA02_1 …….	European
RA02_2 … …	Austrian
RA02_3 … …	Nigerian

1 Not at all 2 3 4 5 Very strongly
-96 Prefer not to say
-97 Don't understand the question
-98 Not applicable
-99 Don't know

Source: FRA Fundamental Rights Survey

The Agency's Fundamental Rights Survey – **a survey of the general population in the EU** – included a question about self-identification as an ethnic minority. In addition to this, the survey asked respondents about their country of birth and their parents' country of birth.

r29 Whether belongs to an ethnic minority
ASK ALL
Question wording: **Do you consider yourself to be part of an ethnic minority in [COUNTRY]?**
SINGLE RESPONSE ALLOWED

1 Yes
2 No
888 Prefer not to say
999 Don't know

Measuring discrimination in FRA's EU-MIDIS surveys

The EU MIDIS surveys and the follow-up surveys on Roma, Roma and Travellers and immigrants and descendants ask respondents if they had felt discriminated against on one or more grounds – skin colour, ethnic origin or immigrant background, religion or religious beliefs, sex, age, disability, sexual orientation, and 'other' grounds – in different domains and activities: when looking for work, at work, in education or when in contact with staff at their children's school, in access to healthcare, in connection with housing, and when using public or private services (such as public transport, administrative offices, when entering a night club, restaurant or a hotel, and when being in or entering a shop).

Respondents who reported discrimination on at least one of three specific grounds – skin colour, ethnic origin or immigrant background, and religion or religious beliefs – were asked further details about the incident, applying the generic term 'ethnic or immigrant background'. The generic term 'ethnic or immigrant background' indicates racial/ethnic discrimination as discussed in this guidance note.

Question wording and sequence

The survey includes a definition of discrimination, which is shown to respondents before they are asked about their actual experiences of discrimination: *"I would like to ask you a few questions about human rights. A basic right is to be treated equally. Still, some people might experience discrimination. By discrimination we mean when somebody is treated unfavourably compared with others because of their skin colour, age, sex, sexual orientation, disability, ethnic origin, religion or religious beliefs."*

Question wording for respondents who had been jobseekers in the five years preceding the survey:

"When looking for work in the past 5 years[53]*, have you ever felt discriminated against for any of the following reasons? Tell me all that apply."*

1 Skin colour or racial origin
2 Ethnic or immigrant background
3 Religion or religious beliefs
4 Age (such as being too young or too old)
5 Sex/gender (such as being a man or a woman)
6 Disability
7 Sexual orientation (such as being gay, lesbian or bisexual)
8 Gender identity or gender expression (this includes for example transgender, transvestite or non-binary people) - IF CAPI: INTERVIEWER: READ THE EXPLANATION / IF CASI/ONLINE: INFO BUTTON: For example, someone who was born as a boy but later feels like a girl/woman or born as a girl and later feels rather like a boy/man. Or someone who wears clothes that are usually designed for the opposite sex.
9 Other (please specify): OPEN TEXT BOX PLEASE SPECIFY THE REASON
10 I haven't felt discriminated against for any reason when in this situation
-96 - Prefer not to say
-97 - Don't understand the question
-99 - Don't know

Respondents who reported discrimination on at least one of the three specific grounds – (1) Skin colour or racial origin; (2) Ethnic or immigrant background; (3) Religion or religious beliefs – are asked further details about the incident, whether a complaint has been made and if so, to which institution/ body.

Conclusions

In its Opinion on the Equality in the EU[54], FRA underlines, EU Member States should ensure the systematic collection of reliable, valid and comparable equality data, disaggregated by sex, racial and ethnic origin, religion or belief, disability, age or sexual orientation. Member States should reinforce regular and comprehensive collection of equality data through their national statistical institutes and other relevant government agencies. This includes systematic

[53] FRA's surveys ask about experiences of discrimination for two periods: 12 months and 5 years preceding the survey.

[54] FRA (2021), Equality in the EU 20 years on from the initial implementation of the Equality Directives. Opinion. Luxembourg: Publications Office, p.20.

compilation of equality statistics based on population and household censuses, administrative registers and official EU-wide surveys, such as the European Union Statistics on Income and Living Conditions, the Labour Force Survey and other representative surveys. To enable the monitoring of equality outcomes, such data sources should (i) cover under-represented groups at risk of discrimination and (ii) include information on experiences of discrimination on the grounds of sex, racial and ethnic origin, religion or belief, disability, age or sexual orientation.

FRA will continue to (1) conduct periodic surveys on the lived experience of discrimination and hatred of different population groups across the EU. These surveys provide stakeholders in the field with a comprehensive source of reliable and comparable equality data on the extent and nature of discrimination and hatred in the EU, including as regards experiences of online hatred; the Agency will continue to (2) facilitate the work of the Equality Data Subgroup and support Member States and Eurostat in improving the collection and use of equality data.

References

[38] FRA (2021), Equality in the EU 20 years on from the initial implementation of the Equality Directives. Opinion. Luxembourg: Publications Office.

[39] FRA (2021), Equality in the EU 20 years on from the initial implementation of the Equality Directives. Opinion. Luxembourg: Publications Office. p.19.

[40] European Commission (2016), European handbook on equality data, Luxembourg, Publications Office, p. 15.

[41] Council Regulation (EC) No 168/2007 of 15 February 2007 establishing a European Union Agency for Fundamental Rights, OJ L 53, 22.2.2007.

[42] See FRA (2021), Equality in the EU 20 years on from the initial implementation of the equality directives, Publications Office, Brussels; European Commission (2021), Round table on Equality Data in September 2021.

[43] FRA (2021), Equality in the EU 20 years on from the initial implementation of the Equality Directives. Opinion. Luxembourg: Publications Office, p.20.

Beyond the gender pay gap

Keywords: non-discrimination, equality, gender pay gap, equal pay, equal opportunities.

Introduction

The principle of ‘equal pay for male and female workers for equal work or work of equal value’ has been enshrined in the European treaties since 1957. It is currently laid down in Article 157 of the Treaty on the Functioning of the European Union and subsequent secondary legislation such as Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)[55]. The European Commission has undertaken a number of initiatives in this field as part of the Gender equality strategy 2020–2025.

Methods and results

Gender Pay Gap

To monitor gender equality, Eurostat publishes annually the unadjusted Gender Pay Gap, an indicator that belongs to the scoreboards developed under the European Pillar of Social Rights and the Sustainable Development Goals of the United Nations.

This indicator calculates the relative difference between the average gross hourly earnings of women and men and gives an overall picture of differences between women and men. However, it measures a broader concept than discrimination in the sense of "equal pay for work of equal value".

It can be explained by different components, on the one hand differences in the average characteristics (i.e. age, occupation, education level, working time…) of male and female employees and on the other hand by differences in the wages paid to women versus those paid to men with the same characteristics.

To single out the contribution of each observed characteristic to the indicator Eurostat developed an improved methodology based on regression models, the Oaxaca–Blinder decomposition[56].

[55] OJ L 204, 26.7.2006, p. 23.

[56] Oaxaca, R. (1973), ‘Male-Female Wage Differentials in Urban Labour Markets’, International Economic Review, Vol 14, No 3, pp. 693–709.

Bazen, S. (2011), Econometric Methods for Labour Economics, Oxford University Press, Oxford.

To perform the decomposition, Eurostat used the Structure of Earnings Survey 2018 data[57] collection. In this collection of data, we have information about characteristics of the employer and the employee that can explain part of the gap.

In particular, we have related the earnings to the following variables: sex, educational level, age, age squared, occupation, job experience, working time, principal economic activity, enterprise size, employment contract and geographical location of the enterprise.

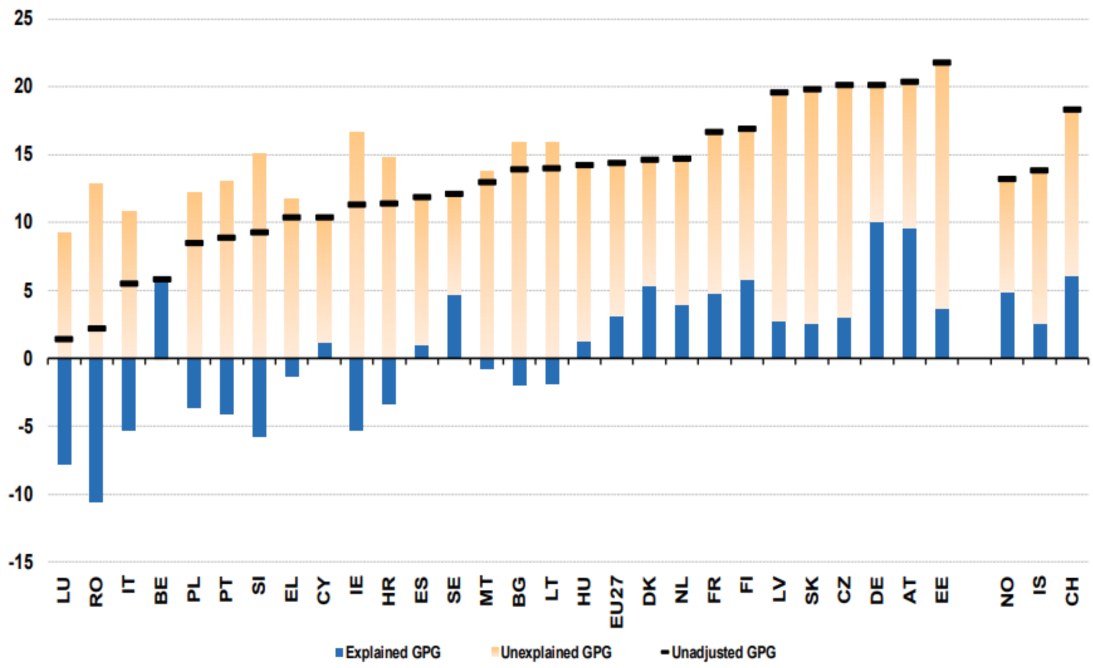
The decomposition methodology is run in two steps:

Firstly, regression analysis equations for men and women are run separately.

The gross hourly earnings are related to characteristics mentioned above of the employee and the employer. Secondly, there is a comparison analysis based on differences on both regressions for men and women that then allows to decompose the pay gap in the explained part and the rest that remains unexplained. Although it is tempting to interpret the unexplained component as a measurement of a possible discrimination through 'unequal pay for equal work', it is not recommended though, as important variables, such as total work experience, are not collected in the SES. Including such additional variables in the regression analysis may change the results.

The results have allowed splitting the 2018 unadjusted gender pay gap into the overall explained part due to the difference in characteristics considered and what remains unexplained by the models.

Figure 1: Gender pay gap adjustments for characteristics, 2018



[57] Eurostat (2021b), SES 2018 implementing arrangements (See item 10.6 at https://ec.europa.eu/eurostat/cache/metadata/en/earn_ses_main_esms.htm).

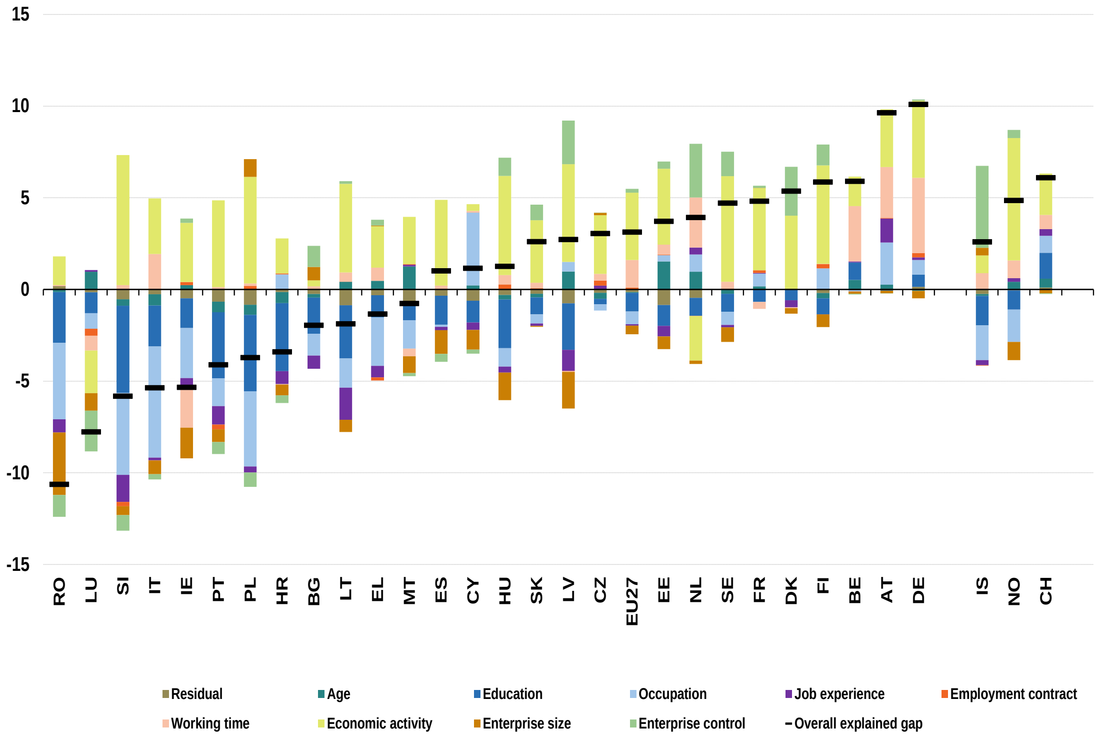
For some countries, the decomposition enables to explain even more than half of the unadjusted Gender Pay Gap. While, at the other extreme, some other countries record a negative explained Gender Pay Gap. This means that female employees present average characteristics on the labour market that are better paid than those of men. This happens for countries where women with lower education and skills refrain from engaging in the labour market, especially when there are few job opportunities.

The decomposition allows also looking more closely to each explanatory factor. As shown in Figure 2. Among the Member States, the explained part is mostly driven by the following three factors: economic activity, education and occupation. However, these factors have different explanatory effects.

The economic activity is playing a role too. For most of Member States, men tend to be employed in better-paid economic activities than women (sectoral segregation).

For education and occupation, countries generally record negative gaps, illustrating the impact of self-selection in the labour market: women who engage in the labour market tend to have a higher education level and take better-paid occupations than men.

Figure 2: Decomposition of the explained gender pay gap, 2018



The decomposition also allowed to analyse whether women and men had different wages for the same characteristics. Working part-time or having temporary contracts was generally more penalising for men than for women. The analysis of the coefficients of regressions for men and women for age and age^2 showed the impact of career breaks on the average earnings of women.

Conclusions

There are clear policy and statistical reasons to go beyond the unadjusted gender pay gap, both understanding the composition of it as well as looking to other indicators in the labour market that could complement the information we have today. By identifying and interpreting the causes of the GPG, policy actions in favour of gender equality can be better targeted.

With the new Regulation (EU) 2019/1700, information about earnings in LFS would be collected too, so it may open the possibility to explore gender pay gaps taken from household survey and consider variables such as care responsibilities, total working experience, household composition. However, this information needs first to be checked for quality and plausibility. They are likely to complement rather than substitute business surveys, which remain the reference source for earnings information hence GPG analyses.

EUROSTAT statistics offer an important tool to measure different types of segregation (occupational segregation, activity segregation, self-selection of women that decide whether they engage or not the labour market…) with a high quality based on Structure of Earnings Survey. Nevertheless, the measurement of possible discrimination through statistics remains a challenge, as it should be ideally analysed case by case.

References

[1] Gender pay gaps in the European Union — a statistical analysis — Revision 1, 2021 edition - Products Statistical working papers - Eurostat (europa.eu)

[2] Gender pay gap statistics - Statistics Explained (europa.eu)

New statistics on discrimination and gender based violence at EU level

Keywords: non-discrimination, equality, gender based violence, disability.

Introduction

Article 19 of TFEU[58] grants the EU the competence to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation ('the six grounds of discrimination'). The Commission has come forward with several policy initiatives related to these areas: Gender Equality Strategy, LGBTIQ Equality Strategy, EU Anti-racism Action Plan, EU Roma strategic framework, EU Strategy on combating antisemitism and fostering Jewish life, pay transparency, Strategy for the Rights of Persons with Disabilities and combatting violence against women.

Equality data' in the current policy framework refers to the population sub-groups characterised by the six grounds of discrimination, including their intersections. Adequate and sound equality data plays a vital role for effective policy monitoring and the fight against discrimination. However, policy makers pointed out that the current data framework limits the application of the core legal EU frameworks in this area and EU policies call for more and improved statistics on equality.

The European Statistical System (ESS) has for several years already been highly active in producing and disseminating equality statistics. It is important to note that significant progress is bound to be made in the coming years according to regulations and planned data collections.

In the short-term, Eurostat will increase its offer of the equality statistics and indicators in order to further promote and consolidate statistics based on the new datasets. In the medium and long term, Eurostat will work together with National Statistical Institutes (NSIs) and other stakeholders to increase further the offer of equality statistics and close the data gaps where possible. This article focuses on the first strand. Substantial methodological work and pilot studies were undertaken or are ongoing to support the extension and improvements of equality data in several areas:

1) ***a better harmonisation of concepts and terminology***. For sex / age / disability / country of birth or citizenship: definitions, taxonomies and implementing guidelines were already discussed extensively with NSIs, researchers and are already available in the core standard variables related to the Regulation (EU) 2019/1700. Special efforts were made to include common harmonised variables across domains with strong guidelines for better comparability. Further pilot studies will address a better coverage of collective households or extended

[58] Treaty on the functioning of the European Union

definitions for disability; improvements in the concepts for migrant background and ethnic origin as well as the clarification of terminology and taxonomies concerning 'sex' and 'gender'.

2) ***a more detailed dissemination of new indicators and multiple breakdowns*** at the appropriate level of granularity. An important step is the extension of demographic indicators via the population grid via Regulation (EU) 2018/1799 with detailed data broken down by age, sex and place of birth/residence one year ago. In addition, regulation (EU) 2019/1700 that covers data collected from samples will extend considerably equality data by disability and foreign background[59], including intersectional aspects. Further work will address issues of small sample size and disclosure control in order to extend the dissemination of multiple relevant breakdowns.

3) ***the collection of harmonised and comparable data on new indicators on experience of discrimination.*** The reasoning for not including certain sensitive variables are often justified by the sensitivity of the questions, cultural considerations, quality of the data and implementation costs. It is recommended, therefore, that countries should undertake a rigorous testing programme before attempting to collect such information in social surveys. The "doing no harm" principle is extremely relevant in this context that means that "data on personal characteristics should be kept safe and used only for the benefit of the groups they describe and society as a whole. Do not harm also means that nothing in the guidance notes should be interpreted as an invitation, encouragement or endorsement of any initiative or practice that seeks to discriminate against population groups and expose them to risks of serious human rights violations (or which has this effect)".

One important step forward to respond to these needs, is the Gender Based Violence Survey (EU-GBV) that will provide comparable data across Europe on the prevalence and dynamics of violence against women and other forms of inter-personal violence. Further progress in this area of indicators is foreseen in the eight-yearly LFS module on "the labour market situation of migrants"; the module on access to services in EU-SILC in which all six grounds of discrimination are collected as well as new statistics on experience of discrimination on-line (ICT).

4***) innovative methods for further consolidating existing indicators.*** As an interesting example, a new methodology has been applied to decompose the gender pay gap into different components to better understand the differences in pay between men and women. This work contributes to relevant policy indicators under the "equal pay for work of equal value principle".

The next section of the paper focus on the methodology and first results for the survey on gender-based violence very relevant in the framework of gender equality and the fight against discrimination. It presents some first results in this area as well as follow-up actions.

Methods and results

[59] Measured via country of birth and country of birth of parents

Gender Based Violence

The Council of Europe adopted, on 11 May 2011, the Convention on preventing and combating violence against women and domestic violence ([60]), the so-called 'Istanbul Convention'. The Istanbul Convention entered into force on 1 August 2014 and has been signed by all Member States and the EU. Up to now, 21 Member States have ratified it making it binding in most Member States of the EU. The EU's accession to the Istanbul Convention remains a key priority for the Commission. According to Article 11 of the Istanbul Convention, Member States acceding to the convention 'shall endeavour to conduct population-based surveys at regular intervals to assess the prevalence of and trends in all forms of violence covered by the scope of this Convention'.

Until recently, only the European Union Agency for Fundamental Rights (FRA) had collected survey-based data on Gender-based Violence with support from private companies. The survey was implemented in 2011-2012 on a randomly selected sample of 40,000 women across the 27 EU Member States and Croatia.

In order to answer the requirements of the Istanbul Convention and to also better answer policy requirements at EU level, the European Statistical System (ESS) composed of Eurostat and of the National Statistical Institutes (NSIs), started in 2016, for the first time ever, development work for an EU-GBV survey. Member States and experts from a range of relevant organisations and disciplines supported the development of the survey methodology resulting in a 'Methodological manual for the EU survey on gender-based violence against women and other forms of inter-personal violence' ([61]) published in 2021. The manual provides information and guidance on all the technical and methodological aspects of the EU-GBV starting with introduction of the main concepts as gender-based violence, interpersonal violence, different types of violence, relationship between victim and perpetrator and timeframe as well as main definitions used in the survey. The chapter on survey preparation and implementation has specific guidelines and methods for pre-testing the survey questions. Taking into account the sensitivity of the topic, more focus is put on recommendations for interviewers' training, managing interviewers' emotional distress and ensure the safety of the survey participants. The EU-GBV survey questionnaire is provided in the Manual accompanied with the description of each variable that is a valuable source of information for translating questions, providing additional instructions for interviewers and respondents, and constructing datasets. The chapter on indicators and dissemination provides guidelines on disseminating the EU-GBV data and describes the steps from validation to dissemination, data analysis and calculation of indicators. Finally, the chapter on quality reporting and assessment provides guidance how to assess and report the survey quality via the metadata handler tool for this survey. After testing the methodology in 2017-2019, the main survey implementation started in 2020, on the basis of methodological manual, including a common list of variables and associated questionnaire, common guidelines for data collection and common indicators to be produced.

The main purpose of the EU-GBV survey is to assess the prevalence of all forms of violence covered by the survey as well as frequency, intensity, and severity of

[60] Available at: https://rm.coe.int/168008482e

[61] Available at: https://ec.europa.eu/eurostat/documents/3859598/13484289/KS-GQ-21-009-EN-N.pdf/1478786c-5fb3-fe31-d759-7bbe0e9066ad?t=1633004533458

experienced violence. The survey includes questions about threats, physical, sexual violence and stalking by any person; psychological violence by partner; sexual harassment at work; and violence experienced in childhood. The target population of the EU-GBV survey is defined as people aged 18-74 who live in private households, with a focus on women. Men can be included in the target population by countries willing to do so.

The NSIs of 18 EU Member States (BE, BG, DK, EE, EL, ES, FR, HR, LV, LT, MT, NL, AT, PL, PT, SK, SI, FI) have implement / are implementing the survey nationally based on the EU methodological manual while the Italian NSI agreed to share the data based on its national Violence Against Women survey.

In order to get results for all EU Member States, the European Institute for Gender Equality (EIGE) and the European Union Agency for Fundamental Rights (FRA) will implement a separate data collection following the Eurostat methodological manual for the remaining eight Member States (CZ, DE, IE, CY, LU, HU, RO and SE).

Concerning the survey implementation in other countries, the NSI of IS, ME, MK, RS and XK will implement the EU-GBV survey nationally while BA and AL conducted some methodological work.

The first EU-GBV survey results from 7 NSIs (BG, FR, LV, LT, NL, AT, SI) have been issued on 25 November 2022. It is planned for data from all NSIs to be available during summer 2023 and for EIGE/FRA data to be released at the end of 2023/beginning of 2024.

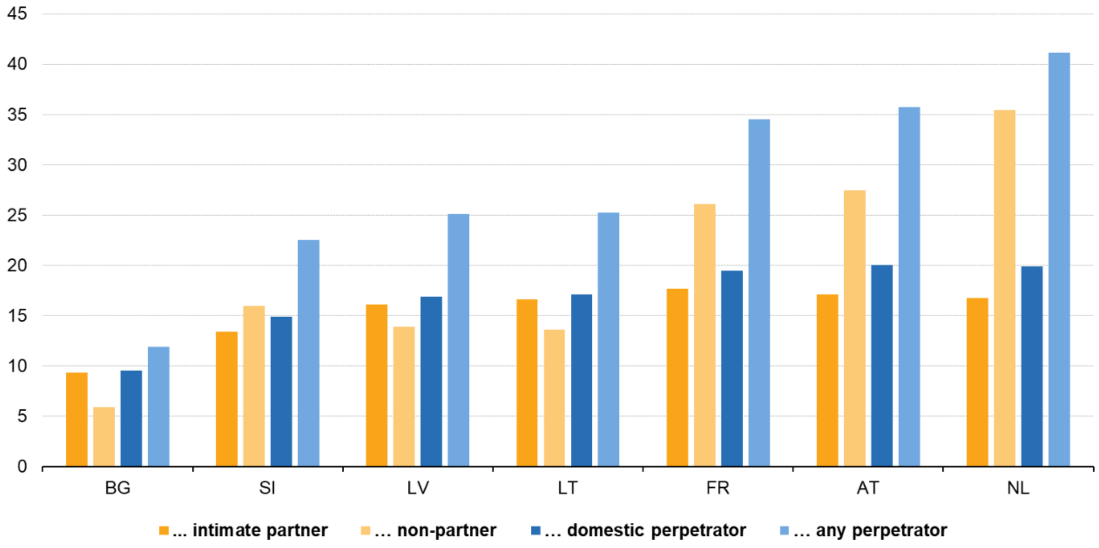
Based on previous research ([62]) one in three women has experienced physical or sexual violence from the age of 15 years old and only one third of victims of partner violence and one quarter of victims of non-partner violence contacted police or a support organisation following the most serious incidents of violence.

Data for the 7 countries for which data are disseminated show similar trend: at least 12 % of women in Bulgaria up to 41 % of women in the Netherlands have said that they have experienced physical violence (including threats) or sexual violence during their adulthood (see Figure 2). More victims of partner violence contact the police compared to victims of non-partner violence, but still less than one quarter of women reported to the police at least one incident of partner violence.

The extent to which violence is tolerated in the wider community might influence the number of women who are ready to share their violent experiences in the survey. Therefore, survey data itself might only be a close proxy to real prevalence. In France, Austria and the Netherlands, women are more ready to disclose violent experiences, especially violence from a non-partner. The analyses of non-partner violence by type shows that the higher prevalence in these three countries is due to a higher prevalence of degrading or humiliating sexual acts other than rape. Therefore, the prevalence of violence might be higher in some countries because women in these countries are more aware and more ready to share their experiences. The first results of the EU-GBV survey also show that younger women are more ready to share violent experiences they have faced. This trend is visible for all countries when comparing the prevalence of non-partner violence by age group.

([62]) FRA, Violence against women: and EU-wide survey. Main results, FRA 2014 (available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2014-vaw-survey-main-results-apr14_en.pdf)

Figure 2. Proportion of women (18–74) who have experienced physical (including threats) or sexual violence during adulthood by type of perpetrator



Source: EU-GBV survey, wave 2021

Conclusions

The EU-GBV survey is implemented on voluntary bases in 19 Member States based on common methodology. It is important to outline a strategy for the medium to long term in order to further consolidate statistics in this area. The Commission adopted this year a legislative proposal on combating Violence against Women and Domestic Violence[63]. This includes an article on data collection and research to require the implementation of a regular population-based survey using the harmonised methodology.

Further work in Eurostat and at ESS level will focus on the dissemination of new indicators and breakdowns, including intersectional. Quality and disclosures issues due to low sample size should be considered and analysed. In specific areas, where several obstacles persist in the data collection (e.g. sexual orientation, gender identity, religion and ethnic origin) further progress requires engaging with data producers, national statistical institutes and stakeholders. Guidelines with harmonised concepts and taxonomies on the six grounds of discrimination should be proposed at EU level building on current practices in several Member States as well as progress at international level. Finally, additional sources and innovative methodologies should be considered with the aim to mainstream equality data across relevant domains beyond social statistics.

References

[63] EUR-Lex - 52022PC0105 - EN - EUR-Lex (europa.eu)

[1] Eurostat, *EU survey on gender-based violence against women and other forms of inter-personal violence (EU-GBV) — first results*, Eurostat, 2022 (available at https://ec.europa.eu/eurostat/product?code=KS-FT-22-005)

Machine learning 1 (MANS2A.1)

Session Chair: **Sarah Bohnensteffen** (Destatis)

Early provision of economic short-term indicators using Machine Learning
Elena Rosa-Perez, David Salgado, Sandra Barragan *(National Statistical Institute-INE)*

Machine learning methods to support a modernized household budget survey
Boriska Toth, Ruben Mustad, Susie Jentoft *(Statistisk sentralbyrå)*

Machine learning and wealth measurement : an experiment on housing wealth of French households
Olivier Meslin *(National Institute of Statistics–INSEE)*

Early provision of economic short-term indicators using Machine Learning

Keywords: Nowcasting, short-term business statistics, machine learning, microdata, accuracy

Introduction

Short-term business statistics (STS) are the earliest official statistics released to show emerging trends in the European economy. The European Statistical System (ESS) is interested in improving its timeliness and we propose as an illustrative example an early estimation of the monthly Industrial Turnover Index (ITI) [1].

Our approach [2] to nowcasting focuses on the traditional survey microdata taking into account (i) that the reference period has to be over in order to measure (not to predict) the economic activity, (ii) that the samping units may need some time to gather the data thus extending the data collection period, and (iii) that the data editing phase (see e.g. [3]) requires also some time consuming tasks, including a minimal amount of interactive editing necessary to guarantee the accuracy and the quality of the data.

Methods

The ITI aim is to measure the evolution of the industrial activity through the turnover, provided by the establishments whose main activity is included in Sections B or C of CNAE-2009 (Spanish adaptation of the NACE Rev.2). The survey provides aggregates at national and NUTS2 levels and by economic activity. Sampling units are selected according to a cut-off sampling design with a sample size of around 12000 units per month. The indices follow a fixed-base Laspeyres index.

This sampling design is not probabilistic and provides zero design-based variance for the traditional estimator $\hat{Z}_{U_d}^{my} = \sum_{k\in s_d^{my}} \frac{z_k^{my}}{\pi_k^{my}} = \sum_{k\in U_c^{my}} z_k^{my}$, where m denotes the month, y the year, z_k is the turnover of the k-th unit, and we can always write $s^{my}{}_d = U_{c,d}{}^{my}$, showing the dependence on the cut-off values (subscript c). Data collection for reference month m starts at $m + 1d$, and data are batch-processed by the survey managers at $m + 20d$, $m + 27d$, and $m + 37d$. Post-collection data editing is conducted upon these batches, and press release takes place at $m + 51d$.

At time t, a concrete subsample of respondents has provided their responses $r_d(t) \subset s_d(t)$. Editing tasks are carried out from the data collection activity to the final estimation phase. These editing tasks may change the value z_k^{my} of a given unit during this editing phase. We shall denote by $z_k{}^{my,\mathrm{val}}$ the final validated value entering into the computation of the first official release of the ITI. Similarly, we shall denote by $z_k{}^{my,\mathrm{val}}(t)$ the value of variable z for unit k at the time t of the editing strategy for the reference month m and year y.

Estimators

The population total, $Z_{U_d}^{my}$, estimated for each reference month m and year y after collecting and editing the whole sample, can be decomposed at any time t as

$$\mathrm{Zb}U_d = Z_{U_{c,d}} = \mathrm{X}\ z_k\mathrm{val} = \mathrm{X}\ z_k\mathrm{val} + \qquad \mathrm{X} \qquad z_k\mathrm{val}. \qquad (1a)$$

$$k\in U_{c,d} \qquad k\in r_d(t) \qquad k\in U_{c,d}-r_d(t)$$

This decomposition can only be actually computed after finishing the collection and editing phases, since we need the final validated values z_k^{val}. Our goal is not to wait until both data collection and data editing are concluded to produce an early estimation of the ITI with the ongoing collected and edited information. Taking into account the values already known and predicting these yet unknown, we decompose this estimate as follows:

$$Z_{U_{c,d}} = \sum_{k\in r_d(t)} \left[z_k^{\mathrm{ed}}(t) - e_k^{\mathrm{meas}}(t)\right] + \sum_{k\in U_{c,d}-r_d(t)} \left[\widehat{z}_k^{\mathrm{val}}(t) - e_k^{\mathrm{pred}}(t)\right], \qquad (1b)$$

where $e^{\mathrm{meas}}{}_k \quad (t)$ denotes the measurement error $e^{\mathrm{meas}}{}_k \quad (t) = z_k^{\mathrm{ed}}(t) - z_k^{\mathrm{val}}$ and $e^{\mathrm{pred}}{}_k\,(t)$ denotes the prediction error $e_k^{\mathrm{pred}}(t) = \widehat{z}_k^{\mathrm{val}}(t) - z_k^{\mathrm{val}}$.

The proposed estimator for the population total $Z_{U_d}(t)$ with data collected up to time t is given by

$$\mathrm{Zb}U_d(t) = \mathrm{X}\ z_k\mathrm{ed}(t) + \qquad \mathrm{X} \qquad \mathrm{Zb}_k\mathrm{val},\xi_p(t), \qquad (2a)$$

$$k\in r_d(t) \qquad k\in U_{c,d}-r_d(t)$$

where $\widehat{Z}_k^{\mathrm{val},\xi_p}(t)$ is the random variable representing the prediction for value $\widehat{z}_k^{\mathrm{val}}(t)$according to prediction model ξ_p. This estimator produces estimates of the form:

$$zU_d(t) = \mathrm{X}\ z_k\mathrm{ed}(t) + \qquad \mathrm{X} \qquad z\mathrm{b}_k\mathrm{val},\xi_p(t). \qquad (2b)$$

$$k\in r_d(t) \qquad k\in U_{c,d}-r_d(t)$$

Notice that, when compared with decomposition (1b), this amounts to neglecting measurement errors $e^{\mathrm{meas}}{}_k\,(t)$ and considering $e^{\mathrm{pred}}{}_k\,(t) \approx 0$. This way, we just need to build only one prediction model ξ_p.

It is important to notice that this formula is specific for cut-off sampling and the generalization to other sampling designs needs a modification, e.g. for the HT estimator, for the ratio estimator, for the GREG estimator, or for the Sanguiao-Zhang estimator.

Regressors and the prediction model

All the regressors $x_k^{(p)}\ {}_{p=1,2,\ldots,P}$ used in the prediction model ξ_p have been constructed computing on survey microdata and/or paradata from the ITI survey itself, with the exception of some aggregates from the Industrial Price Index and Industrial Production Index surveys. Neither administrative data nor new digital data source have been used at all. The regressors can be classified in terms of their semantic content as:

Geographical variables: Edited and validated values of the NUTS2/NUTS3/LAU codes of the enterprise and the establishment.

Time variables: Month and year of the reference period, batch, and number of months verifying some property.

Economic Activity variables: Edited and validated values of the 4-digit, 3-digit, 2-digit and 1-digit codes of the CNAE-2009 of the enterprise and the establishment, and binary related variables.

Target-Related variables: Validated values of the turnover and moving averages thereof, 0.95 quantiles of the moving averages at NUTS2 level and for several economic activity breakdowns, coefficients of variation, minimum, maximum and mean values, standard deviation for different breakdowns, and rates for different variables and different breakdowns, etc.

External Survey variables: monthly Industrial Production Index and monthly Industrial Price Index Indices and rates for different breakdowns.

Regarding the prediction model, our first choice was to use random forest. Next, to improve accuracy and, especially, to deal with outliers, we have chosen to use boosting. Finally, among the different choices we focus on the gradient boosting algorithm and, in particular, on the LightGBM version.

The production process for the early estimates of the ITI has been designed by following international production standard models and the approach about the use of functional modularity stated in the working paper with the aim of producing an industrialised standardized production process.

Results

The main results of this pilot study comprise the series of early estimates of the ITI broken-down according to usual production conditions as well as their corresponding yearly and monthly variation rates for the three batches together with their respective root mean squared error. Figure 1 contains a representation for the evolution of the nowcasted indices as the data collection and data editing field work is conducted, specifically for the three aforementioned process batches.

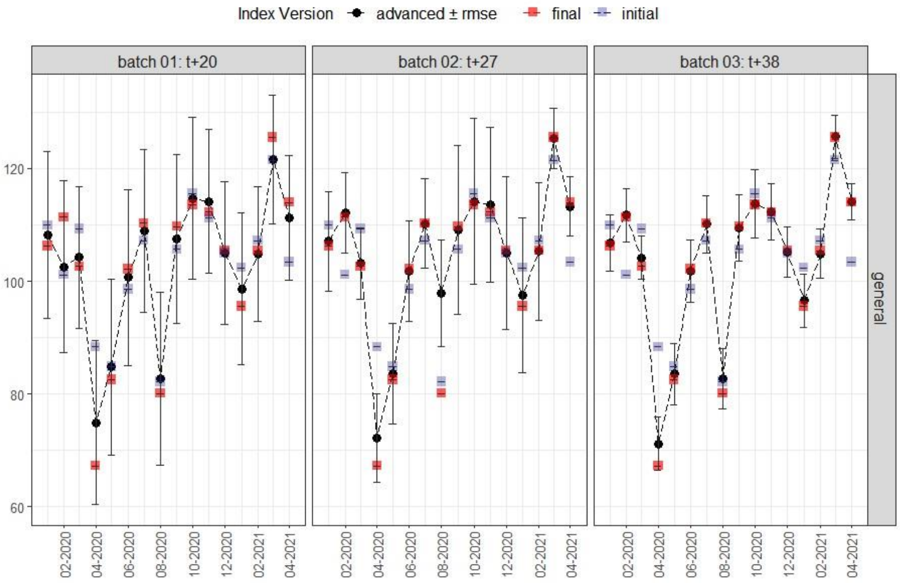


Figure 1: General Index Series from Jan, 2020 to April, 2021.

We have also computed the (nearly) daily evolution of the national (general) index for the successive reference time periods in 2020 as data are collected and edited (batches are especially visible at days $t+20$, $t+27$, $t+39$). Four screenshots of an animation is depicted in gure 2.

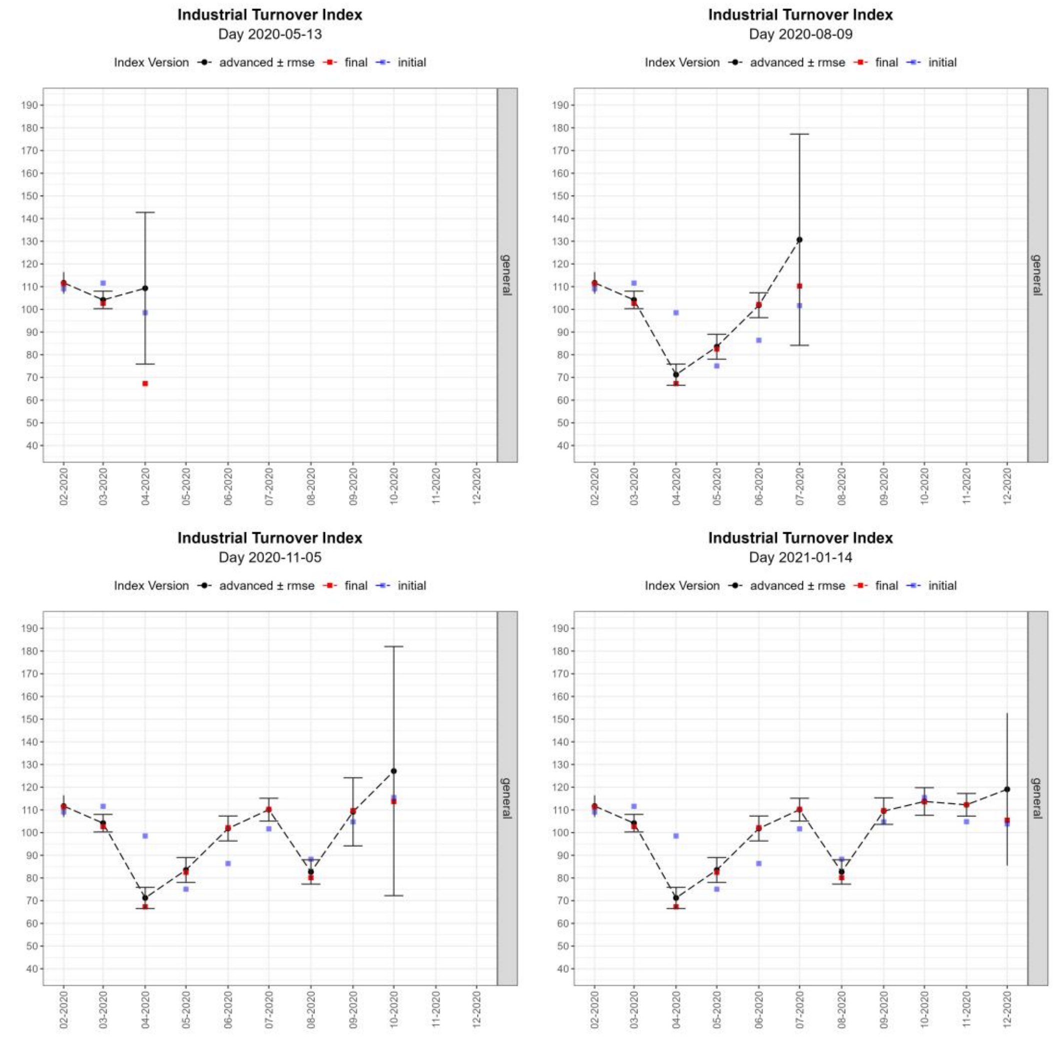


Figure 2: General Index Series from Jan, 2020 to April, 2021. Screenshots from an animation.

Conclusions

Early conclusions from this experimental exercise are:

Statistical learning algorithms with high predictive capacity on early data allow us to improve timeliness under a controlled compromise of accuracy. Subject-matter knowledge is crucial for its incorporation into the statistical model, especially on the selection and construction of regressors. Both microdata and paradata are relevant for high-quality predicted values. There is still a wide range to further investigate improvements on the predictions both on exploring more algorithms and further complementary data. Representative outliers are extremely hard to model and predict. They clearly need management and processing by subject-matter experts.

Measurement errors have an important role in a good quality estimation so they have to be taken under control.

High-quality experimental statistics can be produced with traditional survey data by using novel statistical methods in Official Statistics.

References

[1] INE. Industrial Turnover Indices & Industrial New Orders Received Indices. Base 2015, 2018. https://www.ine.es/en/metodologia/t05/t0530053_2015_en.pdf.

[2] S. BarragÆn, L. Barre nada, J.F. Calatrava, J.C. GÆlvez SÆenz de Cueto, J.M. Mart n del Moral, E. Rosa-PØrez, and D. Salgado. Early estimates of the industrial turnover index using statistical learning algorithms, 2022. Working Paper.

[3] UNECE. Generic statistical data editing model - version 2.0, 2019. URL https://statswiki.unece.org/display/sde/GSDEM. Technical Report.

Machine learning methods to support a big data-based household budget survey

Keywords: supervised machine learning, new big data sources, imputation, matching and data linking, household budget survey.

Introduction

A household budget survey (HBS) is a survey of consumption and expenditure of households. Eurostat requirements dictate that EEA countries publish HBS statistics where household expenditure is broken down into different COICOP (Classification of Individual Consumption According to Purpose) categories at the 5-digit COICOP level, and that total consumed weight is also reported for each 5-digit COICOP category that represents groceries. For instance, the COICOP category 01.1.4.7 refers to “milk-based desserts”. Consumption is reported across all goods and services.

Norway’s 2022 household budget survey (hereby referred to as HBS 2022) is a highly innovative household budget survey for Norway and more generally, as it combines a sample survey with novel big data sources [1]. Respondents in the modernized survey use a smartphone app to enter all their purchases over a one-week period, by choosing to either scan in receipts or manually enter items. In addition, two major big data sources from private enterprise are being incorporated: 1) *store transaction records*, which give detailed purchase transaction information recorded at cash registers for purchases made in the year of the survey from all of Norway’s major supermarket chains and retailer stores, and 2) *bank transaction records*, which give transaction records for all purchases made by debit card in Norway in the year of the survey. Purchases can be linked to the demographic group of the purchaser through a privacy preserving, pseudonymized process of matching store records of transactions to bank records of debit card transactions occurring at the same store with the same timestamp. The use of these massive new data sources in an automatized workflow can allow for far more frequent publishing of HBS statistics as well as more fine-grained information on purchasing habits in specific demographic groups, regions, and times of year, due to the vast amount of data in each stratum. [2] describes methods for making statistical inference in the context where a potentially biased big data source is available alongside a traditional sample survey.

HBS 2022 uses a sample size of 12,000 respondents, each tasked with reporting every purchase made during the course of one week. In most (~90%) cases, the respondent chose to scan in the receipt instead of manually entering items. However, the scanned receipts often have poor image quality and errors result from the optical character recognition process. An advantage of the big data transaction sources is that for roughly half the scanned survey receipts, the matching receipt could be found in store transaction records, which is free of scanning errors and has a richer set of variables. All receipts from the survey include the names and prices of items purchased, demographic variables for the respondent, date and time, and store information.

Machine learning is vital to automatically processing data from such a large survey, and we describe two main workflows we are developing and testing at Statistics Norway: supervised learning for predicting COICOP categories from the text of an item (and possibly other variables); and imputation for imputing values for weight, price, or COICOP code when these were not identifiable from a purchase record.

Once the data has been processed, we have 3 major downstream sources for producing statistics:

- Manually coded purchase information from the survey
- Automatically transcribed scanned receipts from the survey, about half of which can be replaced by a clean store purchase transaction record
- Store purchase transaction records for a whole year in Norway linked to non-identifying demographic information of the purchaser

There is currently some uncertainty regarding the timing of the big data sources' arrival, so in this work we describe the machine learning problems as applicable to the two survey-based sources.

Methods

Supervised learning for predicting COICOP

There are roughly 33000 unique purchased items out of 150,000 total items from scanned receipts data that have arrived by midyear of the survey. We have developed and are continuing to improve a machine learning workflow for predicting COICOP codes from the name of an item (and possibly other variables).

Manual labelling of large quantities of survey data is prohibitively expensive, so we constructed a training set from 8 different sources. This includes 3000 manually labelled survey items, items users manually coded in a pilot survey, a highly accurate dataset of COICOP-coded items from the consumer price index group at Statistics Norway that were automatically or manually coded [3], dictionaries that associate COICOP codes with keywords, and a 1.5-million item dataset of imported goods that could be mapped to a unique COICOP code by translating between several classification schemes. We use a test set of 1000 manually labelled, randomly chosen items from scanned survey receipts.

After some preprocessing steps, the names of items need to be converted into numerical feature vectors that can be input to algorithms. A typical procedure from natural language processing is used to vectorize items using a "bag-of-words" model. The item names were converted into feature vectors either using unit weighting or tf-idf that gives more weight to rarer, more descriptive terms. We implemented several common algorithms for supervised learning (XGBoost, SVM, Random forest and Logistic regression) in Python using the scikit-learn and xgboost modules [4,5].

Imputation for variables that are not identified in the purchase record

It often happens that the price, weight, or COICOP code corresponding to an item cannot be identified but must be imputed from the rest of the data. Imputation methods can range from simpler methods like matching to items having similar names or taking a mean value in a stratum, or they can be more involved, such as creating a data-generating model that models the missingness mechanism explicitly.

- **Missing weight**. The weight must be imputed if the name of a food item does not give a weight and the item is sold by count (vs by weight). We have implemented an algorithm for finding a list of best matches of an item name to other items in the survey or transaction data, measured by edit distance between preprocessed item names. These matches can be used to suggest either weights directly, or suggest the weight/price ratio for the matching food item to calculate weight. Since we often cannot ascertain the actual weight of an item, supervised learning will not work. We propose a rule-based method to calculate an imputed weight from the list of suggestions.

- **Missing price**. Users manually coding often enter multiple items as a comma-separated list when prompted for a single item in the survey. This means that when we separate the items, the total price of the list must be distributed among the individual items. We propose finding a list of the best matches and rule-based imputation again, this time manually checking when the sum of suggested prices is far from the total price over list.

- **Item name entered is associated with multiple COICOP codes**. This is the most complex situation, when a user summarizes multiple purchases by entering for instance “various foods” with a total price attached. We have no reference for what was actually purchased (manually entered receipts cannot be linked to transaction records), so supervised learning is not possible. However, we can fit a model for $f_C(X)$, the distribution of spending that users did not list correctly over some set of COICOP codes, given regressors X (for instance, store, time of year, and demographic characteristics), through observing spending on correctly registered items. We estimate $Pr_C(X)$ by making usual assumptions on the conditional independence of the missingness mechanism.

Results

Through cross-validation, we found unit weights and 2- and 3-character grams to work the best for predicting COICOP on scanned survey data. As Table 1 shows, accuracy was around 60% for predicting COICOP codes at the 5-digit level when working with transcribed item names (not using any matching transaction records). It is pertinent to compare this to a double-coding experiment where two humans labelled items independently, in which the COICOP codes only matched 87% of the time. [5]

Table 6. Performance of COICOP predictors on scanned survey data (matching transaction records not used)

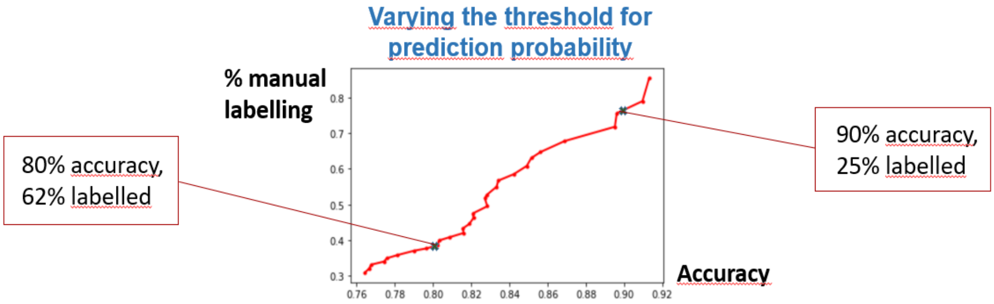
	Extractor	F1-weighted	Accuracy
Logistic regression	good's name, CV-ch33	0.585	0.592
Random forest	good's name, CV-ch23	0.573	0.583

One very useful feature of most supervised learning algorithms is that they produce a prediction probability representing the algorithm’s confidence in the prediction. This value can be used as a threshold to pick a subset of predictions having some expected desired level of accuracy. For example, for the scanned survey data, if we pick a threshold so that accuracy is around 90% for items having prediction probability above the threshold, we can label about 25% of the data using machine learning. Figure 1 shows the tradeoff between accuracy vs percentage automatically labelled. We can use this idea of taking a subset of predictions having prediction probability above a threshold for a human-in-the-loop implementation of automatic prediction that we’re developing at for HBS 2022: machine learning is used to give a human coder the low-confidence cases for which labelling has the most value in updating the model, and the human's labels are fed back into the training set for retraining the model.

Clearly, performance improvements are called for if we are to save a substantial amount of the burden of human labelling. One promising result we saw was that predictors trained and tested on the store purchase transaction data have considerably higher accuracy (80-90% depending on the features used), and that using a grouping variable available in the transaction data as a feature improved performance by about 6% as compared to just using item names. This suggests that we can expect significant performance gains once we use the clean item names and extra variables appearing in the transaction data.

For imputation, we found that 60% of survey items had a very close match in the rest of the survey or transaction data, while about 70% had multiple good matches.

Figure 15. The tradeoff between accuracy vs. percentage automatically labelled



Conclusions

Statistics Norway's Household Budget Survey 2022 is a highly modernized HBS that is dependent on machine learning to be able to process the volume of data. We described our work on supervised learning for COICOP classification and on imputation for missing values. The classifiers have shown modest performance when tested on noisy data from scanned survey receipts. However, experiments on COICOP classification on store purchase transaction records are promising and suggest that when the scanned receipts are linked with transaction records we can expect improvements. This work is active ongoing work at Statistics Norway and results are steadily improving. Various other proposals are underway for improving performance, such as using large and noisy auxiliary sources in a selective way in the training data. For imputation problems in HBS2022, we have seen good results in being able to find multiple reasonably close matches in the survey data for food items, and we are implementing several proposals.

References

[44] M. Runningen Larsson and L. Zhang, Using non-survey big data to improve the quality of the household budget survey, Proceedings of NSM 2022.

[45] L. Zhang, Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society), pp.571-588, 2021.

[46] K.H. Myklatun, Utilizing Machine Learning in the Consumer Price Index. Proceedings of NSM 2019.

[47] D.M. Müller, Classification of Consumer Goods into 5-digit COICOP 2018 Codes. Master’s thesis, Norwegian University of Life Sciences, Faculty of Chemistry, Biotechnology, Food Science, 2021.

[48] S. Jentoft, B. Toth, and D.M. Müller, From manual to machine: Challenges in machine learning for COICOP coding. Proceedings of NSM 2022.

Machine learning and wealth measurement : an experiment on housing wealth of French households

October 16, 2022

1 Introduction

Accurately measuring wealth, wealth inequality and its evolution is of considerable importance for researchers, policymakers and the general public, particularly in a context of rising inequality. Data on wealth, however, is patchy and sometimes unreliable, particularly at the top of the distribution. Household surveys provide detailed information on the assets of most households, but struggle to capture the true distribution of wealth because of the underrepresentation of wealthy households and underreporting of assets. Alternative methods such as the estate multiplier method (Kopczuk et Saez (2004) and Piketty (2011)) and the capitalization method (Saez et Zucman (2016)) tried to overcome these limitations by using another source of data (data on inheritance and data on capital income) to reconstitute the distribution of wealth based on statistical assumptions. This paper departs from this literature by exploring a promising new avenue: the *direct* estimation of wealth *at the household level* based on exhaustive administrative data.

The development of exhaustive databases on wealth must overcome two significant methodological challenges. First, data on wealth is typically scattered in many different sources so that developing databases on households' wealth involves combining large datasets at the individual level. Second, the market value of assets is most often unobserved and can be measured only when a transaction takes place. This may be unconsequential for some classes of assets that can be easily estimated with external information (eg, listed shares), but raises a serious issue for some other classes of assets (real estate, unlisted companies, unincorporated business assets). As a consequence, the definition of an appropriate valuation method for these assets is a major challenge for the development of exhaustive databases on wealth.

In this project, we show how these two challenges can be addressed in the case of French households' housing wealth. The way we address the first challenge is described in detail in André et Meslin (2021) and is not part of our proposal for NTTS 2023. Our NTTS proposal focuses on the second challenge: our ambition is to demonstrate how machine learning algorithms can be leveraged by economists and statisticians to estimate the market value of dwellings, so as to build a better measurement of housing wealth than the one available in wealth surveys.

In this project we make three distinct contributions. First, it introduces a new exhaustive database on housing wealth, opening the way for further research. Second, we introduce a new methodology to estimate housing wealth using exhaustive administrative data. Third, we show that housing wealth is significantly more concentrated than what survey data suggested so far.

2 Measuring households' housing wealth

Building on the growing literature on the potential uses of machine learning techniques by economists (Varian (2014); Mullainathan et Spiess (2017); Athey et Imbens (2019)), we use machine learning algorithms to predict dwellings' market values. More precisely, we train a valuation algorithm on all real estate transactions over the 2015-2019 period and then use this algorithm to extrapolate the 2017 market value of all dwellings located in France.

2.1 Data sources

In this project, we use three exhaustive administrative data sources:

- Land registry data (*données cadastrales - fichiers Majic*): this data describes all properties located in France and contains personal information on their owners (first and last names, date and place of birth, address). All properties are geocoded.
- Household data (*Fidéli database*) : this database contains detailed information on all resident households: personal information on households' members and detailed data on income (wages, pensions, capital income, social transfers).
- Real estate transactions data (*Demandes de valeurs foncières*): this data contains information on real estate transactions (market value and some property features). All transactions are geocoded.

2.2 Why exactly is there an asset valuation method challenge?

Given that housing prices prediction has been the subject of an overwhelming literature in machine learning (see Zulkifley, Rahman, Ubaidullah, et Ibrahim (2020); Mohd, Jamil, Johari, Abdullah, et Masrom (2020); Pagourtzi, Assimakopoulos, Hatzichristos, et French (2003) for surveys), it may seem surprising that we present the definition of an appropriate valuation method as a serious methodological challenge. The nature of this challenge becomes clearer when we emphasize the two key differences between the literature and our project.

First, whereas virtually all published papers model housing prices at the city, metropolitan area or regional level, we want to model housing prices at the *country-wide* level and derive reliable market value estimates *for all dwellings located in all parts of the country*, even remote and rural ones. This difference raises new questions related to the spatial heterogeneity of housing markets. For instance, dense areas are typically characterized by large price variations with respect to location, whereas the impact of location is much smoother in rural areas. In other words, the scale of the relevant housing market varies considerably over the country. Can a model accurately account for this heterogeneity of market scales? More generally, should the country be partitioned for modeling purposes, and if so, how to define this partition in a non-arbitrary way? Another issue pertains to the fact that transaction data is not a random sample of properties, in particular in geographical terms: areas with frequent real estate transactions (mostly large cities) are overrepresented, whereas areas with significant numbers of dwellings but infrequent transactions (mostly rural areas) are underrepresented. How can we be sure that a model will perform reasonably well on all areas, even underrepresented ones? Last, the determinants of housing prices may vary significantly at the national level (think of the differential effect of a swimming pool on residential prices in southern France and in Northern France). How can a model account for that?

Second, whereas the objective of the existing machine learning literature consists in predicting accurately housing prices at the *dwelling level* (dwelling by dwelling), our project uses housing price predictions only as an intermediate step towards the final goal: measuring housing wealth at the *household level*. As a consequence, standard algorithms may need to be adjusted for at least two reasons:

- Algorithms are likely to to underestimate the market value of properties with high unobserved quality (and conversely, overestimate the market value of low quality properties). This may induce a systematic downward bias in the housing wealth of high income households (and conversely, a systematic upward bias in the housing wealth low income households) if property quality is positively correlated with income.
- Algorithms may perform poorly on luxury properties, as these properties account for a small share of transactions and thus do not have much weight in the training process. Again, this problem may be inconsequential for a dwelling-by-dwelling approach, but having it matters crucially for our project as luxury properties may have very high unit values and are highly concentrated among households at the top of the housing wealth distribution.

2.3 Addressing the asset valuation method challenge

We address the valuation method challenge by defining a three-step machine learning pipeline:

- First, an algorithm is trained to predict the local average price per squared meter at a very disaggregated level, using only geographical coordinates as features.
- Second, another algorithm is trained to learn the difference between the price per squared meter and the local average price predicted by the first step, using dwellings' characteristics (floor area, number of rooms, construction period, garage, swimming pool...), neighbourhood caracteristics (median income, poverty rate...) and owner's income as features.
- Third, two auxiliary algorithms are used to improve the predictions for luxury dwellings. First, a classification algorithm is used to detect top-end properties (dwellings belonging to the top 1% of the distribution of dwelling unit value). Second, a variant of the algorithm presented in the second step is trained with heavy weights on top-end properties, and is then used to predict their market value.

This valuation pipeline aims to address the methodological issues explained above: the first step reflects accurately the geographical structure of real estate prices; the second step relates transaction prices with dwellings', neighbourhoods' and owners' characteristics in a flexible way; the third step estimates accurately the market value of top-end properties. This pipeline is trained separately for flats and houses using all real estate transactions on dwellings observed in France between 2015 and 2019 (3.5 million transactions). All steps use a standard gradient boosting algorithm (XGBoost). Hyperparameters of each step were chosen by cross-validation. The predictive performance of the valuation pipeline valuation procedure is finally measured on a test set unused in the training process.

3 Results

3.1 Predictive performance of the valuation pipeline

Tables 1 and 2 presents some descriptive statistics on the predictive performance of the valuation pipeline, as measured on the test test. The performance is higher for flats than for houses: the predicted price is less than 20% away from the observed price for 72% of flats, but for only 61% of houses. The valuation pipeline performs well for more expensive dwellings, but poorly for cheap dwellings (especially houses).

Table 1: Predictive performance on the test set

log(price per m^2)	Houses	Flats
R_2	0.72	0.87
$RMSE$	0.33	0.24

Sources: Fichiers Majic 2017, RCS, Fidéli 2017, DVF 2015-2019, authors' computations.

Table 2: Share of dwellings predicted with an error larger than 20% of the observed price

		Price quintile				
	Total	Q1	Q2	Q3	Q4	Q5
Houses	39.4%	68.3%	41.7%	30.3%	26.3%	30.5%
Flats	28.4%	43.3%	30.6%	23.0%	21.4%	23.4%

Sources: Fichiers Majic 2017, RCS, Fidéli 2017, DVF 2015-2019, authors' computations.

3.2 Predicting the market value of all dwellings

The valuation pipeline is then used to predict the 2017 market value of all dwellings (19.78 million houses and 17.35 million flats). Figure 1 shows that the average market value of dwellings increases with the owner's standard of living, particularly at the top of the distribution. This finding confirms that accounting for top-end properties is essential to measure accurately the distribution of housing wealth.

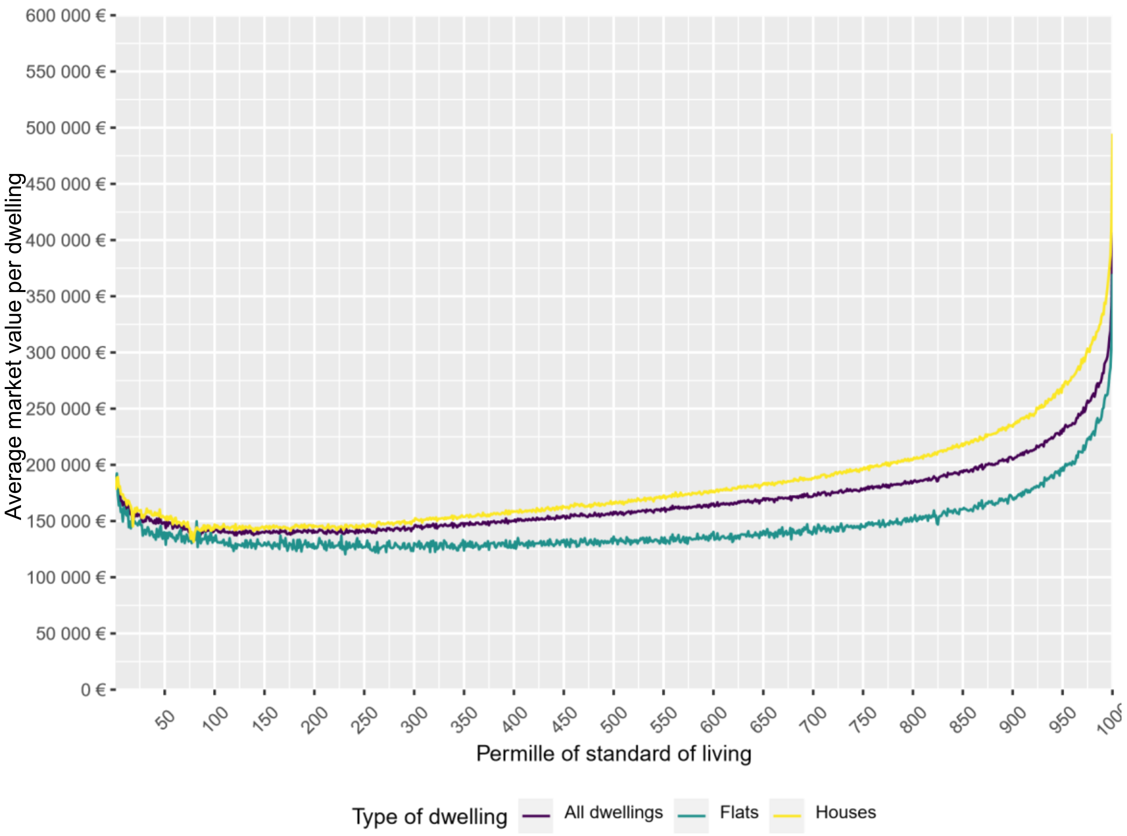
3.3 Concentration of housing wealth

Table 3 compares the concentration of gross housing wealth as measure by the 2017 French wealth survey and by the administrative database. Housing wealth appears to be significantly more concentrated than what the survey suggested: the share of aggregate housing wealth owned by the top 10% of housing wealth is almost 6 percentage points higher in the administrative database. Half of this difference comes from the top 1%, suggesting that the wealth survey does not measure accurately the top part of the housing wealth distribution.

4 Conclusion

Though additional work is still needed to get a fully-fledged valuation methodology, this project demonstrates how machine learning methods can be leveraged by economists and statisticians to improve the measurement of wealth. As a final comment, we want to stress that this kind of exercise is not a straightforward application of existing machine learning algorithms, but must rather be seen as a careful adaptation of these algorithms to new questions they were not designed to answer. In this process, it is essential that practitioners rely on a sound economic and statistical understanding of the final scientific objective.

Figure 1: Average market value of dwellings and standard of living



Sources: fichiers Majic 2017, Fidéli 2017, DVF 2015-2019, authors' computations.

Table 3: Comparison with wealth survey

Group of households	Share in gross housing wealth	
	Wealth survey	Exhaustive database
Top 10% of housing wealth	42.1%	47.9%
Top 5% of housing wealth	28.0%	33.0%
Top 1% of housing wealth	10.1%	12.9%

Sources: HVP 2017, fichiers Majic 2017, Fidéli 2017, DVF 2015-2019, authors' computations.

References

André, M., et O. Meslin (2021): "Et pour quelques appartements de plus: Étude de la propriété immobilière des ménages et du profil redistributif de la taxe foncière," *document de travail, Insee*.

Athey, S., et G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Kopczuk, W., et E. Saez (2004): "Top Wealth Shares in the United States, 1916-2000: Evidence From Estate Tax Returns," *National Tax Journal*, 57(2), 445–487.

Mohd, T., N. S. Jamil, N. Johari, L. Abdullah, et S. Masrom (2020): "An overview of real estate modelling techniques for house price prediction," *Charting a Sustainable Future of ASEAN in Business and Social Sciences*, pp. 321–338.

Mullainathan, S., et J. Spiess (2017): “Machine learning: an applied econometric approach,” *Journal of Economic Perspectives*, 31(2), 87–106.

Pagourtzi, E., V. Assimakopoulos, T. Hatzichristos, et N. French (2003): “Real estate appraisal: a review of valuation methods,” *Journal of Property Investment & Finance*.

Piketty, T. (2011): “On the long-run evolution of inheritance: France 1820–2050,” *The Quarterly Journal of Economics*, 126(3), 1071–1131.

Saez, E., et G. Zucman (2016): “Wealth inequality in the United States since 1913: Evidence from capitalized income tax data,” *The Quarterly Journal of Economics*, 131(2), 519–578.

Varian, H. R. (2014): “Big data: New tricks for econometrics,” *Journal of Economic Perspectives*, 28(2), 3–28.

Zulkifley, N. H., S. A. Rahman, N. H. Ubaidullah, et I. Ibrahim (2020): “House Price Prediction using a Machine Learning Model: A Survey of Literature.,” *International Journal of Modern Education & Computer Science*, 12(6).

Social statistics (GASP2A.2)

Session Chair: **Francesca di Iorio** *(University of Naples)*

Subjective poverty in the European Union: the EU-SILC approach
Estefania Alaminos, Agata Kaczmarek-Firth, Rasim Ryustem (*Eurostat*)

Chatbots to 'talk' with public data: the case of the Luxembourg Income Study *(LIS)*
Marcos Gómez Vázquez (*Open University of Catalonia*), Jordi Cabot *(ICREA - UOC),* Josep Espasa Reig *(LIS)*

Adding a new mode to the mix: Implementation and evaluation of CAWI in EU-SILC
Angela Hammer, *Marlene Blüher (Statistics Austria)*

Tracing respondents' footpaths – How user journey analysis can offer new insight into survey experience
Katharina Roßbach, Anna Lena Keute *(Statistics Norway)*

Subjective poverty in the European Union: the EU-SILC approach

Keywords: subjective poverty, at-risk-at-poverty rate, severe material and social deprivation, EU-SILC.

Introduction

The main focus of European Union statistics on income and living conditions (EU-SILC) is the measurement of objective poverty based on household income and the relative poverty threshold. However, EU-SILC also includes information on subjective poverty such as the ‘ability to make ends meet’.

EU-SILC contains both longitudinal and cross-sectional annual information [1]. Therefore as a panel data collection, it allows the analysis of the household and individual situation before, during and after the COVID-19 pandemic. This gives the possibility to monitor both subjective and objective poverty in times of the global turbulence.

In this article, EU-SILC 2018-2021 data series on both objective and subjective poverty will be analysed. The methods will first be presented before results for subjective poverty are provided. The paper will conclude with comparative results between subjective and objective poverty.

Methods

The analysis of subjective poverty is based on two variables collected in EU-SILC. The first variable is “Ability to make ends meet” collected continuously since the introduction of EU-SILC in 2003. The aim of the variable is to collect respondents’ opinion about the level of difficulty experienced by the household in making ends meet. The variable is categorical and has six answer modalities: 1-with great difficulty, 2-with difficulty, 3-with some difficulty, 4-fairly easily, 5-easily, 6-very easily. The second variable used in the analysis is “Lowest income to make ends meet” included in EU-SILC until 2020. This variable collects information about the amount respondents consider to be the minimum net income that would allow their household to ‘make ends meet’. In the analysis, it is referred to as subjective minimum income.

Subjective poverty is commonly defined in the related literature, as the individual’s perception on his/her financial/material situation. However, there is not a common approach on how to measure it. Therefore, in this work the authors suggest three different approaches based on EU-SILC for describing the subjective poverty phenomena.

Difficulties to make ends meet - SP1

The first approach is to observe the time series of the variable “Ability to make ends meet” and especially the answer modalities 1-with great difficulty, 2-with difficulty.

Inability to make ends meet based on household income – SP2

The second approach will compare the subjective minimum income to make ends meet with the households’ actual disposable income. If the household income is lower than the income considered as minimum to make ends meet (by the household), such a household will be considered in subjective poverty.

Inability and difficulties to make ends meet – SP3

The third approach is a combination of the two above listed approaches. To be considered in subjective poverty, a household needs to receive an income lower than the subjective minimum income necessary to make ends meet and additionally, it should have difficulties to make ends meet.

For the three approaches, the analysis will cover the period of 4 years and will compare the results with the relative at-risk-of-poverty (AROP) rate in Europe and with the severe material and social deprivation (SMSD) rate.

Results

Subjective poverty approaches’ results

Comparing the maps in figures 1-4 for the European Union, it can be seen that when the analysis of poverty relies on relative definitions such as the AROP rate (Figure 1), the higher rates are shown in Eastern Europe countries followed by some Mediterranean countries and the Baltic States. Meanwhile, when subjective poverty definitions are used, higher results are shown mainly for South-Eastern countries (Figures 2 and 4). Figure 3 shows that a high proportion of households across Europe considers to need a higher income than their actual household income.

These results can be interpreted as complementary since the subjective poverty (SP) indicators are able to capture to a certain extend the limitations of household when facing expenses.

Figure 1: At-risk-of-poverty rate 2020	Figure 2: Difficulties to make ends meet 2020 (SP1)

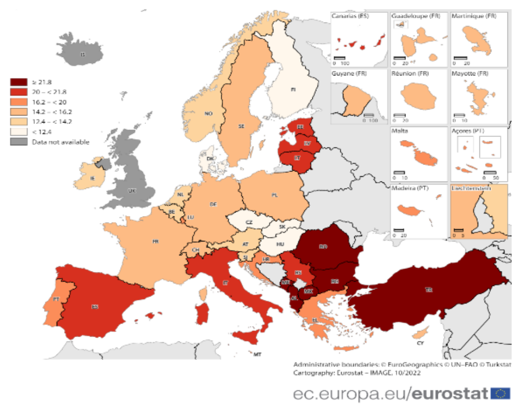


Figure 3:Household income lower than the subjective minimum income necessary to make ends meet 2020 (SP2)

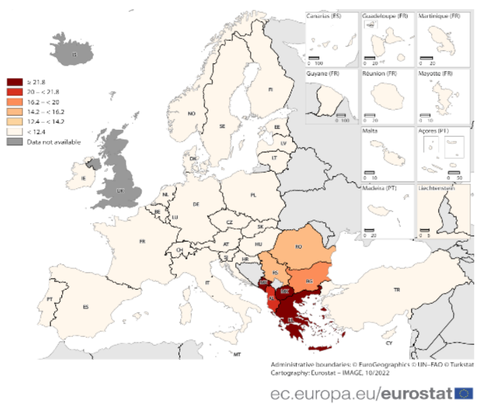
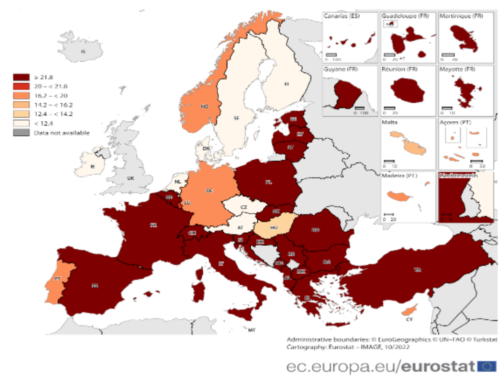
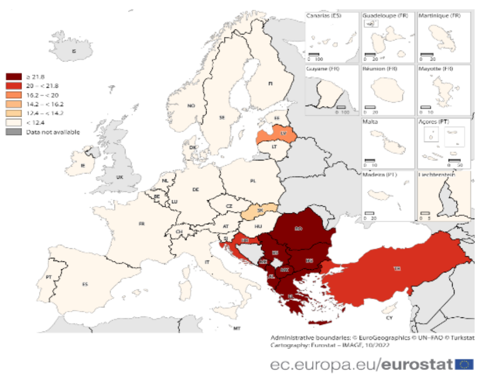


Figure 4: Difficulties to make ends meet and income below the subjective minimum income 2020 (SP3)





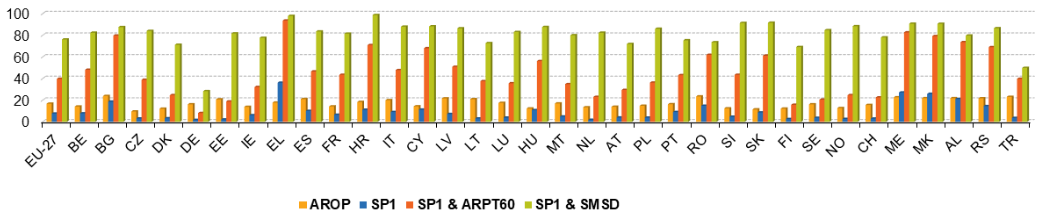
Source: At-risk-of-poverty rate (ilc_li02), Inability to make ends meet (ilc_mdes09) and Eurostat's computations using EU-SILC 2020 data, EU27, EFTA and some candidate countries. Please note that data for Denmark, Germany, Ireland, France and Luxembourg have a break in data series. Figures 3 and 4 present 2019 data for Germany and Poland.

Relative and subjective poverty comparison

Figure 5 compares relative and subjective poverty measures. On EU27 average, the AROP rate is 16.7%, while the subjective poverty rate based on SP1 is only 7.6%. When considering subjective poverty broken down by AROP or by SMSD, it is visible that subjective poverty is much higher among the population severely materially and socially deprived (76.1%) than among the population at-risk-of-poverty (39.7%).

Figure 5: Relative and subjective poverty comparison, EU27 and European countries 2020

Relative and Subjective poverty 2020



Source: At-risk-of-poverty rate (ilc_li02) and Eurostat's computations using EU-SILC 2020 data, EU27, EFTA and some candidate countries. Please note that data for Denmark, Germany, Ireland, France and Luxembourg have a break in data series.

Conclusions

The analysis of EU-SILC data shows that information on subjective poverty can complement usually agreed poverty measures such as the AROP and the SMSD rates. Objective measures are mainly focused on household income, while subjective measures also consider information on consumption and expenditure.

The results of this analysis indicate that when comparing the AROP rate with subjective indicators (based on the "difficulties to make ends meet" approach defined in this article, SP1), the differences are substantial. Relative and subjective poverty measures are more aligned in South and Eastern Europe than for the rest of Europe. The analysis also shows that, across Europe, results based on subjective poverty are more comparable with the deprivation indicators than with monetary relative poverty. This could be explained by the depth of poverty and the in-kind transfers, which are more available in Western-North Europe.

References

[1] European Union Statistics on Income and living conditions (EU-SILC) https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Chatbots to ‘talk’ with public data: the case of the Luxembourg Income Study (LIS)

Keywords : Chatbots, data dissemination, public data, LIS data, user experience, open source

1. Introduction

Dissemination of public data is a core task for organisations producing official statistics. These organisations must ensure that users are able to easily retrieve the right information for them. Retrieving publicly available data can be a rather time-consuming process even for users with ICT skills. They might first need to search the right source of information, such as downloadable files, tables, or dashboards. Then they need to find the desired information within those. Sometimes, the task might also require combining data from multiple sources. Consumers who are not substantive experts might also need to find the relevant metadata and context for the information retrieved. For these types of data users, the growing amounts of data published has made their data consumption processes increasingly difficult. In this sense, we could say public data has not been really opene d to regular citizens .

New technologies can help simplify the processes for retrieving public data for all types of users regardless their ICT skills. Changes should aim to make the experience of most data users of statistical organisations less burdensome and time-consuming. An ideal solution should aim to be:

- User oriented and capable of serving all types of users, with an emphasis on those users whose skills do not allow them to use more complex alternatives;
- Work well for multiple organisations with different capabilities, needs, and domains of data. Adaptation to new contexts should come at low additional cost. This should include the ability to function in multiple languages, both largely spoken ones and those with fewer speakers;
- Capable of learning from the user's experience. It should create a loop where it gathers feedback and improves the interface;
- Useful adding value to the interaction with consumers. It should go beyond simple retrieval of data and also support the user's interpretation of information.

This paper presents a pilot of the BODI (Bots for open data interaction) tool for automatically producing chatbots to explore public data sources. Chatbots offer a natural language interface through which citizens can ask questions in their own language. The bot would then try to find the right answer exploring the data on behalf of the citizen.

BODI aims to offer a solution to the growing complexity of obtaining publicly available data exposed above. We assess the performance of these chatbots using data from the Luxembourg Income Study (LIS). In the current example the pilot bot is thus tailored to the specific case of income inequality and poverty estimates, though the generation process can be seamlessly applied to any tabular data source.

The BODI project might currently be the only one of its kind. To our knowledge, only a couple of Proof of Concepts (PoC) chatbots used for official statistics and public data have been previously published [Sta20][PLM+17]. Both bots were manually created, this is in contrast with the BODI approach where bots are automatically generated. [KSV19] proposed a chatbot to help users find data sources in an Open Data repository by relying on geo-entity annotations. However, the chatbot only suggests the data sources to explore. It does not provide querying capabilities to consult those data sources.

2. Methods

2.1 The BODI chatbots

BODI (Bots for Open Data Interaction) is a tool capable of producing chatbots adapted to a

specific data context. These chatbots are conversational interfaces where users can interact with the data by writing in their own natural language. The bot hides the complexity of finding the right data for the users, as they are able to get the answer they need with no technical knowledge.

Bots are automatically derived based on an analysis of the dataset description and contents. For instance, types of the colums (string, date, integer,...) are inferred from the column values and, together with the column names, used to generate questions users could potentially ask the dataset (e.g. for a date column in cultural agenda dataset, the generated bot will be able to answer questions about the events happening before or after a date).

The process of generating a new bot involves little manual input and can easily be scaled to cover large numbers of applications. Dataset owners can optionally use a web interface to configure the generation process. As an example, they could add synonyms for column names or decide to merge or filter out some columns.

Figure 1 depicts the architecture of the generated chatbots. These chatbots use a double path strategy. The first strategy happens when the bot is confident to have recognized the question (Intent Recognition component tries to match the user question with one of the questions the bot has been trained for). For those questions that the bot can not recognize, the bot applies an advanced fallback strategy. The fallback relies on [VSX20] to automatically translate the user query to SQL and executes such SQL on the tabular data to try to get an approximate answer. Moreover, the chatbots created by the BODI tool use translation models to convert text in different languages to English to be able to use the above strategy (Figure 1 shows the models BODI uses to translate from Catalan and Spanish to English). New languages can be easily incorporated by adding other language translation models.

The BODI infrastructure can be adjusted to particular needs of the organisations and data. This makes them flexible enough to be used in different contexts. In particular, BODI can be used to discover what users want to know from the data to drive future dataset publications. BODI bots monitor their behaviour and keep track of missed questions (or questions users state were incorrectly answered). This information can be used to either improve the bot (if the bot should have been able to answer) or improve the dataset (if the data the user was asking was indeed missing in the dataset).

2 . 2 Current application with LIS data

The Inequality Key Figures are produced by LIS – Cross National Data Center (LIS). They show national-level inequality and poverty indicators. Examples of these are Gini and Atkinson coefficients, relative poverty rates and median equivalised income. It currently covers 51 countries and up to 50 years. The estimates are computed from survey microdata harmonized by LIS and updated on a quarterly basis. They are publicly available through a table at LIS website[64] or a downloadable file.[2]

The Inequality Key Figures dataset was deemed appropriate for a pilot of the BODI chatbot because of the following reasons:

- The indicators and concepts in the dataset are easy to understand, but at the same time belong to a specific scientific domain (i.e. income and inequality studies);
- It is large enough, as it contains over 15,000 cells of data;
- Requires little data cleaning beforehand. Although there are a few columns that should be split to make each variable more informative, the dataset does not contain additional complexities such as duplicated values or data to be parsed from strings.

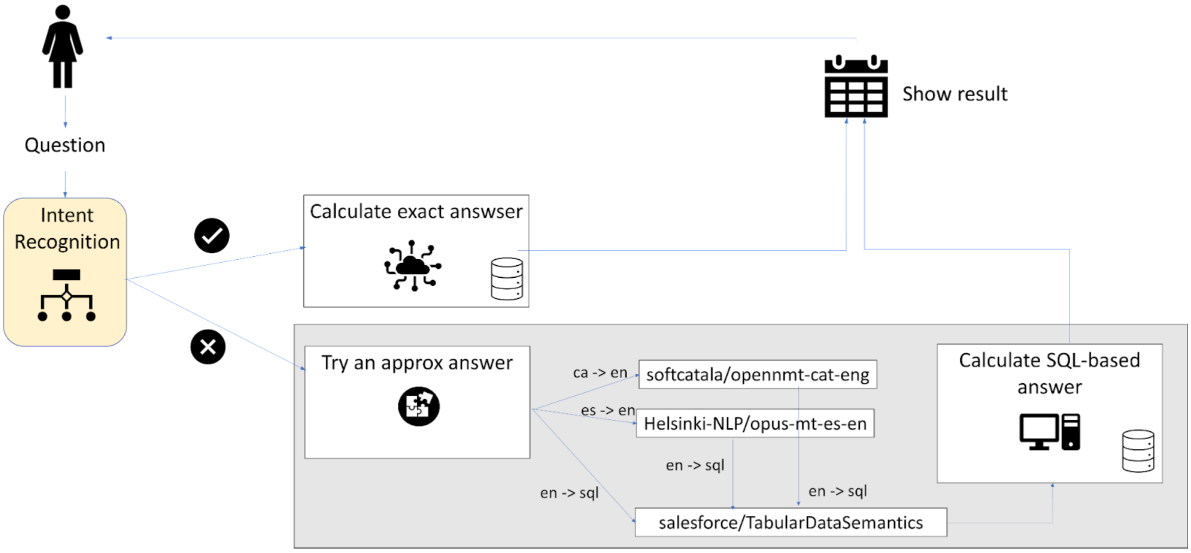


Figure 1: NLP architecture of BODI generated chatbots

2.3 . Applying BODI to other datasets

As described above, BODI is a generic framework useful to any organization publishing datasets. The BODI components will be released as open source for those institutions that want to try it. **3. Results**

We have used BODI to generate a POC chatbot for the LIS data together with a LIS expert. The LIS expert was first in charge of suggesting potentially interesting questions the bot should be able to answer and then evaluating the correctness of the bot answers and their adequacy and usefulness for the LIS environment.

[64] See the search engine: https://www.lisdatacenter.org/lis-ikf-webapp/app/search-ikf-figures [2] Can be downloaded from: https://www.lisdatacenter.org/wp-content/uploads/files/access-key-workbook.xlsx

His feedback is being integrated in a new version of the bot to be released to, first, a reduced selection of users. The feedback is improving not only the LIS specific bot but also triggered new improvements to the BODI infrastructure as a whole to benefit other future bots.

Among others, given that French is a common language for LIS users, a module for the French language has been added to the BODI infrastructure. Support for Luxembourgish, based on [LLV+22] is under study. We are also implementing the option to indicate that a group of columns are semantically related. Then, users can ask questions about values in individual columns or about the group as a whole. Additional simple data cleaning operations were also added to BODI based on the analysis of the dataset content.

4. CONCLUSIONS

BODI is an ongoing initiative expected to reach a TRL5 level over the next months. BODI can benefit any public institution interested in making their data more accessible to non-technical users offering a natural language interface to query the data. Beyond this core basic functionality, BODI has a number of planned extensions to better cover the needs of public organizations based on the feedback we are getting after the few demos and pilot studies.

In particular, we plan to address the massive generation of chatbots for open data portals. Such portals rely on CKAN[65] or similar data management systems to host numerous open data sources. We aim to be able to retrieve them all and automatically generate the corresponding chatbots together with a *metabot* able to redirect user questions to the appropriate bot depending on the topic of the question. We will also look to improve the training of the generated chatbots thanks to the use of ontologies. The idea is to map the metadata describing the dataset (including column names of the tabular data) to semantic concepts in the ontology to generate richer conversations thanks to this better understanding of the dataset.

For the chatbots themselves, we also plan a couple of improvements. First, bots will be able to answer questions with the aid of visual representations (e.g. plots or even maps for datasets with geographical information). Secondly, we will enable the use of other structured data sources, as APIs, as data sources. Similar to the SQL translation, we can hope to have soon open pretrained models for natural language to API queries[66] available. Finally, we will extend the bot functionality with a QA (Question & Answer) pipeline to let the bot look into auxiliar (semi)structured documents for answers that were not part of the input dataset. As one councillor pointed out, citizens ask questions and want answers, but they don't check whether the answer is in the tabular dataset or an accompanying pdf.

[65] https://ckan.org/

[66] GPT-3 like models do already a good job on this but new initiatives like BigCode (https://www.bigcode-project.org/) promise a better and more open solution.

5. R EFERENCES

[KSV19] Keyner, S., Savenkov, V., Vakulenko, S.: Open Data Chatbot. In: Satellite Events of The Semantic Web. pp. 111-115 (2019)

[PLM+17] Porreca, S., Leotta, F., Mecella, M., Vassos, S., Catarci, T.: Accessing Government Open Data Through Chatbots. In: Int. Workshop on Current Trends in Web Engineering. pp. 156-165 (2017)

[Sta20] StatsBot: A Chatbot For Interacting with SDMX Databases. Report to the HLG-MOS Executive Board, October 2020. https://github.com/guillaume-thiry/OECD-Chatbo t

[VSX20] Victoria Lin , X., Socher , R., Xiong, C.: Bridging Textual and Tabular Data for Cross-Domain Text-to-SQL Semantic Parsing. EMNLP (Findings) 2020 : 4870-4888

[LLV+22] Lothritz, C., Lebichot, B., Allix, K., Veiber, L., Bissyande, T. F. D. A., Klein, J., ... & Lefebvre, C. (2022). LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In Proc. of the Language Resources and Evaluation Conference, 2022 (pp. 5080-5089) .

Adding a new mode to the mix: Implementation and evaluation of CAWI in EU-SILC

Keywords: mixed mode, CAWI, expert evaluation, cognitive interviewing, friendly user test, pilot survey, EU-SILC.

INTRODUCTION

EU-SILC is a household survey collecting data on income and living conditions of the population in the European Union. In Austria, the survey has been conducted since 2003, during the past years using CATI (Computer assisted telephone interviewing) and CAPI (Computer assisted personal interviewing) modes. Except for the first wave, which is usually collected via CAPI, there is a mode-mix within as well as between waves. Starting with 2023, the Austrian EU-SILC survey will add another interviewing mode to the currently established CAPI and CATI modes: CAWI – Computer assisted web interviewing.

The decision to implement this new mode was based on several facts: the use of CAWI mode is “on the rise” in National Statistical Institutes (NSIs) in general, CAWI is efficient and flexible for data collectors as well as respondents, and is generally more cost-effective in the large scale. It is also in line with several other NSIs that are already using CAWI for EU-SILC data collection or are planning to do so in the next years [1]. Furthermore, the COVID-19 pandemic has once more accentuated the relevance of mixed-mode surveys and a need for contact-free modes, like CAWI [2].

More specifically, the Austrian EU-SILC CAWI mode will be based on a responsive, mobile first approach to accommodate a variety of devices, with the same questionnaire interface being used for all three modes. The existing questionnaire, including question text, answer categories, instructions and warning texts had to be evaluated, restructured and re-worded to be fit for the new mobile first design and, most of all, self-administration by respondents. For the implementation of the new CAWI mode and the accompanying work and evaluation, a Eurostat Grant was awarded in 2020. The duration of the grant covered a large proportion of the re-wording phase, evaluation and testing of the questionnaire, including a CAWI pilot survey running along the “regular” EU-SILC survey in 2022.

We would like to present the testing process during the SILC CAWI implementation phase with special emphasis on the pilot survey and discuss the implications of adding a new mode to the mix.

METHODS

The methods used to evaluate and drive the re-wording process and the CAWI implementation phase were in part qualitative and in part quantitative. Overall, we followed an iterative process of re-wording, evaluation, and revision with a final evaluation after the pilot survey.

2.1. Expert evaluation and cognitive interviewing

The initial re-wording phase was accompanied by an extensive qualitative and expert evaluation. In-house survey experts provided information and guidance concerning the state of the art in terms of question wording and mobile first questionnaire design. Of the 611 questions in total, around 200 questions were evaluated by two experts from an external service contractor experienced in web interviewing. The questions were evaluated using a checklist [3] and reviewed in discussion rounds with the evaluators and the subject matter experts responsible for rewording the questions. Among those questions, particularly complex questions were selected for cognitive interviewing using an interview guideline [4].There we already employed the new CAWI interface to more closely simulate a real survey situation to test the user experience of the questions in the new design.

2.2. Friendly user test

After programming the entire questionnaire according to the new standards, the questionnaire was tested with nine friendly users, using a combination of observation and, subsequently, qualitative interviews. After incorporating this feedback, a larger-scale friendly user test was conducted with around 90 participants. Qualitative feedback on specific questions and user experience was collected via an integrated feedback function, e-mail or telephone and was then systematized. After the final revision based on this feedback, the CAWI pilot survey entered the field alongside the regular survey. We aimed for 200-300 households completing the CAWI survey.

2.3. Pilot survey

During the pilot survey, we tested the programming of the questions, the new workflow including a specific CAWI reminder strategy, as well as new communication materials. After the field phase, we conducted a quantitative comparison between the new CAWI questionnaire, which is self-administered, and CATI/CAPI questionnaire, which is interviewer assisted. Because the CAWI pilot survey contained a new question flow and re-worded questions, any specific mode effects cannot be extrapolated. Furthermore, the sample differed quite a lot in terms of socio-demographics of the participants to that mode compared to CAPI or CATI. However, by comparing those characteristics as well as paradata, the functionality of the questionnaire as well as the validity of the questions could be tested.

RESULTS

Each phase of testing produced results that fed into the final outcome: a new mode (CAWI) as well as a new questionnaire and survey structure for all modes (CAWI, CAPI, CATI).

External expert evaluations, cognitive testing and the two-part friendly user tests resulted in a new questionnaire with optimized filtering, accurate as well as respondent-friendly questions and during the process continuously improved user experience. This questionnaire and the new survey structure including workflow, communication strategy was finally tested in a real-life pilot survey, enabling a comparison between modes (within the same panel wave); more specifically between interviewer-assisted/old questionnaire (CAWI; households N=411; household members N=900) and self-administered/new questionnaire (CAPI, CATI; households N=756; household members N=1,480).

3.1. Comparison of questionnaire and survey paradata (CAWI versus CAPI/CATI)

Analyses of paradata do not present any evidence that questions of the new questionnaire did not work or have not been understood correctly, which could have been indicated through lots of missing answers or implausible figures.

- The CAWI *response rate* was similar to the usual CAPI/CATI response rate of EU-SILC Austria. Aiming for 200-300 completed household interviews we drew a subsample for the pilot from the pool of households in the second wave of the EU-SILC survey. From around 1,800 households 631 households matched the defined CAWI criteria (e.g. e-mail address available). Of those, 395 households (62.6%) ultimately completed the CAWI questionnaire. This original CAWI net sample size differs from the evaluated sample size (N=411) as it does not include mode changes from CAPI or CATI to CAWI.
- Regarding the *item non-response rate*, it can be said that the CAWI mode worked similarly well as the CAPI/CATI mode. The number of “Don't know” has decreased while the number of “No answer” (which corresponds to a refusal) has slightly increased. The percentage of all valid answers given is in CAWI mode slightly lower than in CAPI/CATI.
- In addition, the *questionnaire duration* was similar and the proxy rate did not rise hinting at the CAWI questionnaire being manageable and accepted by respondents.

3.2. Comparison of socio-demographic characteristics of interviewer-assisted and self-administered modes (CAWI versus CAPI/CATI)

Results show that socio-demographic characteristics between households completing the questionnaire in CAWI versus CAPI/CATI mode differ a lot. This is, of course, also due to the selection of the sample and the assignment to a specific mode. The biggest differences occur in age, educational attainment level and household income:

- *Age:* In CAWI mode, the respondents were on average 8.4 years younger than in CAPI/CATI mode, the median age was twelve years lower. In CAWI, approximately 6%-points more children under the age of 16 appeared, whereas in CAPI/CATI around 18%-points more people over 60 years old were in the sample.
- *Educational attainment level:* Overall, the respondents in CAWI have a higher formal level of education than respondents in CAPI/CATI: The proportion of people with compulsory schooling as their highest level of education is less than half as high as in the CAPI/CATI questionnaire. Fewer people in CAWI have completed an apprenticeship. The proportion of people with a Matura (A-levels) or university degree is also higher in CAWI.
- *Household income:* CAWI respondents reported a higher net household income per month than CAPI/CATI respondents.

The results show the importance of CAPI and CATI. Without those modes, households with lower income and respondents with lower formal education are not adequately represented. These socio-demographics are strongly linked to deprivation, poverty and risk of poverty [5] and are therefore essential to collect. In a pure CAWI design the sample would be substantially biased.

cONCLUSIONS

Conducting different kinds of tests, especially a pilot survey, is inalienable for such a big project like implementing a new mode and creating a new questionnaire and workflow. The process enables to rethink long-established procedures aiming for a more efficient and successful survey.

Further analysis is, of course, needed and descriptive statistics and bivariate analysis can only be the beginning if the aim is to explore the complex issues of mode and questionnaire effects. For our main goal to timely know about the suitability of the new questionnaire and the workflow in CAWI we already learned a lot and will be content to use this design from 2023 as the standard.

rEFERENCES

[1] S. Psihoda, N. Lamei and L. Lyberg, Preventing and mitigating the effects on data quality generated by mode of data collection, coding and editing, In: P. Lynn and L. Lyberg: Improving the measurement of poverty and social exclusion in Europe: reducing non-sampling errors (2022) 337-360.

[2] EUROSTAT, Position paper on mixed-mode surveys (2022), Publications office of the European Union, Luxembourg.

[3] F. Faulbaum, P. Prüfer and M. Rexroth, Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität (2009), VS Verlag für Sozialwissenschaften, Wiesbaden.

[4] T. Lenzner, C. Neuert and W. Otto, Cognitive Pretesting: GESIS Survey Guidelines (2016), GESIS – Leibniz Institute for the Social Sciences, Mannheim.

[5] EUROSTAT, Income and living conditions in Europe - Edited by Anthony B. Atkinson and Eric Marlier (2010), Publications office of the European Union, Luxembourg.

Tracing respondents’ footpaths – How user journey analysis can offer new insight into survey experience

Keywords: user journey mapping; mixed-mode questionnaire; nonresponse; evaluating quality.

Introduction

In Statistics Norway, various methods and techniques are commonly used to improve surveys, such as expert reviews, cognitive and usability testing, and focus groups. For a EUROSTAT grant project regarding the Adult Education Survey (AES) 2022 (EU Grants: Application form (SMP ESS): V1.0 – 15 .04.2021, we attempted a relatively new approach by doing a user journey analysis. User journey methodology has not been used to a large extent in Statistics Norway but is already quite popular in businesses and institutions wishing to improve the user experience of a product or service. It can offer unique insights not only into how respondents experience the survey questionnaire, but the whole process of participating in a survey. For the AES survey, our goal was to map the respondent experience for two demographic groups of specific interest for that survey: Respondents aged 25-34, and respondents with an educational level below ISCED3. This provides us with a better perspective how users in these groups, often underrepresented in net survey samples, experience participation in surveys, what challenges they face, and what influences their motivation to participate.

In this paper, we will present how we managed to construct and analyse a realistic user journey and gain useful insights into users’ experience in participating in the AES. We will discuss challenges we experienced and recommendations for future user journey analyses for surveys in general, but also for the AES in particular. We hope to inspire other statistics offices to conduct user journey mappings since mapping user journeys facilitates understanding of user experience and behaviour when being invited to take part in a survey. It can help to improve the survey itself as well as communication strategies both before and during survey completion, and ultimately contribute to reducing nonresponse. With tailoring communication strategies and offering help to specific groups we hope we can increase positive experience of participating in AES as well as increasing representativity.

Methods

User journey mappings are conducted in the business world to gain insight into users’ experience of a specific product. For Statistics Norway, that specific product is the AES. Mapping the respondent’s journey from being contacted by Statistics Norway to answering the survey or rejecting participation. As described by Kalbach, one essential aspect of user journey mapping is to consider the *touch points* a user has with the product. Touch points are points where interaction between individuals and an organization can take place, for instance as a phone call, through email interaction, or websites (Kalbach, 2020).

Discussing users experience offers useful perspectives on what challenges users face, highlights possible misconceptions or gaps and how touch points are perceived by users: If they experience positive or negative emotions, or if they are indifferent. Moreover, discussing users’ experiences can offer a better understanding how far the goals of an organization are achieved and how to improve the experience of touch points to increase the organization's goals.

Before conducting the user journey mapping, we determined possible touch points between respondents and Statistics Norway. The process was as follows: a survey methodologist identified possible touch points and then presented them to the Division of Media and Culture Statistics and the Division for data collection. The other team members gave useful insights if and how relevant those touch points are and proposed other relevant touch points. One of our strategies was to have a flexible approach to user journey mapping, where updates can be done along the way. This helped us to get better insight into all relevant experiences for respondents and remove aspects of smaller importance for them.

The objective of the user journey analysis is to detect all possible touch points with Statistics Norway, find out if respondents are encountering those touchpoints or not, and improving their experience of those touch points. The survey questionnaire itself was defined as a touch point. As the AES was planned for CATI/CAWI mixed mode, we needed to analyze how both modes were perceived by the two target groups.

Based on the preceding expert reviews and user testing we adjusted the questionnaire to mixed mode, before fielding a pilot for the user journey mapping. The pilot was conducted as follows: we drew a random sample of 200 people for each group, where those invited could answer the survey in a 9-day period. After the 9th day we followed up with focus groups and expert reviews for another week. Focus group and explorative interviews recruits were mostly respondents, but we also conducted one explorative interview with a survey nonrespondent.

Results

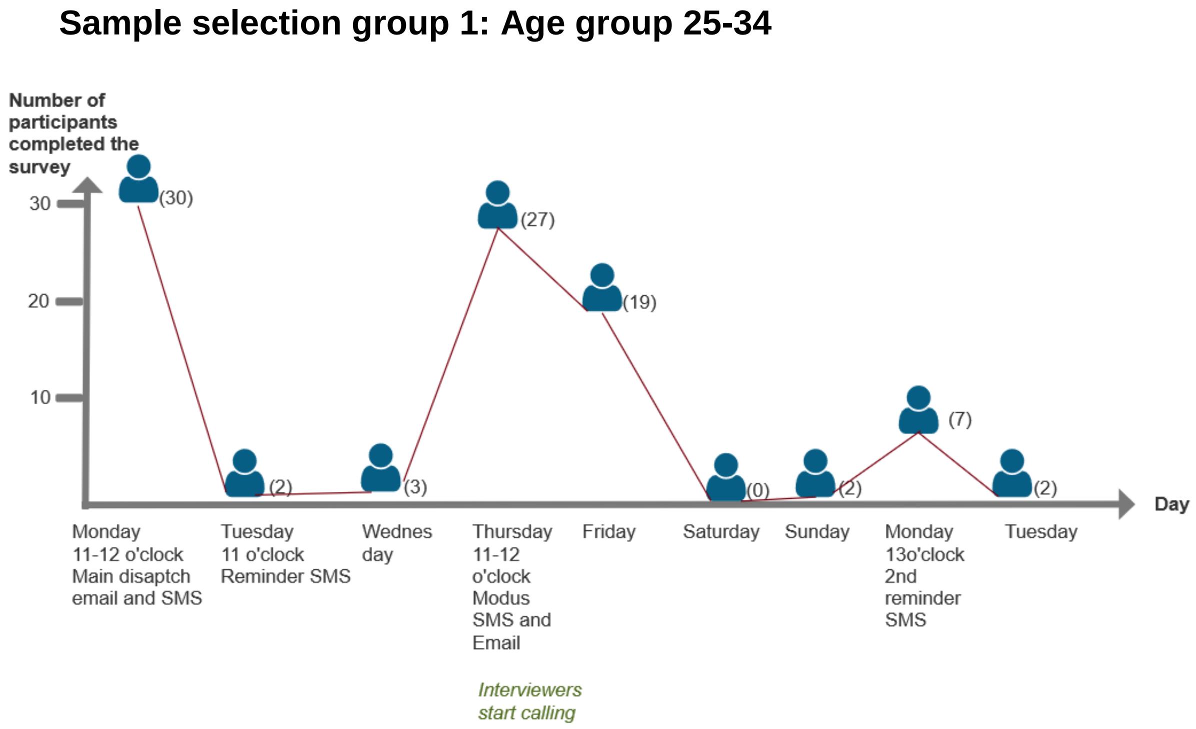


Figure 1. Answer development for group 1

The first group for which we performed user journey analysis was respondents aged 25-34. Figure 1 shows how many participants answered each day, as well as the emails and

SMS sent out. At the start of the 9-day field period, only web response was possible. Towards the end of the week telephone interviews were offered and conducted.

The main findings from the focus groups and explorative interviews for group 1 participating in the user journey are:
1. The respondents perceived the E-mails and SMS reminders as very useful and stated that the content and length is fine.
2. Questions regarding naming courses, seminars and so on were perceived as challenging. The topic about barriers for learning were seen as very relevant by several participants.
3. The invitation letter contains a link to a survey home page providing more information about the AES survey. None of the focus group participants clicked on the link.
4. Regarding mode email, some of the participants proposed to change it. One participant stated that it should contain the exact time when one would be contacted, another proposed that the participants themselves should be able to propose a time that suits them. However, one participant also admitted that knowing Statistics Norway will be calling might lead to not picking up the phone.

3.2. Sample selection group 2: Educational level below ISCED 3

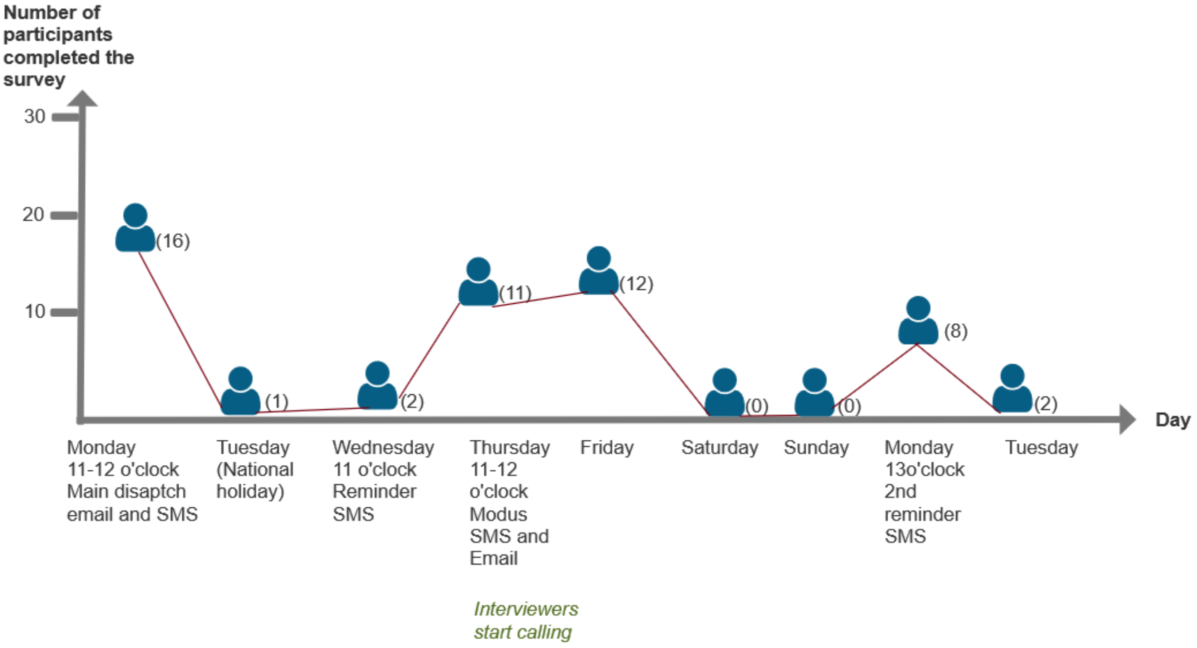


Figure 2. Answer development for group 2

In short, the response rate for the Low Education group was nearly half of that we had in the 25-34 group. This is not surprising. Research findings by Dillman and Messer (2011, p. 445) come to a similar result for mail and web responses, mainly that those who answer on web are younger and have higher educational attainment than compared to those mail follow-up respondents.

The following insights were gained through focus groups and in-depth interviews for group 2:
1. Some respondents read the invitation letter, while others did not.
2. Most survey topics were not problematic for the participants. Yet, questions regarding the names and more detailed information about courses, seminars, private lessons and so on were perceived as difficult. Regarding language questions, respondents did not perceive difficulties but interpreted the question in various ways and answered it

therefore differently. On educational background, which is used in several Statistics Norway surveys, several respondents wondered why we need this information

3. A lower percentage of respondents took part in web survey compared to group 1. Yet, telephone interviews were conducted approximately to the same degree as in group 1.

4. The name of the survey was misinterpreted by some participants. In Norwegian we used the term *voksnes læring* (adults' learning), while we also have the term *voksenopplæring* which refers to specific educational programs. Respondents confused those two terms.

A possible touch point which had not been used by the focus groups participants is Statistics Norway's Information Centre. The Information Centre can be reached via email or phone for more information about surveys, statistics, or other inquiries. The Centre's records indicate that of the 400 in the AES user journey pilot samples, only 2-3 people contacted them.

Conclusions

Recommendations for user journey mappings

- To conduct user tests and an expert review in advance was advantageous in a way that we already reduced some irritations for respondents and had more time to discuss other touch points with the survey such as invitation letter and reminders.
- A pilot has proved to be an extremely useful method to create a realistic user journey. Conducting a pilot survey is more costly than recruiting a specific number of people to participate in a user journey. Still, we recommend following this approach since it allows a more realistic user journey. Thus, funding for this should be allocated.
- The time between the pilot and the user journey focus groups should not be too long, as respondents forget their experience. However, it should also not be too short, otherwise it will be very unrealistic. We found 1 week appropriate for this survey.
- Deciding on whether to do individual follow-up interviews or focus groups depends on target group characteristics. Preferably, focus groups should be conducted which reduces time, costs, and also allows for discussion between participants. Yet, if the participants feel uncomfortable with focus groups, interviews are advised.

Although our setup worked very well, we would like to experiment in the future more with journey mapping. 100 might be enough for an easy recruitable group such as students.

Recommendations for future AES

AES is a very complex and long survey which consists of very detailed questions, especially on non-formal education. Thus, it is very demanding to create a version which fits both modes, web and telephone. Even after conducting expert reviews, user tests and a user journey mapping, we still found challenges regarding nonformal education. We propose to shorten the questionnaire to reduce response burden in future AES to avoid low data quality and drop out. One approach is to reduce the topics on this survey and only focus on non-formal education and leave for instance formal education out of the survey. Another approach is to have the same number of topics as it is up to date but cut out the follow up questions on non-formal education.

As has been stated already by Dillman (2014, p. 47), using multiple modes of communication can be advantageous and increase benefits, decrease costs, and build trust. Especially the latter aspect of building trust was often mentioned in our focus groups. Many focus groups participants stated it is positive that we used different communication strategies such as emails, SMS and phone calls. Some participants found emails more trustworthy, while others had more confidence in e-mails. Yet others had more faith in phone calls.

References

[1] EU Grants: Application form (SMP ESS): V1.0 – 15 .04.2021

[2] Kalbach, J. (2020). Mapping experiences. O'Reilly Media. Canada, Chapter 2: Fundamentals of Mapping Experiences.

[3] Benjamin L. Messer, Don A. Dillman, Surveying the General Public over the Internet Using Address-Based Sampling and Mail Contact Procedures, *Public Opinion Quarterly*, Volume 75, Issue 3, Fall 2011, Pages 429–457, https://doi.org/10.1093/poq/nfr021

[4] Dillman, D.A., Smyth, J.D. and Christian, L. M.: Internet, Phone, Mail and Mixed-Mode surveys- The Tailored Design Method. In Chapter 2: Reducing People's Reluctance to Respond to Surveys. 4th edition, Wiley, 2014.

High frequency Data (JENK2A.2)

Session Chair: **Dominique Ladiray** *(Consultant)*

COVID-19 daily number of diagnoses: an ARIMA analysis
Luis Sanguiao-Sande *(National Statistical Institute-INE)*

Time disaggregation in the Labour Force Survey using statistical learning and state-space models
Sandra Barragan, David Salgado, Miguel Ángel García Martínez *(National Statistical Institute-INE)*

Temporal Disaggregation of the Services Producer Price Index
Maria Novás Filgueira, Carlos Sáez Calvo, Luis Sanguiao-Sande *(National Statistical Institute-INE)*

COVID-19 daily number of diagnoses: an ARIMA analysis

Keywords: time series, outliers, unobserved components, signal extraction, high frequency

Introduction

We analyse the daily number of diagnoses during the COVID-19 pandemic in Spain. As usual, the series has a (big) weekly pattern, which surely is not related to the virus propagation but can be attributed to health services and their decreased activity on weekends. Moreover, the series is quite noisy.

The 14-day incidence rate was introduced to avoid those problems. The indicator was seen many times in the media, and its prediction was considered an important research topic too (see for example [1]). Of course, the 14-day sum removes the week day effect from the daily series, and a smoother series is obtained. However it does have two problems:

- The moving sum is not centered, as information from the future is not available, what means that provides delayed information about the pandemic.
- Smoothing the series by summing up so many observations, tends to mask any short term (less than a week) phenomenon.

The 7-day incidence rate was sometimes used to alleviate both problems, however it is a noisier indicator. Instead, we propose an ARIMA analysis of the series: the weekly pattern and the noise can be removed through Wiener-Kolmogorov filter, the filtered series can also be estimated in the last observation (with worse accuracy, of course), and any short term phenomenon can be included as an outlier or in general as an external regressor.

Methods

As has been already said, an ARIMA model is identified for the series, but some characteristics are expected.

First of all, when the number of diagnoses increases, a higher variance should be expected as the health services become overloaded. That’s why a model in logarithms was used.

Regarding the ARIMA itself, we consider a regular part and a part with lag seven which (through the canonical decomposition [2]) will lead to a cyclic component analogous to the usual seasonal component. Note that a seasonal pattern is expected in most respiratory diseases, but a few years of data would be needed to estimate such component. Anyway, unlike the weekly pattern, it can be considered a real component, and does not affect much to a short term analysis of the series.

At holidays, a decrease in the number of diagnoses is expected, so additive outliers for holidays are included when the statistic $T < -2$. Note that this allows different decreases for different holidays. A joint estimate would of course be possible too, but some holidays (for example Christmas) are expected to decrease a little more the number of diagnoses.

As a final step, an automatic search for outliers is done, mostly following [3], not only to improve the modelling but also to find some interesting short term effects.

The software used to perform the analysis was the SSMMATLAB library [4] under Octave. This library allows to specify an arbitrary integer period for the cyclic component, which we need to set to seven. The software JDemetra+ was also tried, obtaining similar results.

Results

The model identified for the Spanish national series was a (0, 1, 1) × $(0, 1, 0)_7$ with logarithms and no trend, similar to the classical airline passengers model [5]. The holidays finally included in the model were October the 12th, November the 2nd and December the 7th, 8th and 25th in 2020 and January the 1st and 6th in 2021. The less significant was November the 2nd with a T-statistic of -4.9.

Moreover, the following outliers where included:

Table 7. Additional outliers.

DATE	OUTLIER TYPE	T-Statistic
October the 13th, 2020	AO	-5.9
December the 10th, 2020	LS	7.6
December the 17th, 2020	LS	3.6
December the 27th, 2020	LS	6.5
January the 3rd, 2021	LS	5.6

Note the two last outliers are two days after Christmas and New Year respectively. The median incubation time of the virus seems to be closer to five days [6], but people were allowed to travel since December the 23rd and the 30th, so apparently both outliers relate to the holidays.

Now, it is possible to remove the irregular and weekly components, and compare the new filtered series with the 14-day incidence rate. Since the former is expected to anticipate the later in around a week, the graph of the later has been shifted left seven days in Figure 1. Actually, seven more observations are used to elaborate the incidence rate graph. The filtered series has been multiplied by 14, as the 14-day incidence rate is a sum instead an average.

As we can see in Figure 1, both graphs essentially agree, except that:

- The filtered series is slightly smoother. This is because we are removing holiday effects as outliers, but also because the canonical decomposition moves as much noise as possible to the irregular component.

- From December the 10th 2020 to January the 3rd 2021, the effect of the many level shift outliers is distributed by the moving sum. Note that after this period the graphs agree again.

- From January the 28th 2021, the graphs slowly start to separate. Remember that incidence rate does have seven more observations, so the difference is because the revision error. In fact, the filtered series works as a prediction of the future incidence rate.

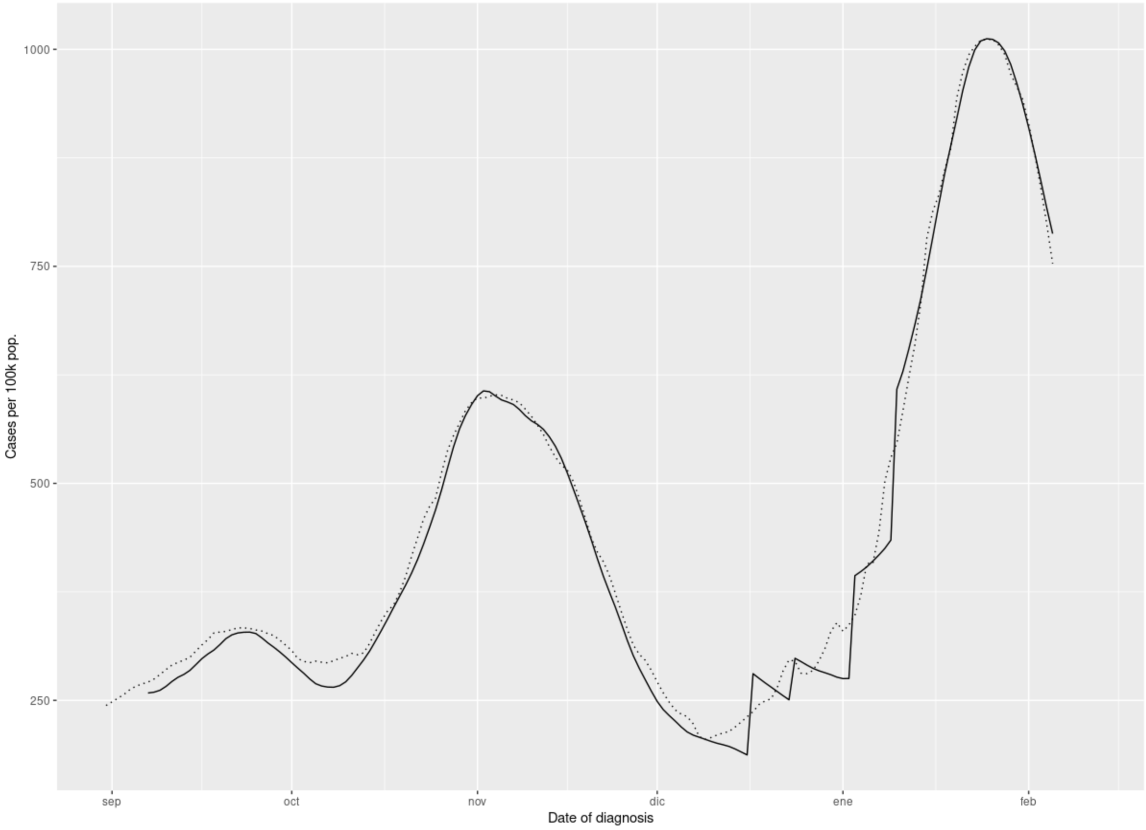


Figure 16. Comparison between seven days anticipated 14-day incidence rate (dashed line) and the series filtered from weekly and irregular components (solid line).

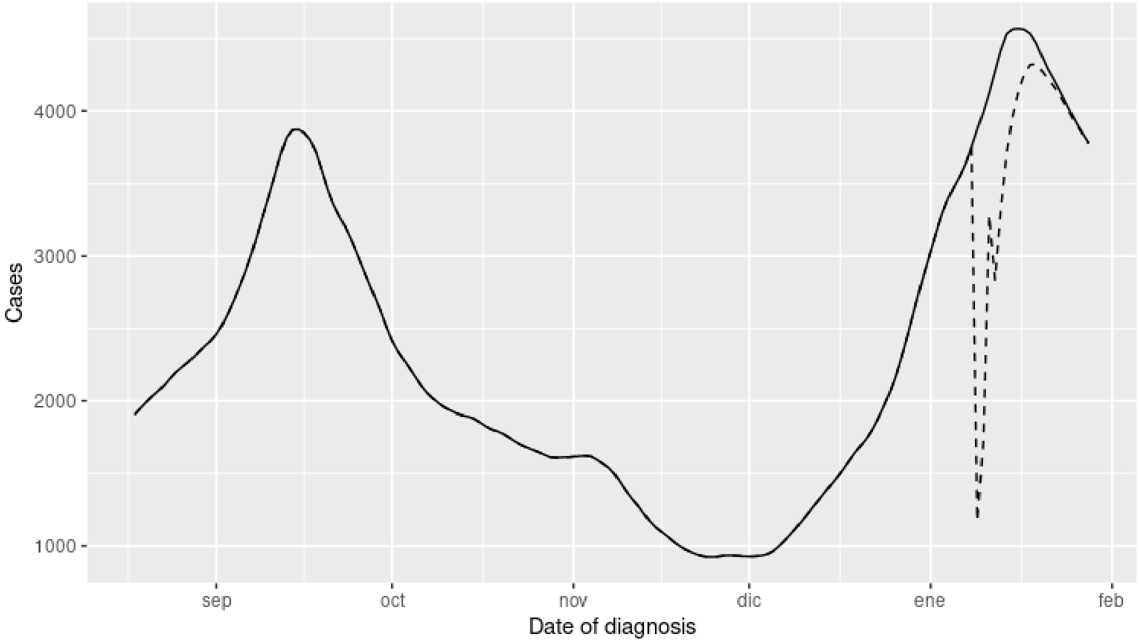


Figure 17. A snow storm in Madrid.

In Figure 2, the filtered series for the autonomous region of Madrid is shown. A heavy snowfall on January the 8th made access harder to hospitals and health centres, causing a sudden decrease in the number of diagnoses that returned to normal in a few days. This was included in the model through a transitory change on January the 9th and an additive outlier on January the 11th. Note that the last outlier was an increase, probably because the delay accumulated the previous days. In the figure, the dashed line contains both outliers while the solid line shows how it would be the evolution without said outliers. This illustrates how the modelling can help to analyse short term external effects.

Conclusions

The ARIMA models allow to extract a filtered series which turns out to be a timelier (and even smoother) indicator for the epidemic current state than the usual 14-day incidence rate. It is also timelier than 7-day incidence rate and much smoother. The only disadvantage is that it requires a careful modelling, while the elaboration of the incidence rates is immediate.

Moreover, the ARIMA modelling allows to include some short term events as outliers, or in general, as regressors. This way, the effect of some external phenomena of interest like holidays, snowfalls and even vaccines, can be estimated.

References

[5] F. Ahouz and A. Golabpour, Predicting the incidence of COVID-19 using data mining, BMC Public Health **21** (2021).

[6] S.C. Hillmer and G.C. Tiao, An ARIMA-Model-Based Approach to Seasonal Adjustment, Journal of the American Statistical Association **77** (1982).

[7] V. Gómez and A. Maravall, Automatic modeling methods for univariate time series, D. Peña, G.C. Tiao, and R. S. Tsay (Editors), A course in time series analysis, Wiley (2001).

[8] V. Gómez, Linear Time Series with MATLAB and OCTAVE, Springer (2019).

[9] G.E.P. Box and G.M. Jenkins, Time Series Analysis: Forecasting and Control. Revised Edition, Holden Day (1976).

[10] Z. Nazar and A.M. Elfadil, The estimations of the COVID-19 incubation period: A scoping reviews of the literature, Journal of Infection and Public Health **14** (2021).

Time disaggregation in the Labour Force Survey using statistical learning and state-space models

Keywords: Time disaggregation, microdata, timeliness, statistical learning, filtering

Introduction

The traditional statistical process using survey data is deemed to fail short for the production of timely official statistics. In this regard, innovation focused on the use of new statistical methods and new data sources, both administrative and digital, has been adopted as the main direction to alleviate this situation and keep statistical offices relevant in society.

We argue that the application of novel statistical methods even with traditional survey data can notably improve the quality of many official statistics. We show that the oft-mentioned necessary trade-off between accuracy and timeliness can be stretched out by using a combination of finite-population methodology and novel statistical methods such as machine learning techniques and state-space models. In particular, we present an experimental combination of sampling weight calibration [1], statistical learning modelling [2], and state-space filtering techniques [3] applied on the quarterly Spanish Labour Force Survey to produce more timely information.

We claim that survey microdata combined with process paradata and these techniques offer a wider and deeper performance of traditional statistical outputs. The incorporation of new data sources, thus, should surpass these potential statistical products.

Methods

We take the quarterly sampling design for the LFS as our starting point, producing quarterly design weights$\widehat{\ }$ $d_k^{[Q]} = 1/\pi_k^{[Q]}$ and calibrated sampling weights$\omega_k^{[Q]}$ for the usual linear estimator$Y_d^{[Q]} = \sum_{k\in s_d} \omega_k^{[Q]} y_k$ for the domain total $Y_d = \mathrm{P}_{k\in U_d} y_k$.

Computation of time-disaggregated design weights

Interviews are weekly, so that every respondent in the rotating panel is assigned a given week w within each calendar quarter Q. This procedure is executed semiautomatically by sampling experts inheriting interview week assignments from preceding waves and assigning new units to fit a balanced data collection field work across the national territory. These assignments involve both the calendar-quarter week $w = 1,\ldots,13$ and the so-called rotation turn τ for field work balancing.

As a first step, we train a random forest model using (w,τ) as the categorical target variable and some identification and design variables as regressors $\mathbf{x}$. The model produces probabilities $\mathrm{P}(w,\tau|\mathbf{x})$ for each value (w,τ) in terms of the regressors, so that we can compute the marginal probabilities $\mathrm{P}_w(w|\mathbf{x})$. Since the regressors can be considered as partial identification variables for each unit, we can write just $\mathrm{P}_k(w)$.

Then, by computing weekly design weights $d_k^{[W]}$ for week *w* we reason as follows:

$$d_k^{[W]} = \frac{1}{\pi_k^{[W]}} = \frac{1}{\pi_k^{[Q]} \times \mathbb{P}\left(k \rightsquigarrow w | s^{[Q]} \ni k\right)} = \frac{d_k^{[Q]}}{\mathbb{P}_k\left(w\right)}.$$

Once weekly design weights are computed we can reason similarly for design weights for month *m*, so that

$$d_k^{[M]} = \frac{d_k^{[Q]}}{\sum_{w\in m} \mathbb{P}_k\left(w\right)}.$$

As a matter of fact, this computation can be generalised to any group of weeks within a given calendar-quarter.

Calibration of sampling weights

Design weights with any time scope (weekly, monthly) are subjected to the same procedure as in the quarterly case, namely (i) we adjust for nonresponse with a ratio estimator identifying response homogeneity groups as the sampling design strata to produce intermediate weights ω_k° and (ii) we use the linear truncated Deville-S rndal calibration procedure with marginal totals $\mathbf{Z}$ to produce the final calibrated weights

$$\omega_k = \omega_k^\circ \cdot \begin{cases} L & \text{si } q_k\lambda^T\mathbf{z}_k < L-1, \\ 1 + q_k\lambda^T\mathbf{z}_k & \text{si } L-1 \leq q_k\lambda^T\mathbf{z}_k \leq U-1, \\ U & \text{si } U-1 < q_k\lambda^T\mathbf{z}_k, \end{cases}$$

where the Lagrange multipliers λ are given by

$$\boldsymbol{\lambda}^T = \left(\mathbf{Z}_U - \sum_{k\in r} \omega_k^\circ \mathbf{z}_k\right)^T \left(\sum_{k\in r} q_k\omega_k^\circ \mathbf{z}_k\mathbf{z}_k^T\right)^{-1}.$$

The weights ω_k are calculated through an iterative process with tolerance ϵ for calibrated equations $|\sum_{k\in r}\omega_k\mathbf{z}_k - \mathbf{Z}_U| < \epsilon$ in each auxiliary variable (see [4, 5]).

Estimation and filtering of time series of estimates

Once weekly and monthly calibrated sampling weights are computed we can produce estimates with those time scopes (week *w* and month *m*), i.e.

$$\widehat{Y}_d^{[W]}(w) = \sum_{k\in s_d(w)} \omega_k^{[W]} y_k,$$

$$\widehat{Y}_d^{[M]}(m) = \sum_{k\in s_d(m)} \omega_k^{[M]} y_k.$$

These raw estimates still contain a lot of variability (noise) since the original sampling design was thought to be quarterly. We propose to use them as building blocks within state-space models such as a Kalman filter. A first trivial elementary model for a moving-quarterly time series can be proposed by setting

$Y_t = F\vartheta_t$, $\vartheta_t = G\vartheta_{t-1} + R\epsilon_t$, with

$$F = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}_{1\times 13}, \qquad G = \begin{pmatrix} \frac{1}{13} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{13} & 0 & \cdots & 1 \\ \frac{1}{13} & 0 & \cdots & 0 \end{pmatrix}_{13\times 13}$$

$$R^t = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}_{13\times 1}, \qquad \epsilon_t \simeq N(0, \sigma^2), \qquad \sigma^2 \approx 0 \, .$$

Basically, this drives us to compute Y_t as a moving average with the 13 most recent weeks, i.e.

$$\widehat{Y}^{[Q_w]}(w) = \frac{1}{13}\sum_{i=1}^{13} \widehat{Y}^{[W]}(w-i) \, . \tag{1}$$

Notice that the reference time period is the moving quarter Q_w comprising the 13 most recent weeks up to week w. Nonetheless the time series is updated weekly (w−dependence).

Neither filtering nor seasonality treatment has been applied in this first proposal. The high-frequency weekly time series and the versatility of state-space models will allow us to provide richer statistical outputs from the same survey microdata.

Variance computation

To assess accuracy, as usual, we shall compute confidence intervals. We need to estimate the variance of $\widehat{Y}^{[Q_w]}(w)$. We reason as follows. Estimator (1) can be considered as a two-fold sampling exercise, namely, firstly the weekly design-based part p and secondly a sequence of 13 independent measurements m upon the population.

Then, we can write $\mathbb{V}\left(\widehat{Y}^{[Q_w]}(w)\right) = \mathbb{V}_m\left[\mathbb{E}\left(\widehat{Y}^{[Q_w]}(w)|p\right)\right] + \mathbb{E}_m\left[\mathbb{V}\left(\widehat{Y}^{[Q_w]}(w)|p\right)\right]$, so that

$$\begin{aligned} \widehat{\mathbb{V}}\left(\widehat{Y}^{[Q_w]}(w)\right) &= \widehat{\mathbb{V}}_m\left[\widehat{\mathbb{E}}\left(\widehat{Y}^{[Q_w]}(w)|p\right)\right] + \widehat{\mathbb{E}}_m\left[\widehat{\mathbb{V}}\left(\widehat{Y}^{[Q_w]}(w)|p\right)\right] \\ &= \mathbb{V}_{JK}\left(\widehat{Y}^{[Q_w]}(w)\right) + \frac{1}{13}\sum_{i=1}^{13}\widehat{\mathbb{V}}\left(\widehat{Y}^{[W]}(w-i)\right), \end{aligned}$$

where $\mathbb{V}_{JK}(\cdot)$ stands for the usual jackknife variance estimation for a point estimator (deleting a week in each jackknife sample) and $\widehat{\mathbb{V}}\left(\widehat{Y}^{[W]}(w-i)\right)$ denotes the design-based variance estimator for the weekly estimator $\widehat{Y}^{[W]}(w-i)$ (which is the finite-population variant jackknife estimator used in the quarterly sampling design). Confidence interval computation is undertaken in the usual way using the normal distribution through the central limit theorem.

Results

The main output from the preceding proposal can be depicted as in figure 1, in this case using the time series of employed females aged 25-74 in the whole national territory. Similar results are obtained for both sexes in age groups 16-24, 25-74, 75-+ for employed, unemployed and inactive resident people.

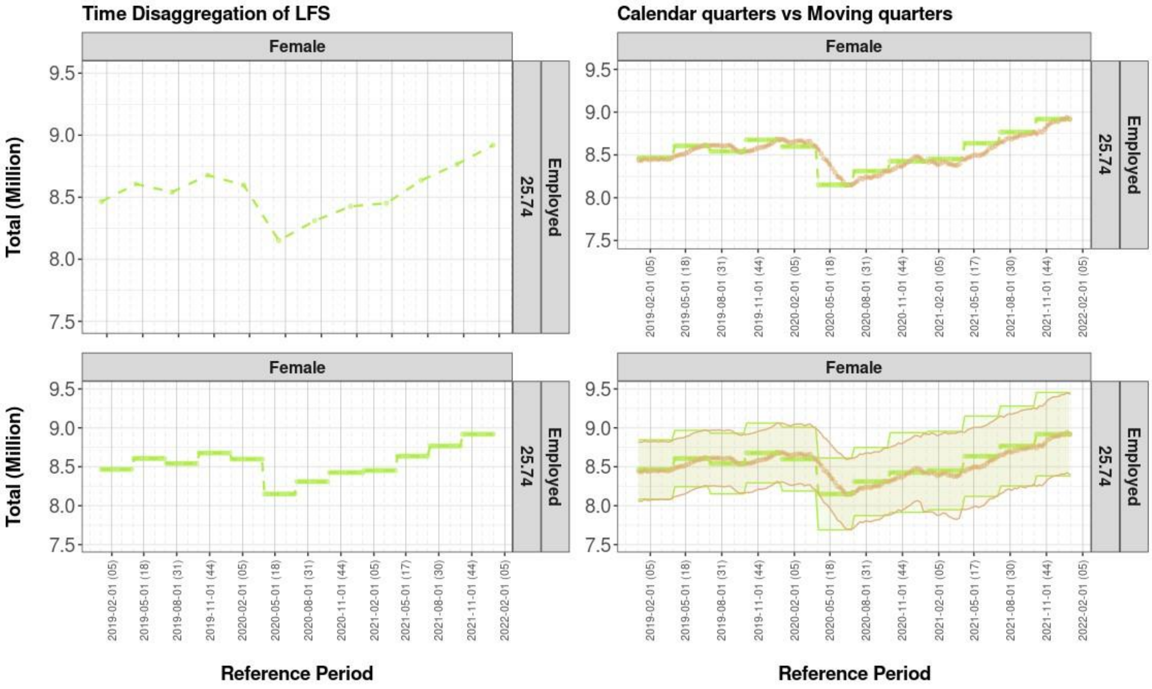


Figure 1: (Top left) Original Spanish quarterly LFS estimates (employed females aged 25-74). (Bottom left) Same original quarterly time series in a weekly time scope. (Top right) Calendar- and moving- quarterly time series updated weekly. (Bottom right) 95% confidence intervals included.

Conclusions

Although many aspects are still under investigation (volatility, turning points, production viability and process adaptation, noise filtering, detailed assessment of timedisaggregated sampling weights, ...), in our opinion some relevant conclusions already arise.

Survey microdata in combination with process paradata and novel statistical methods can notably improve the performance of some classical statistical products by statistical offices. A bottom-up approach focused on the microdata stands up as a serious alternative to the top-down approach based on the application of econometric models upon quarterly times series.

Once the novel statistical methodology is decided for production, the statistical process will need some adjustments. In our opinion, these changes will not be disruptive at all.

Finally, the trade-off between timeliness and accuracy can be stretched out so that traditional data sources are not condemned to sacrifice one of the them. This paves the way for an even greater quality improvement with the integration of new data sources.

References

[1] J.C. Deville and C.-E. S rndal. Calibration estimators in survey sampling. Journal of the American Statistical Association, 87:376 382, 1992.

[2] K.P. Murphy. Machine learning: a probabilistic perspective. MIT Press, 2013.

[3] G. Petris, S. Petrone, and P. Campagnoli. Dynamic linear models with R. Springer, 2009.

[4] D. Haziza and J.-F Beaumont. On the construction of imputation classes in surveys. International Statistical Review/ Revue Internationale de Statistique, 75(1):25 43, 2007.

[5] A.C. Singh and C.A. Mohl. Understanding calibration estimators in survey sampling. Survey Methodology, 22(2):107 115, 1996.

Temporal Disaggregation of the Services Producer Price Index

Keywords: Temporal disaggregation, Chow-Lin, Fernandez, JDemetra+, rjd3bench

Introduction

This document summarizes the steps carried out to disaggregate the Services Producer Price Index (SPPI) of Spain from quarterly to monthly periodicity, using the Consumer Price Index at constant taxes as a proxy for disaggregation.

The relevance of this disaggregation is explained next:

Services Sector Turnover Index (SSTI) at current prices is a monthly index. With the entry into force of the (EU) 2019/2152 Regulation on European Business Statistics and the Commission Implementing (EU) 2020/1197 Regulation, it is established the elaboration of a monthly index of services production (ISP) has to be done. The goal of this indicator is to measure the volume change of value added in close periodic intervals.

Because of the lack of that ISP in a monthly basis we have to use the turnover deflated. An appropriate deflator is the Services Producer Price Index without direct taxes because it is a price producer index, that measures the evolution of prices for households and enterprises sectors. This index has quarterly periodicity, so the option proposed in this document is to apply a temporal disaggregation method to the SPPI series to obtain a monthly SPPI.

Available data

We have the SPPI data series from the first quarter of 2010 until the first quarter of 2018, for the CNAE branches: 50, 51, 53, 61, 62, 63, 71, 73, 78, 80 and 812. On the other hand, we have the CPI at constant taxes series for the same period, but with monthly frequency, for the COICOP products: 0733 (proxy for branch 51 Air Transport) and 0734 (proxy for branch 50 Maritime and Inland Waterway Transport).

Methods

We refer to a related indicator or proxy, as a time series that is linearly related with the time series to be temporal disaggregated. This starting hypothesis is implicit in the definition of the high-frequency model, but can never be verified.

There are two types of temporal disaggregation methods depending on whether a related indicator is available or not:

- There is no high frequency related indicator:
  - Softened methods: such as Boot, Fiebes and Lisman (1967). Identical to Denton considering as proxy a series of ones.
  - Methods of time series: such as: Wei and Stram (1990).

- There is a high frequency related indicator: these methods are known as optimal methods and among them are Chow-Lin (1971), Fernandez (1981) and Litterman (1983), Santos Silva and Cardoso, and Proietti.

The software used is nbdemetra-benchmarking-2.2.2.nbm (a plugin of JDemetra+) and rjd3bench (R package of JDemetra+) to apply Chow-Lin, Fernandez, Litterman and Boot, Fiebes and Lisman, (applying Denton, considering a constant series of ones as proxy) and the Matlab library Temporal Disaggregation to apply Santos Silva and Cardoso, and Proietti.

Temporal disaggregation with a high frequency related indicator

The steps we follow at the INE of Spain to determine the optimal method to use when the series have a high frequency related indicator, are the following:

1. Graphical Analysis
2. Cointegration Test (optional)
3. Chow-Lin
4. Fernandez and Litterman
5. Proietti and Santos Silva and Cardoso

The graphical analysis is complicated and subjective.

Continuing with the next steps, it is assumed that monthly observations (if available) of the series to be estimated satisfy a multiple regression relationship with p related series $x_1, \dots x_p$:

$$y = X\beta + u$$

where y is the observed series, X is the matrix with the regression variables in columns, β is a vector of coefficients and u follows:

- $u_t = \phi u_{t-1} + \epsilon_t$ with $|\phi| < 1$ (Chow-Lin, see [1])
- $\nabla u_t = \phi \nabla u_{t-1} + \epsilon_t$ with $|\phi| < 1$ (Litterman, see [2])
- $\nabla u_t = \epsilon_t$ (Fernandez)

A cointegration test can be done, but it will not be conclusive since this contrast can only be done in low frequency. It must be considered that if the series are cointegrated in high frequency, Chow-Lin should be applied. If the series are not cointegrated in high frequency, we apply Fernandez and Litterman. But we cannot verify this because we do not have the target series in high frequency, and the test conclusions cannot be scaled from low to high frequency, that is why the second step is optional.

Another option can be to go directly to step 3 and 4 and apply Chow-Lin directly, and if ρ is close to 1 then apply Fernandez or Litterman. The ideal model will be determined, also carrying out the diagnosis of the model on the low frequency residuals.

The next step is testing autorregresive distibuted lags (ADL) models. These types of models include a lagged dependent variable that should remove or at least reduce, the autocorrelation in the residuals of the models.

In Santos Silva and Cardoso (2001) the model is:

$$y_t = \phi y_{t-1} + x_t \beta + \epsilon_t$$

where ϵ_t is a white noise process with variance σ^2. This is an ADL(1,0) model.

Proietti (2006) (see [3]) considers the dynamic model:

$$y_t = \phi y_{t-1} + m + gt + x_t\beta + x_{t-1}\gamma + \epsilon_t$$

where ϵ_t is a white noise process with variance σ^2. This is an ADL(1,1) model.

To choose the best method to apply it is necessary to carried out the likelihood ratio contrast between Proietti and Santos Silva Cardoso, and also Proietti versus Chow-Lin, since Santos Silva and Cardoso and Chow-Lin are a particular case of Proietti.

Also, the analysis of the extrapolation of these methods is a good indicator to decide which is the best method to apply.

Temporal disaggregation without a high frequency related indicator

When there is no high frequency related indicator, it is recommended to apply Boot, Fiebes and Lisman, from JDemetra+, running proportional modified Denton, in first differences, considering taking a constant series of ones as proxy.

Results

For simplicity, the results shown here are only refer to the branch 51 Air transport series. We are going to work with the 0733 CPI series and determine if it is a good proxy for the IPS series, branch 51. The first thing we have to do is graphically compare both series.

The graphical analysis is complicated and subjective, it seems that the CPI Air Transport series can be a good proxy for the SPPI Air Transport series (see Figure 1). Now, we must determine what type of method will be applied.

By applying Chow-Lin, in JDemetra+ using the plug-in, it can be verified that $\rho = 0.9819$, with a standard error of 0.017, therefore it seems more convenient to apply Fernandez or Litterman. When executing Fernandez, we can verify that the results are the same as those obtained with Chow-Lin. If we apply Litterman, we can verify that $\rho = 0.2916$ with a standard error of 0.321, therefore, 0 would be within the CI for ρ, which leads us to conclude that it is better to apply Fernandez.

In addition, when carrying out the contrasts on the low frequency residuals, we verify that all the contrasts that indicate that the model is well specified are fulfilled.

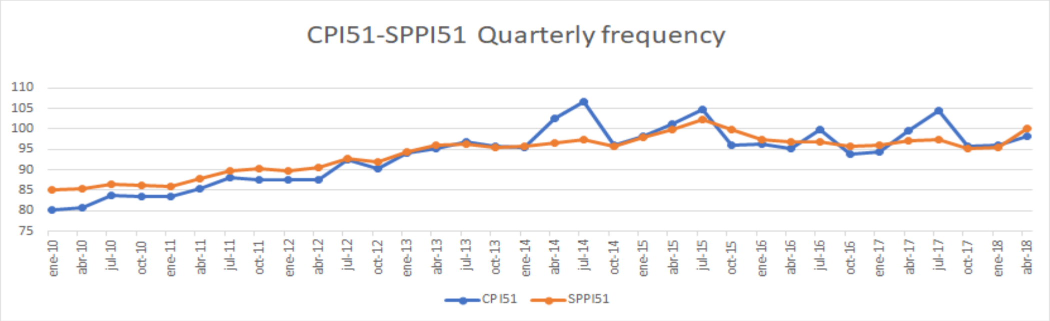


Figure 1: Air transport

The disaggregated series is displayed in the following graph:

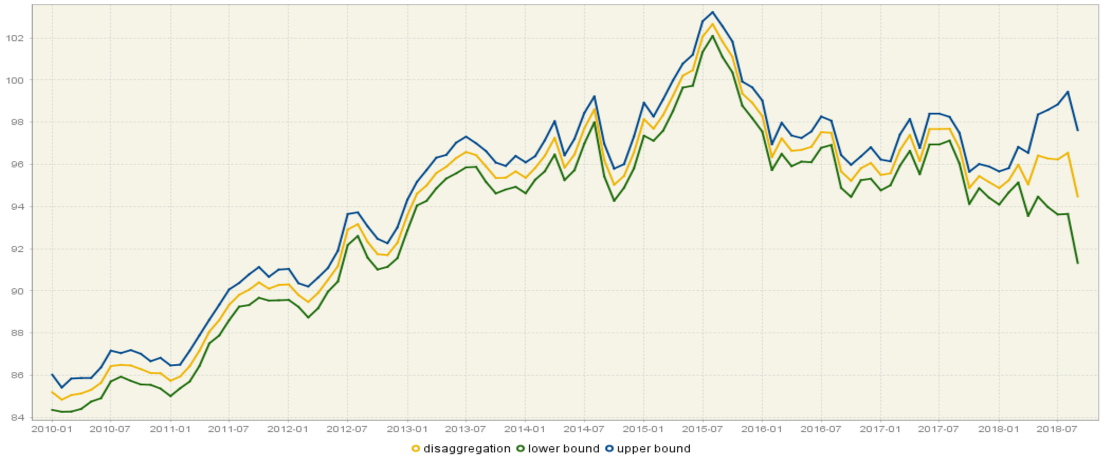


Figure 2: Air transport disaggregated series by Fernandez

We have executed the Santos Silva and Cardoso and Proietti methods on series 51 with Matlab using the Quilis libraries, and we have carried out the likelihood ratio contrast between Proietti and Santos Silva Cardoso, and also Proietti versus Chow-Lin, since Santos Silva and Cardoso and Chow-Lin are a particular case of Proietti.

Since in both cases we reject the likelihood ratio test, Proietti seems to be the most appropriate method. We have also made the diagnosis of the Proietti model on the low frequency residuals. The residuals follow a normal distribution, they are random and uncorrelated. Therefore, the model is well specified, for all the hypotheses that we can test.

It is recommended to disaggregate the 51 series (Air transport) of the SPPI through Proietti, using the CPI 0733 series as a proxy, because the RMSE on the extrapolated data tips the scales slightly to use Proietti instead of Fernandez.

Conclusions

The temporal disaggregation of the SPPI has led to the establishment of a set of steps to determine the optimal method to use when the series have a high frequency related indicator, i.e. to adopt a standard methodology on temporal disaggregation, in accordance with the Guidelines of temporal disaggregation, benchmarking and reconciliation (Eurostat), to save hours of work and also to assure the quality of the process.

For the remaining series there is no proxy so it will be convenient to use Boot, Fiebes and Lisman.

References

[11] Gregory C Chow, An-loh Lin. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series (1971). The Review of Economics and Statistics, Vol. 53, No. 4 (Nov., 1971), pp. 372-375.

[12] R.B. Litterman. A Random walk, Markov model for the Distribution of Time Series. Journal of Business and Economic Statistics, vol. 1(2), pp. 169-173.

[13] T Proietti, Temporal disaggregation by state space methods: Dynamic regression methods revisited, Econometric Journal (2006), vol. 9, pp.357-372.

Machine Learning 2 (MANS2A.2)

Session Chair: **Jean Marc Museux** *(Eurostat)*

Measuring the Italian Sentiment on economy: a Deep Learning application to the sentiment analysis
Elena Catanese *(National Institute of Statistics–ISTAT)*, Francesco Pugliese *(National Institute of Statistics–ISTAT)*

An Experimental Study on Using NLP to Summarize Biotechnology Activities in Türkiye in a Few Words
Bilal Kurban *(Turkish Statistical Institute-TURKSTAT)*

Fuzzy Clustering based Autocoding Methods for Family Income and Expenditure Surveys
Yukako Toko *(National Statistics Center)*, Mika Sato-Ilic *(University of Tsukuba)*

Measuring the Italian Sentiment on economy: a Deep Learning application to the sentiment analysis

Keywords: Sentiment Analysis, Artificial Neural Networks, Deep learning, Twitter data, Word Embedding Models

INTRODUCTION

The purpose of this work is exploring the potentialities of a sentiment classifier built on a long shortterm memory neural network (LSTM). The training stage consists in a mixed two-steps procedure: a) an unsupervised word embedding training model on an unlabelled set of Italian tweets extracted from the Social Mood on Economy Index (SMEI) framework [1]; b) a supervised training step aimed to calculate the sentiment score. This algorithm allows computing a sentiment index about the economy, such as the SMEI. The proposed index has been computed for the first half of 2020 that is the period when the first major Covid-19 outbreak hit the Italian population. The main goal of the present work is analysing the reliability of a general classifier, which has not been trained on data about the pandemic crisis. The paper is structured as follows: in section 2.1, an overview of the neural network will be briefly provided while in section 2.2 the main features of the utilized model and training procedures will be described. In section 3 the results of the analysis will be illustrated, and conclusions will be drawn in section 4.

METHODS

Model

Long short-term memory (LSTM) [2] is a type of Recurrent Neural Networks (RNNs) that can process long sequences of data, therefore it is suitable for text classification and sentiment analysis. A LSTM is usually composed of input and output layers consisting of cells, activated by means of activation functions (e.g, hyperbolic tangents). The cells regulate the flow of information through a system of gates. The proposed model is a neural network built upon a two-layer bidirectional with backward and forward LSTM layers. The input layer of the neural network is an embedding layer, i.e., an embedding space output of a word-embedding model. The embedding model has been built on a corpus of SMEI tweets with the C-BOW method through the fastText algorithm [3]. The output of this model is a vector space, where each word has a semantic vectorial representation. The underlying idea, is that the encoded words that are “closer” in the vector space are expected to be similar in meaning. The dimension of this vector space is set to 300. Then, it is possible to map the input layers into a two-dimensional matrix: one dimension represents the word within the corpus and the other is its vectorial embedding representation. This matrix is the input of the first LSTM layer and the subsequent output is the input of the second LSTM stage. The use of two stacked LSTM layers allows the model to capture the semantic relationships between words and sentences [4]. The first layer has 128 cells while the second 32. Both use a hyperbolic tangent (Tanh) activation function and a dropout rate of 0.5 for regularization. For dimensionality reduction, a 1-dimensional Max Pooling layer [5] is then adopted to convert inputs (with various lengths) into a fixed-length vector. Finally, the output layer is a dense layer, i.e. a single fully-connected layer, which is a binary classifier. It uses a Sigmoid function, which is the predicted sentiment classification of each tweet: if the resulting quantity is higher than 0.5, then the tweet is classified as positive, otherwise negative. All the hyper-parameters have been tuned using a random search algorithm [6].

Data and training process

The training process of the classifier (which allows the sentiment scoring of the output layer) consists of two phases. First, the model is pre-trained on a dataset composed by Italian labelled tweets on various domains. The “pre-training” set is a merge of two popular datasets used for sentiment analysis purposes, Sentipolc [7] and Happy Parents [8]. This dataset consists of 6501 labelled tweets, where the 39.44% are positive tweets. In the second step, the model is fine-tuned by a further training on a balanced set of 900 labelled tweets concerning economic topics used for internal uses in ISTAT. This fine-tuning phase has the benefit of adapting the previously learnt features to the economic domain, which is the objective of the classifier. Both datasets have been split into a training set and a validation set according to a proportion of 80/20. The model classification accuracy has been evaluated in terms of the F1-Score [9]. It is worth to notice that these datasets are not recent (2016 in the best case). The trained model is used to predict the sentiment of a set of 11,979,986 Italian tweets referred to the period January-May 2020 (Covid-Set). The tweets are extracted using a set of keywords related to economy.

The predicted sentiment of each tweet is used to build the daily index, which is computed as

$$I = \frac{Np - Nn}{Np + Nn}$$

where N_p is the share of tweets classified as positive each day while N_n is the share of tweets classified as negative.

The Covid-Set, has been merged with SMEI tweets of the period April-May 2021for a total of 15,115,421 tweets. This set was used to build the embedding space used as input layer.

RESULTS

The proposed model achieves 0.80 as F1-Score on the validation dataset of the first step, 0.79 of the second step. The index created using the predicted sentiment is illustrated in Figure 1. The index records a breakdown since the beginning of the Covid-19 pandemic, when a strong lockdown was imposed to the country. The index shows a level-shift downwards within the period between the 7th of March and the 21st of April (underlined by means of two vertical blue lines in the plot), a period during which the full lockdown was still in place.

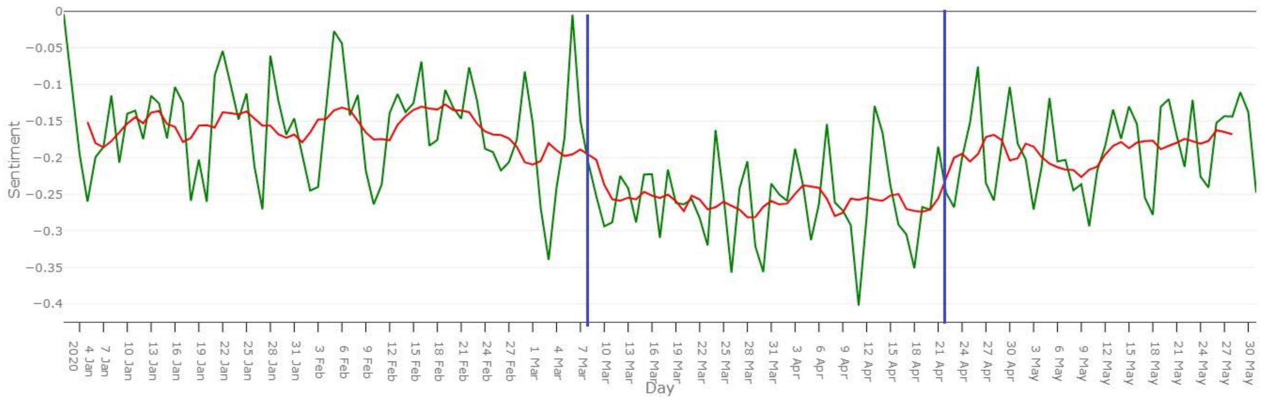


Figure 1: The proposed index evaluated from January to May 2020. The green line represents the daily index, the red line is a 7-days moving average.

To gain more insights on the reasons of this breakdown, we have performed an in-depth analysis on this time interval. Figures 2-A and 2-B display two word-clouds which illustrate the thirty most frequent words between 7th March and 21st April 2020 in the two predicted classes of positive and negative tweets, respectively. A set of stop words has been deleted from the corpus prior to the computation of the frequencies. Neither Coronavirus nor its synonyms are the prominent topics in negative tweets. Indeed, in the negative predicted class, the two most frequent words are *"spesa"* (expenditure) and *"fare"* (do/make), this last one probably linked to the Italian expression *"fare spesa"* (do expenditure related to Italian debt and government expenses).

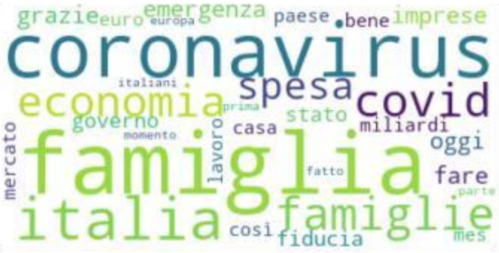


Figure 2-A: Thirty most common words in the tweets classified as period 7 March - 21 April positive in the

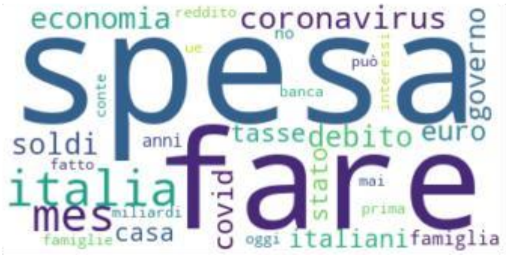


Figure 2-B: Thirty most common words in the tweets classified as negative in the period 7 March - 21 April 2020.

Furthermore, considering tweets that contain the word *"spesa"*, the proportion of negatives increases to 78%, higher than the overall negative proportion in the period (63.25%). Hence, the drop in the sentiment seems to be more related with the word *"spesa"* rather than with coronavirus. Indeed, if we have a look at Figure 2-A, coronavirus is more relevant in positive tweets: the tweets containing coronavirus and its synonyms are less negative than the average (63.25%) in the period, having a percentage of negative tweets of 58.65%. This may be due to a distortion induced by the labelled data used for the training of the model that did not contain any reference to Covid-19. Nevertheless, the embedding space has been built using tweets collected during 2020 and 2021. In this way, the model is able to generate a representation of the coronavirus word in the vector space. However, the embedding space provides a measure of similarity between words, it does not give any indication about tweets sentiment polarity. Therefore, the model may have indirectly classified the polarity of coronavirus using the sentiment assigned to words that are used frequently with it, such as *"famiglia"* (family) or *"Italia"* (Italy), which appear to be associated with a positive sentiment.

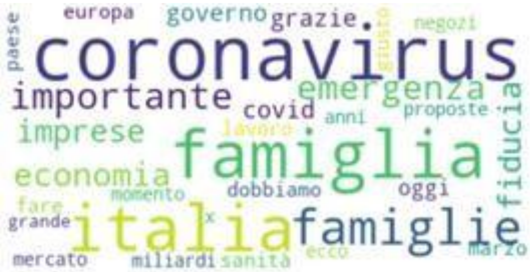


Figure 3-A: thirty most frequent words on the 6th of March 2020 of positive Tweets

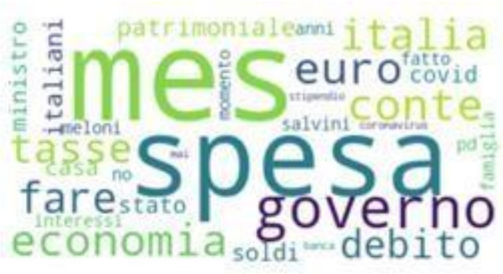


Figure 3-B: thirty most frequent words on the 11th of April 2020of negative Tweets

The index, as shown in Figure 1, has some outliers, which need a deeper analysis. As expected, the maximum of the time-series is observed on the 1st of January. We analyse the second maximum of the sentiment index on the 6th of March and the minimum on 11th of April. In Figure 3, the thirty most common words on the 6th of March for positive tweets are plotted, as well as the most frequent words for negative tweets on the 11th of April. While the minimum value has a consistent meaning, the positive peak seems to be a false positive.

In the minimum value, *"Spesa"* is again among the most common words, a further confirmation that it is correlated with negativity. The debate is focused around *"mes"* (the Italian word for the European Stability Mechanism), which appears to be negatively characterized in the twitter debate, as we observe other words such as *"governo"*(government), euro, *"debito"* (debt), *"taxes"* (tasse) in the conversations about MES.

Concerning the positive peak, we observe again that when Coronavirus appears with the words **"***italia"*, *"famiglia/e"* (family), the tweets tend to be positively classified. The positivity linked to this word seems to confirm the intuition that the classifier assigns it a sentiment according to the words that co-occur with it. Words as *famiglia/e* or *Italy* may have an intrinsic positive meaning, probably due the labelled data set Happy Parents (more likely for family) or Sentipolc (for Italy).

cONCLUSIONS

The proposed index records a breakdown during the Covid pandemic. The model shows a good reliability, mainly because of two reasons: i) the fine-tuning step of the classifier is carried out using SMEI tweets, i.e., an economic annotated dataset; ii) the input of the LSTM model is a word embedding model where Covid tweets have semantic relationships within SMEI tweets. However, we observe some misclassifications due to the training process probably because the labelled dataset did not contain any lexical reference to the pandemic, e.g. Covid19 terms, lockdown. Anyway, results are very encouraging, and the accuracy of the classifier could be improved by using a manually labelled dataset of Twitter conversations about the pandemic for the supervised training step.

rEFERENCES

[1] Zardetto, D. (2018). Using Twitter Data for the Social Mood on Economy Index, Atti della XIII Conferenza nazionale di statistica, Rome, 4-6 July 2018, ISBN 978-88-458-2016-8, (pp. 385-390)

[2] A. Graves, N. Jaitly & A. R. Mohamed, Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (2013), pp. 273-278.

[3] P. Bojanowski, E. Grave, A. Joulin, & T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics (2017), 5, 135-146.

[4] G. Rao, W. Huang, Z. Feng, & Q. Cong. LSTM with sentence representations for document-level sentiment classification, Neurocomputing 308 (2018), 49-57.

[5]Shu, Bo, Fuji Ren, and Yanwei Bao. "Investigating lstm with k-max pooling for text classification." 2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA). IEEE, 2018.

[6] J. Bergstra, & Y. Bengio, Random search for hyper-parameter optimization. Journal of machine learning research (2012), 13(2).

[7] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli V. & Patti, Overview of the evalita 2016 sentiment polarity classification task. In Proceedings of third Italian conference on computational linguistics (2016).

[8] L. Mencarini, D. Hernández Farías, M. Lai, V. Patti, E. Sulis, & D. Vignoli, Happy parents' tweets, Demographic Research (2019), 40, 693-724.

[9] Y. Sasaki, The truth of the F-measure, Teach tutor mater (2007), 1(5), 1-5.

An Experimental Study on Using NLP to Summarize Biotechnology Activities in Türkiye in a few Words

Keywords: NLP, collocation, NLTK, Python, Biotechnology, Top2Vec

Introduction

With the Turkish Statistical Institute (TurkStat) Biotechnology Statistics Survey, it is aimed to create statistical data in the field of biotechnology. All enterprises engaged in biotechnology activities have been covered in the survey and the data is collected directly from enterprises via web survey.

In the 2020 Biotechnology Statistics Survey, unlike previous years, an open-ended question was added to the end of the questionnaire in addition to the routinely asked questions. The purpose of adding this question was not to miss anything about this very technical domain. Because open-ended questions allow the respondent units to freely express the points that the survey designer missed on the subject. The question was: “Briefly inform us about the biotechnology activities carried out by your enterprise and the techniques and applications it uses”. All enterprises carrying out biotechnology activities were required to fill in this open-ended question at the end of the survey questionnaire. Eventually; all 499 enterprises carrying out biotechnology activities in Türkiye filled the free text box provided to them with information about their activities.

This paper presents an experimental study to analyse these free texts and try to understand in a few words what biotechnology enterprises in Türkiye are mentioning about most in these texts in 2020. Of course, these free texts would not be reviewed and analysed one by one, which was not possible anyway. For this reason, a different analysis method based on Natural Language Processing (NLP) and deep learning techniques has been developed using Python to gain insights from these free texts. Natural Language Processing—or NLP for short—in a wide sense to cover any kind of computer manipulation of natural language [1]. The reason for choosing Python programming language is that it has an excellent functionality for processing linguistic data.

Methods

In order to extract insights from the responses provided by enterprises to the open-ended biotechnology survey question as free text, the following processes were carried out in Python programming language:

1. Preparing data by pre-processing,
2. Finding collocations,
3. Finding most common words and
4. Detecting topics present in text with Top2vec algorithm

The second, third and fourth steps mentioned above were repeated once more after both removing stop words and lemmatization.

Preparing data by pre-processing

Before insights could be gained from the free texts, first of all, some operations needed to be done on the raw data. These operations covered deleting nonsense records (the records with only "." or "x"), standardizing some words (such as removing the hyphen between words) and converting all letters to lowercase. After these operations performed on the raw data, the data was ready for pre-processing. Pre-processing stage included tokenization and detecting and removing punctuations.

The analyses carried out to extract meaning from the free texts in this study were also repeated by removing stop words and performing lemmatization. Stop words are basically a set of commonly used words in any language and in NLP and text mining applications, they are used to eliminate unimportant words. As for the lemmatization, it is the method to take any kind of word to that base root form with the context. It groups together the different inflected forms of a word so they can be analysed as a single item [2].

Finding collocations

After the pre-processing was completed, analyses were started to extract meaningful information from the word list obtained. For this purpose, collocations in the word list were searched first. Collocations are expressions of multiple words which commonly co-occur [3]. Natural Language Toolkit (NLTK) library in python has been used to find colocations. NLTK is a leading platform for building Python programs to work with human language data [4]. NLTK contains *collocations* module having tools to identify collocations within corpora.

Finding most common words

One way to find out what is most frequently mentioned in free texts is to look at the most frequently used words in these texts. For this purpose, number of occurrences of each individual of the word/word group were calculated through the *FreqDist* module in NLTK.

Detecting topics present in text with Top2vec algorithm

In this study, it has also been tried to find how many different topics can be produced and what these topics can be by combining similar words in free texts using the *Top2vec* algorithm. Top2Vec is an algorithm for topic modelling and semantic search. It automatically detects topics present in text and generates jointly embedded topic, document and word vectors [5].

Results

First; the bigram and trigram collocations were obtained twice, before and after removing stop words, using three different measure of association, namely Pointwise Mutual Information (PMI), Likelihood Ratio and Raw Frequency.

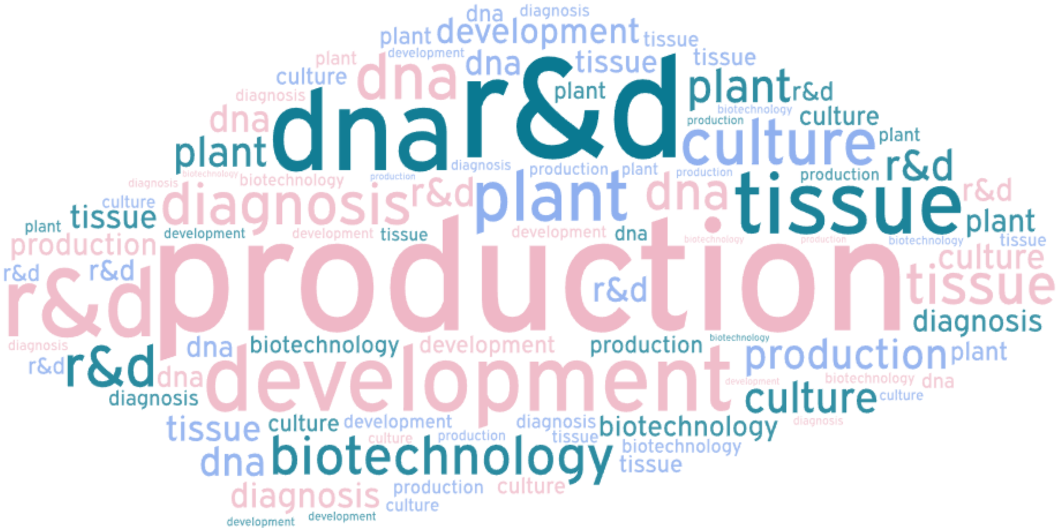
Then, the most used words in the texts entered into the open-ended question, were found as single, double and triple before and after removing stop words. The most important finding here was that after removing stop words, the three-word phrase "plant tissue culture" was repeated six times.

Finally, the top2vec algorithm was used to find similar words used in free texts and to derive topic titles from these word groups. Top2vec derived four topics and also visualized the similar words it grouped.

The words in the free texts were also lemmatized. But this did not work very well in this case as the package used could not find the root words well enough.

Conclusions

Based on their responses to the open-ended question “Briefly inform about the biotechnology activities carried out by your enterprise and the techniques and applications it uses” in TurkStat Biotechnology Statistics Survey, in 2020, the enterprises that carry out biotechnology activities mostly mentioned the words on the below word cloud.



dna diagnosis
plant development tissue
plant development dna tissue tissue
culture dna plant plant r&d
diagnosis dna r&d biotechnology
dna production culture
plant culture plant
diagnosis biotechnology diagnosis production plant production production r&d
culture diagnosis r&d plant dna tissue r&d
plant tissue plant
production development development tissue development dna tissue
culture
r&d r&d production r&d tissue
diagnosis
diagnosis dna biotechnology development production biotechnology
r&d r&d development production plant
dna biotechnology dna
diagnosis culture culture
tissue culture development diagnosis biotechnology diagnosis
culture production tissue
dna biotechnology biotechnology
tissue
diagnosis production culture
culture
development development

Figure 18. Most emphasized single words in open-ended texts

When we dig a little deeper, biotechnology enterprises mention these two words the most in order of emphasis): molecular biology, tissue culture, diagnosis kit, cell culture, R&D activities, test kit, agricultural biotechnology, plant tissue and genetic diagnosis.

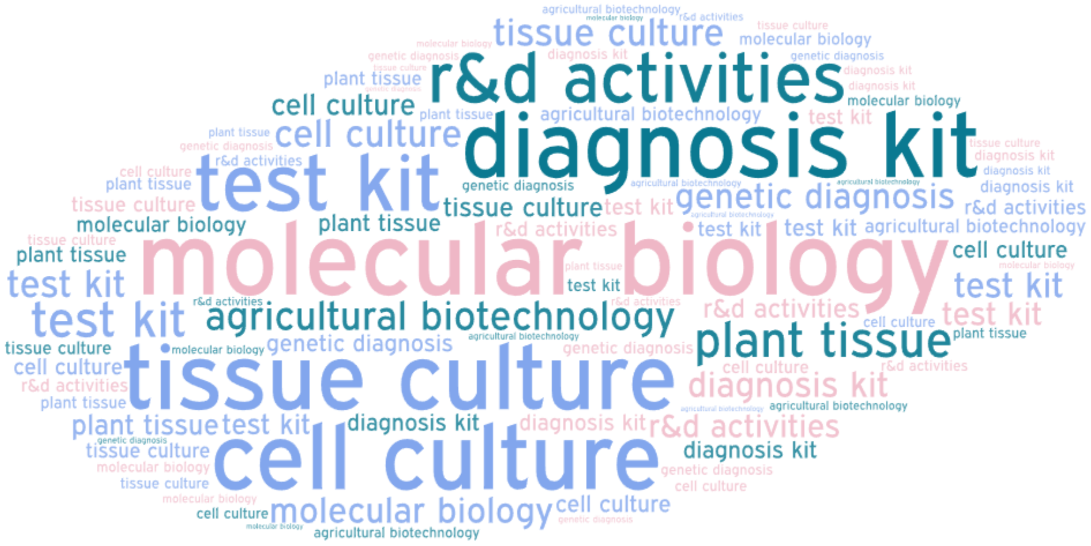


Figure 19. most emphasized two words in open-ended texts

NLP code outputs give a deeper insight into the biotechnology activities of the enterprises. As for three-word expressions, the most emphasized expressions in open-ended texts are: plant tissue culture, research and development, next generation sequencing, real time PCR, high value added, food supplement, diagnosis kit development, microbial fertilizer production, covid19 diagnosis kit, human and animal and development and production.

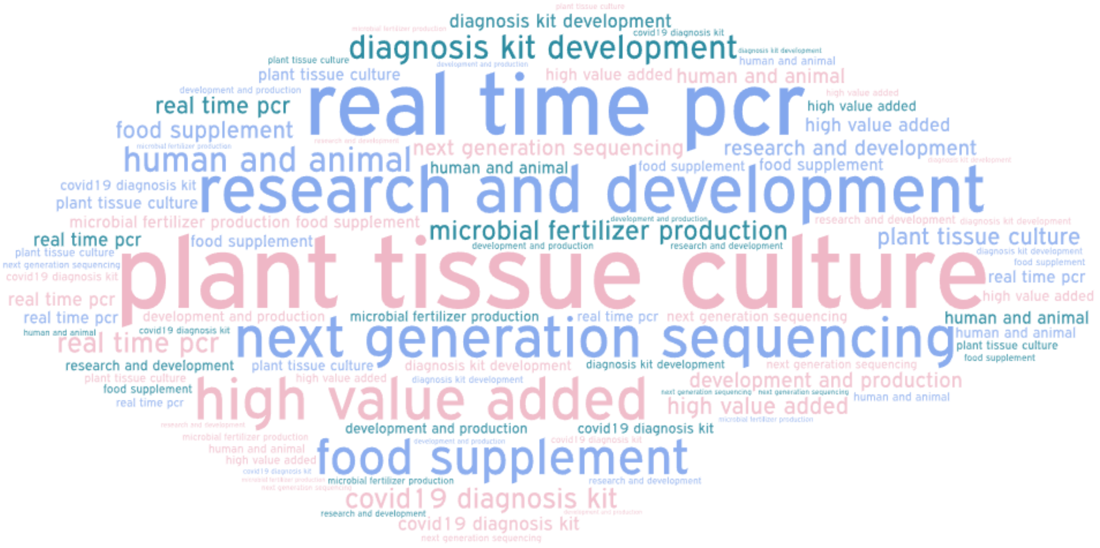


Figure 20. most emphasized three words in open-ended texts

Since the reference period of the Biotechnology Statistics survey is 2020, when the pandemic is more intense, the effect of covid-19 on the results can be easily seen (such as real time PCR, diagnosis kit development, covid19 diagnosis kit etc.).

As a result of this work, it was concluded that NLP and deep learning techniques can be used to reduce human effort in extracting meaning from responses to an open-ended free text. They can automate and speed up an otherwise laborious or infeasible task

References

[14] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, (2009), ix.

[15] https://jaimin-ml2001.medium.com/stemming-lemmatization-stopwords-and-n-grams-in-nlp-96f8e8b6aa6f, accessed 23 September 2022

[16] https://www.nltk.org/howto/collocations.html, accessed 23 September 2022

[17] https://www.nltk.org/, accessed 23 September 2022

[18] D. Angelov, Top2Vec: Distributed Representations of Topics, (2020), arXiv.org, accessed 23 September 2022

Fuzzy Clustering based Autocoding Methods for Family Income and Expenditure Surveys

Keywords: Coding, Machine Learning, Support Vector Machine, Fuzzy c-Means, Reliability Score

Introduction

Fuzzy measure based autocoding methods have been developed and practically implemented for the coding tasks of the Family Income and Expenditure Survey in Japan [1, 2, 3, 4]. These surveys collect data on households' daily incomes and expenditures, such as text descriptions of item names of income and expenditure, their prices, and dates [5]. The data include purchase descriptions extracted from shopping receipt images collected via online survey systems. Each collected text description is assigned a corresponding classification label for data processing (coding task).

For this task, our developed autocoding methods are based on measures obtained by fuzzy clustering methods [6, 7] in Computational Intelligence (CI), which is a part of Artificial Intelligence (AI) known for high performance in dealing with cognitive uncertainty. These autocoding methods are classified into two methodologies. One methodology consists of autocoding methods based on reliability scores defined using both probability measure and fuzzy partition measure, considering human uncertainty when recognizing the autocoding of words for adjusting the complexity of data [1, 2]. Another methodology consists of autocoding methods combined with support vector machine (SVM) [8] and fuzzy c-means (FCM) method [3, 4].

The first methodology-based methods have been implemented in the autocoding system for the Family Income and Expenditure Survey. However, the data are predicted to be more complex with an increase in the number of shopping receipt data, so we must consider robustness, tractability, and low computational cost for obtaining a better performance considering the complexity and large size of data. Therefore, recently, we have developed a combination method of a CI method known as FCM and a machine learning method known as SVM since both techniques are beneficial for obtaining robustness, tractability, and low computational cost when dealing with a large complex amount of data. Although similar projects have progressed in the coding task [9], our proposed methods exploited the CI technique (fuzzy clustering based measure) into the conventional machine learning technique, and this is the special feature of these methods to deal with human cognitive uncertainty with linguistic variables.

Autocoding Method based on reliability scores

The autocoding method based on reliability scores [1, 2] extracts objects and retrieval of candidate classes from the object frequency table provided by using the extracted objects. Then, it calculates the relative frequency of *j*-th object to a class *k* defined as

$$p_{jk} = \frac{n_{jk}}{n_j}, \quad n_j = \sum_{k=1}^{K} n_{jk}\,, \quad j = 1, \dots, J, \qquad k = 1, \dots, K,$$

where n_{jk} is the number of occurrence of statuses in which an object *j* is assigned to a class *k* in the training dataset. *J* is the number of objects and *K* is the number of classes. Then, the classifier arranges $\left\{p_{j1},\cdots,p_{jK}\right\}$ in descending order and creates $\left\{\tilde{p}_{j1},\cdots,\tilde{p}_{jK}\right\}$, such as $\tilde{p}_{j1} \geq \cdots \geq \tilde{p}_{jK}, j = 1,\cdots,J$. After that, $\left\{\tilde{\tilde{p}}_{j1},\cdots,\tilde{\tilde{p}}_{j\tilde{K}_j}\right\}$, $\tilde{K}_j \leq K$ are created. That is, each object has a different number of classes. Then, the classifier calculates the reliability score for each class of each object. The reliability score of j -th object to a class k is defined as

$$\bar{\bar{p}}_{jk} = g(n_j) \times \bar{p}_{jk},$$

where

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, 1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}\right), \qquad j = 1,\dots,J, k = 1,\dots,\tilde{K}_j, \qquad (1)$$

or

$$\bar{p}_{jk} = T\left(\tilde{\tilde{p}}_{jk}, \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2\right), \qquad j = 1,\dots,J, k = 1,\dots,\tilde{K}_j. \qquad (2)$$

These reliability scores were defined considering both probability measure and fuzzy measure. That is, $\tilde{\tilde{p}}_{jk}$ shows the uncertainty from the training dataset (probability measure) and $1 + \sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm} \log_K \tilde{\tilde{p}}_{jm}$ or $\sum_{m=1}^{\tilde{K}_j} \tilde{\tilde{p}}_{jm}^2$ shows the uncertainty from the latent classification structure in data (fuzzy measure). These values of the uncertainty from the latent classification structure can show the classification status of each object; that is, how each object is classified to the candidate classes. *T* shows *T*-norm in statistical metric space [10]. We generalize the reliability score by using *T*-norm, which is a binary operator in statistical metric space. Furthermore, to prevent an infrequent object having significant influence, sigmoid functions $g(n_j)$ were introduced to the reliability score. After the class assignment based on reliability scores, a model improvement process is performed: First, it extracts data whose reliability scores are 0.99 or over from the evaluated dataset. Then, it adds the information of the extracted data to the original training dataset. After that, the evaluation data is performed class re-assignment using the improved training dataset.

The hybrid autocoding system is applied to *the Family Income and Expenditure Survey* dataset: First, the rule-based method is applied to the target dataset of text descriptions, and it assigns classes to clear text descriptions that have clear classifiable features to specific classes. Then the autocoding method based on reliability scores was applied to the remaining uncertain text descriptions. Table 1 shows the classification accuracy of the hybrid autocoding method for each threshold of reliability score. From Table 1, it is found that the developed hybrid method can assign classes with over 0.99 classification accuracy when the threshold of reliability score is set as 0.99, 0.98, or 0.95.

Table 8. Classification Accuracy of the hybrid autocoding system

Total target (a)	Number of text descriptions							Hybrid	
	Rule based method		Machine learning method based on reliability score						
	Assigned (b)	Coverage (b) / (a)	Target text descriptions (a) - (b)	Threshold of reliability score (r. s.)	Assigned (c)	Correctry assigned (d)		Coverage ((b) + (c)) / (a)	Accuracy ((b) + (d)) / ((b) + (c))
881,419	619,545	0.703	261,874	r.s. ≧ 0.99	49,014	47,035		0.759	0.997
				r.s. ≧ 0.98	65,637	62,204		0.777	0.995
				r.s. ≧ 0.95	80,729	75,766		0.794	0.993
				N/A	256,208	188,503		0.994	0.923

autocoding method based on SVM and FCM method

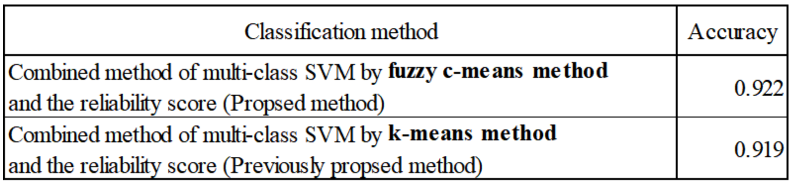
The combined autocoding method of multi-class SVM and fuzzy c-means method [3, 4] performs the following. First, the proposed method tokenizes each text description. Then, it obtains numerical vectors corresponding words utilizing Word2Vec [11] and normalizes each feature of the obtained set of numerical vectors. After that, it produces sentence vectors for each text description based on the normalized vectors. Then, it applies fuzzy c-means method to sentence vectors to classify them into several clusters. After applying fuzzy c-means method, the proposed method assigns corresponding classes applying SVM for each dataset of each cluster. Then, it extracts unmatched data and implements re-classification based on the reliability score to the extracted data.

Table 2 shows the classification performance of the combined method of multi-class SVM and k-means method (hard clustering) and the reliability score. In this table, reliability score (a) is a case when we apply equation (1), and reliability score (b) is the case when we use equation (2). From this table, it can be seen that the classification accuracy of the proposed hybrid method of multi-class SVM and a classification method based on reliability scores obtained a better result compared with both the classification method based on reliability scores and classification by simply applying SVM. The proposed hybrid method obtained a classification accuracy of 0.919, whereas the classification method based on reliability scores obtained 0.888 and simply applying SVM obtained 0.856, respectively. That is, 5,142 text descriptions are increased to be assigned correctly by using the proposed method from the only use of SVM. In addition, Table 3 shows the comparison of the classification accuracy of the proposed fuzzy clustering based and the k-means based hybrid autocoding method. From this table, it can be seen that the fuzzy clustering based method can obtain better results compared with the k-means based method.

Table 2. Comparison of Classification Accuracy of the Proposed K-means based Hybrid Autocoding Method and Conventional Methods

Classification methods	Number of text descriptions			Accurucy
	Training	Evaluation	Correctry assigned	
Classification by the proposed hybrid method (reliability score(a))	733,036	81,455	74,855	**0.919**
Classification by the proposed hybrid method (reliability score(b))			74,851	**0.919**
Classification by SVM			69,713	0.856
Classification based on the reliability score (a)			72,317	0.888
Classification based on the reliability score (b)			72,362	0.888

Table 3. Comparison of Classification Accuracy of the Proposed Fuzzy Clustering based, and the K-means based Hybrid Autocoding Method



Classification method	Accuracy
Combined method of multi-class SVM by **fuzzy c-means method** and the reliability score (Propsed method)	0.922
Combined method of multi-class SVM by **k-means method** and the reliability score (Previously propsed method)	0.919

Conclusions

This abstract presents a summary of recently proposed fuzzy clustering based autocoding methods for the Family Income and Expenditure Survey in Japan. The features of these methods employ a fuzzy clustering measure. Several methods of them are employed practically for the production of Statistics. However, faced with the more complex increasing the number of shopping receipts data, we are in the progress of employing the newer technologies of machine learning. And by combining with the fuzzy technology based on CI, we have proposed further advanced autocoding methods. Several numerical examples show a better performance of the proposed methods.

References

[19] Toko, Y., Sato-Ilic, M., (2020), “Improvement of the training dataset for supervised multiclass classification”, Czarnowski, I., Howlett, R.J., Jain, L. C. (Eds.), Intelligent Decision Technologies, Smart Innovation, Systems and Technologies, Springer, Singapore, Vol. 193, pp. 291-302.

[20] Toko, Y., Sato-Ilic, M., (2021), “Hybrid autocoding method for the Family Income and Expenditure Survey”, in JCS 41st Annual meeting proceeding.

[21] Toko, Y., Sato-Ilic, M., (2021) “A hybrid method of multi-class SVM and classification method based on reliability score for autocoding of the Family Income and Expenditure Survey”, Smart Innovation, Systems and Technologies, Vol. 238, pp. 403-414. Springer.

[22] Toko, Y., Sato-Ilic, M. (2022), “Autocoding based Multi-Class Support Vector Machine by Fuzzy c-Means”, Journal of Romanian Statistical Review, Vol. 1, pp. 27–39.

[23] Statistics Bureau of Japan: Outline of the Family Income and Expenditure Survey. Available at: https://www.stat.go.jp/english/data/kakei/1560.html, last accessed 21 Spt. 2022.

[24] Bezdek, J.C. (1981), Pattern recognition with fuzzy objective function algorithms, Plenum Press.

[25] Bezdek, J.C., Keller J., Krisnapuram, R., Pal, N.R. (1999), Fuzzy models and algorithms for pattern recognition and image processing, Kluwer Academic Publishers.

[26] Cristianini, N., Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press.

[27] Benedikt,L., Joshi, C., Nolan, L., Wolf,. N.D., Schouten, B. (2020), Optical Character Recognition and Machine Learning Classification of Shopping Receipts, Available at: https://ec.europa.eu/eurostat/documents/54431/11489222/6+Receipt+scan+analysis.pdf, last accessed 29 Spt.2022.

[28] Menger, K. (1942), “Statistical metrics”, in Proceedings of the National Academy of Sciences of the United States of America, Vol. 28, pp. 535-537.

[29] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003) “A neural probabilistic language model”, Journal of Machine Learning Research, 3, pp. 1137-1155.

MNO data (GASP2A.3)

Session Chair: **Fabio Ricciato** *(Eurostat)*

Estimation of the length of stay of foreigners in Poland using mobile big data
Maciej Beręsewicz *(Poznań University of Economics and Business)*

Use of mobile network data for official statistics
Dorian Le Jeune, Sandra Hadam, Natalie Rosenski (*Statistisches Bundesamt (Destatis*))

Synthetic mobile networks data generation
Bogdan Oancea *(National Institute of Statistics and University of Bucharest)*, David Salgado *(National Statistical Institute-INE)*, Marian Necula *(National Statistical Institute-INSSE)*, Sandra Barragan *(National Statistical Institute-INE)*, Luis Sanguiao-Sande *(National Statistical Institute-INE)*

Generation of Synthetic Data from actual data of Mobile Network Operators (MNO) through Generative Adversarial Networks (GANs): motivations, techniques and preliminary results
Francesco Pugliese, Massimo De Cubellis *(National Institute of Statistics–ISTAT)*

Estimation of the length of stay of foreigners in Poland using mobile big data

Keywords: non-probability samples, official statistics, measurement error, coverage bias, multiple imputation

1. INTRODUCTION

In this study we utilize mobile big data sources to measure length of stay of foreigners in Poland. Dataset used in the study comes from advertising systems from over 30 mln smartphone users in Poland (almost complete coverage) by Selectivv company. Contrary to Call Details Records (CDR) or signaling data from mobile network providers Selectivv data contains socio-demographic information about the users. While this sounds interesting, background information regarding gender, age, nationality or length of stay is a result of rule-based or machine learning (ML) algorithms which introduce measurement error. Thus, may provide misleading results if misclassification error is high due to low quality input information or algorithms.

In order to assess this error we conducated a validation study using a random sample survey of 500 smartphone users. We asked respondents to provide socio-demographic background to assess the quality of the profiling algorithms. Then, under assumption that survey answers are correct, we correct measurement error using multiple imputation. Unfortunately, at the time of writing the abstract we still waited for the ML estimates of the length of stay of the survey respondens. Thus, we only report length of stay by the corrected demographic variables.

The abstract have the following structure. In section 2, we describe Selectivv data, how it is created and how the length of stay is derived based on users activity. We provide short description how multiple imputation is applied. In section 3 we provide validation study results and selected results of the length of stay by demographic variables. In the final analysis we will correct for the under-coverage of selectivv data by using administrative data.

2. METHODS

2.1. Mobile big data

Selectivv company uses advertisement systems to collect information about the smartphone users. They use Google Advertisement ID (GAID) and Apple’s The Identifier for Advertisers (IDFA). As these IDs refer to devices, Selectivv is using rule-based algorithm to assign multiple devices to users. This approach is based on co-occurance of devices in time and space (based on geolocalisation) as well as information about the wifi connections.

Table 1 presents information about the number of SIM cards for Polish and Ukrainian users for top 4 mobile providers in Poland observed by Selectivv and the postulaed number of mobile users based on Selectivv algorithms. If we compare this number to the population size based on administrative records we notice almost full coverage for the Polish population (here we compare with population register PESEL but it covers only a subset of foreigners) and overcoverage for Ukrainian (here we compare with the insured working population ZUS). Certainly, duplicates are the issue but to unknown factor.

Table 1. Number of SIM cards and estimated number of mobile users in Poland in 2021Q3 (age 18+)

Source	Polish	Ukrainian
Provider 1	15 537 349	413 913
Provider 2	14 865 422	527 565
Provider 3	12 448 975	423 834
Provider 4	10 940 555	492 786
Total	53 792 301	1 858 098
Selectivv estimate of # mobile users	30 988 823	1 258 330
Admin data	31 142 016 (PESEL)	616 901 (working, ZUS)

Source: own elaboration.

Selectivv assigns gender and age using ML algorithms based on smartphone activities and surveys conducted by Selectivv. Nationality (e.g. Polish, Ukrainian) is assigned based on mobile provider, geolocalisation and OS language. Length of stay in Poland (30 days to 3 months, 3-12 months and 12+ months) is based on the geolocalisation and user activity.

2.2. Validation study

We designed a validations study using advertisement systems on smartphones with assumed sample of 500 users. Selectivv send over 55k invites and so the response rate was less than 1% which is inline with existing studies on web/online surveys [1]. Sample was stratified according to gender, age and nationality derived by Selectivv.

2.3. Correction of measurement error

We apply multiple imputation to correct for the measurement error using mass imputation approach [2]. This means that we impute *true* gender, age group, nationality for the Selectivv aggregated data based on validation sample.

3. RESULTS

3.1. Validation study results

Table below presents results of the misclassification error measured in the validation study. We present selected results for gender, age and nationality. In general, Selectivv's classification algorithms provide quite good results with margin of error around 10-12% for gender, 4-15% for age and 4-6% for nationality.

Table 2. Results of validation study

Variable	**Classification Error (in %, sample size provided in brackets)**
Gender	Males (10.6%, n=268), Females (12.8%, n=215)
Age	18-24 (12.3%, n=227), 25-29 (15.1%, n=146), 30-39 (8.2%, n=61), 40+ (4.1, n=49)
Nationality	Belarusian (4%,n=101), Polish (3.5%,n=230), Ukrainian (5.9%, n=152)

Source: own elaboration

3.2. Length of stay

Table 3 provides point estimates of the length of stay after correcting for the measurement error for nationality (Belarisuan, Ukrainians) and age (only for Ukrainians). We see that majority of Belarusians stays in Poland to 3 months or over 12 months. This may be connected with the labour market: short term (to 3 months) and long term (over 12 months) contracts. For Ukrainians we see that majority of young people stay over 12 months which may be connected with studying in Poland. As the age increases the length of stay in Poland decreases. This may be connected to the organisation of labour market, as majority of foreigners in Poland are migrating because of work.

Table 3. Point estimates of the length of stay after correcting for measurement error for selected variables for 2021Q3

Variable	Level	30 days to 3 months	3 to 12 months	over 12 months
Nationality	Belarusian	37.44	26.23	36.33
	Ukrainian	29.58	37.53	32.89
Age (*for Ukrainians only*)	18-24	19.06	23.22	57.72
	25-29	31.26	40.22	28.52
	30-39	23.76	43.10	33.14
	40+	43.51	31.48	25.01

Source: own elaboration.

4. CONCLUSIONS

In this abstract we showed selected results from the length of stay estimation. Presented results are initial and more in depth comparison with other sources such as administrative data is needed. We will provide more details and updated results at the conference.

REFERENCES

[1] Bethlehem, J., & Biffignandi, S. (2021). Handbook of web surveys. John Wiley & Sons.

[2] Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

Use of mobile network data for official statistics

Keywords: Mobile network data, official statistics, mobility indicators, intercensal population updates, static, dynamic

Introduction

As part of the generally increasing use of digital technology, official statistics are facing the challenge to explore and employ new data sources and to organise their processes and procedures accordingly. The use of new digital data, such as mobile network data, provides an opportunity to supplement official statistics, to open up new subject areas and to relieve the burden on respondents. For this reason, various feasibility studies have been conducted to examine the extent to which mobile network data are suitable for mapping the static and dynamic population and to identify potential use cases. In the following, two projects are presented that incorporate static and dynamic considerations.

Experimental georeferenced population figure based on intercensal population updates and mobile network data

Up-to-date small-area population figures are indispensable for political decision-making. Intercensal population updates allow the provision of up-to-date numbers of inhabitants at the geographical municipality level. The number of inhabitants is continuously updated on the basis of the 2011 Census using data from the statistics on births and deaths and migration statistics. A new experimental approach is used in addition to intercensal population updates so that the growing demand for smaller-area population figures can be met in the short term.

Mobility indicators based on mobile network data

Analyses of anonymised mobile network data provide special insights into the dynamics of population mobility. With the appearance of the COVID-19 pandemic, highly frequent mobility data displayed its potential use as a source of public information with high relevance for decision-makers on different levels. Besides this crisis scenario, mobile network data also allows for a quick evaluation of certain policy measures. The so-called “9-Euro-Ticket” introduced in June 2022 in Germany provided cheap access to public transportation for a period of three months. Using mobile network data, the impact of this policy measure was analysed for different means of transport.

Methods

Depending on the feasibility study and the research question, different mobile network data are processed and analysed, which is why different methods are used for the respective analysis. .

Distribution procedure to obtain experimental georeferenced population figure

To meet the growing demand for small-area population data, experimental georeferenced population figures were compiled based on intercensal population updates and using mobile network data. In a distribution procedure, the results of intercensal population updates were redistributed at the municipality level to a 1x1 km grid for the whole of Germany by means of mobile network data.

First, the 1x1 km grid cells of the mobile network data are allocated to municipalities, taking their area coverage as a basis. Based on this, the group-specific probabilities of the smaller-area 1x1 km grid cells to be selected from the potential resident population are determined from the mobile network data. They represent, in a simplified way, the proportions of mobile network activities per grid cell in the covered municipality. The number of inhabitants per municipality as obtained by intercensal population updates is then multiplied by the calculated group-specific probabilities and, on that basis, distributed over the small-area 1x1 km grid cells. Mobile network data thus provide a spatial distribution of the population within a municipality, which can supplement the existing intercensal population updates. In a last step, the population data in a small-area distribution per municipality are rounded using the official population figure of the relevant municipality.

Mobility indicators based on mobile network data

To address different aspects of population mobility the data is aggregated to different spatial and temporal levels. Spatially, aggregation is performed from grid cells to larger geographical units. In a similar fashion, temporal aggregation is used to compare the activities during day- and nighttime. In addition, the data provider applies an algorithm to identify the modes of transport for journeys over a distance of 30 kilometres.

The mobility indicators display the change in recorded activities compared to 2019 per geographical unit. These can differ depending on the level of aggregation. To improve comparability, the changes in activities are calculated in comparison to the mean of the respective weekday in the respective month of 2019. Public holidays are compared to the corresponding days in 2019 since these periods are linked to significant changes in population mobility.

The results of the feasibility studies are published freely accessible and interactive as experimental data.

Experimental georeferenced population figure

The experimental georeferenced population figure determined is shown for every grid cell if the cells are currently filled with mobile network activities and are not subject to anonymisation. The experimental georeferenced population figures are shown in ascending order (Figure 1 **a**). Light coloured cells have a low experimental georeferenced population figure, while dark coloured cells have a higher figure. This additionally allows the current population distribution to be compared between regions. As was expected, there are marked differences in the regional distribution of the experimental georeferenced population figure between urban and rural areas.

To better assess the plausibility of the results, various geodata from German land surveying are used and combined (Figure 1 b). The goal is to determine which grid cells show residential land or residential use and, consequently, whether resident population is to be expected, or can be ruled out, in the relevant grid cells. For the purpose, the datasets “Amtliche Hausumringe Deutschland” (HU-DE) and “Haushalte-Einwohner-Bund” (HH-EW-Bund) from the Federal Agency for Cartography and Geodesy are used. Overall, 67.8% of the grid cells and the allocated experimental population figures are now considered plausible, 22.1% partially plausible, and only 10.1% implausible.

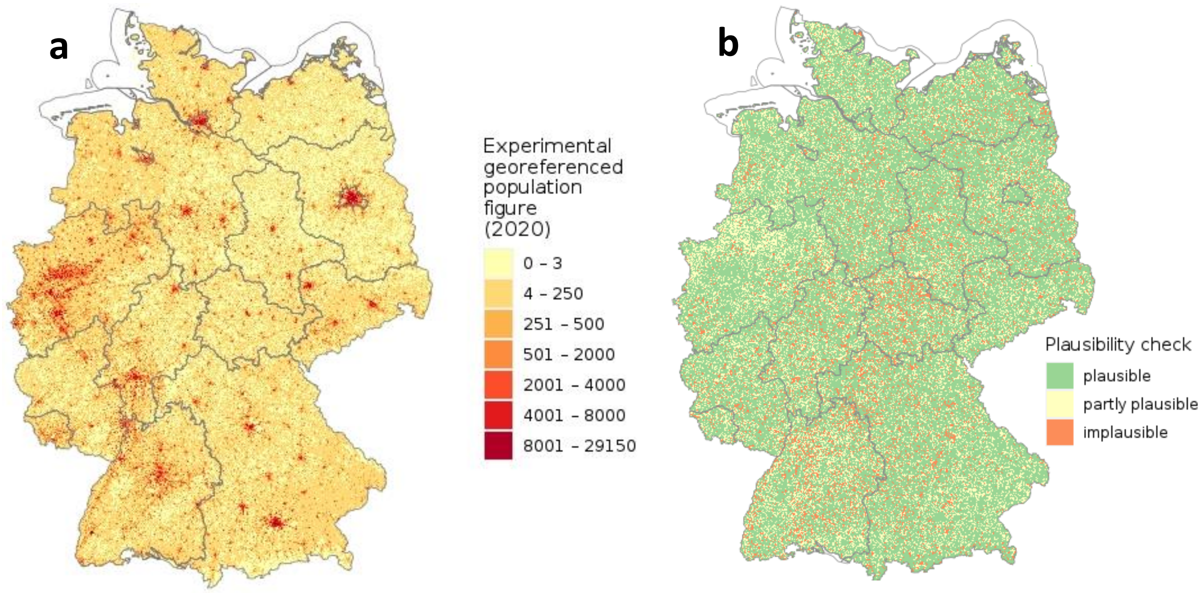
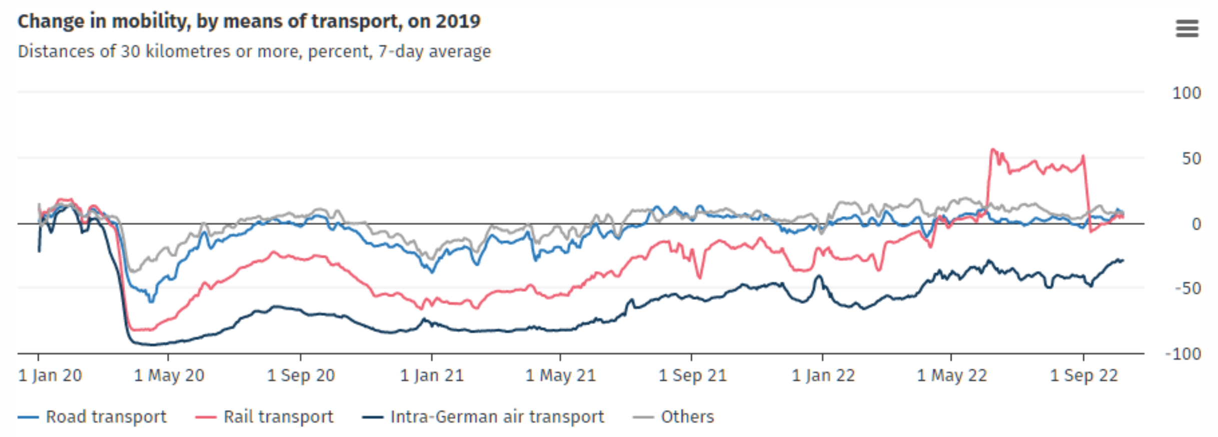


Figure 1. Experimental georeferenced population figure (a) and results of the plausibility check (b)

Mobility indicators based on mobile network data

The mobile network data has proven useful to evaluate the impact of policy measures connected to population mobility. The implementation of time-limited restrictions in reaction to the COVID-19 pandemic is visible in the data and can be analysed on different levels of geographical aggregation. Additionally, the mobility indicators have been combined with several other data sources, e.g. incidence rates, to allow for a more in-depth analysis of the pandemic.

The information on the usage of different means of transport was used to examine the impact of the so-called “9-Euro-Ticket”, which was available in Germany from June to August 2022 (Figure 2). The ticket allowed for cheap access to most means of public transportation. Through the analysis of mobile network data we were able to promptly provide valuable information on the effectiveness of this policy measure.


Change in mobility, by means of transport, on 2019

Distances of 30 kilometres or more, percent, 7-day average

Figure 2. Change in mobility, by means of transport, on 2019

Conclusions

The two experimental products presented, based on mobile network data, show that new digital data can be used to react quickly to urgent information needs and thus support official statistics.

Overall, however, assessing the quality of the results still involves reservations. Using mobile network activities of just one network provider in Germany very likely leads to biased results and uncertainties. This is mainly due to regional market shares and to the data processing methodology applied by the data provider that is not disclosed in detail.

Further feasibility studies, however, especially on the desired utilisation of mobile network data for the production of official statistics, will require anonymised individual data from all mobile network operators in Germany. This would further increase the nationwide representativeness and quality of the data. It is also necessary to create a legal basis in order to permanently secure the access to privately held data and enable their integration into official statistics production in the long term.

Synthetic mobile networks data generation

Keywords: mobile network events data, microsimulation, C++

Introduction

Official statistics is under a continuous process of modernization of the production processes and the incorporation of new digital data sources. One of the most promising data sources for official statistics comes from mobile telecommunication networks, with usage in fields such as the population statistics, transport statistics or tourism statistics. However, this data source is proved to be difficult to access by Official Statistics Bureaus mainly due to reasons related to data confidentiality. Thus, the use of synthetic data allows statisticians to move forward in the development of statistical methodologies before even getting access to real data. In this paper, we present a software tool to generate synthetic mobile network events data. The simulation tool uses an agent-based approach: it places a number of agents on a map, allowing them to move according to some mobility patterns, and records the network events generated by the interaction between the mobile devices and the network. One of the main advantages of using synthetic data is that it produces not only the network events data sets, but also the so-called "ground truth" which is never available in real life. Our simulation tool produces individual level data i.e., data generated due to the interaction between a mobile device and the network (the timestamp of the network events, the cell ID, the timing advance variable, the type of the event, etc.) and make use of telecommunication variables about the configuration of the network such as the position of each Base Transceiver Station (BTS), their emission power, their path loss exponent and other technical characteristics (Salgado et al., 2021). The software is freely available on github (https://github.com/bogdanoancea/simulator) where we provide the source code, a Docker image, an installation kit for Windows and a set of input files.

Methods

We adopted a discrete event system approach to run mobile network data micro-simulations. These micro-simulations provide us with synthetic data, allowing us to immediately proceed with the development of the general framework of producing official statics based on mobile network data and to test these models (Oancea et al., 2022). Checking the real performance of a model in this area of research is almost not possible in real life, because there is no way one can get the real positions of the mobile devices at different time instants, thus simulated data being a very important tool to validate the statistical models.

We started the development process with the identification of the main features of the synthetic data generator tool:

- It should be able to load and use different maps (geographical areas) as a basis for the simulation and define a reference grid overlapped on the map to compute location probabilities and population densities.
- It should support multiple mobile network operators.

- The configuration of the mobile network(s) such as the BTSs locations and their parameters should be configurable.

- The handover process should be flexible and configurable.

- The mobility patterns of the individuals involved in a simulation should be configurable.

- The software implementation should be fast enough to run simulations with large populations.

Secondly, we've made an inventory of the existing software tools, checking if there is one that can be used for our purposes. Thus, we evaluated the *cdr-gen* project (Bordin, 2017), NetSim (Tetcos, 2019), OPNET Network Simulator (Zheng and Hongji, 2012), SUMO (Krajzewicz et al., 2012) and MATSim (Horni et al., 2016) but we found no one to answer all our requests. Therefore, we proceeded to develop our own simulation software using an agent-based model. In terms of algorithmic representation of the agents and their interactions with the network, we found that object-oriented (OO) programming paradigm is the natural choice, and therefore an OO language should be used for this purpose. We considered several OO languages and decided to use C++ mainly because of the performances in terms of the running speed and memory requirements.

Results

Our simulation tool is structured around the following important modules: a *GIS module*, an *agent-based simulation* module and *computational module* to simulate the interaction between mobile devices and the network. The architecture of the simulation tool follows a layered architecture, each layer providing specific functionalities to the other layers above it. Figure 1 shows the five layers of the simulation tool: the basic (runtime) libraries and some utilities (an XML and CSV parser, a random number generator etc.), a GIS layer implemented around the GEOS C++ library, a data encapsulation layer defining the agents (individuals, mobile devices, etc.), a simulation layer responsible with running the actual data generation process and a computation layer that provides a very simple method to compute the location probabilities for each mobile device during the simulation.

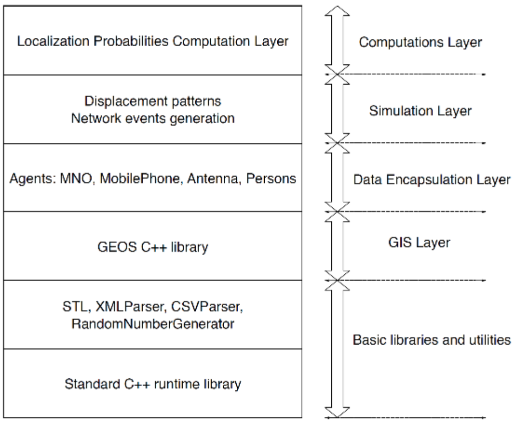

Localization Probabilities Computation Layer
Displacement patterns
Network events generation
Agents: MNO, MobilePhone, Antenna, Persons
GEOS C++ library
STL, XMLParser, CSVParser, RandomNumberGenerator
Standard C++ runtime library
Computations Layer
Simulation Layer
Data Encapsulation Layer
GIS Layer
Basic libraries and utilities

Figure 21. The architecture of the simulation tool

A simulation starts with a synthetically generated population of individuals with the number of the individuals provided by the user together with other personal characteristics (gender, age).

Regarding the mobility of individuals, we implemented two models: individuals moving with a slow speed (it simulates walking) and with a higher speed (it simulates cars/other transportation means) mobility. Each individual involved in a simulation scenario can carry 0, 1, or 2 mobile devices and moves according to a mobility pattern configured by the user. The mobility patterns implemented in the simulation tool are random walk, random walk with drift, Levy flights, Manhattan mobility, home - work (with anchor points), and home – work (with anchor points) using Manhattan mobility. The statisticians provide all the parameters for each mobility pattern through the configuration files. In figure 2 we show three examples of mobility patterns: Levy flights, Manhattan mobility and home-work with anchor points, which means that an individual stays in a location (home), than go to another location (work) where it stays for a period of time, and on the way back home he/she stops in another location (anchor point – a shopping mall for example).

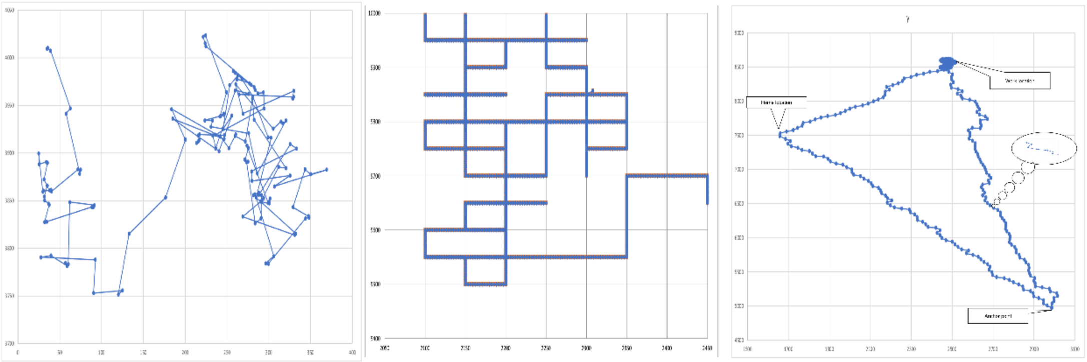


Figure 2. Three examples of mobility patterns: from left to right – Levy flights, Manhattan move, home-work with anchor points.

The interaction between mobile devices and BTSs was modelled using the signal strength and the signal dominance (Tennekes and Gootzen, 2021) and two types of BTSs were considered: omnidirectional and directional. The handover mechanism configured to use either the signal strength or the signal dominance is based on the highest value: a mobile device tries to connect to the BTS that provides the signal with the highest value for the signal strength/dominance.

In figure 3, we show a general data flow diagram of the simulation tool. The simulation software takes its inputs from five configuration files: a *map file*, the *network configuration* file, a general *simulation configuration* file, a configuration file for the *synthetic population* used for simulation, an optional configuration file needed to compute the *location probabilities* of each device. The map file uses the WKT format while the rest of the input files are XML files. As a result of a simulation, our tool outputs a series of information split into several *csv* files. The *network configuration file* contains the cell ID, operator ID, the location and technical parameters and the tile ID for each. The *network events file* contains the details about the network events generated by the interaction between mobile devices and the network: the timestamp of the event, the cell ID, the device ID, the network type (3G, 4G etc.), an event code, the Timing Advance variable, and the exact position (x, y and tile ID) of the device when the event was generated ("the ground truth"). The *persons file* contains the exact position of each person at each time instant, regardless of whether he/she has a mobile device or not. All the input and output files are accompanied by metadata files, which defines the structure and values of the parameters a user can provide in input files and the structure of the output files. They are written in XSD format.

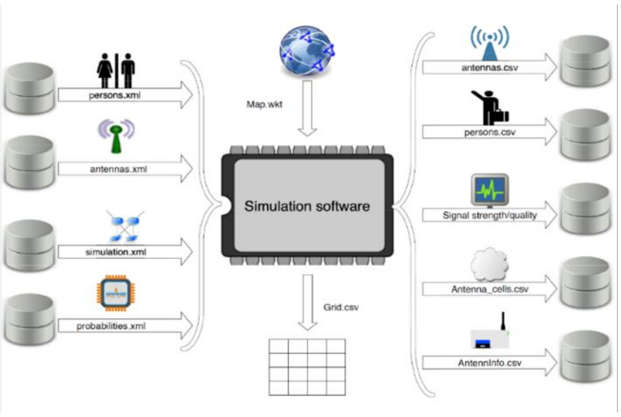


Figure 3. The data flow diagram.

Conclusions

Adoption of the mobile network data in Official Statistics is also bringing the need for new statistical methods to infer about target populations with the traditional high-quality standards. Thus, the development of new statistical production frameworks becomes a challenge, since the use of real data has high costs or is even impossible. To overcome this problem, we developed an agent-based simulation software tool, which can provide a wealth of data to experiment and to test. A strong point of this approach is the provision of the “ground truth” for each simulation scenario, thus allowing the analyst to investigate and test the statistical methods for a wide range of potential situations.

References

[30] D. Salgado, and L. Sanguiao, and B. Oancea, and S. Barragán, and M. Necula. An end-to-end statistical process with mobile network data for official statistics. EPJ Data Science, 2021, 10(20), DOI: 10.1140/epjds/s13688-021-00275-w.

[31] B. Oancea, and D. Salgado, and S. Barragán, and M. Necula. Use of Simulation Models for the Development of a Statistical Production Framework for Mobile Network Data with the simutils Package, 2022, https://arxiv.org/pdf/2201.08171.pdf

[32] Bordin, M. V. (2017). *A Call Detail Record (CDR) generator*. https://github.com/mayconbordin/cdr-gen

[33] L. Zhen, and Y. Hongji. Unlocking the Power of OPNET Modeler. 2012, Cambridge University Press, New York.

[34] Tetcos (2019). *NetSim User Manual*. https://www.tetcos.com/downloads/v12/NetSim_User_ Manual.pdf

[35]D. Krajzewicz, and J. Erdmann, and M. Behrisch, And L. Bieker. Recent Development and Applications of SUMO -Simulation of Urban MObility. Journal On Advances in Systems and Measurements 2012, 5 (3&4), 128–138.

[36] Horni, A., Nagel, K., & Axhausen, K.W. The Multi-Agent Transport Simulation MATSim. 2016, Ubiquity Press, London

[37]M. Tennekes,and Y.A. Gootzen. A Bayesian approach to location estimation of mobile devices from mobile network operator data, 2021, https://arxiv.org/pdf/2110.00439.pdf

Generation of Synthetic Data from actual data of Mobile Network Operators (MNO) through Generative Adversarial Networks (GANs): motivations, techniques and preliminary results

Keywords: synthetic data, generative adversarial networks, gan, telephony data, conditional gan, mno, mnd

INTRODUCTION

In recent years, the generation of synthetic data is becoming one of the most interesting tasks for both public and private research institutions. The advantage of the "Synthetic Data Generation" task is that it produces data that have the same statistical characteristics as actual data. At the same time, this task provides a perfect protection against privacy attacks. In the literature there are many synthetic data generation techniques [1] that differ in terms of trade-off between privacy gain and loss of utility that these synthetic data can achieve. One of the main goals of the synthetic data production is that they can be used to compute statistics using them straightaway. A sample application of the synthetic data generation can be from the Mobile Network Data (MND) which are in general available to develop algorithms for processing commuter travel sections. Some of Mobile Network Data include use cases such as: initiating or receiving voice calls, sending or receiving SMS messages and mobile data usage. These data are converted, by the MNO, into Call Detail Records (CDR), which are then used for billing customers. These data do not contain the actual contents of the communication. Generally, the mobile data are considered private by the provider, this fact has been shown in several works where it is possible to re-identify users also from the aggregated data [2] [3]. For this kind of data, the direct management of the micro-data is entrusted to the telephony providers which are the owners and manage them. In light of the successful outcomes of the experience in the "ESSNET Big Data" projects, the Official Statistics Institutes are oriented towards the creation of a process pipeline that sees the computation on raw data carried out by the providers (Mobile Network Operator). This data processing is generally executed by means of methods and algorithms to yield aggregated data which are critical for the Statistical Institutes that will finally develop the official statistics of interest. The synthetic data generated from the real Mobile Network Data (MND) could support Statistical Institutes in the process of designing and develop some processing algorithms that would guarantee complete transparency of computations that the providers will perform on the raw data [4]. In this, work, to achieve the production of synthetic data from actual telephony data, we have harnessed the power of "GAN" (Generative Adversarial Networks) [5]. Basically, we have focused our attention on the development of a Synthetic Data model specific for tabular data, which is called "CTGAN" (Conditional Tabular GAN) [6]. To do this we have made use of a Framework called "SDGym" (Synthetic Data Gym) [7]. We have evaluated the performance of the model thanks to great support of Utility metrics and Privacy Metrics included in the SDGym.

[1] The deliverables from the ESSnet Big Data II - WPI are publicly available from https://ec.europa.eu/eurostat/cros/content/wpi-milestones-and-deliverables_en and from the ESSnet Big Data I - WP5 are publicly available from https://ec.europa.eu/eurostat/cros/content/wp5-mobilephone-data_en

METHODS

2.1. Generative Modeling

“Generative Modeling” is an unsupervised learning task leading to the automatic discovering of regularities or «patterns» within the actual data in order to generate new synthetic data. In the original formulation, Generative Adversarial Networks represent a way to learn Latent Spaces of Images, namely a simple hidden representation of one data point (observation). Basically, a GAN is made of a "Fake" Network (Generator) and an "Expert" Network (Discriminator) which are adversarially trained trying to overcome one another reciprocally. During the first stage training, the generator takes a random vector as input (a random point in the latent space) and decodes it into a synthetic image. Once it is trained, the generator generates new samples shaping new data over the input data distribution in a totally unsupervised way. In the second stage training, the discriminator (or adversarial) takes a real or synthetic image (they are mixed) as input and classifies it as real or fake. In this project we have adopted SDGym (Synthetic Data Gym) which is a framework to benchmark the performance of synthetic data generators based on SDV (the Synthetic Data Vault project by DataCebo) and SDMetrics (Synthetic Data Metrics). For our study we have chosen a CTGAN as synthesis model, since they are a combination of Conditional and Tabular GANs. The need for tabular GANs is born since the original design of GANs aims to yield synthetic images. On the other side, condition tabular GANs are a class of generative adversarial networks able to generate tabular data with various data types (int, decimals, categories, time, text) and with different shapes of distributions. Furthermore, Conditional GANs have been proposed by modifying the architecture by adding the label y as a parameter to the input of the generator and trying to generate the corresponding data point. It also adds labels to the discriminator input to distinguish real data better.

2.2. The experimentation process

The actual data (which we generate synthetic data from) are data of the Mobile Network Operator and consist of 4 variables (fields like SIM code, antenna/sector code and Time) which are characterized by high volume and low dimensionality. The dataset is made of a random sample of 10,000 Call Detail Records (CDR). CDRs are detailed records of all the telephonic calls that pass through a telephone exchange or any other telecommunication equipment. The dataset is composed of the following four attributes: "SIM code", "Call Date", "Call Time" and "Cell_Call_Code": the *"SIM code"* uniquely identifies the number from which the call has come from, whereas *"call date"* and *"time date"* represent the temporal information of the dataset. Finally, *"cell_call code"* is the Antenna code which the telephone call has come from. We have pseudo-anonymized all the data rows and fixed the the data types of its attributes. We have casted the two symbolic attributes into categorical variables (SIM code, Antenna code) and the other two into continuous variables (date and time of call).

RESULTS

We have synthetized 10,000 new synthetic telephony data, preserving the same data structure as the actual data. This fact is clearly depicted in the Figure 1 by comparisons of the univariate distributions of the categorical variables (SIM code, Antenna code) from the real and the synthetic data. We have selected the first 50 bars in descending order for each variable. From this chart emerges that the univariate distribution of the categorical variables is preserved in the synthetic dataset. In order to estimate the performance of the synthesis model (CTGANs) we make use of two different types of more specific measures for the synthetic data generation task with regard to the distributions plot over the data: a) Utility Metrics; b) Privacy Metrics. The simplest measure available within the SDGym belongs to the family of Statistical Metrics comparing the tables by running different types of statistical tests on them. One of these measures is the “Chi-Squared Test”, since we have achieved **1.0** in this test, it means that our distributions (original and synthetic) are sampled from the same distribution of data, in a significant way. Another statistical metric is a twosample “Kolmogorov–Smirnov Test” to compare the distributions of continuous columns. Since our result has been **0.95375**, it means that the distance between CDF (Cumulative Distibution Function) of original data and CDF of Synthetic Data is low, hence the two distributions are very near in terms of this measure. Instead within the Likelihood Metrics family, the “BNLikelihood” fits a “Bayesian Network” to the real data and then evaluates the average likelihood of the rows from the synthetic data on it. Therefore, the “Bayesian Networks Likelihood” is the Error calculated on Synthetic data after fitting the model on Real Data. We have obtained a very low error (likelihood) like **0.00013183** which means the two datasets are very close in terms of probabilistic models. Finally, “Privacy Metrics” are a kind of metric aimed to estimate the privacy preserving extent of the synthetic dataset. We have performed 2 Privacy Metrics Tests: In the “Privacy Metrics Test 1” we have taken the variable ANTENNA_CODE as first field for the merge and SIM_CODE as second feature for the final match. Viceversa, in the “Privacy Metrics Test 2” we have chosen the SIM_CODE as first field and ANTENNA_CODE as second field. The First Chart (Test 1) means that within the 81 rows obtained with the merge on ANTENNA_CODE there are 80 with 0% of matches on SIM_CODE and 1 with 100% of matches. So there is a small failure in the privacy preserving process. In the second chart there are no matches and failures. These matches will be depicted on a bar plot via histogram, such as the one in Figure 2. The idea behind this Privacy Metric is looking for “values” within the Synthetic Dataset which are also

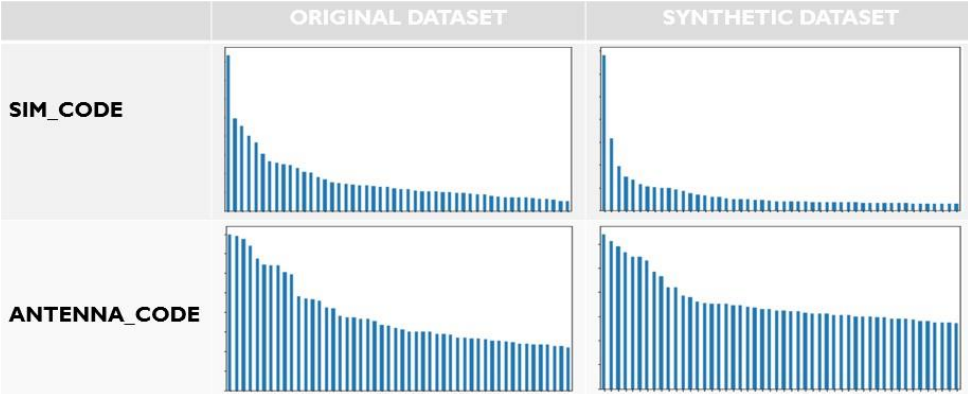


Figure 1. Original and Synthetic Data Univariate Distributions Visualization

inside of the Original Dataset. If there are many of these values, this means that the CTGAN model is capable to generate a Synthetic Dataset similar to the Original one, but not preserving all the values from the Original Dataset, and so not preserving its privacy at all. Eventually, to have a final aggregate measure of privacy we have designed the Aggregate Privacy Metric (APM) which is in the interval [0, 1]. Basically, we have calculated a normalized sum of all matches. More matches mean less privacy, if 1 - (normalized sum) is equal to 1, this represents high privacy, 0 means low privacy. In our first test we have achieved a value close to **0.9876**, namely very high privacy. In the second test we have achieved **1.0** as aggregated value which means maximum privacy. The formula we have adopted is reported as follows:

$$\mathbf{APM} = \mathbf{1} - \frac{\sum_{k=0}^{n} \frac{S_k}{100}}{N}$$

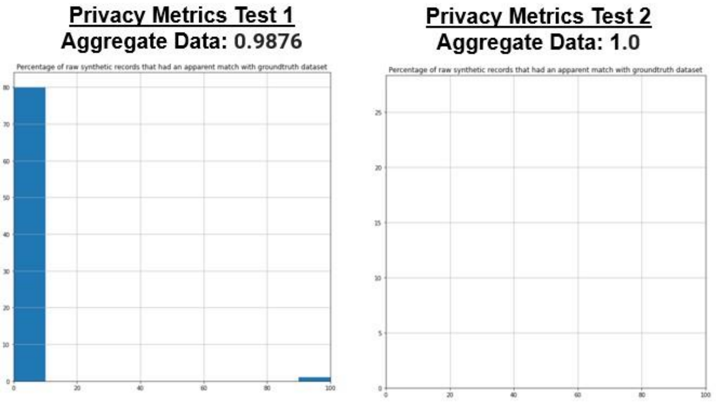


Figure 2. Original and Synthetic Data Univariate Distributions Visualization

cONCLUSIONS

Our experimental setup suggests that Conditional Tabular GAN (CTGAN) are a valid tool for the generation of Synthetic Data which respects the original spatial distribution with a a very high privacy preserving extent. However, reproducing the temporal structure of all the calls couplings can require other models such as Time Gans or Graph Gans. In the future direction we will take more in consideration the temporal dependencies amongst the data. This may foster capturing also the multivariate joint distributions from the original data towards the synthetic data.

rEFERENCES

[1] Soltana, G., Sabetzadeh, M., & Briand, L. C. (2017, October). Synthetic data generation for statistical testing. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 872-882). IEEE.

[2] Xu, Z., Li,Y., Zhang,P., Fu,X., & Jin, D. (2017). Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated Mobility Data, Proceedings of the 26th International Conference on World Wide Web, Pages 1241–1250

[3] F. Ricciato, A. Bujnowska, A. Wirthmann, M. Hahn, and E. Barredo-Capelot. A reflection on privacy and data confidentiality in official statistics. In ISI World Statistics Congress, 2019. https://www.bis.org/ifc/events/isi_wsc_62/ips177_paper3.pdf.

[4] F. Ricciato, J. Grazzini and J.M. Museux, (2021). “Public manuals and open-source code: rethinking methodological documentation for new data sources”, NTTS 2021

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[6] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems, 32.

[7] https://github.com/sdv-dev/SDGym

Transport Statistics (JENK2A.3)

Session Chair: **Evangelia FORD-ALEXANDRAKI** *(Eurostat)*

Port Visits Using New Data Sources
Nele van der Wielen, Justin McGurk, Time Linehan, John Flanagan (CSO)

TranStat - application of maritime data in the traffic intensity and transportation volume
Michał Bis *(Statistics Poland)*

Using Open-source Data to Estimate Mobility Statistics
Laurent Smeets *(GOPA Luxembourg)*

Port Visits Using New Data Sources

Keywords: Big data, New Data Sources, AIS, Geospatial Statistics, Shipping, Maritime Statistics

Introduction

The Automatic Identification System (AIS) is the international system for tracking ship movements. Ships of a certain size and type are required to carry a transponder which transmits their location and other key information regularly, and this is used for real-time ship monitoring. Stored AIS data is an example of Big Data due to the extremely large amount of signal data generated every minute by ships around the world. This data offers an alternative method for compiling statistics on maritime traffic. AIS-based data could significantly reduce the time lag between the reference period and the publication of official statistics [1]. This paper aims to showcase how AIS data can be used to generate reliable and valid port call statistics by comparing AIS data with official data sources in Ireland. Ireland, as an island relying heavily on sea transportation for tourism and transport, is ideally placed to capitalise on the potential offered by faster AIS-based maritime indicators.

Methods

The use of AIS gives statistical offices the opportunity to collect homogenous near real-time and historical data that can be used to produce maritime transport statistical estimates more frequently than traditional methods. For this paper two methods were developed: The Boundary Crossing Method (BCM) and the Stationary Marine Broadcast Method (SMBM).

The AIS Source Dataset

AIS data were supplied by the Task Team on AIS Data of the UN Committee of Experts on Big data and Data Science for Official Statistics (Task Team on AIS Data — UN-CEBD), and accessed through the UN Global Platform — UN-CEBD (UNGP) which holds a global repository of live and archived AIS data. This data is provided by ExactEarth who combine their own satellite data with terrestrial data from FleetMon. In addition to location, bearing and navigation status, the AIS data includes unique identifier information on International Maritime Organisation (IMO) number and Maritime Mobile Service Identity (MMSI) number.

The UNGP not only holds AIS data but also the IHS Shipping Registry. Incorporating SeaWeb and Lloyd’s Register of Ships (published since 1764), the IHS Shipping Registry provides detailed information on all self-propelled and seagoing merchant ships. Among the information included in the registry are IMO number, MMSI number, ship name, ship type, cargo type, ownership, registration, tonnage, dimensions and propulsion. The United Nations Global Platform produced a handbook which provides a snapshot of global AIS analysis [2].

Methodological Notes

This paper compares two methodologies developed for generating AIS-based port calls. The first methods, Boundary Crossing Method, is based on ships entering and leaving the port area, which is defined by a polygon (port polygon). It is a quick and easy to understand method that counts all ships that fall in any of the defined port polygons as port calls. While this method allows macro level port call analysis, a more detailed analysis is needed when looking at more specific shipping activities within a port polygon.

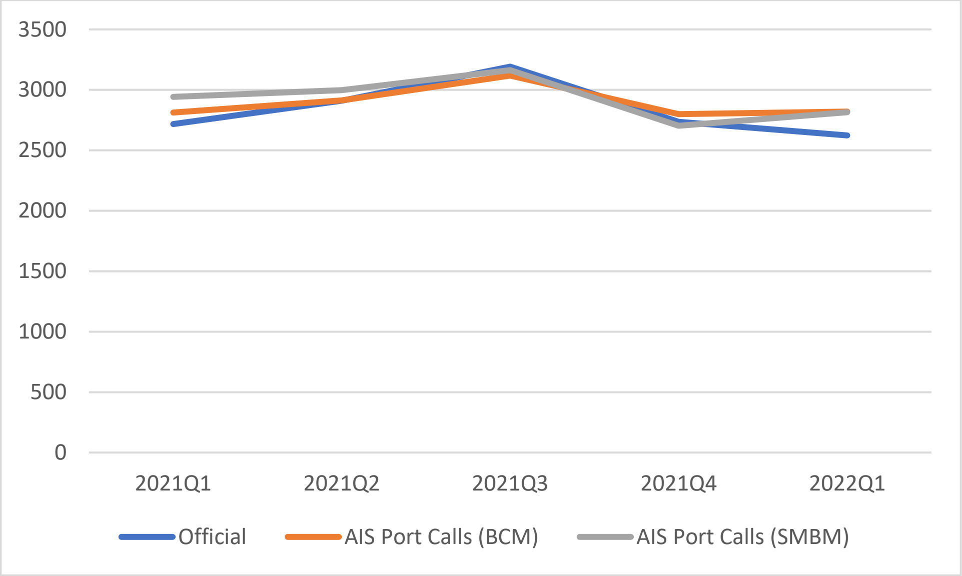
The second method, Stationary Marine Broadcast Method, identifies stopped ships within a port area using the same polygons as the first method. However, it also uses AIS data to estimate a location where a ship is stopped, the upper-lower-time durations for ship stoppage and the number of AIS messages counted while stopped.

Both methodologies aimed to develop robust code that can be implemented by agencies in other countries, in addition to Ireland. As part of validation, both methodologies developed were benchmarked with MarineTraffic.com data

Results

Overall, on a national level, the AIS-based port calls follow closely official number of port visits. Figure 1.1 compares the number of vessels that arrived at six Irish ports using AIS-based port calls and official numbers from the first quarter of 2021 to the first quarter of 2022.

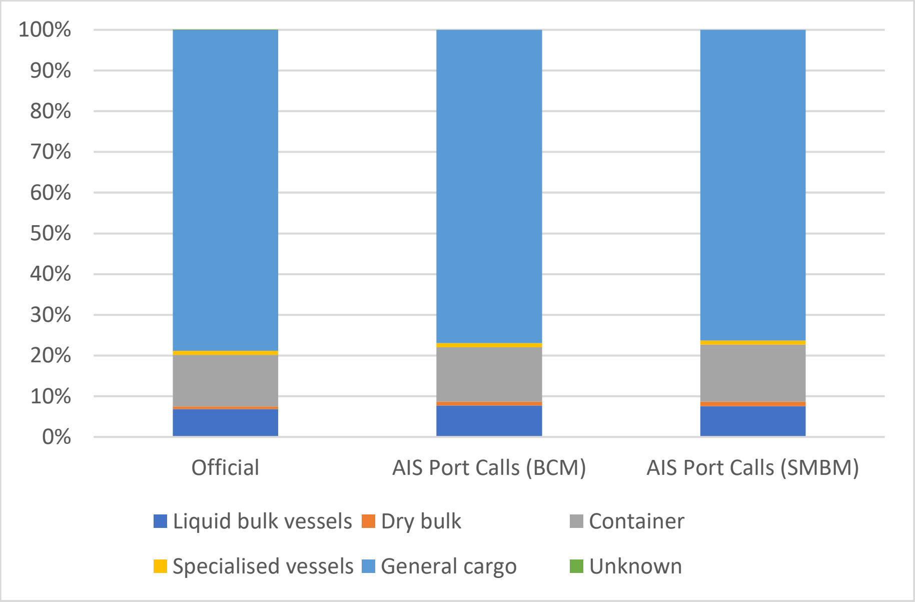
Figure 1.1: Comparison of Official versus AIS-based port calls, Q1 2021 to Q1 2022



Based on the administrative data for 2021, a total of 11,552 vessels arrived in six Irish ports. This number increased to 11,640 when using AIS data with the Boundary Crossing Method (BCM) and to 11,804 when using the Stationary Marine Broadcast Method (SMBM). While the AIS methodology produces a slightly higher number of port calls it is clear from Figure 1.1 that the AIS data mirrors the trends in port calls across 2021 very closely. Similarly, when the AIS data is analysed across the individual ports it produces the same trend as the administrative data across the five quarters.

The AIS data provides a very similar breakdown of ship classification when compared to administrative data. In 2021 cargo ships accounted for 79% of the total port calls in Ireland according to the administrative data. Similarly, both AIS-based port call methods show that cargo ships accounted for nearly 77% of visits. The smallest share of arrivals were dry bulk and specialised ships and Figure 1.2 highlights that AIS data also capture this breakdown correctly.

Figure 1.2: Port calls by type of ship, 2021



Conclusions

Overall, the findings of this project confirm that AIS data could be used to produce valid estimates of port calls and illustrates the potential of Big Data for official statistics. This paper confirms the potential of AIS data as a supplementary source to existing administrative data to enable timelier official maritime statistics going forward. The use of AIS data would allow official statistical offices to publish preliminary estimates of port call statistics and trends within days of the end of the reference period, with these estimates being revised or adjusted subsequently with administrative data.

References

[38] Emmens, T., C., Abdi, A., Ghosh, M. (2021). The promises and perils of Automatic Identification System data. ScienceDirect, 178, https://doi.org/10.1016/j.eswa.2021.114975.

[39] United Nations Global Platform (2020). AIS Handbook Online. Geneve: United Nations.

TranStat - application of maritime data in the traffic intensity and transportation volume

Keywords: Big Data, AIS, maritime data, statistics, traffic intensity, transportation volume, Apache Spark, Scala, Hadoop.

Introduction

In the light of the increasing level of transshipments in sea ports, Statistics Poland and Maritime University of Szczecin developed models for statistics of traffic intensity, transportation volume, emissions in maritime transport using Big Data methods and tools as well as non-statistical sources, e.g. AIS (Automatic Identification System).

The implementation of the developed solution in maritime transport by Statistics Poland concerns four ports of fundamental importance for the national economy in Poland: Gdańsk, Gdynia, Szczecin and Świnoujście.

Methods

Big Data sources, methods and tools

The main source of maritime data for the TranStat system is AIS. It is an automatic tracking system used on ship for exchange data with other ship, AIS base station and satellites. An additional source is the maritime transport data set based on Directive 2009/42/EC of the European Parliament and of the Council of 6 May 2009 on statistical returns in respect of carriage of goods and passenger by sea.

All the statistics (traffic intensity and transportation volume) are presented within the TranStat system consisting of the following functional subsystems:

- Data collection and processing subsystem responsible for the following processes:

- processing of streaming data from sensors,
- AIS data decoding,
- data integration, validation and transformation,
- data aggregation.

- Data presentation and analysis subsystem, intended for an external user, operating on the basis of calculated aggregates and indicators.
Application link - https://transtat.stat.gov.pl

For our work, we have chosen environment based on Hadoop Stack, in order to ensure high scalability and performance. The raw AIS data is collected and stored in real time in the Hadoop File System (HDFS). For the processing data, we developed code in Scala and processed by Apache Spark which is computing platform.

Traffic intensity statistics in maritime transport

Gdansk, Gdynia, Szczecin and Swinoujscie, points (containing geographic coordinates: longitude and latitude) were determined, which form polygons that fall within the boundaries of the ports on the basis of the regulation of the minister competent for maritime economy. These constituted areas for the study of ship traffic volume.

These were areas of study of the traffic intensity. Traffic intensity is understood as the intensity of the stream, defined as the number of transport units passing through the line delimiting a given area in a certain period of time.

In order to develop a methodology for calculating the traffic intensity in a specific area and in a specific time unit, depending on the method of calculating the intensity and the location of the calculation procedure on the timeline, the method of counting units based on reporting times was used.

2.3. Transportation volume in maritime transport

We understand the transportation volume as "The product of the distance performed by the means of transport: the length of the road (number of kilometers) and the number of tons of goods transported (cargo). The measurement unit is tonne-kilometer (tkm) - one tonne-kilometer transports 1 ton of cargo over a distance of 1 km.

In the model for estimating the transportation volume , it was planned to present the possible routes of the vessel in the form of a directed (weighted) graph, where the vertices of the graph are waypoints or quays, and the edges are straight sections between them. Each edge contains the coordinates of the start and end points, and its weight is the length of the segment between individual nodes, calculated by the Haversine formula (distance).

Results

The TranStat application allows to generate fifteen indicators (for breakdowns and data periods) for the traffic intensity and eight indicators for transportation volume including charts, data sets and metadata. All indicator charts you can download as open formats e.g. pdf, jpg, pdf, svg.

Traffic intensity statistics in maritime transport

As a result of the developed algorithms for the traffic intensity, the following variables and breakdowns are obtained, among others:

a) Variables:
- Number of ships at a seaport
- Number of arrivals / departures by maritime ships;

b) Breakdowns:
- Time: day, month, quarter, year
- Spatial - ports located on the coast of Poland
- Means of maritime transportation: by type, by country of flag

The port of Szczecin was selected to illustrate the possibilities of the TranStat application in the scope of generating traffic intensity statistics in maritime transport. Visualization was made for one month (May 2022), broken down into days.

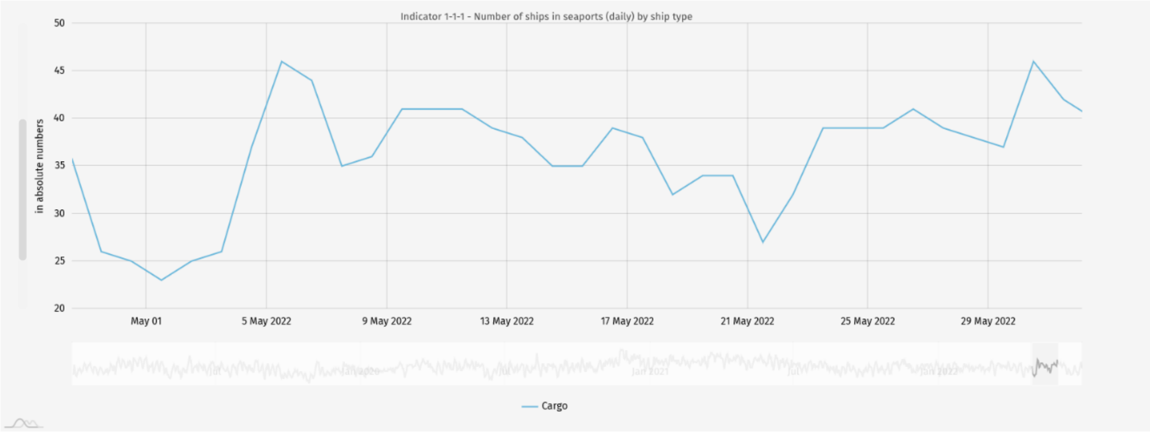


Figure 22. The number of ships in the port of Szczecin in May 2022 by ship type (cargo)

The first chart shows the number of cargo in the port of Szczecin in May 2022. It should be borne in mind here that these are all ships that were in the port that day, they could have entered the day before, so this number is not the same as the number of calls.

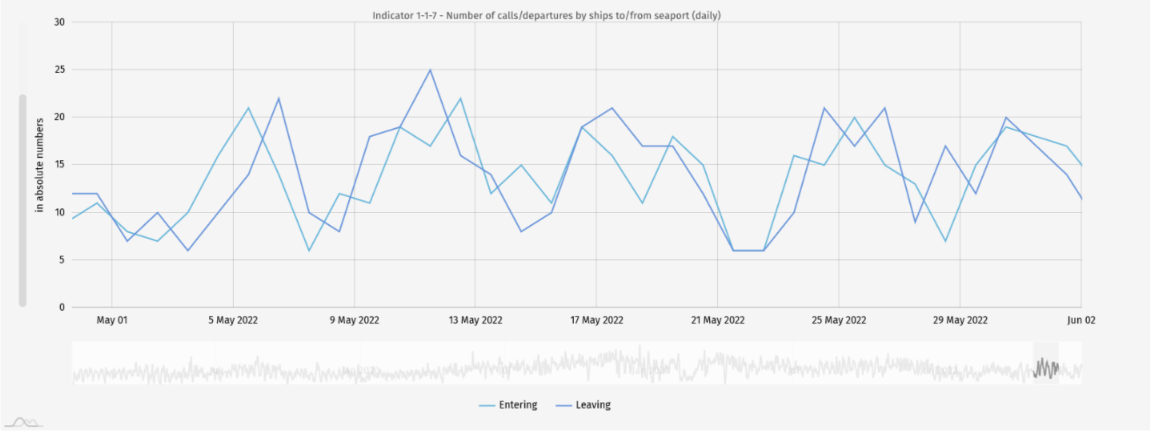


Figure 2. The number of calls/departures by ships to/from the port of Szczecin (daily)

The second chart shows the number of calls/departures for the port of Szczecin in May 2022 by direction. Fluctuations can be noticed - the intensity was not uniform in the analyzed period. The highest number of ship departures was recorded on May 11, 2022, while the largest port calls were recorded on May 12, 2022.

Transportation volume in maritime transport

As a result of the developed algorithms and combined data sources, the following variables and breakdowns were obtained:

a) Variables:

- transportation volume for goods ,
- transportation volume for passengers,
- summary distance - distance travelled by all vessels on arrival/departure routes when carrying goods or passengers;

b) Breakdowns:

- Time: day, month, quarter, year
- Spatial - ports located on the coast of Poland, direction, country
- Means of maritime transportation: by type, by flag country, by gross tonnage,
- Type of cargo - cargo group, commodity group

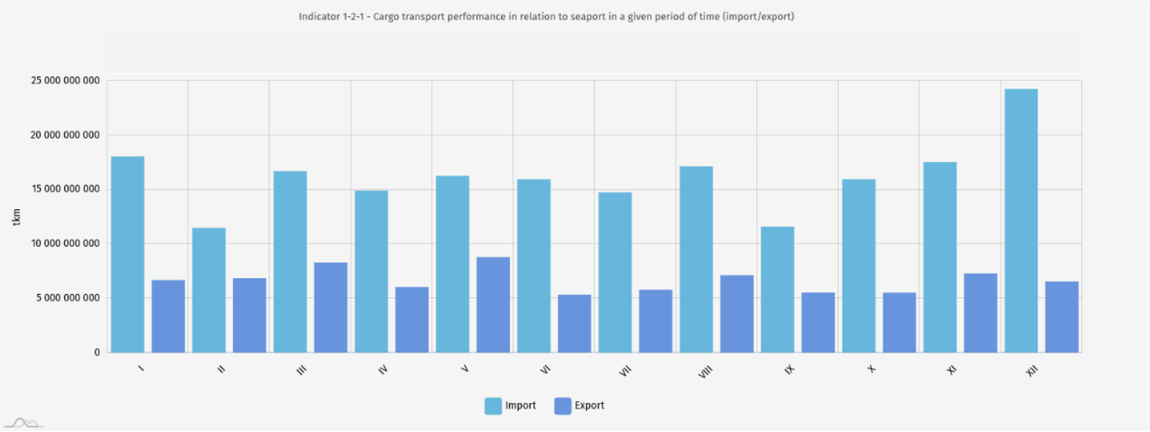


Figure 3. Cargo transportation volume in relation to the port of Gdańsk in 2021 by import/export.

Based on the data for 2021, fluctuations in the volume of transportation volume, which consists of both the number of transported loads and the distances travelled, can be seen. The highest value for transportation volume in relation to the port of Gdańsk was achieved in December 2021 (24.2 billion tonne-kilometers for import and 6.5 for export)..

Conclusions

The implementation of the TranStat project in the field of maritime statistics has enriched the current statistical production carried out by Statistics Poland through:

- access to streaming Big Data source related to maritime transport (AIS);
- implementation of the necessary Big Data technology for sensory data enabling an automatic process of data flow, validation and processing;
- development of traffic intensity and transportation volume models in maritime transport with the use of sensory data;
- development of algorithms enabling generation of new statistics and obtaining new knowledge in the field of maritime transport statistics by using the correlation of multiple data sources;
- reduction of research costs thanks to the use of modern technology in the acquisition and processing of non-statistical sources (AIS);

References

[40] Statistics Poland (2020). Report from the research phase on the implementation of the project under the Program "Social and economic development of Poland in the conditions of globalizing markets" GOSPOSTRATEG.

[41] Statistics Poland (2021). Final report of the project implementation under the Program "Social and economic development of Poland in the conditions of globalizing markets" GOSPOSTRATEG.

Using Open-source Data to Estimate Mobility Statistics

Keywords: open source data, urban passenger mobility, GTFS, access to public transport, big data, data science, R

Introduction

This interactive presentation will show the potential for the use of GTFS, OpenStreetMap and other open-source and freely available data to estimate access to public transport mobility in a timely and easy to scale way. It emphasizes the advantages of GTFS over other data sources and shows that, in addition to simple proximity to public transport stops, GTFS data in combination with population density data can be used to estimate more complex and granular public transport mobility statistics [1, 2].

An experimental method, solely relying on openly available data, will be presented to estimated Sustainable Development Indicator (SDG) Target 11.2.1. This is a United Nation SDG indicator under SDG goal 11 (Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable) and reads: *"Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities"*

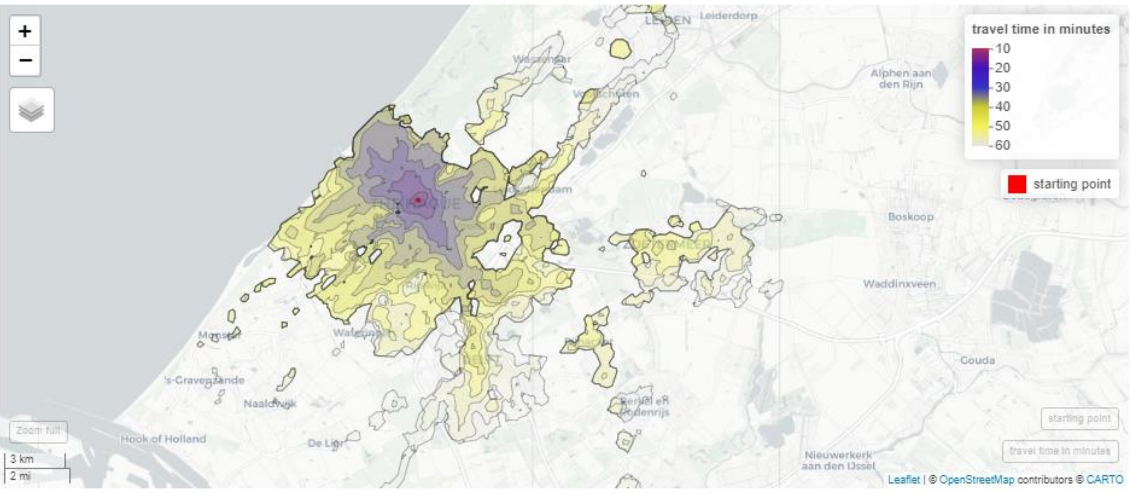


Figure 23: The map above shows an example isochrone map with the different isochrone shapefiles that can be reached within a certain travel time using public transport and walking, starting from the national parliament in the Netherlands.

Methods

The proposed presentation will be an interactive presentation, showing the data different data sources being used and the code developed to estimate these statistics. Data to estimate these statistics comes from General Transit Feed Specification (GTFS) feeds of different countries, OpenStreetMaps, and population density data from Worldpop.

To estimate the statistics a combination of newly developed R code, a locally run OpenTripPlanner instance and a dockerized of OpenRouteService were used.

While there are many and clear limitations to the methods explained here, the advantages of these methods and the reasons for their selection are as follows:

- **Scalable**: this means that once access and the data pipeline are set up the product can be scaled across different urban centres in the EU. This means that one-of data sources that might only be available in a few urban centres are not integrated

- **Passive**: all proposed data sources are so-called passive data sources. They are collected as meta-data from other data sources as passive crowd-sourced data. This means that no interaction with the respondents providing the data is required.

- **Close to real-time:** the proposed data sources allow the production of mobility statistics in near real-time.

- **Easy update:** the proposed data sources allow for timely updating of mobility statistics. Without much effort, mobility statistics can be updated on a daily, weekly, or monthly basis.

- **Geospatial:** all data sources are geospatial in nature.

- **(semi)-structured data:** all proposed data sources are semi-structured data sources that use APIs to retrieve the data. This facilitates the easy and timely generation of statistics. This means that data from unstructured data sources, such as traffic cameras, which require more custom pre-processing to produce mobility statistics, are not considered. While there is real potential for e.g. using traffic cameras o estimate cyclist or pedestrian traffic, the data requires extensive processing and is unlikely to scale easily between cities as there are different types of cameras and recording setups in different urban areas.

- **Open source and free to access**: Somewhat self-explanatory. The fact that all data sources specified here are freely accessible facilitates their scaling and implementation.

Results

The figures below show examples of the statistics that can be estimated using the proposed method for five European capitals. The use case proposes a new method of estimating these statistics, which is easily scalable to any city or Functional Urban Area within the European Union.

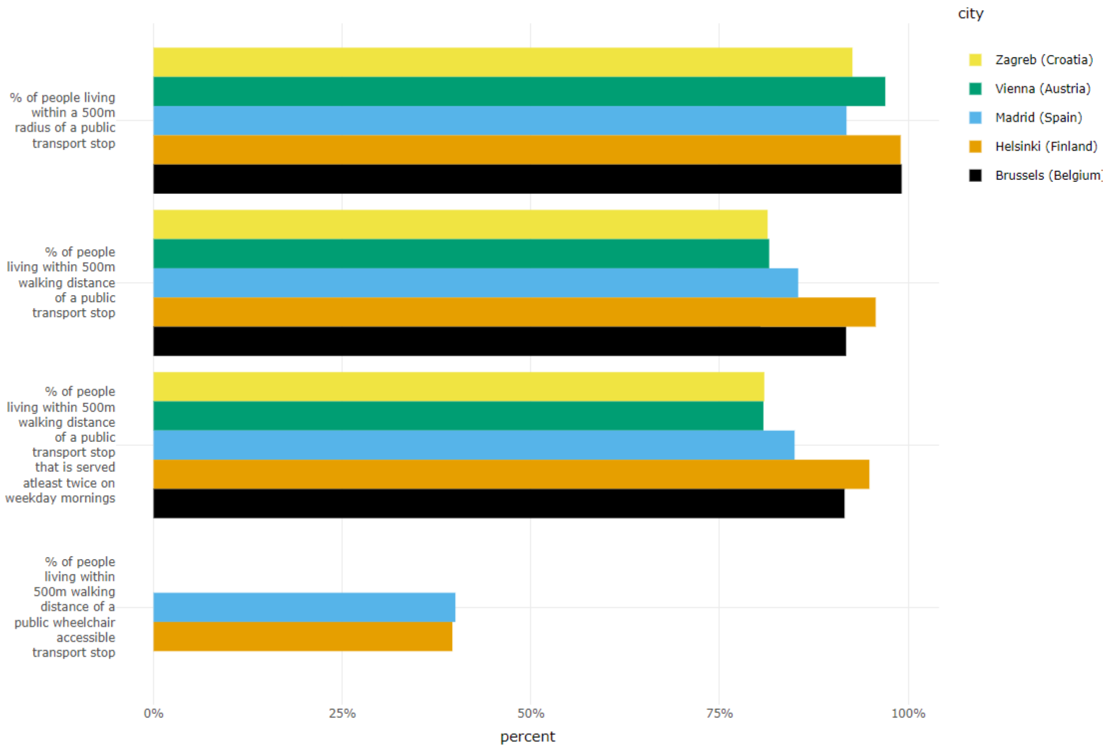


Figure 24: Four of the different accessibility statistics that can be estimated using the presented method

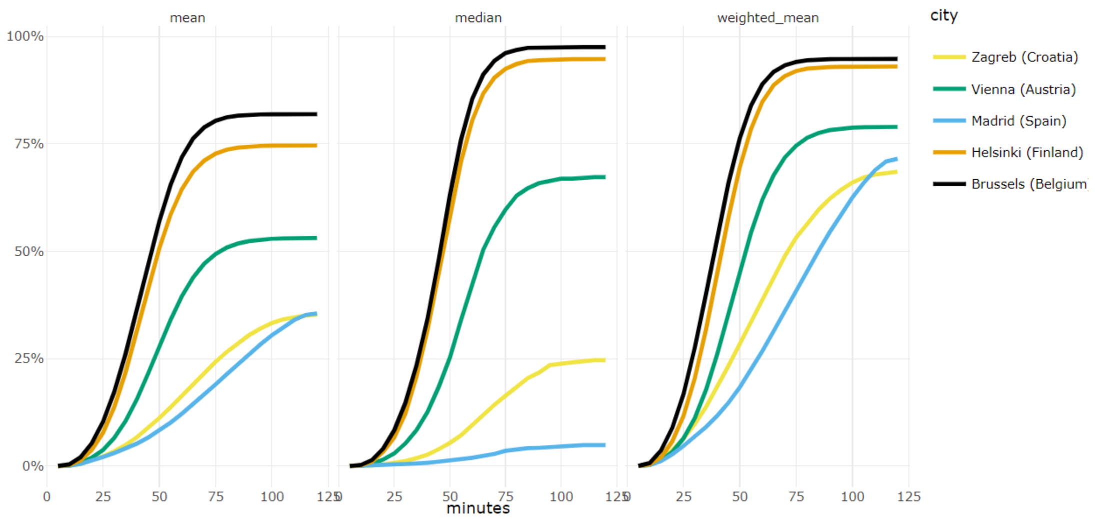


Figure 25: Percentage of population that can be reached using public transport at different time intervals in five European capital cities.

Figure 3 shows both the mean, median and weighted mean (by population of starting point). For instance, in Brussels 65.9% of the population can be reached in 45 minutes on average from any starting point in the city using public transport, while this is only 14.8% in Madrid.

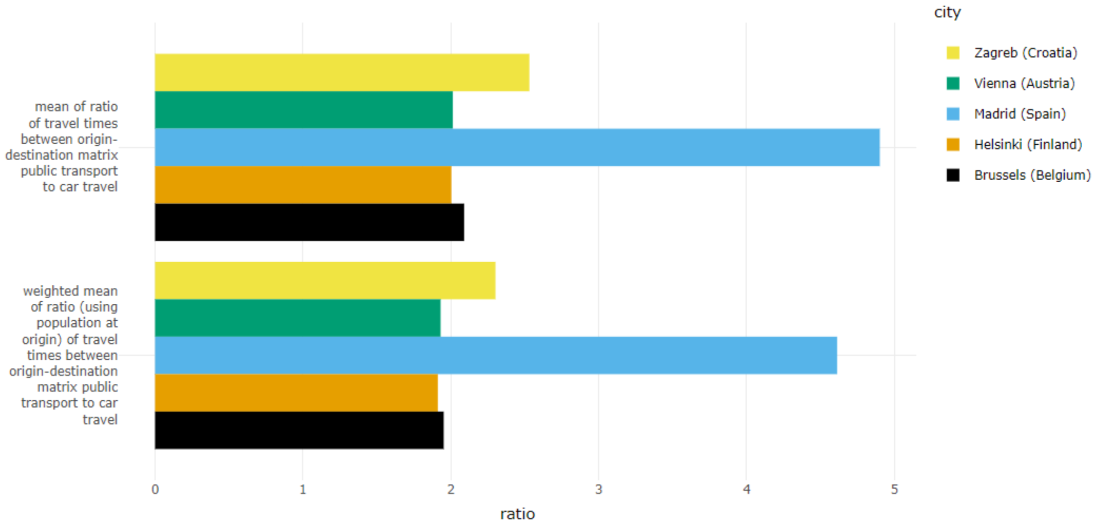


Figure 26: Ratio of travel time public transport to car use for five European capitals.

A value equals to 1 would indicate a perfect parity of travel time by public transport and private car use and a value of larger than 1 means that expected public transport time is longer than car use.

Conclusions

This presentation will show a new way to estimating urban mobility statistics, which does not rely on traditional survey methods or diary-based respondent tracking. It will highlight the advantages over traditional methods, while also showing some of the potential drawbacks. Finally, it will highlight some of the recent developments in the GTFS RealTime data feed standard, which has the potential to estimate more public transport use statistics.

References

[1] Palonen, Tuomas, and Riku Viri. "Benchmarking public transport level-of-service using open data." Transportation Research Procedia 42 (2019): 100-108.

[2] Antunes, H, Figueiras, P, Costa, R, Teixeira, J, & Jardim-Gonçalves, R. "Discovery of Public Transportation Patterns Through the Use of Big Data Technologies for Urban Mobility." Proceedings of the ASME 2019 International Mechanical Engineering Congress and Exposition. Volume 2B: Advanced Manufacturing. Salt Lake City, Utah, USA. November 11–14, 2019. V02BT02A023. ASME. https://doi.org/10.1115/IMECE2019-11415

Blockchain and Artificial Intelligence (MANS2A.3)

Session Chair: **Joao Rodrigues Frade** (*European Commission*)

Exploring the potential of blockchain to support innovation in European statistics - An interactive workshop by the European Blockchain Services Infrastructure
Maxine Lemm *(European Commission)*

webAI - A new type of web-based indicators for economic analyses: The example of sustainability
Sebastian Schmidt *(ISTARI.AI)*, Jan Kinne *(ISTARI.AI)*

The usage of blockchain technology for official statistics
Cristiano Tessitore *(Eurostat)*

Exploring the potential of blockchain to support innovation in European statistics. An interactive workshop by the European Blockchain Services Infrastructure

Keywords: blockchain, ebsi, verifiable credentials, traceability, ledger, dlt, timestamping, workshop, innovation

Introduction

Blockchain is an innovative technology that allows the creation of a tamper-evident, immutable distributed ledger of data that can be leveraged for data analytics.

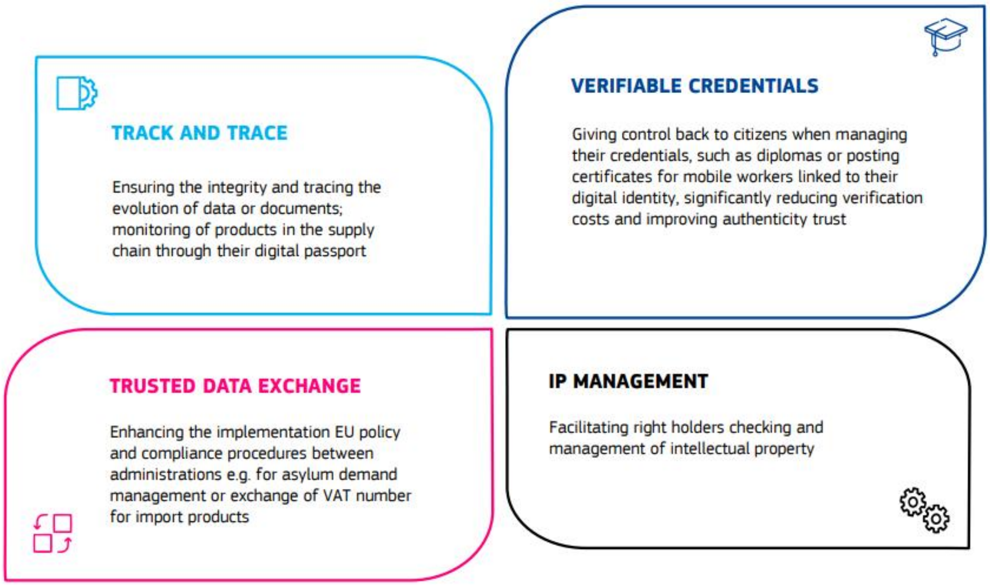
About EBSI

The European Blockchain Services Infrastructure (EBSI) is a joint initiative between the European Commission and the European Blockchain Partnership (EBP) to deliver EU-wide cross-border public services using blockchain technology. Launched in 2019, EBSI aims to build a network of distributed nodes across Europe (the blockchain) while exploring an increasing number of applications focused on specific use cases.

EBSI is made up of two main components:

- A blockchain infrastructure operated by Node Operators from EBP Member Countries, hosting a tamper-evident and immutable ledger, enabling decentralised and trusted verification of information, as well deploying a series of APIs to allow applications to record information on the ledger
- Use Cases, which are functional groupings of capabilities in a sector or domain, defining the business applications of EBSI

There are four main Use Case families being piloted and deployed on EBSI, as outlined in the image below:



TRACK AND TRACE

Ensuring the integrity and tracing the evolution of data or documents; monitoring of products in the supply chain through their digital passport

VERIFIABLE CREDENTIALS

Giving control back to citizens when managing their credentials, such as diplomas or posting certificates for mobile workers linked to their digital identity, significantly reducing verification costs and improving authenticity trust

TRUSTED DATA EXCHANGE

Enhancing the implementation EU policy and compliance procedures between administrations e.g. for asylum demand management or exchange of VAT number for import products

IP MANAGEMENT

Facilitating right holders checking and management of intellectual property

The most advanced Use Case of EBSI is a family of Use Cases known as the Verifiable Credentials framework. This self-sovereign pattern of information sharing allows Data Subjects (Holders) to present claims about themselves, without issuers and verifiers of claims coming in direct communication with each other. The ledger operates as a source of authenticity of Issuers and of these claims.

In 2022, 18 universities in 15 European countries successfully piloted the cross-border exchange of Verifiable credentials in the educational domain using EBSI and EBSI Conformant Wallets.

The potential of EBSI for Statistics and Data Analytics

As part of its Use Case Development, EBSI has developed some Core Technical Capabilities which have enormous potential to enhance the EU’s ability to provide detailed statistics on the performance of the Digital Single Market. As the ledger only contains information about legal entities and no personal data, and is public permissioned, ledger data can be used to extract relevant information about several domains for those use cases already being piloted.

In addition, for use cases that are currently in development, the ability to extract meaningful data to describe the status of the European Single Market (or any other relevant information) can be part of the use case’s business requirements from inception.

One well-known application of blockchain technology concerns traceability. This is why EBSI developed a Timestamp API that is ready to use for European pilot projects to improve the reliability and transparency of EU statistics (in fields such as value chain & product traceability, energy traceability, etc.).

EBSI’s Trusted Issuers Registry also allows to create a tamper-proof trust-chain using the W3C Verifiable Credentials format to make public information about legal entities that issue credentials and claims. In EBSI’s current use cases, these claims concern Natural persons (e.g. a diploma or professional qualification). However, this framework can be applied to any type of claim, with or without the Verifiable Credentials format, to identify authentic issuers of data in the field of European statistics.

Methods

Presentation of EBSI's Core Technical Services, including Timestamp API and Trusted Issuers Registry + Interactive Workshop with audience to understand how EBSI can be used in the field of Statistics.

Results

18 universities in 15 European countries have successfully piloted the cross-border exchange of Verifiable Credentials using EBSI.

Conclusions

EBSI has tremendous potential to increase transparency and reliability of European statistics. Let's find out how together.

References

n/a (not an academic paper)

webAI - A new type of web-based indicators for economic analyses: The example of sustainability

Keywords: artificial intelligence, web analysis, natural language processing, sustainability, greenwashing

INTRODUCTION

Economic research often deals with highly current topics that can show rapid development and high fluctuation. However, the data on which classical econometric analyses are usually based often cannot keep up with the high pace of economic change. Traditional company databases, for instance, are only updated at longer intervals or do not contain information for emerging topics at all. Qualitative expert interviews are more flexible in terms of time and scope, but do not allow for large samples for reasons of effort and cost. A new data source that can guarantee up-to-dateness and adequate scope is therefore of utmost importance for state-of-the-art economic research.

An innovative and as yet not often used source of data for such research are corporate websites. A large proportion of economically active companies, at least in Europe, nowadays have their own website on which they provide information about their products and services [strategies, agendas, personnel and so on [1]. Textual content can therefore be used to derive a great deal of highly topical information about the respective companies. Through hyperlinks to other companies, websites also reveal offline network structures [2].

The webAI methodology presented here was developed to gather fine-grained, highquality, near-real-time information about company activities from website texts. A particular focus lies on the investigation of innovative technologies (e.g. artificial intelligence) and current topics such as ecological sustainability. Based on this methodology, we conducted a study that shed light on the issue of greenwashing in the US metals industry. For this, we combined the analysis of websites with remote sensing data.

METHODS

The webAI developed by ISTARI.AI is based on state-of-the-art machine learning and natural language processing methods. For this, we scrape the website texts of all companies in Europe at regular intervals. For this purpose, a clear heuristic is applied that focuses on the subpages with the shortest URL, as the most up-to-date information is usually found on these top-level pages. Based on topic-specific keywords, text paragraphs are then identified in the respective HTML bodies, which are classified by machine learning in a next step. Therefore, pre-trained transformer models of the BERT family are used, which are specified by manually labelled training data. The application of this methodology can be multilingual. By intersecting the resulting results with classic data from the company database, statements can then be made about the structural and geographical characteristics of the identified companies (e.g. company age, industry, geographical location). ISTARI.AI has already developed several functional agents based on this general methodology, one of which is to be examined in more detail at the NTTS conference.

Ecological sustainability is a topic that is becoming increasingly important for a company's external presentation in times of emerging climate activism such as "Friday's For Future". In order to be able to examine the extent to which companies address the topic of sustainability on their websites, we developed a specific agent of webAI. It is based on keywords that are directly related to the environment. The first study that used this data dealt with the topic of greenwashing, i.e. the immense difference between a company's external image and its actual commitment to environmental protection.

For this purpose, all companies in the US metal industry were divided into sustainable and non-sustainable businesses based on their website texts, using our sustainability agent. In a further step, these analysis results were blended with measurements from the TROPOMI sensor. This is a measuring device located on the European Space Agency's Sentinel 5-P satellite that can approximate the concentration of various air pollutants in the troposphere. In the context of the study, we focused on sulphur dioxide (SO_2), which is one of the most harmful pollutants emitted by the metal industry.

Methodologically, these very different data sources were combined within the framework of a spatial regression analysis, which also took other control variables into account. These included other sources of pollutants such as vehicles or power plants. Our dependent variable was the concentration of SO_2 based on the satellite measurements, which we received on a 7km grid. Our independent variable of interest was the logarithmised count of employees in sustainable metal industry companies per grid cell. The use of a spatial regression, in this case a spatial lag model, was necessary because atmospheric phenomena are usually non-stationary and thus a consideration of the spatial neighbourhood is important. Due to the regular grid structure of our data, we defined the spatial weighting matrix based on queen contiguity, which includes all neighbours sharing vertices.

Using the results of our regression model, we tried to answer the following research question:

> *Does the self-representation of metal industry companies regarding sustainability coincide with findings from satellites on air pollution?*

RESULTS

Based purely on the classification of the website texts, we were able to make various statements on the topic of sustainability in the US metal industry. Only 8.1% of all companies were classified as sustainable by our machine learning model based on their website texts. The proportion of companies that either did not write anything about sustainability or used relevant keywords in a different context, on the other hand, was 51.3%. However, it has to be noted that a large proportion of US metals companies did not have their own website and could therefore not be analysed.

As is standard with machine learning-based approaches, we performed a validation of our classification results, which yielded an overall accuracy of 87%. With a good precision of 81% for the sustainable class, we were also able to state with a fair degree of certainty that companies we had declared as sustainable had also been correctly identified.

A clear focus of our study was on the spatial dimension. Therefore, we looked a little closer at the spatial distribution of sustainable and non-sustainable metal industry companies. We found clear hotspots of sustainable companies e.g. in Toledo, Pittsburgh and north of Chicago. Due to the general distribution of the US metal industry, the majority of these hotspots were located on the East Coast or in the Rust Belt, which are shown in Figure 1.

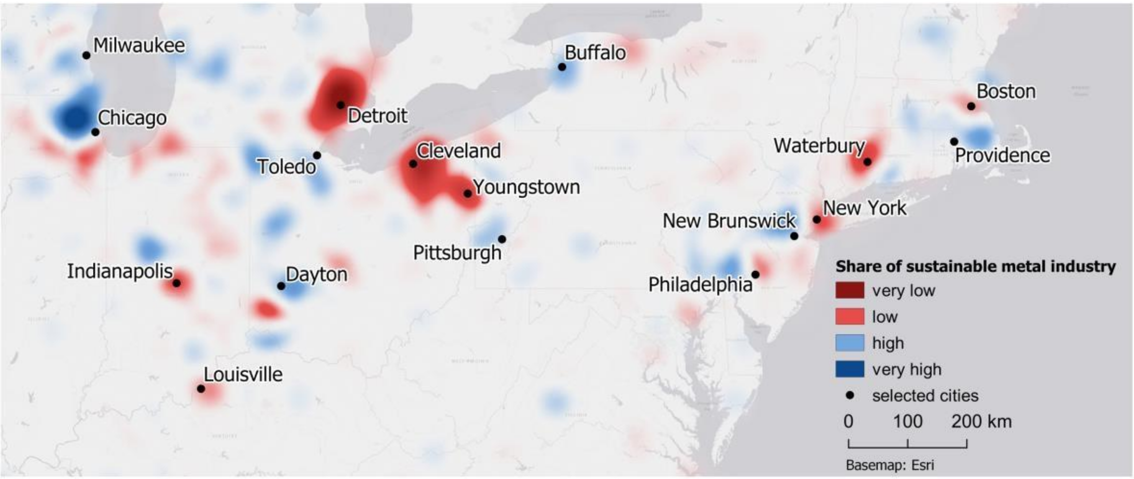


Figure 1. Distribution of the sustainable metal industry on the US East Coast.

In a further step, we performed a spatial regression analysis, the coefficients of which were used to answer our research question. In our best model, the logarithmised number of employees in sustainable metal industry companies had less influence than the number of employees in other companies of the same industry. However, we did not achieve statistical significance for the former variable. We interpreted these results as an indication that there was no fundamental greenwashing in the US metal industry, as there was no detectable effect of the sustainable companies on SO_2 concentrations, while the non-sustainable companies had a statistically significant effect. For more detailed information about our study, please see [3].

cONCLUSIONS

The study presented here on greenwashing in the US metal industry was based on a methodology that is part of ISTARI.AI's webAI. It represents a novel way of quantitatively investigating economic phenomena in near real time by accessing the content of corporate websites and classifying them using machine learning solutions. Due to the international availability of websites, this opens up a cross-national methodology for statistical analyses that ensures the comparability of results.

The applicability of the webAI methodology to diverse subject areas also opens a multitude of doors for future research and collaboration, for which the NTTS conference would provide an ideal platform.

rEFERENCES

[1] J. Kinne and J. Axenbeck, Web mining of firm websites: A framework for web scraping and a pilot study for Germany, ZEW Discussion Papers 18-033 (2018).

[2] M. Krüger and J. Kinne and D. Lenz and B. Resch, The Digital Layer: How Innovative Firms Relate on the Web, ZEW Discussion Papers 20-003 (2020).

[3] S. Schmidt and J. Kinne and S. Lautenbach and T. Blaschke and D. Lenz and B. Resch, Greenwashing in the US metal industry? A novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining, Science of the Total Environment 835 (2022), 155512.

The usage of blockchain technology for official statistics

Keywords: blockchain, official statistics, EBSI, e-Government

Abstract

The interest about blockchains has increased over time, and also this technology has greatly matured since the year 2009, when bitcoin - just an application of this technology and probably its main driver - was released. Industrial, scientific and e-government applications of blockchains are flourishing in recent years, some as a mere exercise of style, some others to stay. This article, after showing simplified basic concepts about blockchain technology, analyses some concrete application of it to the official statistics, leveraging the power of the European Blockchain Services Infrastructure.

Introduction

IBM defines the Blockchain as a shared, immutable ledger for recording transactions, tracking assets and building trust[67]. In recent years, the interest around blockchain technologies has risen, evolving from a mere speculative interest around cryptocurrencies to the industrial adoption of the technology by big players and in future by central banks.

In a nutshell, a blockchain is…a chain of blocks containing data. Imagine a block as a text file in which some lines of text are stored, and the blockchain as a folder containing all the blocks.

Blocks are numbered sequentially; everyone (except the first) contains the “fingerprint” (called *hash*) of the previous one, together with some information to be stored permanently. The blocks are normally stored in several interconnected machines (called nodes) performing the necessary operations to guarantee the integrity of the network. Everyone can access the nodes to see their content.

Blockchain for official statistics

2.1 Do we need yet another blockchain project?

In many industrial and scientific environments, the *blockchain* sometimes represents an hammer looking for a nail: the “powered by blockchain” sticker, applied to a plethora of products, make those latter more attractive for the markets.

[67] https://www.ibm.com/topics/what-is-blockchain

So one of the first questions to ask ourselves is: “do I really need a blockchain to do that”? Often the answer is simple: “no”. One example is Tradelens, a blockchain-enabled global trade platform run by Maersk and IBM, that will be discontinued in 2023 due to the lack of interest of potential customers[68].

On the other hand, the failure of titanic projects by big players of the industry does not imply that the blockchain technology is useless or too complicated to use. Small sized projects can start on public blockchains with few lines of code, a little or zero investment, and impacting the whole world.

2.2 The European Blockchain Services Infrastructure

The European Blockchain Services Infrastructure (hereinafter EBSI[69]) is the ideal infrastructure to host official statistics projects.

EBSI dates back to 2018, when 29 countries (EU member states, Norway and Liechtenstein) and the EU Commission have joined forces to create the European Blockchain Partnership

(EBP), whose vision is to leverage blockchain to create cross-border services for public administrations, businesses, citizens and their ecosystems to verify information and make services trustworthy.

EBSI, contrary to commercial blockchains, offers free transactions to its users and is publicly funded; this, together with a comprehensive documentation and the support of its technicians, makes EBSI a great host for e-government applications: some use cases applied to official statistics are quickly presented in the next chapter.

Use cases

Blockchain technologies can be very useful for the official statistics. At this stage, six main use cases can be identified:

Statistical Tables integrity verification

Hashing a downloaded table to check its integrity in the blockchain

Source verification

Verify a table source directly in the blockchain

Statistics Table Versioning

[68] https://www.maersk.com/news/articles/2022/11/29/maersk-and-ibm-to-discontinue-tradelens

[69] https://ec.europa.eu/digital-building-blocks/wikis/display/EBSI/Home

Keeping track of versions via the blockchain

File linking

Linking files and aggregate indicators with their source (between differents organisations)

Publication of main indicators

Publication of main statistical indicators via the blockchain. Power blockchain oracles for smart contracts.

Consensus between different NSOs

Reach consensus among different organisations via blockchain mechanisms for different scopes, like the limited release of microdata for scientific purposes to known subjects

The detailed description of the use cases being beyond the scope of this long abstract, it is important to point out that some of those use cases are relevant in the context of blockchain technology only if adopted by several subjects.

4. Conclusions

Since 2009, Blockchains have greatly matured and represent today a stable technology to rely on. EBSI (chapter 2) is a public implementation, led by the European Blockchain Partnership, offering several solutions to citizens, researchers and institutions, in particular for e-government. Blockchain technologies may add enormous value to official statistics, as it was presented in chapter 3.

See also:
https://www.mdpi.com/2073-431X/10/12/168

Microsimulation (GASP3M.1)

Session Chair: **Markus Zwick** *(German Federal Statistical Office-DESTATIS)*

A microsimulation model for population projections in official statistics
Pauline Pohl (Statistics Austria)*; Philip Slepecki (Statistics Austria); Martin Spielauer (WIFO)

A combined methodology SimulSTAT for public data indicators by using statistical and simulation means for improving decision-making in administration policies
Sergio Gallego García (UNED)*; Manuel García García (UNED)

A confidentiality concept for a simulation data centre
Martin Palm (*German Federal Statistical Office-DESTATIS)*

A microsimulation model for population projections in official statistics

Keywords: Microsimulation, Demographic projection, Official Statistics, Emigration hazards.

INTRODUCTION

National statistical institutes, governments and international organisations predominantly employ the cohort-component method for the production of population projections [1, 2]. This method is computationally simple, does not require a broad range of input data and is well established in the literature. However, it cannot account for complex and dynamic demographic processes, model interactions or produce results for a variety of individuallevel characteristics. To overcome these limitations, we implement a microsimulation model for demographic projections that builds on the characteristics of individuals instead of entire cohorts and allows for the simulation of realistic life-courses. To mitigate the effect of this methodological break on the comparability of projection results over time and to enable users to track changes step-by-step, we start by replicating the results of the cohort-component method in a microsimulation and gradually develop and extend individual model elements. As a first step, we incorporate a more realistic model of emigration behaviour, which explicitly accounts for the relationship between the emigration risk and the duration of stay in the host country. In the future, the microsimulation model can be extended with additional modules for education, employment, health and other socioeconomic characteristics.

In the following, we present the model and compare its results with those of the cohortcomponent method, using data for the Austrian population.

METHODS

In the microsimulation, individual life-courses are simulated over time. Since our aim is to produce a population projection, we focus on simulating demographic events (births, deaths, migration). The simulation occurs in continuous time, so events can be realised at any point and are simply aggregated by simulation year to produce results for each projection year. The microsimulation is implemented using Modgen [3], a freely available software from Statistics Canada.

The model

The cohort-component method uses event rates to determine the projected paths of fertility, mortality and migration. In the microsimulation, these rates are converted into waiting times using the inversion method (inverse transform sampling). Each person in the microsimulation is assigned waiting times based on their individual characteristics. The event with the shortest waiting time is realised. As soon as an event occurs, one of the characteristics of the simulated person changes. Based on the new characteristics, new waiting times are assigned for all events. If a person dies, moves abroad or reaches the end of the forecast horizon, the simulation ends for this person and the next person is simulated.

Moving from the cohort-component method to a microsimulation represents a fundamental change in the way a population projection is computed. It requires a deeper understanding of model building as well as more advanced statistical programming and data analysis skills. As a first step in extending the model, we propose a more disaggregated and realistic model of emigration that does not require a wide range of additional model variables or data sources. In lieu of emigration rates, we estimate piecewise constant hazards [4] for emigration for 17 different country groups, using age, sex, the current federal province of residence and the duration of stay in the host country as input variables. For the nativeborn, age- and sex-specific emigration rates are used instead of these estimated hazards.

The data

We perform the microsimulation using administrative data for the Austrian population and starting from the base year 2012. This allows for an out-of-sample evaluation of the projection results until 2020. The data include the population size at the start of the base year, migration stocks and flows, births and deaths as well as demographic indicators (e.g. life expectancy) derived from these data. The 17 country groups used to estimate the emigration hazards are established through a cluster analysis, which uses additional data from the Asylum Statistics and the Register-based Labour Market Statistics. Due to the availability of register data, the microsimulation is performed using micro data for the entire Austrian population. However, the model could also be applied to a large, representative sample or a synthetic population.

RESULTS

Estimated emigration hazards

Figure 1 plots the emigration hazards and the corresponding survival rates for a man who immigrates at the age of 18 and lives in Vienna, for two different country groups. The left pane shows the results for a man born in a high-income EU member state in Northern or Western Europe, the right pane for a man born outside of Europe, in a country whose natives have a long duration of stay and a high number of asylum applications in Austria. The figure shows substantial differences in emigration behaviour, with immigrants from the Northern and Western EU member states experiencing much higher emigration hazards in the first 10 years following immigration. Using the cohort-component method, individuals from both country groups would have been assigned the same, durationindependent emigration rates.

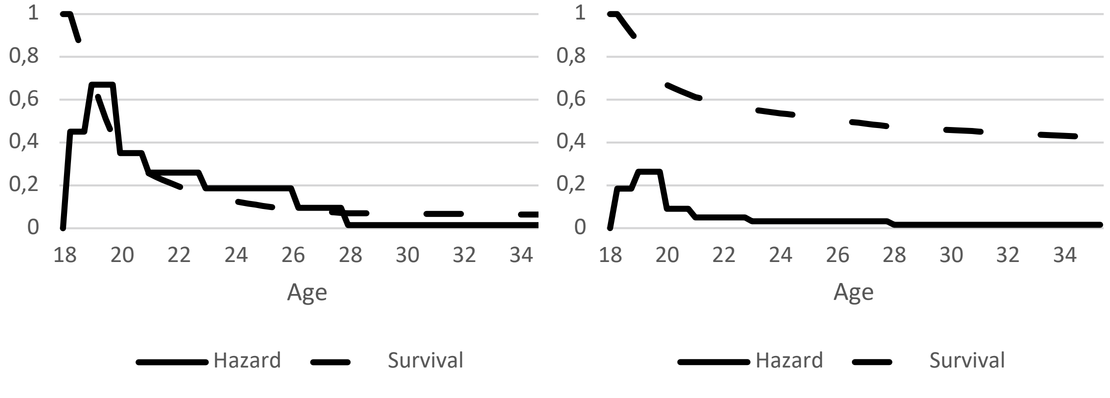


Figure 1. Emigration hazards and survival rates for a man who immigrates at age 18, lives in Vienna and was born in a high-income EU member state in Northern/Western Europe

(left pane) vs. a non-European country with a long duration of stay and a high number of asylum applications in Austria (right pane), depicted until age 35

Performance of the microsimulation vs the cohort-component method

Using 2012 as the base year, we compare the projection results of the microsimulation with the cohort-component method and find that the diagnostic test results for the microsimulation are consistently better. In particular, the forecast error of the microsimulation is lower than that of the cohort-component method. The measures for bias and MAE are the same, as both projections were consistently higher than the actual figures in the given time span. Migration flows are also better represented by the microsimulation model and the success rate is higher. This means that the direction of migration flows (rising or falling) can be mapped better. However, both models do not pass the independence test. The test statistic is not significant, i.e. the null hypothesis, according to which the signs of the change in the forecast and the realisation are independent of each other, cannot be rejected. A possible explanation for this result is that only nine years were validated ex-post and this small number of years reduces the significance of the tests.

When comparing the results of the forecasted emigration paths, the differences between the two models become evident (Figures 2 and 3). While the cohort-component model does not capture the increased emigration following the years of high immigration (2015/2016), the microsimulation captured this pattern fairly well. In addition, with the cohortcomponent method, it is not possible to capture the duration-dependent patterns of emigration. In the microsimulation, using the estimated emigration hazards, the most recent immigrants are most likely to emigrate. Accordingly, those who immigrated to Austria in 2015/2016 are also most likely to emigrate in the following years. Since only static emigration rates can be used in the cohort-component model, realistic emigration patterns cannot be captured.

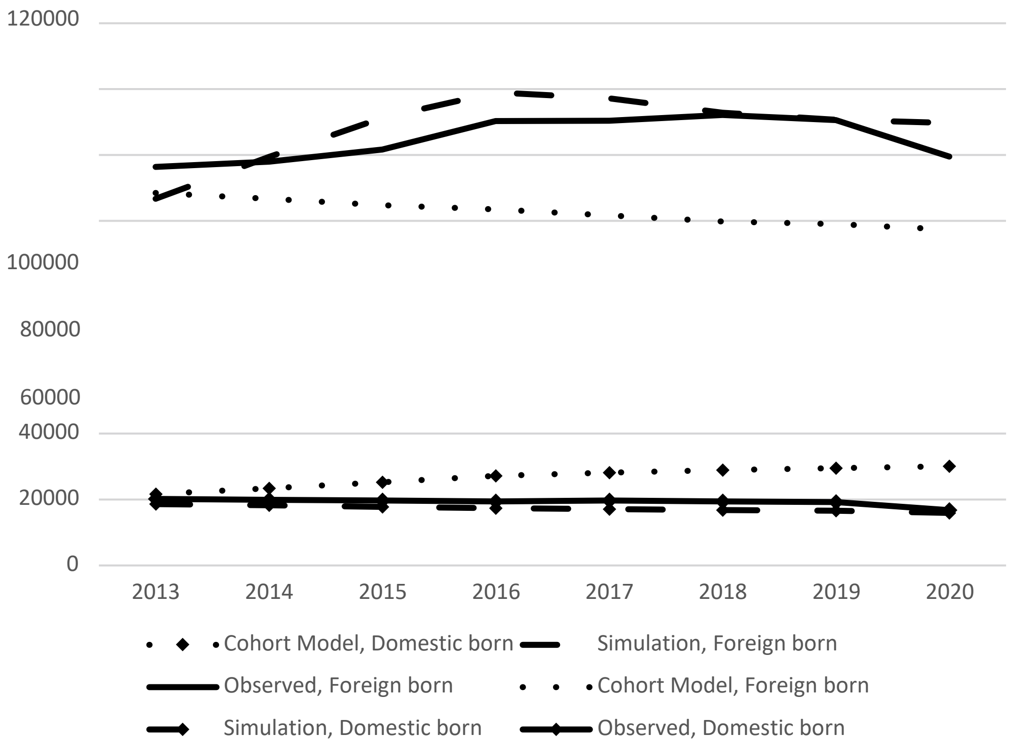


Figure 2. Projected annual emigration for the native and foreign-born population in Austria until 2020, based on the standard cohort-component method vs. the proposed microsimulation model, using 2012 as the base year

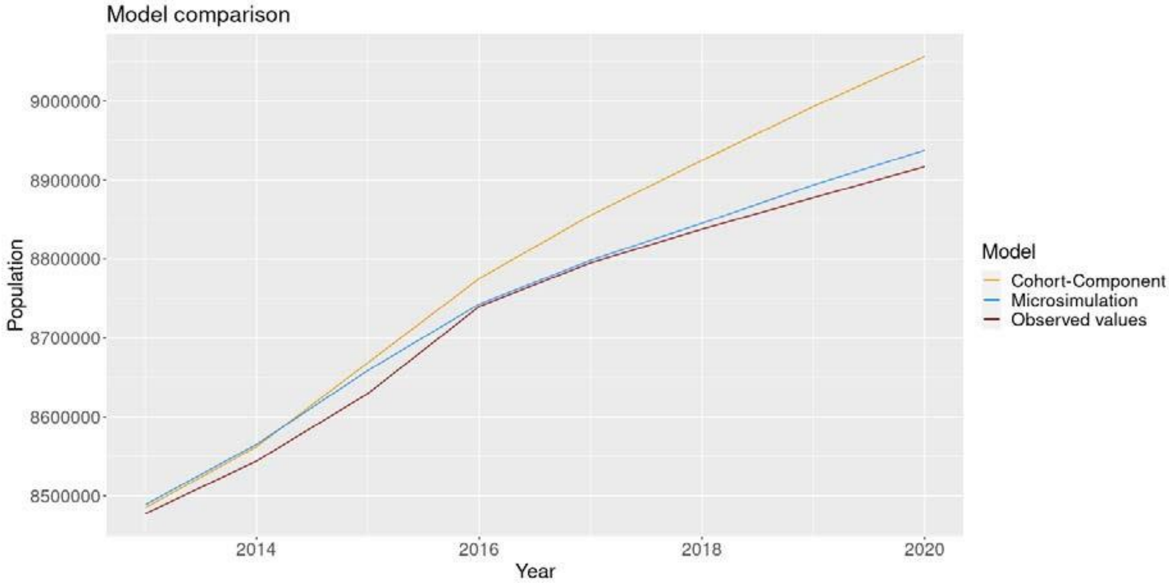


Figure 3. Population projections for Austria until 2020, based on the standard cohortcomponent method vs. the proposed microsimulation model, using 2012 as the base year

CONCLUSIONS

In conclusion, compared to the standard cohort-component method, a microsimulation model can account for more complex and dynamic patterns in demographic processes and produce results for a variety of individual-level characteristics. Using data for Austria and starting with 2012 as the base year, we show that the proposed microsimulation model produces more accurate population projections.

In the coming years, we plan to extend the model from its demographic core to include modules for education and employment. This will enable us to derive projections for the education level and employment status of the population and its subgroups within the microsimulation framework. Furthermore, we will be able to model demographic processes dependent on the education and employment characteristics of the individuals, e.g. modelling women’s fertility dependent on their education level and employment status.

REFERENCES

[1] S.K. Smith, J. Tayman and D.A. Swanson, Overview of the Cohort-Component Method. In: State and Local Population Projections, Dordrecht: Springer, Chapter 3 (2002), 43-48.

[2] M. Spielauer and O. Dupriez, A Portable Dynamic Microsimulation Model for Population, Education and Health Applications in Developing Countries, International Journal of Microsimulation 12(3) (2019), 6-27.

[3] https://www.statcan.gc.ca/en/microsimulation/modgen/modgen

[4] Willekens, F. "Bridging the Micro-macro Gap in Population Forecasting." Netherlands Interdisciplinary Demographic Institute (2006).

A combined methodology SimulSTAT for public data indicators by using statistical and simulation means for improving decision-making in administration policies

Keywords: Simulation, Digital Twin, Data Analytics, Artificial Intelligence, Public data, Administration policies, Human-Computer Interaction, Decision-making, Process Mapping, Use cases, European Semester.

INTRODUCTION

In many cases decision-making processes in public administration are driven by political, ideological reasons exposed through debates and discussion. In addition, it is also common to use point data values or some evolutions of certain parameters to influence the decisionmaking process. However, many of these processes lack a holistic and dynamic assessment with interrelationships for supporting the different decision-making process they face. For this purpose, this paper aims to provide a framework of simulation with statistical data analysis tailored for specific needs of public administration purposes. Based on the approach several use cases are described to validate the approach and support decisionmaking and recommendations for the related use cases.

METHODS

The development methodology is based on the following methods and use cases:

Process mapping related to the decision-making process

Process mapping has its origins in industrial engineering. Industrial engineering comprises a group of techniques that can be used to eliminate waste, inconsistencies, and irrationalities, and provide high-quality goods and services easily, quickly and inexpensively. The technique is known by several names, although process analysis is the most common. The first step of this general approach is the study of the flow of processes. The completed mapping can then be used as the basis for further analysis and subsequent improvement. Often this is achieved using techniques such as the 5W1H (asking: Why does an activity occur? Who does it? On which machine? Where? When? and How?) [1]. As applied for different sectors such as the mining and metallurgical sector, process mapping is the basis for analysing the current state and allowing a systematic process optimization considering related activities, decisions, indicators, and flows [2, 3].

Decision-makers need to deal with uncertainty and delays from multiple sources. In today’s organizational environments, humans and computers interacting in almost all decision-making processes. Therefore, for the successful realization of managerial positions, a methodological approach is key to consider the impact of decisions in the target criteria of a system according to the strategic, tactical, and operational levels. For this purpose, a methodology that consists of eight steps in order to guide managers and planning employees. This methodology starts from the relevant factor (the decision-related factors and criteria) that must be considered from a process perspective, followed by the mapping of the processes related to it. Then, the interrelationships can be described, and the different tasks carried out within the process can be classified attending to the nature of the interaction between the elements involved. As a result, the interrelationships of the tasks with other elements, and their place in the process are identified. From the analysis of the tasks, the potential bottlenecks and related measures and decisions can be determined. Based on this analysis, models based on mathematical formulation, simulation, and AI can be designed to assess the impacts of decisions on the system's performance. Moreover, a sensitivity analysis can be performed to show the impact of variation of certain parameters. The methodology supports the determination of priorities and increases the frequency with which certain decisions are considered in the global system, as well as enabling the use of lessons learned, such as by planning actual deviations to develop AI models and creating recommender systems for future planning periods [4].

Digital Twin and Simulation

A digital twin model and the link with information system databases are the basis for performing simulations that enable one to analyse policies based on what-if analysis and statistical techniques. All these models share the capabilities of facing dynamic environments and attenuating their volatility by predicting and managing changes through enhanced adaptability. By doing so, any change in any part of the whole digital twin model can be identified and assessed, thus calculating the impacts and risks associated with any change and improving the adaptation capabilities [5]. According to the "VDI-Richtlinie", a simulation is "the reproduction of a system with its dynamic processes in an experimental model capable of acquiring knowledge that can be transferred to reality. In particular, the processes that develop through the time" [6]. Simulation models are mainly used to support decision making since they show the dynamic behaviour of a system [7]. System dynamics is an efficient method to obtain useful information about situations of dynamic complexity. It has been increasingly used to design more successful corporate policies and public policy adjustments [8]. System dynamics deals with the behaviour of a system and how it influences its own future evolution, which can be considered as the strategic issues that affect the top management of organizations [9].

Statistical methods and techniques

The suitability of a forecasting method depends on the pattern [12]. Therefore, several models were selected. The forecasts of the different methods in the conceptual model were compared by means of the forecast error. There are different methods to measure and provide conclusions regarding the accuracy of the used forecast method [13]. The mean absolute deviation (MAD), the mean square error (MSE), and the mean absolute percentage error (MAPE) as methods that produce consistent results when comparing different forecasting methods [11]. The methods used are [10, 11]:

1. Moving average and cumulative moving average.
2. Linear regression.
3. Exponential smoothing of first, second, and third order.
4. Croston method.

In this paper, the methods are applied to forecast budgets, deficits, and growth, among others. Moreover, hypothesis test is applied to identify changes in the data series analysed. **2.4. Decision-making in administration policies and Key Performance Indicators**

Decision-making processes affect public policies and relevant key indicators for society and sustainable development in the European Union. In this context, the paper aims to describe the application of the combined approach to different use cases as for decisions of The European Semester. It is the framework for integrated surveillance and coordination of economic and employment policies across the European Union. Since its introduction in 2011, it has become a well-established forum for discussing EU countries' fiscal, economic and employment policy challenges under a common annual timeline. One of the use cases is the country-specific recommendations that are documents that provide an analysis of each member state's economic situation and recommend measures that each country should take over the coming 12 months [14].

RESULTS

First result is the combined methodology SimulSTAT for public data indicators by using statistical and simulation means based on the digital twin of the process mapping of the current state for improving decision-making in administration policies:

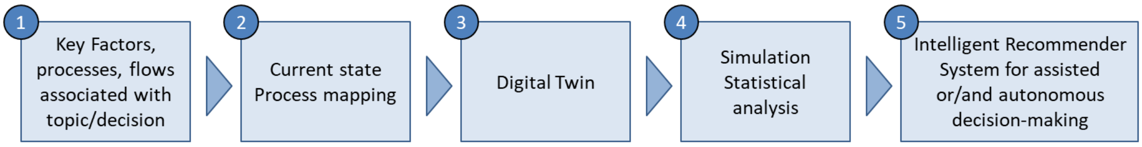

1
Key Factors, processes, flows associated with topic/decision
2
Current state Process mapping
3
Digital Twin
4
Simulation Statistical analysis
5
Intelligent Recommender System for assisted or/and autonomous decision-making

Figure 1. Extract of results: combined methodology SimulSTAT.

Secondly, the results provide descriptions of use cases in which the methodology can be applied such as for deficit forecasting and economic growth among others. In this context, the approach is able to determine and improve existing forecasts with the simulation of different scenarios so that recommendations are derived depending on selected influencing factors. In this context, countries could receive dynamic recommendations depending on internal and external factors as well as an information system and methodology for monitoring the development of recommendations that enabling the adherence to target indicators and therefore increasing the capability for sustainability. Third, the approach is able to provide tailored recommendations for different scenarios and a system for monitoring the development that can be assisted by methodologies to implement relevant improvement strategies as well as industry 4.0 technologies [15, 16] to identify potential measures for optimization and closing the related functional gaps [17].

CONCLUSIONS

The combined methodology SimulSTAT is able to support the improvement of forecasts, the development of scenario-based recommendations, as well as a tool for managing and monitoring targets enabling the assisted and/or autonomous decision-making towards goals and capabilities for sustainability. The approach can be applied initially for specific use cases, and later roll-out as a global methodology.

REFERENCES

[1] Hines, P., & Rich, N. (1997). The seven value stream mapping tools. International journal of operations & production management.

[2] Pérez, S., Gallego, S., & García, M. (2021, October). Production optimization oriented to value-added: from conceptual to a simulation case study. In IOP Conference Series: Materials Science and Engineering (Vol. 1193, No. 1, p. 012100). IOP Publishing.

[3] Gejo Garcia, J., Gallego-García, S., & García-García, M. (2019). Development of a pull production control method for ETO companies and simulation for the metallurgical industry. Applied Sciences, 10(1), 274.

[4] Ren, D., Gallego-García, D., Pérez-García, S., Gallego-García, S., & García-García, M. (2021, December). Modeling Human Decision-Making Delays and Their Impacts on Supply Chain System Performance: A Case Study. In International Conference on Intelligent Human Computer Interaction (pp. 673-688). Springer, Cham.

[5] Gallego-García, S., Ren, D., Gallego-García, D., Pérez-García, S., & García-García, M. (2022). Dynamic Innovation Information System (DIIS) for a New Management Age. Applied Sciences, 12(13), 6592.

[6] Brunner, A. Simulationsbasierte Bewertung von Supply-Chain-ManagementKonzepten Apprimus; Apprimus Verlag: Aachen, Germany, 2011.

[7] Reggelin, T. Schneller Entscheiden. Log.Kompass; DVV Media Group: Hamburg, Germany, 2012; p. 5

[8] Sterman, J.D. Business Dynamics: Systems Thinking and Modeling for a Complex World; Irwin/McGraw-Hill: New York, NY, USA, 2000.

[9] Coyle, R.G. System Dynamics Modelling: A Practical Approach; Chapman & Hall: London, UK, 2008.

[10] Schuh, G.; Stich, V.; Wienholdt, H. Logistikmanagement.; Springer: Berlin/Heidelberg, Germany, 2013.

[11] Meyer, J.C.; Sander, U.; Wetzchewald, P. Bestände Senken, Lieferservice Steigern-Ansatzpunkt Bestandsmanagement; FIR: Aachen, Germany, 2019.

[12] Gallego-García, S.; Reschke, J.; García-García, M. Design and simulation of a capacity management model using a digital twin approach based on the viable system model: Case study of an automotive plant. Appl. Sci. 2019, 9, 5567.

[13] Schönsleben, P. Integrales Logistikmanagement: Operations und Supply Chain Management Innerhalb des Unternehmens und Unternehmensübergreifend; Springer: Berlin/Heidelberg, Germany, 2011.

[14] https://www.consilium.europa.eu/en/press/press-releases/2022/06/17/europeansemester-2022-country-specific-recommendations-agreed/

[15] Groten, M., & Gallego-García, S. (2021). A Systematic Improvement Model to Optimize Production Systems within Industry 4.0 Environments: A Simulation Case Study. Applied Sciences, 11(23), 11112.

[16] Gallego-García, S., Groten, M., & Halstrick, J. (2022). Integration of Improvement Strategies and Industry 4.0 Technologies in a Dynamic Evaluation Model for TargetOriented Optimization. Applied Sciences, 12(3), 1530.

[17] Winkler, M., Gallego-García, S., & García-García, M. (2022). Design and Simulation of Manufacturing Organizations Based on a Novel Function-Based Concept. Applied Sciences, 12(2), 811.

A confidentiality concept for a simulation data centre

Keywords: Microsimulation, confidentiality, open and reproducible research, microdata, public use data, statistical disclosure control.

INTRODUCTION

MikroSim is a multi-sectoral and regional dynamic microsimulation model of the German population at the level of persons and households. In absence of an appropriate dataset, within the MikroSim-project a synthetic population of Germany was created. For this purpose, data from the preparatory studies for the Census 2011 and numerous crosssectional and longitudinal datasets from official statistics and empirical social research has been used. Based on this dataset, in each simulation year, individuals and households run through a set of simulation modules such as mortality, fertility, regional mobility, education and employment (Münnich et al., 2021).

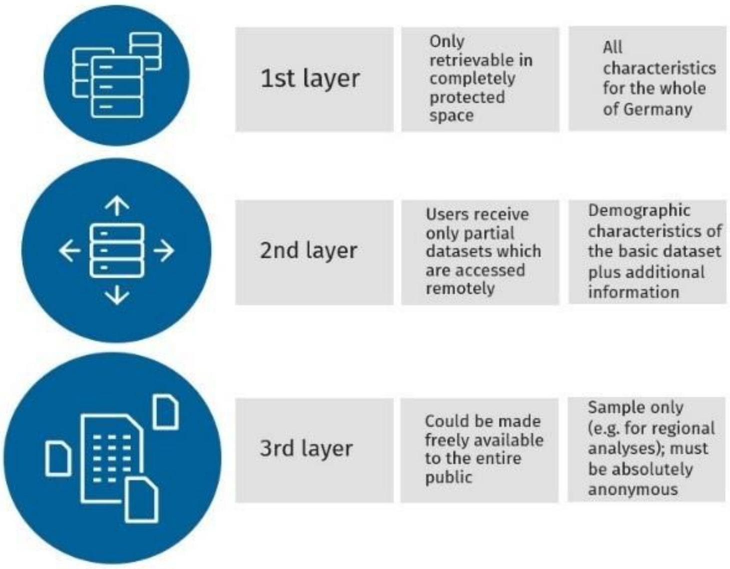
The MikroSim-project aims to make the simulation infrastructure and the basic data available to researchers by implementing a simulation data centre. This requires anonymization of the input data to protect against re-identification and, due to the spatial granularity, an independent confidentiality concept for output data.

Anonymization and data protection of the individuals providing information on the one hand are naturally in conflict with the usability and the scientific gain of knowledge of the researchers on the other hand. Synthetic data are only considered anonymous if they cannot be related to individuals (Drechsler and Jentzsch, 2018). The task of suitable confidentiality procedures and (partially) synthetic data is to guarantee the necessary anonymity while maintaining a high quality of results. For (partially) synthetic data, no corresponding concepts exist so far. But Synthetization itself is used as an anonymization method and different measures for its efficiency exist (Burnett-Isaacs et al., 2021).

In order to arrive at a confidentiality concept for a simulation data centre, this paper will discuss whether the synthetic data set at hand poses disclosure risks. This requires both measures to quantify this risk and a suitable basis of comparison to which they can be applied.

METHODS

The primary goal of confidentiality procedures and the resulting confidentiality concept for the construction of a simulation data centre is to ensure the anonymity of the individuals providing information. Here, a distinction has to be made between different degrees of anonymity, some of which are prescribed by law (absolute, de facto, formal anonymity). With regard to the planned simulation data centre, a distinction must be made between input and output anonymization. Input anonymization is about what data may be used for simulation within MikroSim. Output confidentiality is about which data may be released at the end, taking into account re-identification risks. The proposed architecture of the simulation data centre provides for a multi-level access system with 3 layers made available to different user groups.



1st layer	Only retrievable in completely protected space	All characteristics for the whole of Germany
2nd layer	Users receive only partial datasets which are accessed remotely	Demographic characteristics of the basic dataset plus additional information
3rd layer	Could be made freely available to the entire public	Sample only (e.g. for regional analyses); must be absolutely anonymous

Figure 1. 3-Layer-Architecture for different user groups.

Functionality of the MikroSim dataset

Within the individual sub-modules of MikroSim, data sets are added to the basic data. On the basis of e.g. official statistics, analyses are carried out and assigned to the individual feature carriers in the data set. This is not a record linkage, where data sets are linked on the basis of unique identifiers, but a prediction on the basis of the transmitted coefficients or benchmark values. This raises the question of whether these synthetically generated data sets may be made available to the public under confidentiality aspects, whether they are already sufficiently protected or modifications are necessary, because even partially or fully synthetic data harbor re-identification risks, such as the risk of random hits and that data actually occur partially in this way in reality (for partially synthetic data).

Analogous to the difference of confidentiality of input and output, the question arises for the input, which confidentiality rules and resulting values as basis of the simulation lead to which synthetic results and for the output, which synthesization methods lead to which results?

Risk measures

The basic problem in choosing appropriate measures is to find a suitable basis for comparison. Research on anonymity aspects has so far been conducted mainly on data sets that have been linked via record linkage - i.e. that occur exactly in reality - but not on data sets that have already been synthesized. As a result, measures of the quality of the synthesis and of the risk of disclosure are not readily applicable. For datasets synthesized for the purpose of confidentiality, the possibility of a comparison to the original data exists; for synthetically created datasets, this possibility does not exist or does not exist in this form.

Thus, we need 1.) a basis of comparison against which we can measure whether or not there is a disclosure risk, since no original dataset is available, and 2.) for this comparison, a suitable risk measure that is applicable to the basis of comparison.

Various approaches are discussed as methods for measuring disclosure risk, as proposed e.g. for data sets, that have been linked via record linkage (Drechsler et al., 2008; Reiter, 2005). Since these are only partially transferable - since the possibility for record linkage of the original data is not available here - the following approach will be used as a measure:

Checking whether additionally addressed characteristics simulated in the base data set and present in the original statistics are identical. Idea: Base data set D with x1, x2, x3, ... is extended by official specialized statistics D* with x1*, x2*, ...

D* consists of characteristics of interest used for D as x4, ...; but also contains demographic variables x2*, x3*. Not all variables contained in D* are used to extend (synthetic) data set D; interesting characteristics (x4...) are added to base data by simulation.

Characteristics 2 and 3 are therefore available in two variants: in original x2*, x3* and in simulated form x2, x3. Comparison original/simulated data is now possible: Did the simulation match the originals: x2* = x2? Is there a case found in the originally collected statistics that happened to be exactly replicated in the simulation, given the characteristics that were collected in the specific statistics (but could not/was not used for simulation)?

Basis of comparison

As a basis for comparison, to which the proposed risk measures can be applied, the following possibilities can be considered in principle:

1.) Direct transfer of approaches based on real datasets: If additional characteristics of the base population are simulated based on coefficients, one could back-calculate benchmarks after simulation and evaluate how far they differ from the original estimates. Problem: tautological result, since the coefficients are based on these values.

2.) Perform record linkage between original datasets in the secured research data centre (RDC) infrastructure to determine if there are any problematic cases within the simulated data. Problem: The original database is not completely available in this form, otherwise synthesization would not be necessary as a method to construct the MikroSim dataset in the first place. This approach would also not be within the legal framework, even if it were theoretically possible.

3.) We simulate original data sets to use common procedures: Simulated original data are checked for confidentiality according to common rules. These are then combined by simulation and thus synthesized. Now statements can be made about how well synthetic data match the original data. In addition, risk measures can be applied to datasets that have been connected via record linkage. Thus, the characteristics the data generated regarding confidentiality rules for input/output can be evaluated. Conceivable: Comparisons of how often synthesized data hit original data (risk of re-identification) – if this is significantly frequent, underlying confidentiality rules must be adapted. The determined scheme can now be transferred to the real simulation infrastructure.

rESULTS

The procedure demonstrated here corresponds to a complete simulation of the entire confidentiality and synthesis process.

Existing data sets of the simulation population are used as the data basis. For the present work, only a part of the entire modules/data sets of the MikroSim project is used, as well as a section of the base population. This is to first generate a single data source from two unrelated data sets that cannot be merged via record linkage, which will serve as a simulated original source in the following.

In order to simulate suitable data, it is first put into a form that is suitable to provide the desired insights, i.e., that a confidentiality check would be necessary in principle, in order to get as close as possible to the non-existing original data. To achieve this, existing rules used to generate the input must be flipped, e.g., by simulating low numbers of cases that would otherwise have to be blocked. Specific rules corresponding to the confidentiality of the input are applied to these simulated originals. From the simulated originals, data are now synthetically generated as they might occur in the simulation population. Now it is possible to measure how often data from the fictitious original dataset was hit and to measure re-identification risks by comparing the simulated original data with the synthesized data.

The proposed approach shows that simulations are generally suitable for measuring disclosure risks. The effect of different input rules as well as synthesis procedures on the probability of disclosure risks can thus be evaluated and serve as an estimate for the overall data set.

The synthesized results correspond on average to the regression results used to generate the synthetic data, but do not allow any conclusions to be drawn about the underlying individual data. Data attackers can thus only estimate probabilities for whether they could be real units with great uncertainty.

cONCLUSIONS

As an important part of the confidentiality concept, it is necessary to increase the number of partial data sets assumed as the data basis as well as the number of statistical units in simulation studies to a number close to the actual data sets used.

Important research needs would be especially in the construction of tests that index reliable disclosure risks and in an increase of the amount of partial data sets for the simulation studies.

Perspectively, selected methods will be used to show how the application of different methods affects different data sets with different structures. Quality criteria will be used to evaluate which methods are suitable for which types of data sets.

rEFERENCES

[1] R. Münnich, R. Schnell, H. Brenzel, H. Dieckmann, S. Dräger, J. Emmenegger, P. Höcker, J. Kopp, H. Merkle, K. Neufang, M. Obersneider, J. Reinhold, J. Schaller, S. Schmaus, P. Stein: A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model, methods, data, analyses (2021), Vol. 15, No. 2, pp. 241-264.

[2] Drechsler, J., Jentzsch, N.: Synthetische Daten. Innovationspotential und gesellschaftliche Herausforderungen. Stiftung Neue Verantwortung e. V. (2018).

[3] Burnett-Isaacs, K., Girar, C., Ramsden, A., Sallier, K. & Slokom M.: Synthetic Data for National Statistical Organizations – A Starter Guide, UNECE (2021).

[4] J. Drechsler, A. Dundler, S. Bender, S. Rässler, T. Zwick: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel, Transactions on Data Privacy (2008), Vol. 1, No. 3, pp. 105-130.

[5] J. Reiter: Estimating Risks of Identification Disclosure in Microdata, Journal of the American Statistical Association (2005), Vol. 100, No. 472, pp. 1103-1112.

Privately Held Data (JENK3M.1)

Session Chair: **Michail Skaliotis** *(European Commission)*

(Re-)using privately held data : experimental statistics on online collaborative economy platforms
Simon Bley, Christophe Demunter *(Eurostat)*

Practical Privacy-Aware Data Linkage and Statistical Aggregation based on Privacy Enhancing Techniques
Chris Dugdale (Statistics Canada); Benjamin Santos (Statistics Canada)*; Zachary Zanussi (Statistics Canada)

Towards a Taxonomy for Business-to-Government Data Sharing
Serena Signorelli *(European Commission - Joint Research Centre)*

(Re-)using privately held data : experimental statistics on online collaborative economy platforms

Keywords: platforms, tourism, privately held data, accommodation, tourism statistics, collaborative economy, experimental statistics.

Note:

The authors submitted a proposal for a Special Session for the NTTS 2023, with the same title as above. The session will take a multidisciplinary view on the enablers and the risks of using privately held data (from online collaborative economy platforms) to produce data on short-stay accommodation booked through these platforms.

The set-up of the session (duration 60' or 90' – to be determined) is a panel discussion initiated by short presentations (of 5-7 minutes) by each of the panellists. The panel will consist of speakers from Eurostat (the authors), the industry (representative of one of the platforms), an NSI (speaker to be determined), the policy side (DG GROW) and a local authority (city planner or regulatory body using the data).

The current abstract serves two objectives:

- *If the special session will have a duration of 60 minutes only, it can be considered as a submission for a spin-off presentation to be added to a relevant session on innovation and/or privately held data, discussing this project in a more general way (ideally on day 1 or 2 of the conference). This additional presentation would serve as a general overview of the project and to introduce the issues discussed at the Special Session; by spinning off these topics we save time for the in-depth discussion on one of the following days;*
- *It can considered as a background paper for the Special Session (ideally scheduled on day 2 or 3 of the conference).*

Introduction

In the 21st century, the digitalisation of society and businesses helped statisticians to transcend the limits of traditional surveys and to address user needs better and more quickly. Re-use of privately held data is one of the more promising paths. Since 2020, Eurostat and the NSIs have been exploring and using data obtained from Airbnb, Booking.com, Expedia Group and Tripadvisor to produce data on short-term accommodation booked via these online collaborative economy platforms.

The authors' contribution to the NTTS2023 (a stand-alone presentation as well as a Special Session – see introductory note above) will take a multidisciplinary look at the enablers of the project and of the main methodological and institutional challenges. The current contribution is a follow-up of earlier interventions at the IAOS 2022 that took place in Krakow and the Q2022 Conference that took place in Vilnius ([1]).

The project originates from the need to better cover a traditionally ignored segment of the accommodation sector. Smaller businesses, often family businesses that are not always captured in statistical registers or in tourism registers, used to stay under the radar of official statistics. The trade-off or balance between additional quality (completeness) and cost or complexity of data collection, changed with the rise of online platforms, now over a decade ago. The digital footprint that service providers leave behind on the platforms, offered unseen opportunities to capture and measure their activity. Furthermore, the accessibility and global nature of the platforms, made it easier for service providers to reach potential guests and motivated households and businesses to enter the market. This growth of the market made it more important to cover this previously under-covered segment of official tourism statistics. As a secondary benefit, the project serves as a lighthouse project to explore the use of privately held data in the specific setting of the ESS (European Statistical System).

Methods

The project on using privately held data from online collaborative economy platforms to estimate tourism flows in short-stay accommodation offered through such platforms, explored a range of unknown, innovative territories, thus faced (and still faces) many challenges to deal with.

In the next paragraphs, we will briefly discuss the governance in relation with the platforms and with the partners in the ESS, the data processing procedures, and the outstanding methodological challenges that need to be solved as a condition for becoming official statistics.

Governance

An early incarnation of the project (2016) was not successful, most likely due to a too narrow set-up: statisticians talking to data scientists, but not sufficiently embedded in the decision-making structure on both sides. In 2018, a more multi-disciplinary approach led to agreements between the European Commission and four major international platforms (Airbnb, Booking.com, Expedia Group and Tripadvisor) arranging the sharing of data on capacity and occupancy of accommodation facilities offered and booked through the platforms. This time, the project team consisted of a multi-disciplinary group, including data scientists and statisticians, as well as experts in contracts and legal affairs, experts in investor relations, etc. Furthermore, negotiations covered four platforms, offering a level playing field for the sector.

Secondly, the governance of the project also involved managing a reverted data flow in the ESS. Whereas traditionally, and in line with the subsidiarity principle, data collection happens decentralised by the Member States, who transmit aggregate data or microdata to Eurostat, in this project the data collection happens centrally at Eurostat, where the data is validated, merged and transmitted to the national statistical institutes (NSIs).

Quality assurance framework

The quality assurance consists primarily of traditional format and plausibility checks, but given the specific nature of the project, with Eurostat serving as the central hub for receiving the data from the platforms, a two layered approach was installed.

Firstly, Eurostat runs format and plausibility checks on the individual datasets transmitted quarterly by each of the four platforms. Once the data is validated, the data for the four platforms is merged into a total value – data at individual platforms level cannot be disclosed as per the non-disclosure agreements concluded with the platforms – and shared with the NSIs at the same level of geographical granularity (namely broken down at the level of local administrative units). The Member States can in this second phase run additional checks, but also use their local knowledge of the tourism market or rely on auxiliary external sources, such as registers or third party data, to assess the plausibility of the volume or trends observed in the data.

Eurostat is exploring other approaches to validate the data against external input, as part of the quality assurance framework. This is not only relevant in the context of assessing the completeness or coverage of the data, but also in the context of the principles of impartiality laid down in the Code of Practice for European statistics, given that a handful of companies provide data that will ultimately be used by policy makers to monitor or evaluate the sector’s impact on the economy and on local communities.

Methodological challenges

A key objective of the project is to enhance existing tourism statistics with the data on platforms tourism. However, two key methodological issues have to be dealt with first, both relating to double counting in the data.

Firstly, the capacity data can include double counting of listings because a host can (and often does) advertise an apartment or dwelling on several platforms simultaneously to maximise the changes that a possible guest sees it. This means that the capacity data from the four platforms cannot be added to obtain the correct total capacity. Eurostat, and a number of Member States, are currently looking into deduplication methods that could produce the necessary factors to correct for such double counting.

Secondly, the occupancy data can in some cases overlap with data already transmitted by service providers or property managers to the NSIs in the context of the regular collection of accommodation statistics. To avoid that stays or nights are counted twice, deduplication methods also need to be developed to identify those establishments that are included in the platforms data as well as in the statistical register or tourism register used by the national authorities for collecting and producing accommodation statistics.

Results

Eurostat agreed with the platforms that the data published should be similar (in terms of variables and breakdowns) to the data published by Eurostat relating to other actors in the accommodation sector, pursuant to Regulation (EU) No 692/2011 concerning European statistics on tourism.

Eurostat has been publishing the data on platform tourism as experimental statistics since June 2021. The releases cover data on occupancy, broken down by NUTS 2 and NUTS 3 regional level, as well as for selected cities. Due to the double counting issue outlined in paragraph 2.3 above, it is not yet possible to publish capacity data.

Eurostat collects the data centrally and shares the data with the NSIs who can disseminate the data about their country via their usual channels. However, because NSIs might wish to publish data at a more granular level than the data published by Eurostat, for instance for additional cities or for specific delineations of tourism destinations that differ from the (administrative) breakdown into NUTS regions, a procedure was put into place to have these requests approved by the platforms.

Conclusions

The project on measuring the short-stay accommodation offered through online collaborative economy platforms entered a more mature stage, but still an important number of issues relating to methodology or quality need to be addressed.

Notwithstanding these pending question marks, the ESS is already regularly publishing experimental statistics. Since autumn 2022, this happens at quarterly intervals. The take up by users of these additional accommodation statistics stemming from this innovative data source is good and underlines the relevance of the project in improving the completeness of tourism statistics.

Beyond the relevance of the domain of tourism statistics, the project is also functioning as a test case or proof-of-concept that privately held data can be a sustainable and high-quality source for producing official statistics.

References

[42] S. Bley and C. Demunter, From cradle to production – measuring the collaborative economy, paper presented at the Q2022 European Conference on Quality in Official Statistics (Vilnius LT, 8-10 June 2022).

Practical Privacy-Aware Data Linkage and Statistical. Aggregation based on Privacy Enhancing Techniques

Keywords: Privacy enhancing techniques, secure multiparty computation, data integration, record linkage, privacy preservation

INTRODUCTION

Record linkage is the process of combining information about entities contained in multiple data sources into a single linked dataset. It is an important part of many statistical programs and can lead to reductions in cost, time, respondent burden, and may be the only feasible way to obtain certain statistical information. Linking provides richer and more complete datasets that allow analysts to study more complex features and trends. In most statistical linkages, the desired product is not the linked data itself but some set of aggregates computed on the linked table. The aggregates can be very simple, such as the cardinality of the linked set, or more complex, such as weighted sums based on some combination of categorical data.

A common data linkage project for a National Statistics Organization (NSO) involves linking survey data with data held by a secondary, in this case federal, organization. The latter party has a dataset consisting of identifiers with numerical microdata--for example, if this was a federal department, they might have a dataset of individuals identified by a country's universal identifier number with some numerical value relating to a program they run. The department wants to evaluate the effectiveness of the program, without exposing any sensitive information regarding the people in the datasets.

In this work we consider the case where the secondary organization seeks to enrich their dataset by linking it to the respondents of a sample survey taken by the NSO. We perform the linkage on the country's universal identification number and perform exact matching. In order to address privacy concerns, we make use of Privacy Enhancing Techniques (PETs).

Privacy Enhancing Techniques are being developed to facilitate computation on sensitive data, opening the possibility of performing record linkage while ensuring privacy. Indeed, the well-studied problem of Private Set Intersection (PSI), where two parties in possession of private datasets aspire to compute the intersection of their datasets without revealing unnecessary information about their elements, is applicable in many scenarios where data sharing is desirable but prohibited by data privacy laws. An extension of PSI is Privacy Preserving Record Linkage (PPRL). Associated to each identifier is auxiliary microdata, called a payload. In addition to computing the intersection, the parties share their payloads, resulting in an enriched data table consisting of only records that both parties possess and the payloads held by either party. For an overview of PPRL, consult [1].

Recently, it has been noted that several protocols for PPRL can be extended to protocols for PPRL with Aggregation ([2], [3], [4]), where the goal is not to compute the linked table but rather to perform some sort of analysis on it. This is done by combining the base protocol with secret sharing, where sensitive values are split up into shares and distributed to be computed on by the data holders. After the computation, the shares can be returned and recombined into the results of the desired calculation.

For such work, we apply the PPRL protocol outlined in ([2], [3]) to the problem outlined above. Though both works describe how one can extend their protocols for PSI to facilitate payloads and aggregation, to our knowledge neither work implemented nor benchmarked this extension. While these works considered very simple computations on the intersection (computing the cardinality, and only returning it if it is greater than an agreed upon threshold), we extend their methods to accept payloads as well as generic computation on these payloads. We also consider a simple way to pack multiple variables together into a single payload on both sides, allowing computations with more than one input from each side. Benchmarking suggests that this protocol is not only possible but also practical for organizations with modestly sized datasets (tens of thousands to hundreds of thousands of elements).

The protocol runs in two stages. First, the data are input into an Oblivious Programmable Pseudo-Random Function (OPPRF), which allows for data to be securely obfuscated in a controlled way that facilitates linkage. Next, the parties make use of Secure Multi-Party Computation (SMPC) to compute the aggregates. Here, the obfuscated values are split into secret shares and used to compute a suite of aggregates based on the attributes present in the NSO's survey dataset. We consider both numerical attributes, where the value represents some quantity, and categorical attributes, which assign the identifier to one of a limited number of discrete classes. The goal is to combine the microdata from the two sources and use the linked dataset to compute the desired aggregates.

METHODS

An important preliminary step in many ([4], [3], [5], [2]) PSI protocols is cuckoo hashing, and is reminiscent of the bucketing protocols that are used in standard statistical linkages. In a naive linkage protocol between two sets with □ points each, to find all potential matches, one must compare every point in each set to every point in the other set, resulting in a protocol with □(□) comparisons. In bucketing, the sets are first sorted into buckets of points that are similar so that we need only compare points within the buckets, resulting in a significantly more efficient protocol. For example, in a linkage on numerical identifiers, one can sort one of the sets into buckets based on the first few digits of the identifier, and at linkage time one must only make comparisons within the appropriate buckets.

Since the intersection containing the common identifiers is itself sensitive, the parties involved want to replace them with random ones. One way to ensure consistent randomized identifiers across datasets is to use a pseudo random function. To deidentify their datasets, the two parties can agree upon a key □ and then use the PRF to randomize their identifiers. Then if one party sees the identifiers of the other, they are not able to distinguish them from random values. After deidentification, the parties can exchange their datasets freely.

This simple method is vulnerable to brute-force attacks, where one party computes the PRF for many input values, storing the outputs. They can then link their outputs to the ones present in their counterpart's dataset, and reidentify any that match. This is a problem in a semi-honest security model, where both parties will try to recover any sensitive information possible. We need some way of restricting access to the PRF once the datasets change hands.

This can be achieved using an oblivious PRF (OPRF). An OPRF is a two-party protocol where one party, the sender, holds a PRF F and a key □, and the other, the receiver, holds a binary input y. As output, the receiver learns the value □(□, □) and the sender learns nothing. A simple protocol for PSI can be constructed directly from the OPRF. The sender can apply the PRF to their own dataset, and the sender and receiver can collaborate to apply the PRF to the receiver's dataset. Then, the sender can forward their obscured dataset to the receiver, who can perform the matching. Notice how the receiver is not able to evaluate the PRF on their own, thus eliminating the possibility of the brute-force attack. Implementations of the OPRF can be found in [6].

Next, our method incorporates Secure Multiparty Computation (SMPC) techniques. The SMPC techniques that we employ are based upon secret sharing. The key idea here is that a secret is split into multiple components, and then distributed amongst the parties. The parties then perform computation upon the components that they have access to, and then combine them to determine the result. For examples of secret sharing schemes, see [7], [8], [9]. In our work, we are only interested in the two-party case, however extensions to multiple parties exist.

Our protocol combines the above three methods in the order that they have been presented.

RESULTS

We are modelling our problem based upon a NSO having a dataset with 60000 entries and 12 survey attributes, and a federal organization having 380000 entries to match against the survey. They hope to compute 132 statistical aggregates.

To our knowledge, the most performant open-source implementation of the type of OPPRF our protocol uses are [3], [2]. We refer to these papers for benchmarking, as we make use of them. We need two invocations of one of these protocols to obfuscate the identifiers and payloads. Using the protocol of [2], both obfuscations can be completed on the full datasets in less than 20 seconds and with about 20 MB of communication. This is with the NSO functioning as the receiver and the secondary organization filling the role of the sender. The parties use three hash functions for to construct their tables

Table 1. Performance of our Protocol

# Aggregates	Time (s)		Data (Mb)	
	Setup	Online	Setup	Online
1	2.26	0.378	175	3.03
10	8.36	1.09	614	11.3
132	70.2	11.8	6,561	122

cONCLUSIONS

In this brief report, we have outlined a protocol for Privacy Preserving Record Linkage with Aggregation. This allows privacy-conscious organizations to link datasets without compromising data privacy, thus facilitating the integration of data from various sources. These methods can possibly reduce expensive legal and administrative agreements which normally precede data sharing.

The protocol we have outlined is efficient enough to be run on commodity computing hardware in a reasonable amount of time. Our method works with both numerical and categorical payloads held by either organization, and indeed the computation done on the intersection can be arbitrary and is only limited by the computational and communication complexity that the organizations are willing to handle.

The next steps in this line of research could be extending the number of parties to be greater than two, or to considering administrative data files which can have numbers of records in the millions. One could also consider more complex or more specialized aggregates to compute.

rEFERENCES

[1] R. Hall, S. Fiaenberg. Privacy-Preserving Record Linkage. Privacy in Statistical Databases 2010, 269-283.

[2] N. Chandran, D. Gupta, A. Shah. Circuit-PSI With Linear Complexity via Relaxed Batch OPPRF. Proceedings on Privacy Enhancing Technology. 2022, 353-372

[3] B. Pinkas, T. Schneider, O. Tkachenko, A. Yanai. Efficent Circuit-Based PSI with Linear Communication. Advances in Cryptology—Eurocrypt 2019. 2019, 122-153

[4] H. Chen, Z. Huang, K. Laine, P. Rindall. Labeled PSI from Fully Homomorphic Encryption with Malicious Security. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

[5] K. Cong, R. Moreno, M. DaGama, W. Dai, I. Iliashenko, K. Laine, M. Rosenberg. Labeled PSI from Homomorphic Encryption with Reduced Computation and Communication. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021.

[6] M. Freedman, Y. Ishai, B. Pinkas, O. Reingold. Keyword Search and Oblivious Pseudorandom Functions. Theory of Cryptography. 2005, 303-324.

[7] A. Shamir. How to Share a Secret. Communications of the ACM. 1979. Vol 22, Issue 11, 612-613.

[8] G. Blakley, G. Kabatianskii. Linear Algebra Approach to Secret Sharing Schemes. Selected Papers from the Workshop on Information Protection, Error Control, Cryptology, and Speech Compression. 1993, 33-40.

[9] A. Yao. How to Generate and Exchange Secrets. 27th Annual Symposium on Foundations of Computer Science. 1986, 162-167.

Towards a Taxonomy for Business-to-Government Data Sharing

Keywords: Business-to-Government, B2G, data sharing, taxonomy

INTRODUCTION

Business-to-Government (B2G) data sharing is represented by a collaboration between a private company or organisation and the public sector (conceived at different levels: local, regional, national or supra-national), where the former makes available its data to the latter[70]. The aim of this sharing of data should be a public interest purpose, like, for example, protection of the environment or the response to a public emergency.

The last few years have seen this phenomenon growing, as research has shown how privately held data could have a huge potential when used to tackle societal policy issues. Many examples are available in literature, especially in the official statistics domain, often based on one-off voluntary kind of cooperation between the two actors (private and public sector) (as an example, see [1]).

One of the aims of the European Strategy for Data[71] is the adoption of "legislative measures on data governance, access and reuse". Two legislative Acts have been proposed along this line: the Data Governance Act[72] , already adopted by the EU [73], and the Data Act, still under negotiation. In particular, the Data Act contains specific provisions concerning B2G data sharing in exceptional circumstances, for example in case of a public emergency, or when needed to implement a legal mandate, if data are not otherwise available. In those cases, private companies shall be asked to share data with the public sector in order to allow it to respond quickly and securely, but at the same time minimizing the burden on businesses[74].

B2G data sharing can be employed in different situations: from emergencies (among the ones ruled by the Data Act), to the construction of official statistics and to the use in research, just to name a few. In all these circumstances, the quality level required to the data could be different, as different principles could prevail upon others (e.g., timeliness in the case of emergencies is the key parameter).

This heterogeneity in possible use cases motivates the present work. Indeed, our objective is to understand and classify the different circumstances in which B2G data sharing may happen. In practice, we aim at creating a taxonomy of B2G data sharing, in which we identify all the different situations where B2G could occur, and afterwards we add as attributes some identified quality principles that characterise the different data sharing situations. The work aims at

[70] https://digital-strategy.ec.europa.eu/en/faqs/business-government-data-sharing-questions-and-answers

[71] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

[72] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868

[73] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R0868

[74] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113

providing further information that can help clarify specificities and requirements of B2G data sharing in order to ultimately enable and make more dynamic relevant data flows.

METHODS

We first need to identify all the situations when B2G can occur, and, as a second step, define the quality attributes that are needed in each situation with a view to cluster different instances of B2G into subgroups. In the 'Results' section we will then put together all these identified pieces to build the taxonomy.

Different situations of B2G

In order to identify the different circumstances when B2G occurs, we decided to start from the European taxonomy of public services [2], that has been published in 2019 with the aim of helping public administrations in harmonising their catalogue of services.

The EU taxonomy of public services is built as the combination of two elements: themes and patterns. Examples of themes are defence, education, environmental, while examples of patterns are financing, control & monitoring, taxation. For our work, we decided to consider all the themes and only a few patterns, with some adjustments to make them fit for our purpose. We will start the analysis with patterns, as it needs a bit of reflection and explanation.

Patterns

Patterns are public service types, or the core services of governments. When public services are broken down to their core by removing any context, the detail of a core service remains. The overarching concept of all these details is defined as a pattern.

The EU taxonomy of public services is conceived with the following nine patterns: framework, information, registration, certification, financing, production, feedback, control & monitoring, taxation, and each of them is proposed with a list of sub patterns [2] that can help in understanding better the various cases.

Starting from our own considerations and reflections on B2G data sharing, we imagined the following four patterns as fit for purpose: control & monitoring, official statistics, planning & management, research.

Only the first one is explicitly a pattern from the EU taxonomy of public services, while for the other three we needed to check the sub patterns in order to find some links. Indeed, we think that official statistics could be included into the pattern 'Information', and planning & management is part of the 'Framework' one. Research is the only pattern that we added, as a match in the existing taxonomy could not be found. The final list of patterns of our taxonomy is the following: *control & monitoring*, *information*, *framework*, *research*.

Themes

For the themes, we decided to consider them in their totality as they appear in the EU taxonomy of public services, as they fully cover all thematic situations in which B2G data sharing could be applied (even if some cases could appear more theoretical than others – see for example 'Religious'). The list is composed by 31 themes (see Table 1).

Table 1. List of themes from the EU taxonomy of public services

Themes		
Agriculture & food	General government	Public space management & heritage
Animal	Health care	Religious
Border control	Housing & building	Retail
Culture, sport & leisure	Legal	Stock market
Defence	Life event & identity	Tourism & travelling
Digital	Manufacturing	Transportation & Transportation infrastructure
Education	Media	Utilities
Emergency	Monetary policy	Voluntary organisation & charity
Environmental	Money and debt	Welfare & social care
Family	Natural resources	Work
General business		

Source: [2]

Attributes

Beyond the taxonomy of B2G data sharing, we thought that it could be useful to identify all the characteristics that such data sharing initiatives would have to comply with in the different B2G situations. We identified the following three broad dimensions to be covered: *spatial and temporal*, *methodological*, *legal and governance*.

In literature there are some existing sets of principles (not specifically targeted to B2G data sharing) that cover some of these dimensions, but not all of them. As an example, consider the UN quality principles of official statistics [3], that are composed by six elements: relevance, accuracy, timeliness, accessibility, interpretability and coherence. Similar approaches have been followed by other statistical organisations (UNECE and OECD, among others), adding a few components (like reliability and clarity) to the UN's set of principles (for a detailed review, see [4]). Other sets of principles that could be used are the FAIR [5] and / or the CARE [6] ones, but as already anticipated, they do not cover the whole set of aspects that we envisioned.

To our knowledge, the only example that was developed with the specific aim of identifying principles for B2G data sharing is contained in [7]. It is the result of a working group named "Facilitating the use of new data sources for official statistics" set up by Eurostat, the statistical office of the European Union.

We decided then to adapt some of the existing principles to our own needs, adding the ones that we think were missing. Following the three dimensions previously stated, the list of principles that we identified is available in Table 2.

Table 2. Attributes

Dimensions	Principles

Spatial and Temporal	Timeliness and punctuality
	Time series / temporal coverage
	Coherence and Consistency (comparability) over time, space and provider
	Spatial and statistical unit (e.g., population) coverage
Methodological	Accuracy (representativity, reliability, veracity)
	Continuity / Sustainability over time
	Transparency
	Interpretability/clarity
Legal and Governance	Partnership scheme (marginal cost, donation, preferential treatment as well as type of access)
	Purpose limitation, data products distribution, data reuse and IP/rights over the data
	Accountability of data provider
	Governance
	Stakeholders
	Relevant data sources (broad typology, e.g., mobility data, consumer prices and behaviours, energy consumption etc.)

Source: own elaboration

RESULTS

After the showcase of the elements that will constitute our taxonomy of B2G data sharing, it is time to put them all together. Our taxonomy is composed by a combination of the themes and identified patterns, and the single user can identify him/herself in a specific B2G data sharing situation, combining one theme with one pattern (for example, “Control & monitoring” as a pattern, and “Health care” as a theme). In this way, the taxonomy is composed by 124 different situations, that is the result of the four patterns multiplied by 31 themes.

Each single B2G data sharing situation can then be characterised by the list of 14 attributes that we identified, with the aim of defining different levels of importance of the attributes themselves. As an example, in emergency situations (pattern: “Control & monitoring”, theme: “Emergency”) “Timeliness” would constitute one of the most important characteristics of the data, while “Comparability over time” could be considered marginal, as well as “Sustainability over time”. These last two attributes, by contrast, could result in being fundamental in other situations (like in all the ones that involve “Information” – read it as official statistics). Of course, there could be some common approaches to different combinations of patterns and themes, that would allow to group together some of the B2G data sharing identified situations.

CONCLUSIONS

This paper presents a first proposal of a taxonomy for B2G data sharing, therefore is expected to be further consolidated and integrated with new insights, especially in the “Attributes” part. Moreover, after the structure of the taxonomy will be consolidated, the following step of the work will be to identify the importance of each attribute in the different B2G data sharing situations, in order to group together cases where situations are different but the quality principles required are the same. This work aims at providing insights into the differences of B2G settings, commonly grouped as a single instance of data flows whereas they exhibit a broad and relatively diverse range of goals, contexts and therefore requirements.

rEFERENCES

[1] UN Global Working Group on Big Data for Official Statistics, *Handbook on the Use of Mobile Phone Data for Official Statistics*. 2019. [Online]. Available: https://unstats.un.org/bigdata/task-teams/mobilephone/MPD%20Handbook%2020191004.pdf

[2] DIGIT Directorate-General for Informatics and ISA2 Programme, 'European taxonomy for public services', European Commission, 2019. [Online]. Available: https://joinup.ec.europa.eu/sites/default/files/custom-page/attachment/2020-12/ISA2_European%20taxonomy%20for%20public%20services%20%281%29.pdf

[3] G. Brackstone, 'Managing Data Quality in a Statistical Agency', *Surv. Methodol.*, vol. 25, no. 2, pp. 139–149, 1999.

[4] S. Vale, 'Statistical Data Quality in the UNECE', 2010.

[5] M. D. Wilkinson *et al.*, 'The FAIR Guiding Principles for scientific data management and stewardship', *Sci. Data*, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18.

[6] S. R. Carroll *et al.*, 'The CARE Principles for Indigenous Data Governance', *Data Sci. J.*, vol. 19, p. 43, Nov. 2020, doi: 10.5334/dsj-2020-043.

[7] European Commission. Statistical Office of the European Union., *Empowering society by reusing privately-held data for official statistics: a European approach: final report prepared by the high level expert group on facilitating the use of new data sources for official statistics, 2022 edition.* LU: Publications Office, 2022. Accessed: Oct. 12, 2022. [Online]. Available: https://data.europa.eu/doi/10.2785/948477

Agriculture Statistics (MANS3M.1)

Session Chair: **Jose Domingo Martinez Solano** *(Eurostat)*

Strategy to Modernise Agricultural Statistics: New Pathways for the Future
Nicolas Lampach *(Eurostat)*

Strengthen the statistical capacities in the developing country to close the agricultural data gap
Neli Georgieva Mihaylova *(FAO)*

The Modern 2020 Agricultural Census in Spain: a Massive Use of Registers and Alternative Data Sources
Antonio Martínez Serrano (*INE Spain*)

Strategy to Modernise Agricultural Statistics: New Pathways for the Future

Keywords: Automatisation, Agricultural statistics, Content analysis, Dissemination, Data infrastructure, EU law, Modernisation

INTRODUCTION

Agricultural Statistics Strategy 2020 aims on producing data on agriculture that meets the current and future user needs [1]. Recent transformation of regulatory environment by reforms of Common Agricultural Policy (CAP), considerable shift in farm structures and the turmoil of agricultural markets carry over further requirements for agricultural statistics. Alongside these changes in user needs, new data sources (e.g administrative data, earth observation, modelling) and big data have become readily available to reduce administrative burden and to enhance data collection methods [2]. Although, new data sources and flexible ways of data collection seem promising to fill data gaps, it requires the design of new data processing systems integrating automatic cross-domain checks and improved timeliness.

We present an innovative strategy to modernize agricultural statistics pursuing to meet emerging user needs, enhance the quality, accuracy and timelines of data and improve the coherence between statistical domains. Furthermore, we propose a modern approach to disseminate the enormous bulk of data by reviewing the structure of data along with the design of more advanced visualization tools. While some attempts have been made to assess the collaboration between statistical and agricultural domains [3], few scholars describe the underlying structural changes of European statistical apparatus and the necessary adaptation of legal framework to cater future needs [2]. This work contributes to the agricultural statistics literature by presenting the modernisation pathway of the European Statistical System (ESS).

METHODS

The modernisation Strategy for Agricultural Statistics 2020 and beyond builds up from the legislative framework and methodologies until the dissemination, covering all phases of the statistical production business process. The introduction of two new legal frameworks, Integrated Farm Statistics (IFS) and Statistics on Agricultural Input/Output (SAIO); and the amending of an existing one, Economic Accounts for Agriculture (EAA); aimed to allow the use of new data sources and innovative approaches and the integration of the different statistical domains.

Data collection

Agricultural statistics are collected from new data sources such as administrative registers, earth observation, and modelling or farm management software. Consequently, this results in a substantial reduction of the statistical burden to farmers, but also in lower costs to the statistical providers and higher quality of the data.

Data processing

Eurostat has introduced an automated data processing system which pre-validates the transmitted dataset by performing all validation checks between the same dataset, that is, all checks in levels 0 and 1 and part of level 2 checks[75] following the ESS validation levels [4]. This reduces the time between transmission from national data providers and dissemination to the public but also results in a more efficient data processing flow, where potential basic errors are systematically detected (mostly) without human intervention and timelier feedback is provided to the data sender.

Data analysis

Agricultural statistics are integrated into more coherent and interoperable IT systems improving the efficiency of data flows and the quality of statistics (timeliness, internal coherence). Easier cross-domain checks would be possible resulting in higher coherence between the different agricultural domains.

Dissemination

In light of the challenge to produce a considerable number of dissemination tables at the end of every survey year and to fulfil user needs, we identify the most prevalent tables based on web monitoring reports and apply content analysis methods to explore the strength of co-relationship between tables. This allows us not only to reduce the sheer number of tables, but also to restructure and design new tables accommodating emerging user needs.

RESULTS

User Statistics

To identify the user’s need and interest in farm statistics, we retrieved user statistics of all dissemination tables related to farm structure survey[76] from web monitoring reports between January 2020 and February 2022 published on monthly basis at an internal website of Eurostat. Displayed in Figure 1 are the monthly variations of user clicks on Eurofarm tables across distinct groups. While external users, such as national administrations, researchers, NGOs, general public, tend to be mostly affected by COVID-SAS-19 pandemic showing a steep drop in April 2020, the European Commission internal users follow a more consistent pattern with peaks in most busy months (January, June, and September). It can also be seen that external users represent the largest share in the demand for European farm statistics.

[75] While level 0 and 1 refer to structural and content validations, level 2 comprises the validation of the consistency with other datasets within the same domain and data source.

[76] Currently, there are 213 tables related to farm structure survey published at Eurostat website.

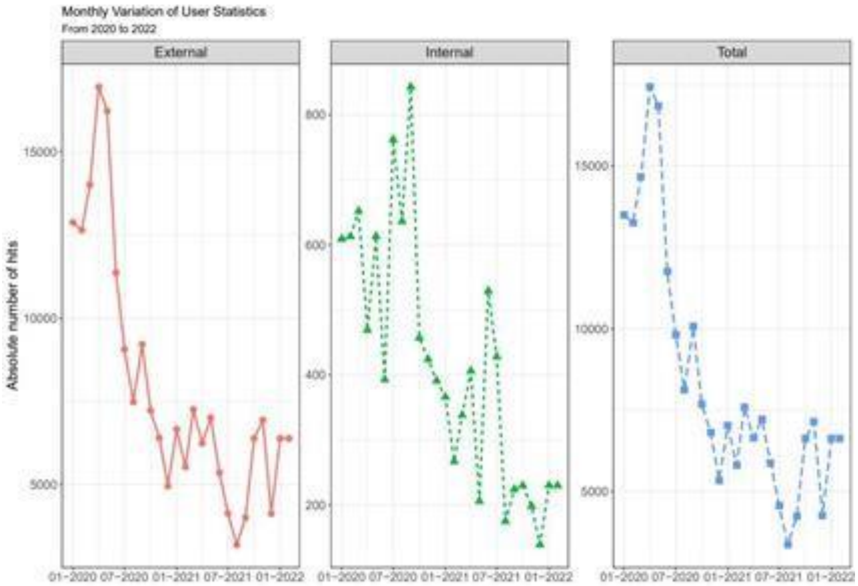


Figure 1. Monthly records of user statistics for distinct user group

Ranking the ten most salient tables from farms structure domain for each user group reveals that users prefer mostly generic tables summarizing general information on agricultural area, type of farm, standard output, legal form, crops and animals and tend to favour less so specialized tables with detailed information on a specific topic (see Figure 2). Preferences between external and internal users seem not to diverge considerably in this respect.

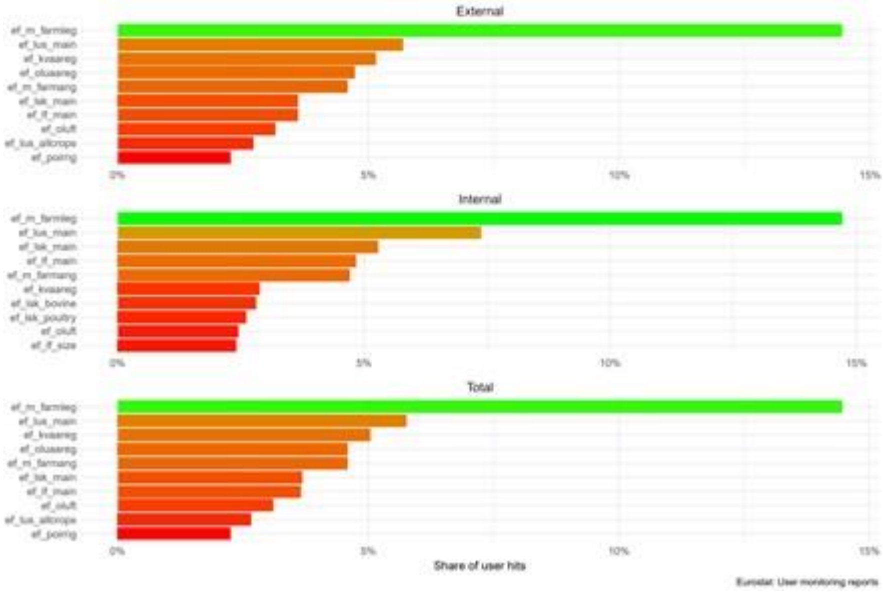


Figure 2. Most salient Eurofarm tables across distinct user groups, aggregated

Content analysis

We design a table-dimension matrix of 30 most salient tables to determine the degree of similarity between tables based on matching coefficient of [5] calculating distance matrices for binary data for each possible pair of combination.

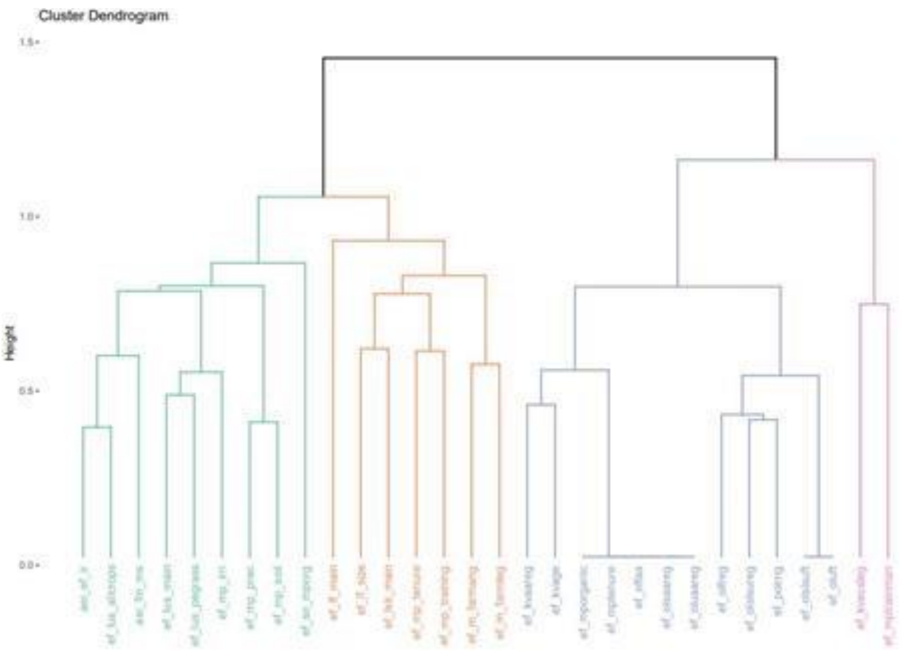


Figure 3. Most salient Eurofarm tables across distinct user groups, aggregated

Depicted in Figure 3 is the divisive hierarchical clustering also known as DIANA (Divisive Analysis) applying the inverse of agglomerative clustering. It reveals that the content of the tables can be divided into four main groups: "Crops and Management Practices" (greenish), "Main Structural Indicators" (orangish), "Agricultural indicators" (bluish), (purplish) "Other". While the "Agricultural indicators" accounts for the largest number of tables across the groups and it is formed by two subgroups, the group of "Crops and Management Practices" seems to more heterogeneous in terms of subgroups such as organic farming, agricultural practices, land use and crop production.

cONCLUSIONS

We elaborate further illustrations of the data collection, processing, analysis and dissemination of the modernisation strategy. Furthermore, we strive to provide comprehensive guidelines and hands-on examples on how to improve statistical systems and enhance the overall quality, timeliness and accuracy of data.

rEFERENCES

[1] European Commission, 2019. Strategy for European statistics and beyond 2020. Retrieved on 07.10.2020 at: https://ec.europa.eu/eurostat/documents/749240/749310/Strategy+on+agricultural+statistics+Final+version+for+p ublication.pdf/9c7787ca-0e00-f676-7a64-7f56e74ec813

[2] Selenius, J., Wirtz, C., Florescu, D. and Lazar, A.C., 2021. Agricultural census 2020–how to reduce costs and burden? The European statistical system approach. Statistical Journal of the IAOS, 37(1), 327-332.

[3] Tóth, K, 2018. Georeferenced agricultural data for statistical reuse. Geosciences 2018, 8, 188, 1-19.

[4] European Commission, Collaboration in Research and Methodology for Official Statistics. Validation in the ESS. Retrieved on 14.10.2022 at: https://ec.europa.eu/eurostat/cros/content/validation-levels_en

[5] Sokal, R.R., & Michener, C.D. 1958. A statistical method for evaluating systematic relationships. University of Kansas science bulletin, 38, 1409-1438.

Strengthen the statistical capacities in the developing country to close the agricultural data gap

Session title: *Innovation in agricultural statistics: From new data sources to integrated and automated data processing systems*

Keywords: developing counties, capacities, agricultural statistics

Introduction

The concept of “Agriculture” is changing and covers wider set of activities, thus the agricultural statistics should follow this trend to meet the emerging needs for statistical information. Agricultural statistics “describe agricultural land use, production of crop and animal products, farm structures, prices, economic inputs and outputs and the impact of agriculture on the environment, health and wellbeing.” [1,p.2]. This enlarged scope call for more integrated approach, complex data bases, new technlogies and inevitably improved statistical skills.

In the same direction, all United Nations Member States in 2015 adopt the 2030 Agenda for sustainable development, setting an ambitious objective for transforming the world [2] to end poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development. The 2030 Agenda “provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership” [3].

At the same time, critical gaps on data production persist in many countries. Some countries have yet to successfully leverage the technical and institutional innovations available for the industrialization of statistical production. Indeed, the majority of IDA countries have not conducted agricultural censuses or annual surveys for the past 15 years. Summarizing user needs, most countries still lack the databases required to run the economic models that provide the essential inputs into policy-making, market management and research support. Generally, the available data are more than five years old, a factor that weakens any analysis. Coherence among data series is often poor, due to the lack of integration of data sources. [4] These problems require urgent measure for improving the statistical knowledge, technological platforms and data collection methodologies in developing countries.
In response to these challenges and seizing the opportunities set by the International community, FAO , together with other development partners, prepared programmes of activities covering various aspects, towards the objective of closing the agricultural data gaps in developing countries.

Methods

Global Strategy to Improve Agricultural and Rural Statistics – continuous efforts to strengthen the countries capacity in agricultural statistics

The Global Strategy to improve Agricultural and Rural Statistics (GSARS) was developed in 2009 as a blueprint for a coordinated and long-term initiative to address the decline in the agricultural statistical systems of many developing countries [5] and endorsed by the United Nations Statistical Commission (UNSC) in 2010. It is designed as a long-term process to provide a framework for national and international statistical systems that would enable developing countries to produce the necessary data in the 21st century. The implementation of Phase 1 of GSARS (2012-2018) significantly impacted the agricultural statistical systems of many developing countries and demonstrated its ability to respond to the needs of the evolving international and regional agendas.

The second phase of the Global Strategy (GSARSII) builds upon the successful achievements of and lessons learned from Phase 1. Although solid foundations have been laid and momentum created at international, regional and national level for significantly improving agricultural statistics during phase 1, **the investment made in creating the methodologies, starting technical assistance, and initiating a generation of new agricultural statisticians must be transformed into more concrete capacity and increased data production and dissemination activities at country level.** [6]

Development of methodology for Integrated surveys in the agricultural sector

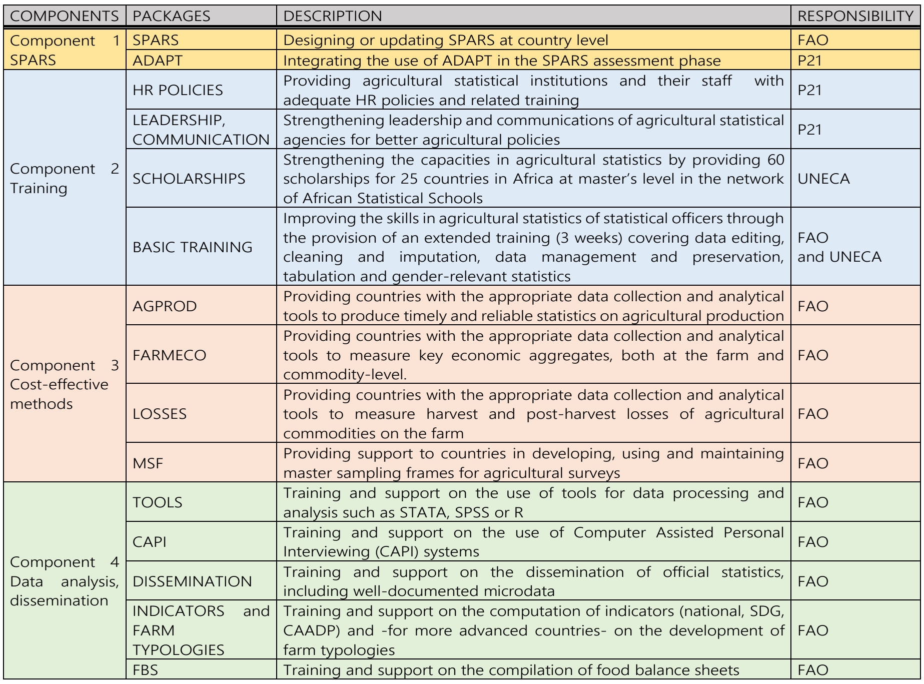
The Agricultural Integrated Survey (AGRIS) methodology was developed as part of the GSARS I research component, providing a methodology for producing basic agricultural data for responding to national policy demand and the monitoring and reporting of several Sustainable Development Goal (SDG) indicators.

As one of the main features of cost-effective methods, AGRIS is designed to help national agencies accelerate the production of quality disaggregated data on the technical, economic, environmental and social dimensions of agricultural holdings, in line with the Eurostat's Strategy for agricultural statistics for 2020 and beyond [1, p.13]. AGRIS builds on the previous work of the Global Strategy to present a unique opportunity to channel these methodological innovations and achieve real impacts on data systems on the ground.

Results

GSARS II project (2021- 2023) operational in 25 countries in Africa

The main overall technical focus for Phase 2 is based on several main principles: (i) driving country use of existing, innovative tools developed during Phase 1; (ii) training on better use of data; addressing gaps in terms of skills and knowledge required to process data and informing and sensitizing policy-makers on how to read, interpret and use statistics; (iii) promoting innovative capacity development strategies based on experience gained from Phase 1 and integrating new approaches to capacity development; and (iv) improving advocacy, communication and dissemination, to build greater awareness of the activities and impact.



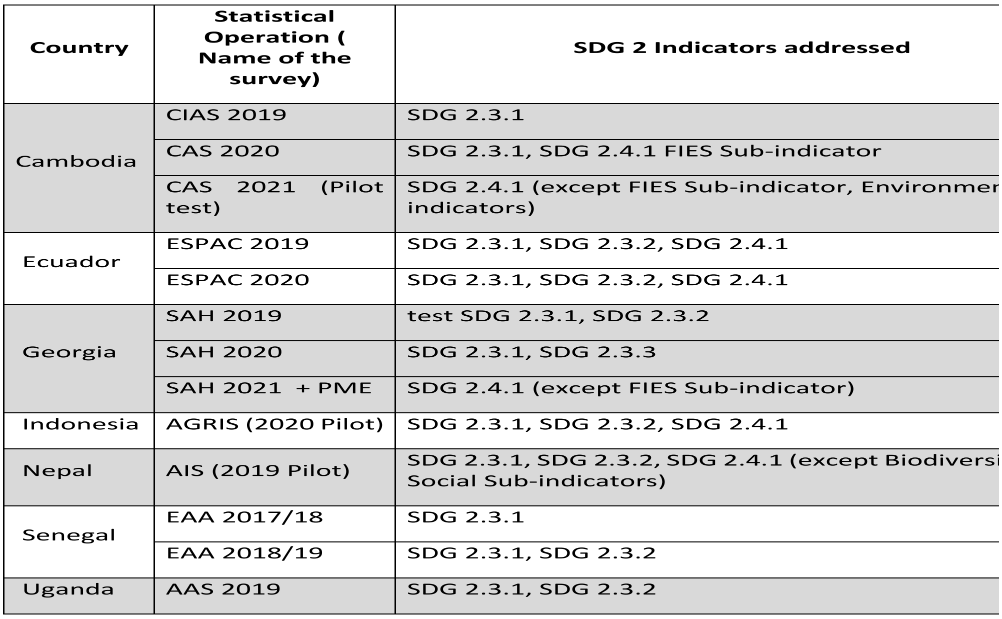
COMPONENTS	PACKAGES	DESCRIPTION	RESPONSIBILITY
Component 1 SPARS	SPARS	Designing or updating SPARS at country level	FAO
	ADAPT	Integrating the use of ADAPT in the SPARS assessment phase	P21
Component 2 Training	HR POLICIES	Providing agricultural statistical institutions and their staff with adequate HR policies and related training	P21
	LEADERSHIP, COMMUNICATION	Strengthening leadership and communications of agricultural statistical agencies for better agricultural policies	P21
	SCHOLARSHIPS	Strengthening the capacities in agricultural statistics by providing 60 scholarships for 25 countries in Africa at master's level in the network of African Statistical Schools	UNECA
	BASIC TRAINING	Improving the skills in agricultural statistics of statistical officers through the provision of an extended training (3 weeks) covering data editing, cleaning and imputation, data management and preservation, tabulation and gender-relevant statistics	FAO and UNECA
Component 3 Cost-effective methods	AGPROD	Providing countries with the appropriate data collection and analytical tools to produce timely and reliable statistics on agricultural production	FAO
	FARMECO	Providing countries with the appropriate data collection and analytical tools to measure key economic aggregates, both at the farm and commodity-level.	FAO
	LOSSES	Providing countries with the appropriate data collection and analytical tools to measure harvest and post-harvest losses of agricultural commodities on the farm	FAO
	MSF	Providing support to countries in developing, using and maintaining master sampling frames for agricultural surveys	FAO
Component 4 Data analysis, dissemination	TOOLS	Training and support on the use of tools for data processing and analysis such as STATA, SPSS or R	FAO
	CAPI	Training and support on the use of Computer Assisted Personal Interviewing (CAPI) systems	FAO
	DISSEMINATION	Training and support on the dissemination of official statistics, including well-documented microdata	FAO
	INDICATORS and FARM TYPOLOGIES	Training and support on the computation of indicators (national, SDG, CAADP) and -for more advanced countries- on the development of farm typologies	FAO
	FBS	Training and support on the compilation of food balance sheets	FAO

…

Figure 27. GSARS II capacity development areas

Technical assistance projects to roll out the integrated survey methodologies

Since 2017 FAO has started to assist several developing countries in the implementation of the Agricultural Integrated survey (AGRIS), to improve the data availability for national agricultural policies and to produce several Goal 2-related SDGs.



Country	Statistical Operation (Name of the survey)	SDG 2 Indicators addressed
Cambodia	CIAS 2019	SDG 2.3.1
	CAS 2020	SDG 2.3.1, SDG 2.4.1 FIES Sub-indicator
	CAS 2021 (Pilot test)	SDG 2.4.1 (except FIES Sub-indicator, Environmer indicators)
Ecuador	ESPAC 2019	SDG 2.3.1, SDG 2.3.2, SDG 2.4.1
	ESPAC 2020	SDG 2.3.1, SDG 2.3.2, SDG 2.4.1
Georgia	SAH 2019	test SDG 2.3.1, SDG 2.3.2
	SAH 2020	SDG 2.3.1, SDG 2.3.3
	SAH 2021 + PME	SDG 2.4.1 (except FIES Sub-indicator)
Indonesia	AGRIS (2020 Pilot)	SDG 2.3.1, SDG 2.3.2, SDG 2.4.1
Nepal	AIS (2019 Pilot)	SDG 2.3.1, SDG 2.3.2, SDG 2.4.1 (except Biodiversi Social Sub-indicators)
Senegal	EAA 2017/18	SDG 2.3.1
	EAA 2018/19	SDG 2.3.1, SDG 2.3.2
Uganda	AAS 2019	SDG 2.3.1, SDG 2.3.2

Table 2. Countries covered by AGRIS projects, SDG2 indicators produced, Source: [7]

50 x 2030 Initiative

An important scale up of the methodological development on the Integrated surveys is the 50 x 2030 Initiative to Close the Agricultural Data Gap, with the objective to empower partner countries' statistical systems for evidence-based decision-making, especially to achieve Sustainable Development Goal (SDG) 2. It is being implemented through a unique partnership between the World Bank, Food and Agriculture Organization of the United Nations (FAO) and the International Fund for Agricultural Development (IFAD).[8]

Conclusions

The long-term capacity development support being provided by FAO and its partners to developing countries allow their national statistical system to produce better quality agricultural statistics, draft evidence-based agricultural and rural development programme and investment plans, align to the international agenda objectives. Continuous efforts in this direction are needed now more than ever to cope with the new challenges in terms of food security and quality, climate change, environmental protection and human and animal wellbeing. FAO, together with other international development partners is committed to continue to provide this support and to strengthen the statistical expertise and human resources capacity in its member countries in need.

References

[43] Eurostat. (2015). Strategy for agricultural statistics for 2020 and beyond. Available at: https://ec.europa.eu/eurostat/web/agriculture/methodology/strategy-beyond-2020.

[44] UN, Transforming our world: the 2030 Agenda for Sustainable Development , A/RES/70/1 , 21 October 2015, available at https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/89/PDF/N1529189.pdf?OpenElement

[45] UN, Department of Economic and Social Affairs, Sustainable development, https://sdgs.un.org/goals .https://sdgs.un.org/2030agenda

[46]Global Strategy to improve agricultural and rural statistics. AGRIS Handbook on the Agricultural Integrated Survey. 2017. Available at: http://www.fao.org/in-action/agrisurvey/resources/resource-detail/en/c/1198081/

[47]World Bank, FAO & UN. 2011. Global Strategy to improve Agricultural and Rural Statistics. Report 56719-GLB. World Bank Publication: Washington, D.C. http://www.fao.org/3/am082e/am082e00.pdf

[48]GSARS II, Second Global Action plan, 2018, https://www.fao.org/in-action/global-strategy-agricultural-statistics/en

[49]Bolliger, Flavio *et all*, Statistical-journal-of-the-iaos/sji210913, Mar 21 2022 , The 50x2030 Initiative and production of SDG 2 indicators: Country challenges and experiences, available at https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210913?resultNumber=2&totalResults=15&start=0&q=50x2030&resultsPageSize=10&rows=10

[50]The 50x2030 Initiative. Available from: https://www.50x2030.org/about.

Communication and dissemination (GASP3M.2)

Session Chair: **Jean Pierre Poncelet** *(Eurostat)*

The release package of the Eurostat regional yearbook 2022 - Something for everyone?
Bianka Fohgrub *(European Commission)*

ESA interactive: towards a new generation of statistical standards
Nicola Massarelli *(Eurostat)*

Visualization and communication of statistics
Martin Tennekes *(Statistics Netherlands-CBS),* Marco Puts *(Statistics Netherlands-CBS)*

The release package of the Eurostat regional yearbook 2022 – Something for everyone?

Keywords: data storytelling, interactive visualisations, statistical maps

Introduction

The EU Member states and EFTA countries collect data and compile statistics for national and EU purposes. However, European citizen are increasingly interested in detailed information at regional level as national figures alone cannot reveal the full picture of what is happening within the EU member states.

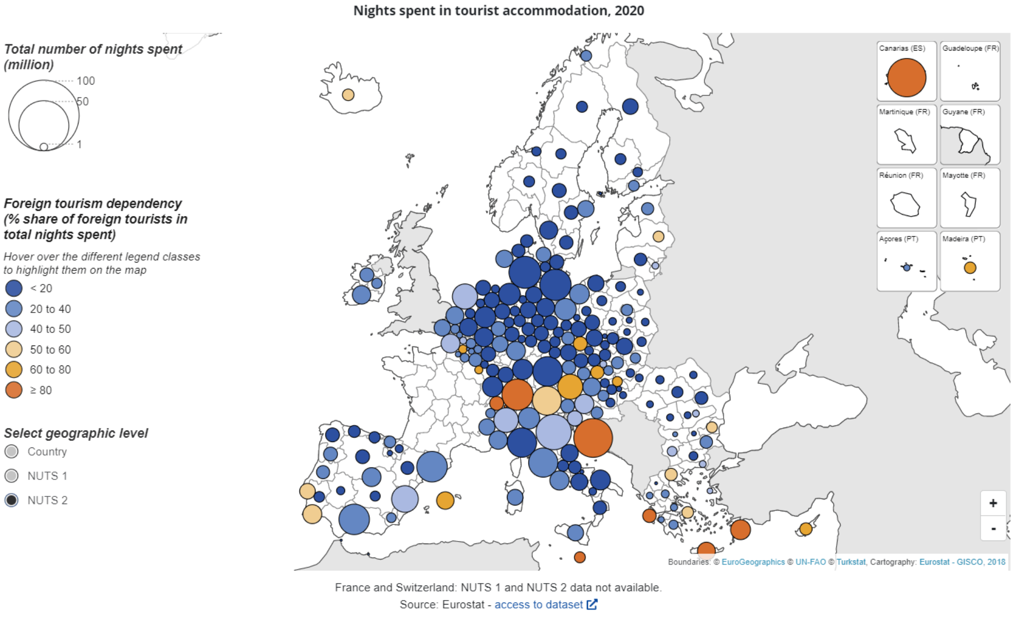
In the last decades, we have seen a transformation from large offset printed-paper publications towards digital first, each with its own advantages and disadvantages. This paper offers insights into how Eurostat addressed this challenge and took the opportunity through the example of the "digital transformation" of the *Eurostat regional yearbook.*

Methods

The *Eurostat regional yearbook* [51] (RYB) is a very popular Eurostat publication with a long-standing tradition. This publication was started in the 1970's and since 2000, has been published on an annual basis. Over the last couple of years, the publication concept has been expanded to reach a wider audience. In 2012, an accompanying geospatial map viewer – the *Statistical Atlas* [52] (SA) has been introduced, followed by the *Statistics Explained articles* [53] (SE) in 2018, concluding with the introduction in 2020 of the *Regions in Europe – interactive edition* [54] (REI). These tools complement each other to address different focus groups and reach a wider and inclusive audience down to the citizen level.

Results

The examples below show different visualisations of the same indicator (here: Nights spent in tourist accommodations, 2020) in various publication formats to illustrate the variety of approaches to communicate regional statistics to the various user groups and communities. Certain elements are ensured at the various dissemination channels (e.g. graphical style guide, link to source datasets), while each of them has its own specificities and focus (e.g. embed visualisation, interactivity, overlay usage with other datasets).



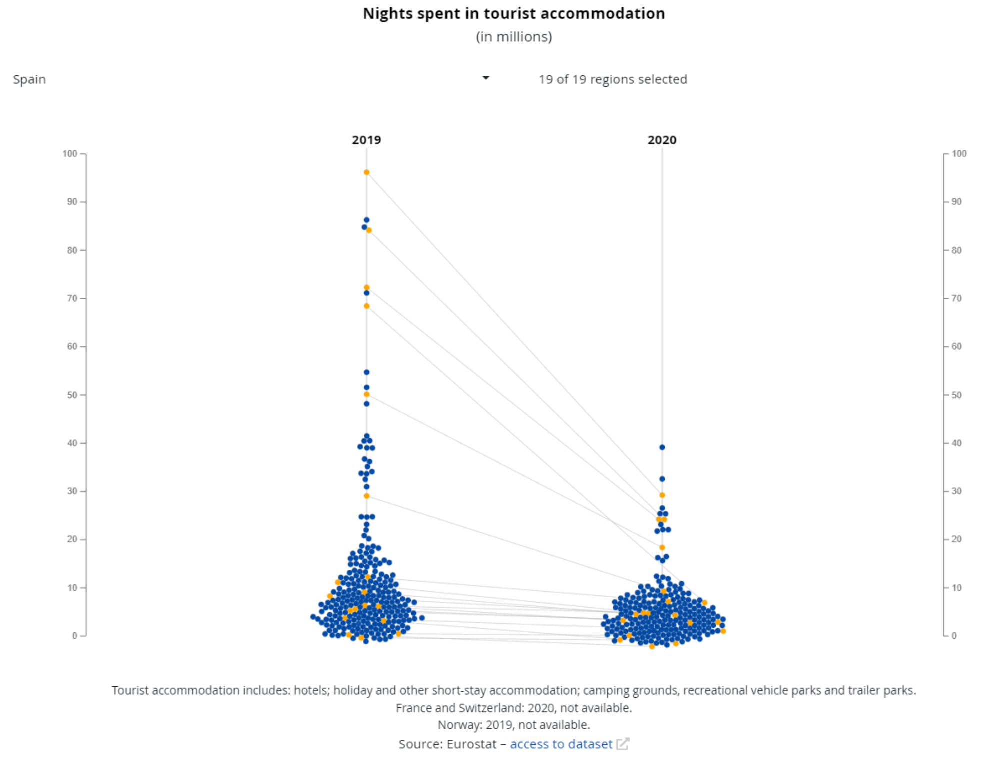
Nights spent in tourist accommodation, 2020

Total number of nights spent (million)
100
50
1
Foreign tourism dependency (% share of foreign tourists in total nights spent)
Hover over the different legend classes to highlight them on the map
● < 20
● 20 to 40
● 40 to 50
● 50 to 60
● 60 to 80
● ≥ 80
Select geographic level
○ Country
○ NUTS 1
● NUTS 2
Canarias (ES)
Guadeloupe (FR)
Martinique (FR)
Guyane (FR)
Réunion (FR)
Mayotte (FR)
Açores (PT)
Madeira (PT)
+
-
Boundaries: © EuroGeographics © UN-FAO © Turkstat, Cartography: Eurostat - GISCO, 2018

France and Switzerland: NUTS 1 and NUTS 2 data not available.

Source: Eurostat - access to dataset

Figure 28. Map example from the Regions in Europe – interactive edition



Nights spent in tourist accommodation

(in millions)

Tourist accommodation includes: hotels; holiday and other short-stay accommodation; camping grounds, recreational vehicle parks and trailer parks.

France and Switzerland: 2020, not available.

Norway: 2019, not available.

Source: Eurostat – access to dataset

Figure 29. Example of a Bee swarm graph from the Regions in Europe – interactive edition

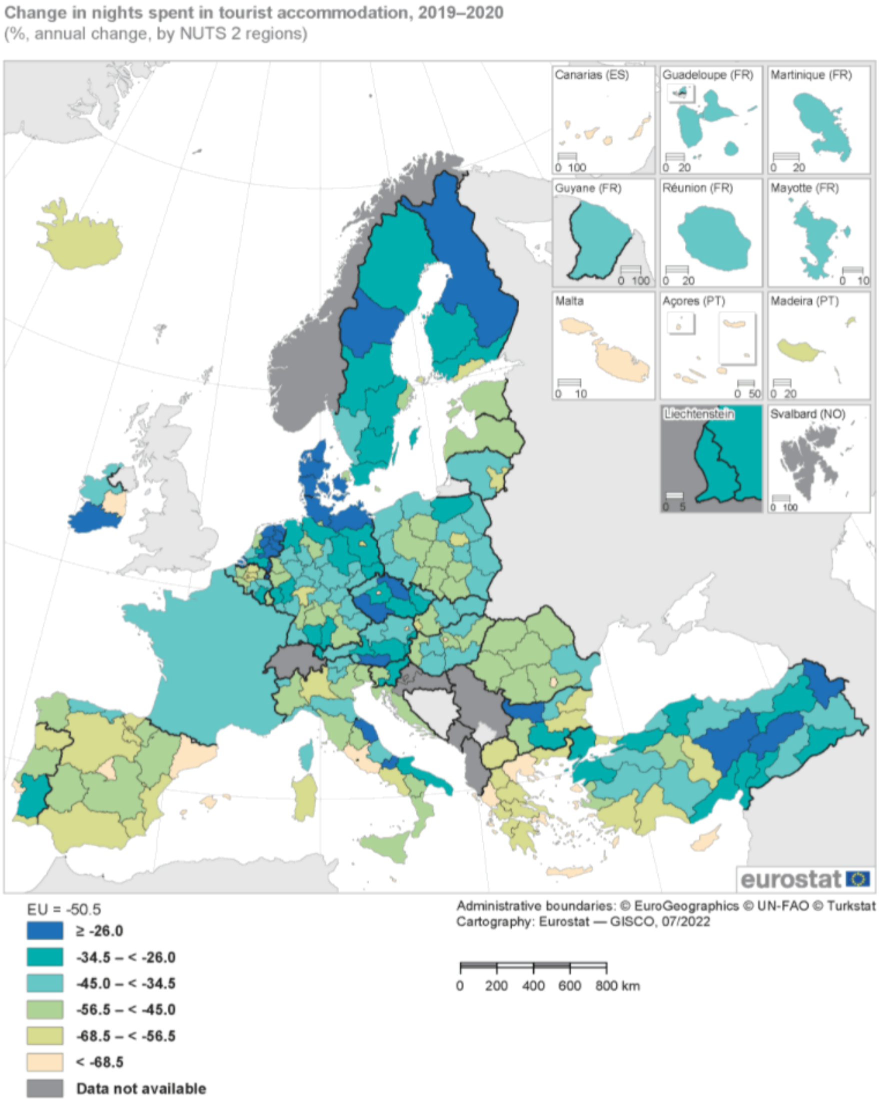


Figure 30. Map example from the Eurostat regional yearbook

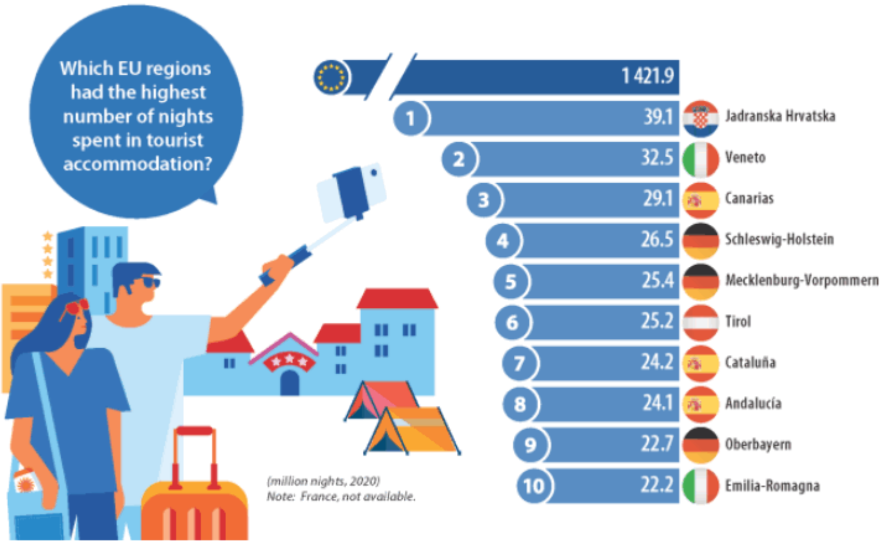


Figure 31. Example of an Infographic from the Statistics explained articles

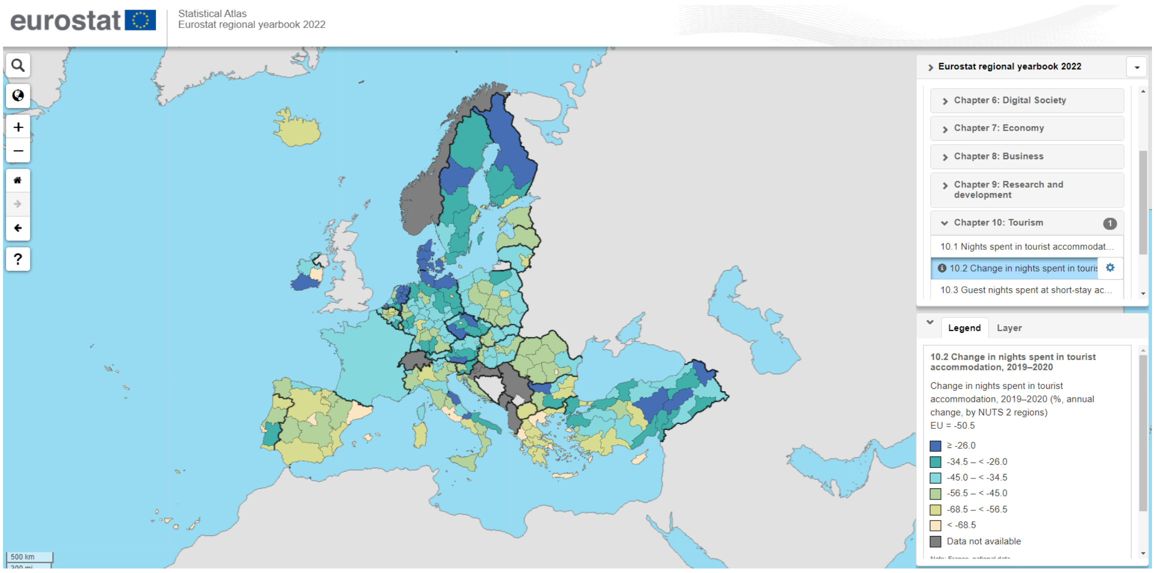

eurostat
Statistical Atlas
Eurostat regional yearbook 2022
?
500 km
200 mi
Eurostat regional yearbook 2022
Chapter 6: Digital Society
Chapter 7: Economy
Chapter 8: Business
Chapter 9: Research and development
Chapter 10: Tourism
1
10.1 Nights spent in tourist accommodat…
10.2 Change in nights spent in touris
10.3 Guest nights spent at short-stay ac…
Legend
Layer
10.2 Change in nights spent in tourist accommodation, 2019–2020
Change in nights spent in tourist accommodation, 2019–2020 (%, annual change, by NUTS 2 regions)
EU = -50.5
≥ -26.0
-34.5 – < -26.0
-45.0 – < -34.5
-56.5 – < -45.0
-68.5 – < -56.5
< -68.5
Data not available

Figure 32. Map example from the Statistical Atlas

Conclusions

We believe that with this range of publication elements, we can meet the constantly evolving user expectations and needs in various forms: people preferring the haptic feeling of a printed publication in their hand can use the RYB in its traditional form while users searching for a digital access to the full content can study the complete publication using the online PDF or SE articles. Users focused on an easy and quick digital access to prepared visualised data can use the REI to interact with regional statistics in a playful and easily understandable way. Map enthusiasts can explore those with the SA interactive map viewer. The key to accomplish these wide varieties of data storytelling are motivated, highly skilled, well-managed, multi-disciplinary team members containing web developers, communication experts, data visualisation experts, graphic designers, GIS experts and statisticians. The support and commitment given by the hierarchy was also essential.

References

[51] https://ec.europa.eu/eurostat/en/web/products-statistical-books/-/ks-ha-22-001
[52] https://europa.eu/!cpdQCv
[53] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Eurostat_regional_yearbook
[54] https://ec.europa.eu/eurostat/cache/digpub/regions/

Visualization and communication of statistics

Keywords: Data visualization, perception, cognition

Introduction

Statistics Netherlands (CBS) is responsible for compiling and publishing reliable statistics of the Netherlands serving as a quantitative backbone of the Dutch society. Policy makers, journalists, and teachers are among the groups that actively use output produced by CBS. As intended by the European Code of Practice [1] CBS puts much effort in producing high quality statistics and developed expertise in assessing the quality of surveys, registers, data editing, estimation and aggregation. The production of high quality statistics is essential, but not sufficient to accomplish the mission of a statistical office: statistics have to be published and communicated to be useful for society.

Current approaches of visualizing and communicating official statistics are often based on convenience, user experience and common journalistic principles and practices. Quality aspects with respect to the understanding of the message of statistical communication are less known, and have not received much attention yet. It is unclear how information that CBS publishes comes across and how it effects the understanding and interpretation of the published statistic. Naïve statistical communication may introduce the risk of visual bias in a chart or prime a reader to draw premature conclusions. Furthermore effectively communicating (in)accuracy of statistics is important: it indicates the value for society.

This paper suggests several research directions that can improve the effectiveness of (visual) communication. We will focus on visualizations, since those are effective means to summarize data, but the research can be extended to textual communication. We plan to conduct in-depth user studies go obtain insights about the effectiveness of our communication methods, and propose alternatives when needed. Furthermore, our aim is to make checklists and guidelines that help visual designers.

The central research question that we propose is: *How can statistical information be effectively visualized so that it is understandable and usable by the intended audience?* Although it may seem a trivial question, surprisingly little research has been done to tackle it. Research has been done in several academic disciplines, such as information visualization, statistics, computer science (information theory), and psychology, but a holistic, interdisciplinary, approach is missing. We propose a model for data visualization that can be used as a guide to develop relevant research questions as well as user experiments to meet the needs of official statistics.

A data visualization model

Rensink [2] views visualization as a process in which graphical representation of data is converted via a visual percept to a conceptual representation. Chen and Golan [3] have an information theoretic perspective on data visualization, where the amount of information (Shannon entropy) is purposely reduced in order improve the efficiency of the visualization. Questionnaire design is similar to data visualization in many respects as the way a respondent perceives the response task together with the visual design of a survey question and his answer options. Tourangeau [4,5,6] breaks this cognitive question-answer process down to four stages: comprehension, retrieval, judgement and response. In Figure 1, a model of data visualization is proposed where these perspectives are united.

Data visualization has emerged as part of statistics and computer science [7], and less as part of visual perception and cognition. Therefore, it is common that large organizations only recruit data visualization designers who have a background in statistics, computer science, of graphic arts. However, little is known how users perceive and interpret visualizations. Therefore, extra knowledge and research is needed from visual perception and cognition.

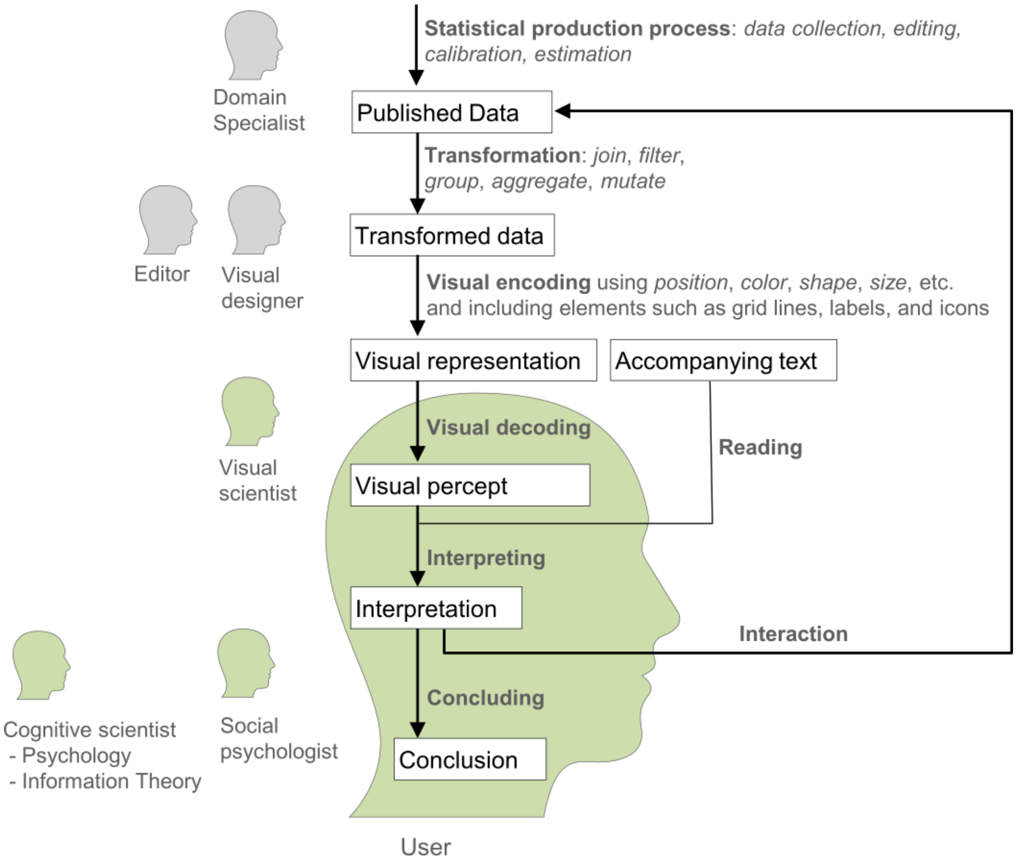


Figure 33. A data visualization model. The process is depicted on the right hand side. The required roles for designing data visualizations are shown on the left hand side.

Quality Guidelines

The research that we propose should gradually and eventually result in guidelines that can be used in practice to improve overall quality of official statistics. Table 1 illustrates the relationship between quality dimensions for Official Statistics [8] on the one hand (rows) and communication (text publications) and visualizations on the other hand (columns).

Table 1. Relationship between quality indicators and communication / visualization

	communication	**visualization**
relevance		
accuracy	perception biases based on opinions and culture, leading to ambiguities.	non-linear behavior of the visual system (see next paragraph)
timeliness		
accessibility	e.g. readability of the text	easiness vs. attractiveness

interpretability	e.g. definitions	a good visualization should guide the user in interpreting the statistics
coherence		tool to integrate different statistics and gives the ability to compare.

Relevance refers to the degree to which the statistical output is able to meet the real needs of the client. When it comes to accuracy, it refers to how well the phenomena are described. This dimension includes different types of errors. When it comes to timeliness, it's about how long it takes for a phenomenon to reach its final publication. Accessibility refers to the ease of obtaining information from the NSO. Interpretability refers to the availability of metadata and additional information to assist in the interpretation of the data. Lastly, coherence describes the extent to which the statistics are able to be integrated consistently.

Visual decoding

Research should be performed that focusses on bringing the theoretical knowledge on visual decoding into practice, and thereby providing information on how reliability and validity of visualization can be improved. As an example, Cleveland and McGill [9] studied the effect of using different visual variables on the accuracy of the perceived data. The findings of this empirical study, which are verified [10] are summarized in Figure 7.

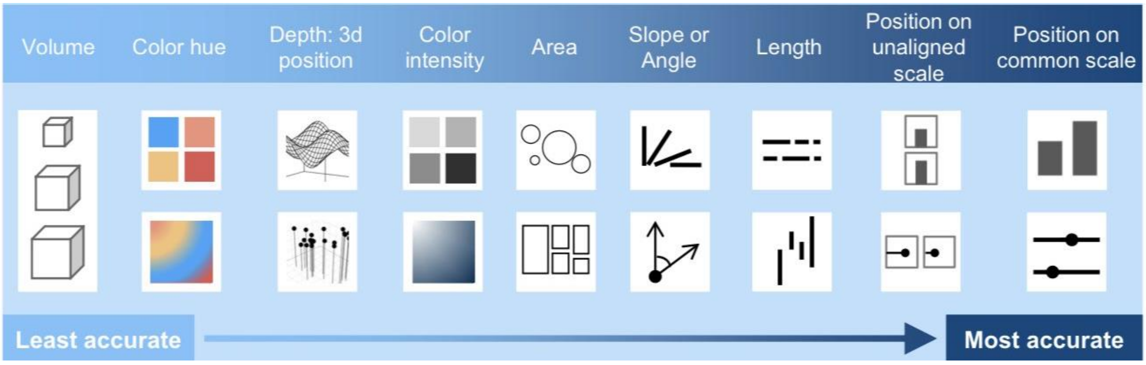


Figure 2. Accuracy of visual variables according to Cleveland and McGill (1984)

It is important to realize is that the human visual system evolved over millions of years in a natural environment in which 3D objects were present. Consequently, Gregory [11] and Marr [12] suggest that every image, including a 2D picture, appears as a certain view in a 3D environment. Visual elements, such as angles orientation, and colors are used by the human visual system as depth cues, leading to the creation of visual illusions. The presence of these factors could introduce biases into the perception of charts, even the simplest of them.

Another source of bias is shown in Figure 3. A five color blue scale is used in this choropleth, but in some cases is it hard to compare non-neighboring regions, for instance the two regions marked in red. Although the two marked areas have exactly the same value, they are perceived differently due to contrast effects. This is due to the fact that the perception of colour is often relative to its surroundings.

Interpretation

Visualizations could definitely be improved so that they truly ease the understanding and adequate use of statistics. But visualizations could also misguide the user. Most statistical output consist of summary statistics such as means and totals and bar charts are often used to visualize these. Kerns and Wilmer [13] suggested that proper mean bar charts can nevertheless be misinterpreted, because users may not realize that the underlying data distribution may be less smooth than it may seem. Just using a visualization may not be sufficient, and it is unclear whether such side-effects occur with other visualizations of standard statistics.

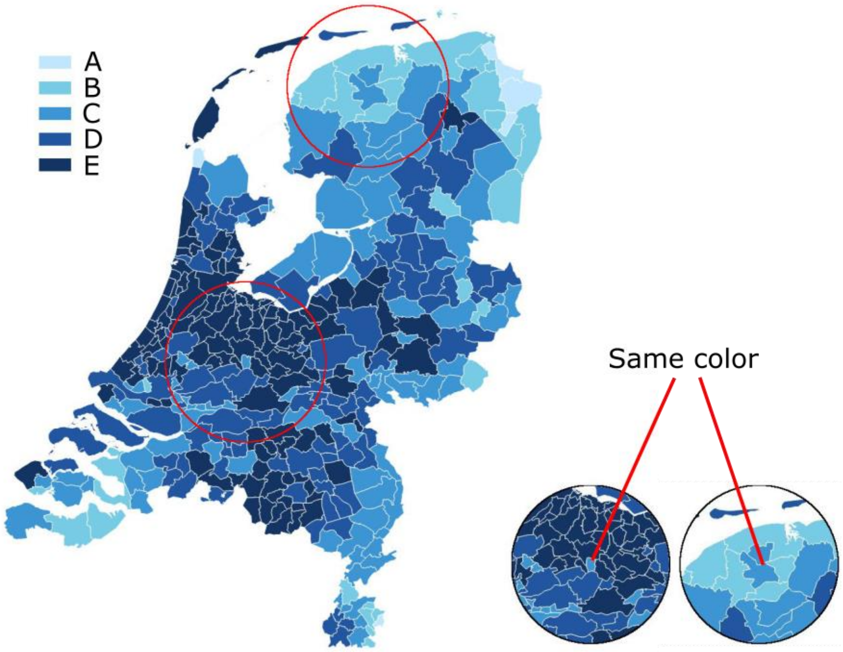


Figure 3. Choropleth showing that color perception depends on the surroundings

While visualizing standard statistics may already cause problems, things may even get more complex with more advanced statistical output like uncertainty. Official statistics has the aim to produce accurate, precise and valid estimates, so each statistical estimate should include an indication of its accuracy, precision and validity. Knowledge with respect to communicating uncertainty is scarce and proper guidelines are lacking [14, 15]. A review of different visualization methods for communication uncertainty in Official Statistics is given in [16].

Conclusions

Effective communication of Official Statistics is an essential part of the mission of NSI's. However, while the quality of official statistics is highly regarded, the quality of visualization and communication of those statistics is often overlooked. Communicating statistics has be done for many years, but perceptual, cognitive and other presentation errors are abound. Visualizations may suffer from all kinds of perception biases and misinterpretations. Surprisingly, the practice of communication of statistics lacks scientific rigor. Furthermore, statistics have limited precision, but accuracy of statistics is seldom communicated.

Therefore, we encourage the official statistics community to investigate how users perceive, interpret, and use our statistical output and in particular visualizations. We propose to develop guidelines based on latest insights from several scientific fields and based on user studies to elevate the quality of communication above “best practices”.

As a start, we plan to conduct user studies at CBS, both in a controlled environment (for studies on visual perception), and in a free environment (for qualitative studies). In this way, different visual designs and components, such as colour schemes, can be compared systematically. The results should eventually lead to checklists and guidelines for visual designers.

References

[55] Eurostat (2017). European Statistics Code of Practice (https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice). Luxembourg: Eurostat.

[56] Rensink, Ronald. (2014). On the Prospects for a Science of Visualization. 10.1007/978-1-4614-7485-2_6.

[57] Chen, M. & Golan, A (2016) What may visualization processes optimize? IEEE Transactions on Visualization and Computer Graphics 22:12, 2619–2632.

[58] Tourangeau, R. (1984). Cognitive science and survey methods. *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, (Eds., T.B. Jabine, M.L. Straf, J.M. Tanur and R. Tourangeau). Washington, D.C.: National Academy Press, 73-100.

[59] Tourangeau, R., Conrad, F. and Couper, M. (2013). *The Science of Web Surveys.* Oxford University Press, New York.

[60] Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. New York: Cambridge University Press.

[61] Tufte, E.R. (1983) The Visual Display of Quantitative Information

[62] Brackstone, G. (1999), Statistics Canada, Survey Methodology, Catalogue No. 12-001-XPB, 1 Vol. 25 No. 2, December 1999

[63] Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387), 531-554.

[64] Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 203-212). ACM.

[65] Gregory, R.L. (1968). Perceptual Illusions and Brain Models. Proceedings of the Royal Society of London, Series B, Biological Sciences, Vol. 171, No. 1024. A Discussion on the logical Analysis of Cerebral Functions (Dec. 13, 1968), pp. 279-296.

[66] Marr, D. (1982), Vision: A Computational Approach, San Francisco, Freeman & Co.

[67] Kerns, S.H. & Wilmer, J.B. (2021) Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs. Journal of Vision, 21: 17, 1-36.

[68] Petersen, A.C., Jansen, P.H.M., Van der Sluijs, J.P., Risbey, J.S., Ravetz, J.R., Wardekker, JA., Martinson Hughes, H. (2013) Guidance for Uncertainty Assessment and Communication 2nd Edition, PBL

[69] Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R. et al. Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. 2010.

[70] De Jonge, E. (2020), Communicating uncertainties in official statistics — A review of communication methods, Eurostat, Communicating uncertainties in official statistics — A Review of communication methods — 2020 edition - Products Statistical working papers - Eurostat (europa.eu)

Environmental Statistics (JENK3M.2)

Session Chair: **Arturo De La Fuente** *(Eurostat)*

Satellite altimetry to water level monitoring: Italian and Finnish case studies
Cristina Deidda *(Politecnico di Milano)*

Degree of urbanization in Catalonia: high-resolution grids and commuting times
Clara Hormigos Feliu *(European Commission-Joint Research Centre), Eduard Suñé, Daniel Ibáñez, Mireia Farré, Cristina Rovira (Statistical Institute of Catalonia)*

Remote sensing application in official statistics. Urban green index for Romanian cities
Marian Necula *(National Statistical Institute-INSSE),* Bogdan Oancea *(National Statistical Institute-INSSE and University of Bucharest)*

Satellite altimetry to water level monitoring: Italian and Finnish case studies

The water level monitoring of rivers and lakes is a fundamental issue to assess the health status of water bodies and to know the availability of freshwater resources. The presence of in-situ water level stations is quite rare, and when available it may have discontinuous functioning, or lack of maintenance, making this kind of information difficult to be used in hydrological modelling. In the water level monitoring problem, recently, remote sensing data collected through satellite altimeters have showed great capability, not only in the case of rivers with wide sections (>100 m), but also in the case of rivers with narrow sections (< 100 m).

Here, we have assessed the feasibility of satellite altimetric measurements (Jason, Sentinel 3A/3B, Envisat) in rivers and lakes, in Italy and Finland, making a comparison to in-situ data. We have used four different methods of comparison between in-situ and remote water level measurements: (1) comparing the differences between successive times of satellite and in-situ measurements, (2) comparing the satellite and in-situ measurements assuming that the initial value of the joint measurements is equal (specifically the value of satellite water level is equal to the value of in-situ one), and using the temporal differences between two successive satellite measurements, (3) comparing the satellite and in-situ measurements calibrating the initial water level of satellite measurements minimizing the RMSE, (4) comparing the satellite and in-situ measurements calibrating the initial water level of satellite measurements minimizing the MAE.

Degree of urbanization in Catalonia: high-resolution grids and commuting times

Keywords: Degree of urbanization, high-resolution grid, urban-rural continuum, remote rural areas

INTRODUCTION

The collection of statistics for urban and rural areas constitutes an important tool to inform territorial policy decisions. In order to harmonize the description of the urban-rural continuum, the European Commission (EC) has put forward a methodology, endorsed as well by the United Nations, which allows to compute the degree of urbanization of any given territory [1]. This classification defines three categories (cities, towns and semidense areas, and rural areas) which are assigned to administrative divisions, based on population clusters computed in a 1 km grid. In this work, we implement the computation of the degree of urbanization in Catalonia (Spain), using the standard European 1 km grid as well as higher-resolution (500 m and 250 m) grids. Our motivation is twofold: firstly, we aim to update the degree of urbanization regularly and observe yearly changes (Eurostat's latest update is based on the 2011 census population [2]), and we also want to investigate the impact of grid size when classifying the urban-rural continuum. Furthermore, in order to improve the characterization of rural areas, we implement the concept of remoteness. Following again EC guidelines, we identify remote rural areas by studying median travel times from rural municipalities to cities and discuss their distribution in Catalonia.

METHODS

Our study is based on geocoded population data from the Statistical Population Registry (REP) maintained by Institut d'Estadística de Catalunya (Idescat), as well as 2021 municipality geometries from Institut Cartogràfic i Geogràfic de Catalunya (ICGC) [3] and the standard European grid [4]. In what follows, we briefly describe the computation methods we employ to determine the different territorial classifications.

Degree of urbanization with 1 km and higher-resolution grids

In order to compute the degree of urbanization, we first divide the territory of Catalonia according to the standard European grid, with a cell size of 1 km, and compute cell population totals and densities using 2018 REP data. Following EC methodology [1], we then classify each cell by identifying three types of clusters: urban centres (cells with population density higher than 1,500 inhab/km^2, and with a total cluster population of 50,000 inhabitants or more), urban clusters (cells with population density higher than 300 inhab/km^2, and with a total cluster population of 5,000 inhabitants or more), and rural grid cells (cells which do not belong to urban centres or urban clusters, with population density usually below 300 inhab/km^2). Once grid cells are classified, we superpose the geometries of the Catalan municipalities and classify them. If at least 50% of the municipality's population resides in urban centre cells, then it is classified as a city or densely populated area, while if at least 50% reside in rural grid cells, it is considered to be a rural area or thinly populated area. Finally, if none of those conditions are met, municipalities are classified as towns and semi-dense areas or intermediate density areas.

The standard degree of urbanization methodology specifies that cell classification must be based on the standard European 1 km grid. Going beyond that, we apply the same algorithm to higher-resolution grids, with cell sizes of 500 m and 250 m, each computed by consecutively dividing the standard European grid.

Remote rural areas and road networks

In order to provide a finer characterization of rural areas, EC guidelines define remote rural areas as those rural areas from which travel to the closest city takes longer than 45 minutes by car [5]. We implement this calculation for the rural municipalities of Catalonia, using geospatial transport network data from Open Street Maps (we include networks for Catalonia, Aragon and the south of France) and modelling it as a graph using postGIS extension pgRoute. Using 2020 data, we compute travel times for all the population in each rural municipality, and classify as remote rural municipalities those where the median travel time exceeds 45 minutes.

RESULTS

Degree of urbanization in Catalonia

Our results for the degree of urbanization in Catalonia (based on the 2018 REP population) are shown in Table 1 together with Eurostat's determination, which employs 2011 census data [2]. Using the 1 km base grid we find that, out of 947 Catalan municipalities, 691 (74.7%) are rural areas, 201 (20.9%) are classified as towns and semi-dense areas and 55 (4.4%) as cities. These results differ somewhat from Eurostat's classification, which shows 707 rural areas (74.7%), 198 (20.9%) towns and semi-dense areas and 42 (4.4%) municipalities classified as cities. Population growth in the 2011-2018 period accounts for these differences: rural areas decrease while higher-density areas grow. Even if these differences lead to changes in the categorization of a small number of municipalities, being able to frequently update the classification can have a substantial impact on local policies where degree of urbanization is an input. An example of these is the allocation of LEADER European funds for rural areas [6], which are managed at local level.

Table 1. Number of municipalities in Catalonia by degree of urbanization according to our results (Idescat), which use 2018 population data, and Eurostat (based on the 2011 census).

Degree of urbanization	Eurostat (2011)	Idescat (2018)		
Grid size	1 km	1 km	500 m	250 m
Densely populated (cities)	42	55	40	35
Intermediate density	198	201	196	183
Thinly populated (rural areas)	707	691	711	729

Our results using higher-resolution grids also show significant differences. We observe that increasing grid resolution leads to more areas being identified as rural, both at the cell and municipality level, while less municipalities are classified as cities or towns and semidense areas.

This is a consequence of cell-level cluster identification differences: as resolution increases, urban centres reduce their size, break up or are not identified altogether, while urban clusters break up as well and their total number increases. This can be seen in Figure 1 for the Barcelona area, where we observe that extensive regions along the coastline between the cities of Barcelona and Mataró, which are identified as urban centres with the 1 km grid, are no longer identified as such when switching to the 500 m grid. Therefore, the results indicate that the determination of the degree of urbanization is not resolution invariant, and that the underlying grid size of 1 km should be born in mind when classifying small spatial units according to this methodology.

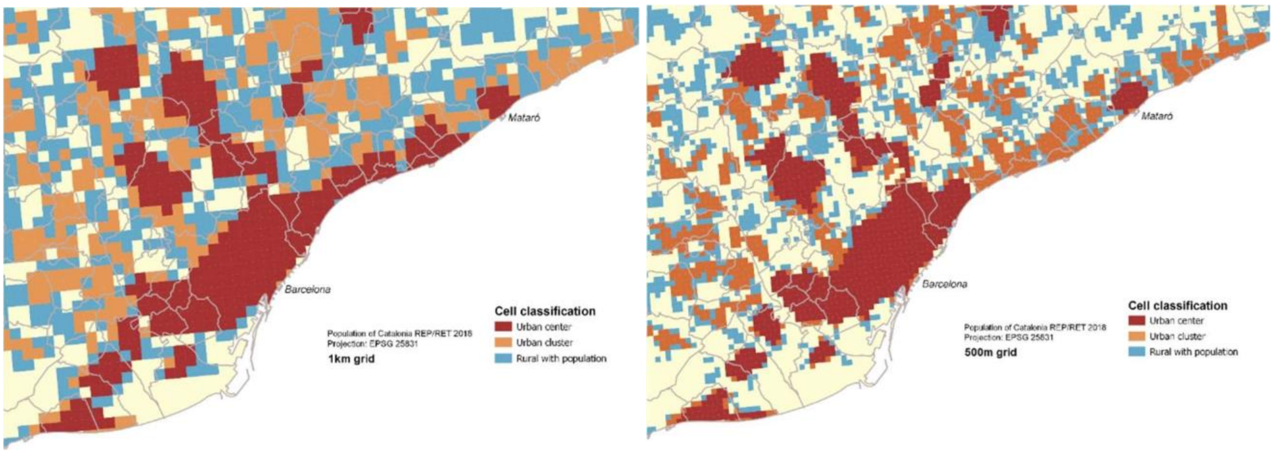


Figure 1. Grid cell classification around the Barcelona area employing a 1 km grid (left) and a 500 m grid (right). In the latter there is a visible reduction of the area covered by urban centres along the coastline between Barcelona and Mataró.

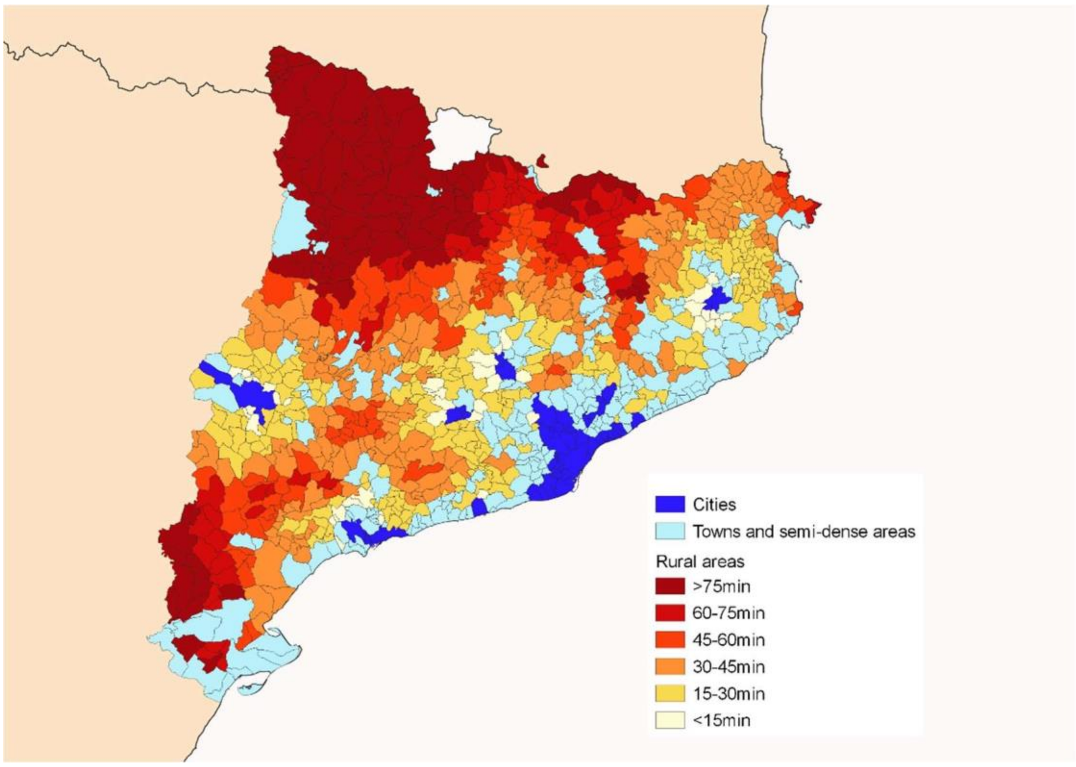


Figure 2. Median traveling times from rural municipalities in Catalonia to the closest city. 3.2. Identifying Catalonia's remote rural areas

The results of section 3.1 indicate that almost three quarters of Catalonia's municipalities are rural areas. However, these might present markedly different realities: for instance, rural municipalities in the Pyrenees could find more difficulties reaching certain services than other rural municipalities that are close to large cities. The EC's guideline to use travel times to define remote rural areas presents an opportunity to provide a distinction among rural areas without depending on external parameters (e.g. grid size, density cuts).

In order to determine which rural municipalities should be considered remote, we compute median travel times to the closest city for all rural municipalities in Catalonia, which are shown in Figure 2. We observe that longest travel times arise in the northern regions at or close to the Pyrenees, as well as in municipalities on Catalonia's south-western corner, in the province of Tarragona. Out of the 947 municipalities of Catalonia, 234 (24.7%) are rural municipalities which are above de 45 minute travel time threshold and are therefore classified as remote rural areas. The population of these areas reaches 162,730 inhabitants, which constitutes only 2.1% of Catalonia's population.

cONCLUSIONS

We have studied the urban-rural continuum of the region of Catalonia by computing the degree of urbanization of its municipalities and identifying its remote rural areas. Our results, based on the 2018 population, show differences with Eurostat's calculation based on the 2011 census, which highlights the importance of performing frequent updates of the degree of urbanization in order to better inform regional and local policymaking. Apart from using the standard 1 km European grid, we have also studied the degree of urbanization using higher-resolution (500 m and 250 m) grids, finding that an increase in resolution leads to larger rural areas, and therefore indicating a grid-size dependence on the methodology, which should be taken into account when interpreting results.

The delimitation of remote rural areas does not present the same dependence on external parameters, as it is based on travel times between rural areas and cities, with the threshold being set at 45 minutes. We have seen that remote rural areas account for 24.7% of the municipalities of Catalonia, while its population amounts only to a 2.1% of the total. The inclusion of travel times not only to cities, but also to a variety of services, can provide an insight on territorial disparities which we plan to explore in the future.

rEFERENCES

[1] Eurostat, Applying the degree of urbanisation : a methodological manual to define cities, towns and rural areas for international comparisons : 2021 edition, Publications Office (2021), https://data.europa.eu/doi/10.2785/706535

[2] Eurostat, NUTS – Nomenclature of territorial units for statistics. Local administrative units (n.d.), https://ec.europa.eu/eurostat/web/nuts/local-administrative-units

[3] Institut Cartogràfic i Geològic de Catalunya, Divisions administratives (2021), https://www.icgc.cat/Descarregues/Cartografia-vectorial/Divisions-administratives

[4] European Environment Agency, EEA reference grid (November 2017), https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2

[5] European Commission, Staff Working Document (part 1): a long-term vision for the EU's rural areas (2021), https://ec.europa.eu/info/sites/default/files/strategy/strategy_documents/documents/ltvrac2021-345-documents-part1_en.pdf

[6] European Network for Rural Development, LEADER/CLLD (consulted: October 2022), https://enrd.ec.europa.eu/leader-clld_en

Remote sensing application in official statistics.Urban green index for Romanian cities.

Keywords: official statistics, remote sensing, land cover, urban vegetation.

Introduction

Land cover and land usage statistics are of core importance in official statistics product portfolio, increased by a shift in public agenda on climate change. Within the Sustainable Development Goal 11.7 regarding universal access to green public spaces, indicator 70 was proposed regarding the proportion of green spaces relative to the total area of urban settlements, in order to monitor the quality of life within human settlements classified as cities and to ensure equitable access to green infrastructure. At the European level, there are many initiatives regarding the sustainable development of cities and the conservation/expansion of green infrastructure and ecosystems, among which we mention the Urban Agenda [1], the European Strategy on Green Infrastructure [2] or the decision of the European Parliament on the General Union Environment Action Programme to 2020 [3]. Among the European initiatives for implementing monitorisation of urban surfaces Land Surface Monitoring Service, within the Copernicus program of the European Commission developed in partnership with the European Space Agency stands among the top initiatives [4]. The Copernicus program provides users, free of charge, with access to data obtained through remote sensing, through the Sentinel space missions, but also to data obtained through in situ collection [5]. The paper at hand describes the process of obtaining urban green index for county capital cities based on vegetation share from total surface.

Methods

Data input consists from Terra MODIS [6] and Sentinel 2 [7] datasets, collected between 1st of June and 31st of July. As a method to discriminate between surfaces covered with vegetation and other types of land use classes, we employed the normalized difference vegetation index (NDVI). The normalized difference vegetation index is a popular measure in remote sensing data analysis applications for the identification/visualization of vegetated areas [8]. The index is a dimensionless quantity, with values in the closed interval [-1, 1], quantifying the degree of vegetation coverage of the earth's surface. The normalized difference vegetation index was created based on empirical observations of the interaction between electromagnetic waves, light in the red visible spectrum and near infrared spectrum, and vegetation [9]. Through the measurements, an increase in the intensity of radiation in the near infrared zone and the absorption of red light from the visible spectrum was observed. NDVI has some ability to discriminate between the types of vegetation covering an area (agricultural vegetation, forest, shrubs, etc.), as well as the quality of that vegetation (dry, green, etc.). The sensor-independent formula is:

$$NDVI = \frac{NIR-RED}{NIR+RED},$$

where NIR is reflectance near infrared and RED is reflectance of visible red light.

An issue reported in research papers and applications is selecting the threshold value of NDVI in order to discriminate between pixel associated with vegetation (see Table 1.)

Table 1. Some NDVI cutoffs for vegetation discrimination.

Category	STATCAN [10]	NOAA AVHRR[9]	[11]
Non-vegetation	[-1, 0.5]	[-1, 0]	<= 0.4
Vegetation	[0.5, 1]	(0, 1]	>0.4

In order to select an optimal threshold for the discrimination between areas covered with vegetation and those with non-vegetation, we performed manual comparisons between the images represented in natural (visible) colors provided by the Google Maps service (satellite image layer) and different intermediate discrimination thresholds starting from literature data, respectively the 0.5 to 0.7 NDVI threshold for MODIS, and the 0.3 to 0.6 NDVI threshold for Sentinel-2, both thresholds built with a step of 0.05.

The data collection, pre-processing and analysis was carried out in free and open source R programming language [12], using free distributed libraries such as getSpatialData [13], sen2r [14], raster [15], ggplot2 [16], sf [17] and others.

Results

In figure 1

the computed share of vegetated areas is presented for 41 Romanian cities, which encompass public and private domains. The agreement between MODIS and Sentinel 2 estimations is problematic, mainly due to different spatial resolution, although the two estimates are strongly correlated ($r_{Pearson}$= 0.9275), also strongly correlated with city surface area ($r_{Pearson-MODIS}$= 0.8802; $r_{Pearson-Sentinel\ 2}$= 0.9722932).

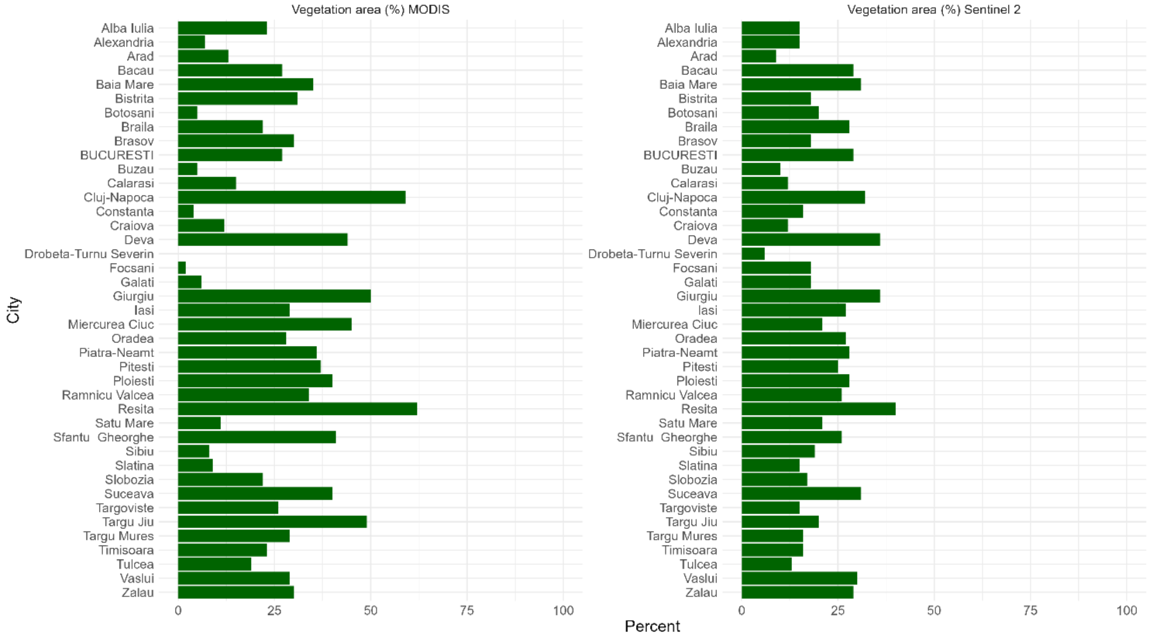


Figure 1. Romanian cities percentage of total area covered by vegetation.

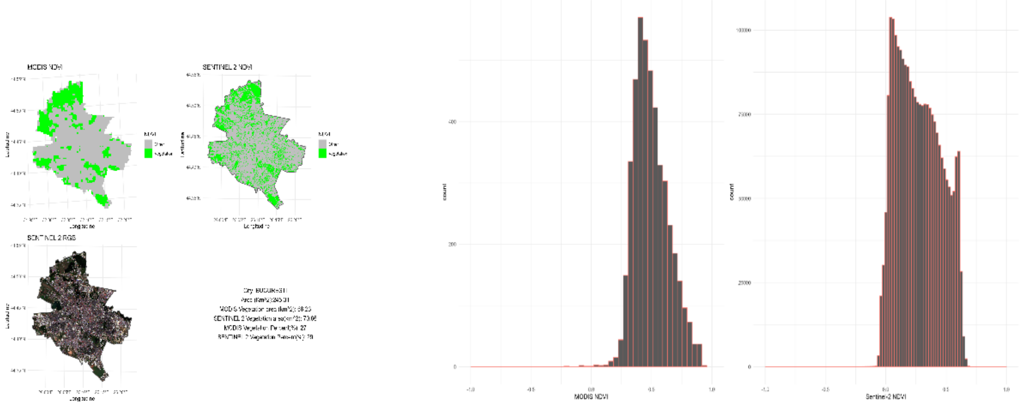

MODIS NDVI
SENTINEL 2 NDVI
SENTINEL 2 RGB
Latitude
Longitude
NDVI
Other
Vegetation
City: BUCURESTI
Area (km^2):245.31
MODIS Vegetation area (km^2): 68.23
SENTINEL 2 Vegetation area(km^2): 70.06
MODIS Vegetation Percent(%): 27
SENTINEL 2 Vegetation Percent(%): 29
count
MODIS NDVI
Sentinel-2 NDVI

Figure 2. Bucuresti (capital city) vegetation cover

estimated using MODIS and Sentinel 2 NDVI and histograms.

Figure 2 presents differences between the MODIS and Sentinel 2 NDVI results, mainly due to different spatial resolution (MODIS 250m, Sentinel 2 10m), considering that spectral resolution is similar for the problem at hand, i.e. computing NDVI. Also, from figure 2 (NDVI histogram for both sensors) we can also observe differences in sensor sensitivity. The main advantage of MODIS consists in a longer span of remote sensing observations (starting with year 2000), compared with Sentinel 2 (starting with year 2016), translated into a more robust and versatile dataset with respect to machine learning applications.

Conclusions

An advantage of using remote sensing data is translated into timeliness and zero response burden on statistical respondents, with arguably, lower costs in terms of statistical production. Potential disadvantages consist in pre-processing costs and calibration validation hurdles, if in-situ data is used as input, given the size and specificity of this type of data which imply an interdisciplinary knowledge base (GIS, remote sensing, statistics). Other major disadvantage with passive remote sensing, i.e. using solar reflected electromagnetic radiation, is an increased likelihood of data unavailability due to atmospheric phenomena (heavy cloud presence) which requires some forms of spatiotemporal interpolation techniques or fusion with data from active sensors (synthetic aperture radar). Some future prospects of statistical research will include some statistical models to combine remote sensing data with statistical office data in order to explore different hypothesis regarding spatial distribution of green capital. Results can be used as a test in the experimental statistics section to check the potential interest of statistical data users relative to the advantages/disadvantages of the data source.

References

[1] European Commission, “The Urban Agenda for EU”, (2016). Available at: https://futurium.ec.europa.eu/en/urban-agenda/pages/what-urban-agenda-eu

[2] European Commission, “The EU Strategy on Green Infrastructure”, (2013). Available at: https://ec.europa.eu/environment/nature/ecosystems/strategy/index_en.htm

[3] European Parliament, “7th Environmental Action Programme”, (2013). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013D1386

[4] European Commission, “Copernicus Earth Observation Programme”, (2022a). Available at: https://www.copernicus.eu/en/about-copernicus

[5] European Commission, “Copernicus Land Monitoring Service”, (2022b). Available at: https://land.copernicus.eu/.

[6] National Space Agency, “Moderate Resolution Imaging Spectroradiometer.” (2022). Available at: https://modis.gsfc.nasa.gov/

[7] European Spatial Agency. “Open Access Hub” (2022). Available at: https://scihub.copernicus.eu/dhus/#/home

[8] Weier, J. Herring, D. “Measuring Vegetation (NDVI & EVI)” (2000) Available at: https://earthobservatory.nasa.gov/features/MeasuringVegetation

[9] Huete, A. Justice, C. Van Leeuwen, W.J.D. “MODIS vegetation index (MOD13)” Available (1999) at: https://www.researchgate.net/publication/268745810_MODIS_vegetation_index_MOD13

[10]Lantz, N. Grenier, M. Wang, J. “Urban greenness, 2001, 2011 and 2019” (Statistics Canada). Available at: https://www150.statcan.gc.ca/n1/pub/16-002-x/2021001/article/00002-eng.htm

[11] European Spatial Agency, “Level-2A Algorithm Overview”.(2022) Available at: https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm

[12] R Core Team. “R: A language and environment for statistical computing” (2022). Available at: https://www.r-project.org/

[13] Schwalb-Willmann J, Fisser H _"getSpatialData: Get different kinds of freely available spatial datasets_. R package version 0.1.2" (2022)., Available at: http://www.github.com/16eagle/getSpatialData/

[14] Ranghetti, L., Boschetti, M., Nutini, F., Busetto, L. sen2r: An R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. Computers & Geosciences, (2020) 139, 104473.

[15] Hijmans R "raster: Geographic Data Analysis and Modeling_. R package version 3.5-21" (2022)., Available at: https://CRAN.R-project.org/package=raster

[16] Wickham, H. "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York, 2016.

[17] Pebesma, E., "Simple Features for R: Standardized Support for Spatial Vector Data." (2018), The R Journal 10 (1), 439-446.

Energy data spaces and statistics (MANS3M.2)

Session Chair: **Carola Carstens** *(Eurostat)*

PLATOON- Digital PLAtform and analytical TOOls for eNergy
Valentina Janev, (*The Mihajlo Pupin Institute, University of Belgrade)*

Smart Statistics from Smart Meters
Soren Andersen, *(Statistics Denmark)*

Big Data for OPen innovation Energy Marketplace
Jordi Jené Jaume (*Aněll*)

POST03: Poster session

Eurostat's GISCO services at your disposal
Hannes Reuter *(Eurostat)*

From a glossary to a controlled vocabulary: a preliminary Istat experience
Emanuela Recchini *(National Institute of Statistics–ISTAT)*

An open source data science platform to foster innovative and reproducible statistical projects
Romain Avouac *(National Institute of Statistics–INSEE)*, Frédéric Comte *(National Institute of Statistics–INSEE)*

Mapping circular economy EU funded projects using text mining
Anabela Santos *(European Commission-Joint Research Centre)*

Integrating Big Data in the traditional survey of maritime transport statistics: open challenges
Angela Pappagallo *(National Institute of Statistics–ISTAT)*, Luca Valentino *(National Institute of Statistics–ISTAT)*, Francesco Amato *(National Institute of Statistics–ISTAT)*

How well-behaved are revisions to quarterly fiscal data in the euro area?
Krzysztof Bankowski *(European Central Bank),* Thomas Faria *(Insee)*

An Android app for continuous census management
Eleonora Sibilio *(National Institute of Statistics–ISTAT)*, Giuseppe Giuliano *(National Institute of Statistics–ISTAT)*

Bridging independent taxonomies using AI: an application to skills mismatch using PIAAC, ESCO Skills and Online Job Ads
Francesco Trentini *(University of Milano-Bicocca)*, Fabio Mercorio *(University of Milan Bicocca)*, Mario Mezzanzanica *(University of Milan-Bicocca)*, Filippo Pallucchini *(Università di Milano Bicocca)*, Yuchen Guo *(University of KU - Leuven)*

Use of multilevel grids to release protected grid data from the French 2021 Census
Julien Jamme *(INSEE)*

Hedonic House Price Index for Poland based on listings
Radoslaw Trojanek *(Poznań University of Economics and Business)*

Geospatial Data Store - centralization of spatial data in Statistics Poland
Amelia Wardzinska-Sharif *(Statistics Poland)*

Eurostat’s GISCO services at your disposal

Keywords: Geospatial information, GSGF, corporate level services, GISCO, IMAGE

Introduction

The integration of Geographical Information (GI) and statistical data play an increasingly important role in policy-making, administrative planning processes, policy assessment and quantitative monitoring of the effects of policy decisions [1]. It can provide information with the right spatial resolution from regional to European and global level, allowing for policy interventions at the level where they are most effective. Eurostat’s Geographic Information System Coordination (GISCO) is currently offering technical and coordination services on geographical information to the European Commission and internationally, supports the development of methodology for the integration of geospatial and statistical data, promotes the further use of GIS in the ESS and offers training and geospatial analysis.

The Global Statistical Geospatial Framework (GSGF) contains five different principles which are depicted in Figure 1, which addresses a range of issues in statistical and geospatial data management. With due respect the GISCO team aims to provide a range of corporate level services to the European Commission (EC) as well to the European Statistical System (ESS) to support these. The abstract showcases a selection of these for various application examples to support the integration of statistical and geospatial information

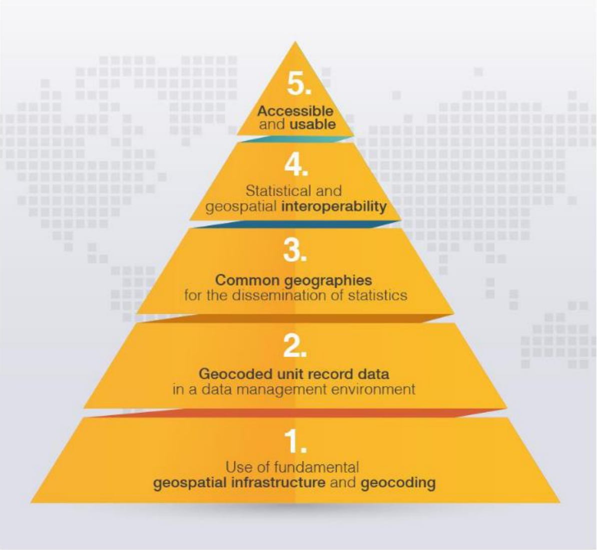

5.
Accessible and usable
4.
Statistical and geospatial interoperability
3.
Common geographies for the dissemination of statistics
2.
Geocoded unit record data in a data management environment
1.
Use of fundamental geospatial infrastructure and geocoding

Figure 34: Five principles of the GSGF

Methods

Within Eurostat, GISCO is responsible for meeting the European Commission's geographical information needs. Each of the services outlined below represents an example of supporting this and in turn supports the implementation of the GSGF, its European adaption GSGF: Europe [2] and the ESS in general. These services are based on standard based implementation (e.g. Open Application Interface (OpenAPI), OpenGeospatialConsortium(OGC), INSPIRE) aimed at facilitating robust data handling.

Fundamental geospatial infrastructure and geocoding - the case of the Address-API

The European Register of Addresses (or short address-API[77]) addresses the issues that in many business processes addresses are recorded in free form fields without any validation. These creates issues at further processing. Example of issues observed are: a) fantasy (non-existing) streets or house numbers, over 20 different variants of one street name, cities allocated in the wrong region or even country. In the world of Total Quality Management [3], this is a standard example. The European Register of Addresses aims to create a geocoding infrastructure to solve this issue. It's based on authoritative information obtained from Member states, updated at least annually and implements the provide-only-once principle. Some other services are also available internally to the European Commission services.

Geocoded unit record record data – the case of the ID service

Geocoded unit record data are increasingly implemented in the various surveys, algorithms and datasets in the ESS ecosystem. With this, coordinates of records become available – being it point, line or polygon features. For point locations the question arises quite often in which given statistical (e.g. NUTS, grids), administrative (Local Administrative Units, Countries) or ecological area (e.g. sub river basins) these has been in the past or will be in the future. The IDentity service[78] allows addressing this for machine-to-machine communication or via a graphical user interface. The currently recorded maximum usage are 9 million request during a 24h period.

Common geographies and interoperability -

The third level in the GSGF are the common geographies and datasets. The various user communities are supported via data dissemination endpoints addressing different user requirements. Desktop user can obtain the data via bulk downloads for human interaction[79] or machine-to-machine communication[80], while the web community can consume OGC compliant services (e.g. vector tiles[81], feature services[82], TopoJSON[83] (dedicated for visualisation purpose) to represent the geographies in the web browser. To ensure interoperability we aim to have consistent IDs to ensure statistical and geospatial interoperability. On average up to 3TB per month are served online to the various user communities.

Accessible and Usable – the Interactive MAp GEnerator

[77] https://gisco-services.ec.europa.eu/addressapi/

[78] https://gisco-services.ec.europa.eu/id/

[79] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data

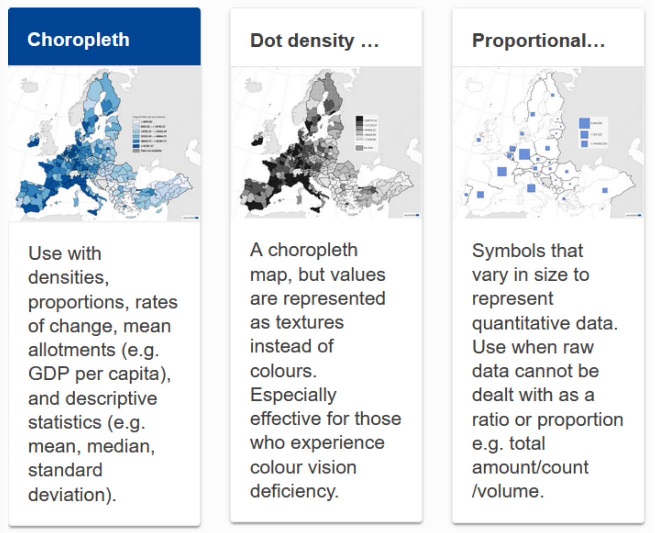
[80] https://gisco-services.ec.europa.eu/distribution/v2/

[81] https://gisco-services.ec.europa.eu/vectortiles/

[82] https://gisco-services.ec.europa.eu/features/

[83] https://gisco-services.ec.europa.eu/pub/nuts2json/v2/

The officers in the European Commission are in the need for a self-serving tool to generated maps for example to be included in reports, presentation or even on social media. Therefore, the IMAGE tool[84] has been developed for the production of thematic maps for a pre-defined map extent of the EU/EFTA/Candidate Countries. It is intended to enable users to quickly make professional, statistical maps for publication that adhere to Eurostat's Style guide. Based on file uploads or a direct connection to Eurostat's database, it allows to generate on the fly Choropleth, Dot Density and Proportional Symbol maps as seen in Figure 2 above. On a monthly basis over 250 standardized maps are made with this tool.



Choropleth

Use with densities, proportions, rates of change, mean allotments (e.g. GDP per capita), and descriptive statistics (e.g. mean, median, standard deviation).

Dot density ...

A choropleth map, but values are represented as textures instead of colours. Especially effective for those who experience colour vision deficiency.

Proportional...

Symbols that vary in size to represent quantitative data. Use when raw data cannot be dealt with as a ratio or proportion e.g. total amount/count /volume.

Figure 35: Example Maps for the IMAGE tool and their proposed usage

Conclusions

The integration of the geospatial and the statistical domain implies collaboration and knowledge exchange between the statistical and the geospatial domain. Any data producer being it in the statistical domain or outside should use the same sources of geospatial information. This leads to an efficient use of resources and improved integration of data. The selected examples shown above support the outlined five principles as described in the introduction – providing them to the ESS's daily Generic Statistical Business Process Model (GSBPM) production processes.

References

[71] UN (2019), *The Global Statistical Geospatial Framework*, 9th session of the GGIM committee, URL: https://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf

[72] Geostat4 (2022) *Global Statistical Geospatial Framework: Europe*, Report, Eurostat GA945503 - 2019-FI-GEOSTAT4 URL : https://www.efgs.info/wp-content/uploads/2022/03/GSGF_Europe.pdf

[73] Labovitz,G. Sang Chang,Y. Rosansky,V (1992) *Making quality work : a leadership guide for the results-driven manager*, OMNEO, Essex Junction, VT

[84] https://gisco-services.ec.europa.eu/image/

From a glossary to a controlled vocabulary: a preliminary Istat experience

Keywords: metadata, controlled vocabulary, semantic web.

ıNTRODUCTION

The management of information and knowledge relating to one or more domains and the communication exchanges among the actors who operate within them cannot be separated from a shared use of terminology, making it necessary to define tools and resources that allow for an organization of terminology as much as possible consistent and unambiguous. These are specialized lexicons (unstructured lists of terms pertaining to a knowledge sector), glossaries (that add the relative definition to each item), classification systems, controlled vocabularies (that integrate the terminology by making explicit the semantic relationships existing among concepts), finally conceptual maps and ontologies [1]. The underlying principle of these resources is terminological control, implemented at an increasing level of semantic structuring and formalism [2]. The use of a controlled vocabulary (CV), focus of the present work, enhances consistency and efficiency in the production of metadata, gives precision in their searching and allows semantic and technical interoperability among different repositories. A CV is an organized list of terms used for metadata and information retrieval. It may be a flat list or more complex construct where semantic relationships are made explicit by identifying broader, narrower and related terms, synonyms and scope notes. Ideally, the terms in a CV should be exhaustive (covering the whole dimension of the issue), mutually exclusive (no overlaps between terms) and clearly defined (definitions/scope notes given for the meanings of the terms). CV is often used in specific contexts, and definitions/scope notes clarify and disambiguate the meaning of a term in a particular context as it may differ from the meaning in natural language [3].

1.1. Controlled vocabularies and international metadata statistical standards

Since statistical metadata describe a wide range of information (concepts, processes and methods involved in the process, as well as description of variables and units) they can be categorized using several frameworks. Among the different international metadata statistical standards, the most used by national statistical offices worldwide are: the Generic Statistical Business Process Model (GSBPM), describing and defining the set of business processes needed to produce official statistics; the Generic Statistical Information Model (GSIM), providing a set of standardized, consistently described information objects, which can be used as inputs and outputs in the design and the production of statistics; the Data Documentation Initiative (DDI), a standard designed to document and manage different stages in the research data lifecycle. CVs play a critical role in metadata standards, which have two basic components: semantics (definition of the meaning of metadata elements); content (declaration of instructions for what and how values should be assigned to elements). Using CVs has several advantages (e.g. for DDI), many of them related to overcoming difficulties caused by natural language in documentation and information retrieval [3]. As regards GSBPM, every task is identified with the syntactic construction “verb+name”: the verb identifies the phase, the name the element of production upon which the action (identify, design, develop, execute, monitor) is exerted. This construction finds its limits in the use of natural language itself so that some combinations of those verbs with the different elements are not appropriate and alternative verbs must be used with the same general meaning. In this case, the adoption of an information model standard (e.g. GSIM) is useful for the analysis of all the objects of the statistical production chain. This normalization of language came up against the general usage of diversified jargon in each statistical domain, but it constitutes the first step towards a CV [4].

METHODS

The Italian National Institute of Statistics (Istat) regularly updates and publishes a single Glossary of the terms used within official statistics. This Glossary is drawn from a much richer information Terminology Collection (TC) that documents the terminology related to metadata. Starting from this, it has been possible to start an experiment aimed at developing a CV limited to a specific domain. It has been identified as connected to the creation of a thematic statistical register on disability based on the individuals register.

2.1. Glossary and controlled vocabulary

In any sectoral language, such as that of official statistics, the terms are often used in a peculiar way, even in the context of extremely restricted thematic subdomains. A glossary of terms, therefore, is the tool that can make the meaning of the language used unambiguous and shared and ensure the quality of the communication provided. It allows the use even by non-experts of complex concepts strictly related to the sector, thus facilitating the dissemination and correct use of statistical production, data collection and storage, the exchange among heterogeneous sources. A consistent TC is an essential support for the construction of CV. The current Istat TC has an average degree of structuring: each term is associated with a unique definition (the lemma-definition pair is uniquely determined) and the exhaustive list of synonyms. A CV requires, in addition, a complete documentation of the semantic relationships between concepts. Three standardized types of relationships need to be made explicit: equivalency; hierarchy; association. The equivalence relationship allows the management of synonymy, quasisynonymy and linguistic variants. The hierarchical relationships is based on degrees or levels of superordination/subordination, where the superordinate term represents a class or a whole, and subordinate terms refer to its elements, parts or individuals. The associative relationship covers associations between terms that are neither equivalent nor hierarchical, yet the terms are semantically or conceptually associated to such an extent that the link between them should be made explicit in the controlled vocabulary, on the grounds that it may suggest additional terms for use in indexing or retrieval [5]. In order to manage the semantic relationships in a machine-actionable format, the Simple Knowledge Organization System (SKOS) seems to be a suitable framework for processing the Istat CV. SKOS can act as a modular expansion of many of the metadata systems already in existence, conferring upon them greater interoperability, and expanding efforts to link data between divergent information repositories.

2.2. The Istat case study

The disability register (DR) has been designed and developed by Istat to respond to an increasingly complex and articulated information demand. An ever higher level of data granularity is required. A strong encouragement in this direction comes from the UN Convention on the Rights of Persons with Disabilities, which requires countries to collect statistical information to support the monitoring of inclusion policies implemented by governments. CVs respond to the demand for accessibility to information. In a statistical register metadata have to be organized in such a way as to make the information as accessible as possible in an increasingly machine-to-machine perspective. In this regard, information retrieval, reuse and interoperability are crucial aspects to be guaranteed. DR is the product of the integration of, at the moment, two administrative sources from the Italian National Social Security Institute (INPS): the archive of disability and invalidity certificates and the archive of pension benefits. The DR unit, understood as a single record, is represented by the person holding a certification with or without related pension benefits. A person can have more than one certification: in this case, the person is associated with as many records as there are certifications he/she has. The severity degree is indicated for each certification. The main pathology is not currently processed. Pension benefits are always linked to a certification. A person can be the holder of several pension benefits, in which case, the person is associated with as many records as there are benefits from which he/she benefits. Within the activities carried out for DR, an in-depth analysis of administrative sources and a complete documentation of the metadata involved in the creation of the register has been performed, leading to the definition of a list of terms. Starting from this list, a multilingual domain glossary has been developed. It consists of about 50 terms documented with: type of metadata (according to the GSIM standard); definition; source; person responsible for updating; reference legislation. The experiment consisted in transforming the glossary into a CV which provides a structured hierarchy of preferred terminology and captures semantic relationships across terms. Built upon international standards for CV development the Istat experiment includes all the semantic relationships, definitions, notes outlining the sources for the establishment of preferred terminology.

RESULTS

The two tables below document the preliminary implementation of the Istat CV associated with the disability register.

Table 1. Disability CV (Term: Person with ascertained limitations)

Id Glossary	***
Lemma	Persona con limitazioni accertate
Lemma EN	Person with ascertained limitations
Definizione	Persona con una (o più) patologie di tipo fisico, mentale, intellettivo e/o sensoriale determinanti difficoltà a svolgere i compiti e le funzioni proprie dell'età nella vita quotidiana, accertate da una commissione medico legale ai fini dell'erogazione di specifici benefici, servizi e/o prestazioni di tipo monetario diretto.
Definition	Person with one (or more) physical, mental, intellectual and/or sensory pathologies determining difficulties in carrying out the tasks and functions proper to age in daily life, ascertained by a medical examiner commission for the purpose of providing specific benefits, services and/or monetary benefits.
Statistical process code	***
Fonte	Registro disabilità
Source	Disability Register
Person responsible	***
Reference legislation	-
Validation date	04/10/2022
End of validity	
Note	Disability task force 2021/2022
Equivalence Relationship	
Synonym	-
Lexical variant	-
Near synonym	Disabile, Persona disabile, Diversamente abile, Inabile, Invalido, Persona con handicap

Hierarchical Relationship	
Broader Term Generic (BTG)	Persona
Narrower Term Generic (NTG)	Invalido civile Cieco Sordo
	Invalido per servizio
	Invalido di guerra
Associative Relationship	
Related Term (RT)	Persona disabile

Table 2. Disability CV (Term: Person with disability)

Glossary Id	***
Lemma	Persona con disabilità
Lemma EN	Person with disability
Definizione	Persona che presenta durature menomazioni fisiche, mentali, intellettive e/o sensoriali che, in interazione con barriere di diversa natura, possono ostacolare la sua piena ed effettiva partecipazione nella società su base di uguaglianza con gli altri. Tale definizione è quella fornita dalla Convenzione ONU sui diritti delle persone con disabilità (Art. 1), entrata in vigore nel 2006 e ratificata dall'Italia con la legge 3 marzo 2009, n. 18. Si noti che rispetto a tale definizione, nel Registro delle disabilità si documenta la persona con limitazioni accertate, per la quale le menomazioni fisiche, mentali, intellettive e/o sensoriali sono state accertate da una commissione medico-legale.
Definition	Persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others. This definition is the one provided by the UN Convention on the Rights of Persons with Disabilities (Art. 1), which entered into force in 2006 and ratified by Italy with Law 3 march 2009, n. 18. It should be noted that with respect to this definition, the Disability Register documents the person with ascertained limitations, for whom the physical, mental, intellectual and/or sensory impairments have been assessed by a legal medical commission.
Statistical process code	***
Fonte	Registro disabilità
Source	Disability Register
Person responsible	***
Reference legislation	UN Convention on the Rights of Persons with Disabilities (6 December 2006)
Validation date	04/10/2022
End of validity	
Note	Disability task force 2021/2022
Equivalence Relationship	
Synonymy	Diversamente abile
Lexical variants	Disabile, Persona disabile
Near synonymy	Inabile, Invalido, Persona con handicap
Hierarchical Relationship	
Broader Term Generic (BTG)	Persona
Associative Relationship	
Related Term (RT)	Persona con limitazioni accertate, Invalido civile, Cieco, Sordo, Invalido per servizio, Invalido di guerra

CONCLUSIONS

For the metadata standardization process CVs are essential: their use ensures consistent description of resources and their attributes and enables effective information retrieval. CVs, usually determined during the metadata design phase, address ambiguities of natural language at different levels of semantic control. The disability domain, here considered, consists of a standardized metadata system and associated experimental CV. In order to share and link this particular knowledge organization system via the Web, the use of a data model like SKOS is appropriate. A CV represented in SKOS enriches the traditional standard metadata system allowing cross collection searching, in so overcoming the limits due to the different languages of the multiple users.

REFERENCES

[1] A. Folino, Tassonomie e thesauri, in Documenti Digitali, a cura di R. Guarasci, A. Folino, Iter, Milano (2013), pp. 387-444.

[2] M. L. Zeng, A. Salaba, Toward an International Sharing and Use of Subject Authority Data, FRBR Workshop, OCLC (2005).

[3] T. Jääskeläinen, M. Moschner, J. Wackerow, Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability, IASSIST Quarterly Spring - Summer (2009).

[4] D. Salgado, A.I. Sánchez-Luengo, Process metadata development and implementation under the GSBPM v5.0 at Statistics Spain (INE), European Conference on Quality in Official Statistics, Madrid, 31 May-3 June (2016).

[5] ANSI/NISO Z39-19:2005 (R2010), Guidelines for the construction, format, and management of monolingual controlled vocabularies (2010).

An open source data science platform to foster innovative and reproducible statistical projects

Keywords: New Statistical tools and software (Python, Julia, Shiny, etc.), Reusing tools and services

1. Introduction

Following the *European path towards Trusted Smart Statistics* conference of 2018, the European Statistical System has adopted an ensemble of principles aiming at providing capacities to handle new data sources and statistical methods [1]. These principles involve simultaneously the need for new technical skills as well as innovative IT solutions. Not incidentally, an increasing number of public statisticians trained as data scientists have joined NSIs in recent years. However, these new profiles often find themselves isolated in national statistical systems, and their ability to deliver value is limited by several challenges.

The first challenge is related to a lack of proper IT infrastructures to tackle the new data sources that NSIs now have access to as well as the accompanying need for new statistical methods. For instance, big data sources, such as mobile phone data or receipts data, have been experimentally used to provide new statistical indicators (e.g. present population) or to refine existing ones (e.g. price indexes) [2]. However, such data requires huge storage capacities and distributed computing frameworks to be processed, which generally cannot be provided by traditional IT infrastructures. Similarly, the adoption of new statistical methods based on machine learning algorithms often require IT capacities (graphical processing units - GPUs) to massively parallelize computations [3].

Another challenge is related to the difficulty of transitioning from innovative experiments to production-ready solutions. Even when statisticians have access to development environments in which they can readily experiment (e.g. build an interactive R Shiny app for some data visualization), the step towards putting the solution in production is generally very large. Traditionally, this step cannot be performed by statisticians themselves as it requires the intervention of IT teams who manage production infrastructures. However, this organization appears inefficient in many cases. It is for instance often the case that statisticians and IT teams do not use the same programming languages (e.g. R and Java), so that the source code of experimental applications often has to be partially or completely rewritten to produce a production-ready solution. Such examples highlight the need to make statisticians more autonomous regarding the orchestration of their computations as well as fostering a more direct collaboration between teams, as advocated by DevOps and DataOps approaches.

A third challenge is to foster reproducibility in official statistics production. This quality criterion involves devising processing solutions that can produce reproducible statistics on the one hand, and that can be shared with peers on the other hand [4].

The last challenge is related to the difficulty of building proper environments for training programs on innovative statistical languages (e.g. big data frameworks) or tools (workflows frameworks, databases, etc.). Since these are still rarely used in the actual production of official statistics, developing environments providing them are not available for training purposes, which in turn limit their widespread adoption. There is thus a need to provide reproducible environments in which statisticians can hone their skills by experimenting with innovative languages and tools.

2. Methods

Against that background, we developed a data science platform built upon state-of-the-art IT components to provide statisticians with scalable and reproducible environments to experiment with new data science methods and data sources.

2.1. An open platform scaled for innovative data science experiments

The platform we develop is best described as a Datalab: it aims at providing statisticians with both the physical and logical resources necessary to properly prototype and test a data science pipeline from end-to-end. Contrary to conventional IT infrastructures, access to these resources is immediate: there is no need for instance to ask beforehand to provision a storage space or an execution environment. It is open on the internet and can thus be accessed from everywhere, independently of the NSI infrastructures.

On the physical side, the platform is a private cloud based on a cluster of about 20 servers, for a total capacity of 10 TB of RAM, 1100 CPUs, 34 GPUs and 150 TB of storage. These resources enable the platform to scale to most data science experiments. Big data sources can be handled using frameworks such as Spark to distribute computations over the servers. Furthermore, the availability of graphical processing units (GPUs) allow for the training and use of large deep learning models.

2.2. Architectural choices aimed at fostering scalability, autonomy and reproducibility

The platform is based on three deeply structuring choices: cloud computing, containerization and object-storage. These technologies have become the new standards in modern data science infrastructures.

In a cloud environment, the computer of the user becomes a simple access point to perform computations on a central infrastructure. This enables both ubiquitous access to and scalability of the services, as it is easier to scale a central infrastructure — usually horizontally, i.e. by adding more servers. This rationale has also influenced the choice of the data storage technology. The platform uses MinIO, an open-source and S3 compatible object storage framework. In this model, users can store data as “objects” (data and metadata) in their own “buckets” (data store). This storage model is optimized for scalability and intensive computations. Data access is also quite easy and cloud-native as buckets can be queried through a REST API.

In line with the cloud approach, the cluster is running on the open-source framework Kubernetes to deploy and manage containerized services. The choice of containerization is fundamental as it tackles the two main issues pertaining to data processing environments: managing concurrency in access to processing resources (RAM, CPUs, GPUs..) while properly isolating the running services from one another. As a result, users can both freely tailor services to their needs (programming language, system libraries, packages and their versions, etc.) while scaling their applications to the computing power and storage capacities it demands, e.g. by distributing computations over several containers.

Besides autonomy and scalability, these architectural choices also foster reproducibility of statistical computations. Contrary to traditional IT infrastructures — either a personal computer or a shared infrastructure with remote desktop access — the user must learn to deal with resources which are by nature ephemeral, since they only exist at the time of their actual mobilization. This fosters the adoption of development best practices, notably the separation of the code — put on an internal or open-source forge such as GitLab or GitHub — the data — persisted on a specific storage solution, such as MinIO — and the computing environment. The projects developed in that manner are usually more reproducible and portable — they can work seamlessly on different computing environments — and thus also more readily shareable with peers.

2.3. A catalog of services which covers the entire lifecycle of a data science project

The aim of the platform is to provide statisticians an environment to prototype their data science projects end-to-end. To do so, it offers a wide range of services which cover the entire lifecycle of a data science project:

- Data services: as aforementioned, the platform offers an object-storage service, but also various databases (PostgreSQL, MongoDB, ElasticSearch..)
- Execution environments : RStudio for R processing, Jupyter and VSCode for Python processing, distributed computation engines (Spark, Dask, Trino)
- Automatization tools : a batch deployment service (argo-workflow), a GitOps deployment service (argo-cd) and a MLOps service (MLFlow)
- Dissemination tools : visualization/BI services (Redash, Superset) and an API management platform (Gravitee)

2.4. Providing reproducible environments for training programs

The fact that deployed services are simply containers running on a centralized infrastructure makes it very easy to provide reproducible environments for training programs. Teachers can provide trainees with environments tailored to the specific needs of their training program, e.g. by pre-downloading necessary data, packages, etc. Besides, these environments can be deployed directly in the training catalog of the platform, so that trainees can launch them using a simple URL. This ensures a seamless training experience for both teachers and trainees.

2.5. A fully open-source project aimed at fostering reusability

In order to improve reusability, an open-source project was developed in order to make the deployment of similar state-of-the-art data science platforms possible in other organizations. The full code-source is available on GitHub and an accompanying documentation website thoroughly details the steps needed to instantiate a platform.

3. Results

The platform is now widely used in the national statistical system and even beyond, with about 500 unique users per month. These users form a dynamic community which, through the use of a centralized discussion canal, help improve the experience by reporting bugs and suggesting new features or even directly contributing to the codebase. It is also used in several data science schools and universities to host courses on statistical languages and frameworks. Finally, it is used to host innovative events such as hackathons, both at the national and international level.

The open source project has also been getting a lot of attention. Multiple organizations already have instantiated a platform or plan to do so, both at the national and international level, and also in the private sector. Some of these organizations are also interested in contributing to the code base in the near future.

4. Conclusion

We developed a modern data science platform which aims at making data scientists in NSIs more autonomous by providing them with scalable computing resources and a wide range of modern data science services to prototype their projects from end to end. In order to encourage reusability, an open-source project is also developed to facilitate the instantiation of similar data science platforms in other organizations.

References

[1] EUROSTAT, Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics) (2018). https://ec.europa.eu/eurostat/fr/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted

[2] UNECE, Big data and modernization of statistical systems, Report of the Secretary-General, 45th Statistical Commission (2013). https://unstats.un.org/unsd/statcom/doc14/2014-11-bigdata-e.pdf

[3] UNECE, HLG-MOS Machine Learning Project (2021). https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project

[4] S. Luhmann, J. Grazzini, F. Ricciato, M. Meszaros, K. Giannakouris, J.M. Museux, & M. Hahn, Promoting reproducibility-by-design in statistical offices, Proceedings for New Techniques and Technologies for Statistics (NTTS) (2019).

Mapping circular economy EU funded projects using text mining

Keywords: Circular economy; Text mining; EU funds; Europe.

INTRODUCTION

Circular economy refers to an economic and operating system where the linear industrial model (taken, use and dispose) is re-thought to reduce (or eliminate) waste and to ensure a more efficient, rational and sustainable use of resources. It implies to make smarter use of products, resources and materials, as well as to extend the lifespan of products and its parts [1]. Transition to a circular economy is one of the main priority areas of the European Green Deal, as part of the Commission's objective to become a climate neutral economy by 2050. Closing the loop will allow to make a more efficient use of the natural resource to ensure a more sustainable supply of raw materials and to contribute to the preservation of ecosystems and biodiversity.

To support policymaking and the design of a circular economy agenda, a clear understanding of the territorial performance in using public money to support circular economy is needed. However, existing universal taxonomies to classify EU funded projects, e.g. EUROSTAT classification[85] or European Regional Development Fund's (ERDF) thematic objectives or categories of intervention, do not allow having a full picture of the circular economy projects funded by EU grants. An example is the development a new operating system or software able to reduce or avoid products defects, which leads to waste reduction and a more efficient use of resource. Other examples can be related to the development of a biodegradable plastic packaging or using the residues of cork industry on building construction or renovation to improve the insulation of houses and then reduce energy consumption. One way of mapping these activities is using text-mining analysis. This technique refers to the process of extracting knowledge from text documents [2].

The aim of the present paper is twofold. First, it aims to identify ERDF projects in the field of circular economy. Second, show the utility of text mining analysis to support policy monitoring, by providing a more accurate analysis compared to a NACE code taxonomy. The originality of the paper lies on being the first to provide a clear mapping of circular economy ERDF investments in the EU28, in the period 2014-2020, as well as to test the accuracy of text-mining methodology.

[85] The EUROSTAT has developed a methodological approach using NACE classification codes related to recycling, reuse and repair to delimitate circular economy-related economic activities. However, this classification does not permit to identify circular economy projects when the development or adoption of more eco-efficient technologies is happing in manufacturing sector.

We focus on ERDF because this EU instrument aims to reinforce economic and social cohesion by correcting inequalities between its regions. The transition to a low-carbon economy is one of the key priority areas for ERDF investment.

DATA AND METHODS

To conduct our analysis we use the JRC-WIFO ERDF database [3]. This database comprises around 600,000 observations on ERDF project beneficiaries during the 20142020 period providing a unique coverage and level of details on the ERDF operations. Based on a list of keywords, several text algorithm runs are made on the text of the projects descriptions to identify those subsets of investments related to "circular economy". Ex-post quality checks are then made on the resulting sample via an iterative process to refine the final list of keywords. We also combined text-mining analysis applied to projects' description together with the NACE codes of beneficiaries operating in circular economy sectors, according to the EUROSTAT classification.[86] Both criteria are not mutually exclusive, meaning that a project is classified as related to the circular economy using one of the two criteria alone or together.

The list of keywords used to identify projects in circular economy includes around 170 words or expressions coming from different sources: (i) European Science Vocabulary (EuroSciVoc); (ii) synonyms of the dimensions' described in Table 1 (extracted from scientific literature) using Thesaurus; (iii) keywords or/and taxonomies provided by external experts [4].

Table 1. Dimensions of Circular Economy's concept

Dimensions	Description
Reduce	▪ Reduction of waste or residue on production process (e.g. re-thinking the design or packaging of product) and by improving the use of products by consumers (e.g. re-thinking ownership, by sharing or renting)
Recycle	▪ Using waste or residue to produce a new product, after its processing (recycle) ▪ Employing parts of a discarded product in a new product with the same function (remanufacture) or with a different one (repurpose)

[86] **Recycling**: 38.11 - Collection of non-hazardous waste; 38.12 - Collection of hazardous waste; 38.31 - Dismantling of wrecks; 38.32 - Recovery of sorted materials; 46.77 - Wholesale of waste and scrap; 47.79 - Retail sale of second-hand goods in stores. **Repair and reuse**: 33.11 - Repair of fabricated metal products; 33.12 - Repair of machinery; 33.13 - Repair of electronic and optical equipment; 33.14 - Repair of electrical equipment; 33.15 - Repair and maintenance of ships and boats; 33.16 - Repair and maint. of aircraft and spacecraft; 33.17 - Repair and maint. of other transport equip.; 33.19 - Repair of other equipment; 45.20 - Maintenance and repair of motor vehicles; 45.40 - Sale, maint. and rep. of motorc. and related parts and acc; 95.11 - Repair of computers and peripheral equipment; 95.12 - Repair of communication equipment; 95.21 - Repair of consumer electronics; 95.22 - Rep. of household appliances and home and garden equip.; 95.23 - Repair of footwear and leather goods; 95.24 - Repair of furniture and home furnishings; 95.25 - Repair of watches, clocks and jewellery; 95.29 - Repair of other personal and household goods.

Reuse	▪ Reuse of a product by another consumer (e.g. selling of second-hand good) ▪ Repair or maintenance of malfunctioning equipment/machinery or damaged product to extend its lifespan ▪ Restore an old product to bring it up to date
Recover	▪ Recovering the energy produced by waste elimination (e.g. through incineration of materials)

Source: Authors' own elaboration based on [1] and [5].

RESULTS

Identifying circular economy ERDF project: text-mining *versus* NACE codes taxonomy

Table 2 reports the frequency of keywords-related to different circular economy taxonomies as the result of text-mining analysis. Most of the ERDF projects in circular economy are associated with resource efficiency and input optimisation, as well as recycling, repair and reuse.

Table 2. Frequency of keywords-related to circular economy by taxonomy

Taxonomy	Frequency
Green building design (energy efficiency or saving)	21,926
Recycling (reuse by-product, waste)	9,514
Repair and restore (life extension)	8,157
Resource and/or production efficiency (waste minimisation, input optimisation)	6,910
Bio-based product (biomass, biogas, biofuel)	4,224
Resource recovery (energy, water, heat, carbon)	946
Circular design (eco-design and packaging)	106

Source: Own elaboration based on the results of text-mining analysis. Note: A project can be classified in more than one taxonomy.

If we compare the results of project identification using text-mining analysis and NACE codes taxonomy (Table 3), we can observe that using the latter leads to strongly underestimate and identify ERDF projects in circular economy (311 Million versus 23 billion).

Table 3. ERDF projects in circular economy, 2014-2020 (EUR amount of funds), by categories

Categories	EU fund share	% Total
Total ERDF	187,735,241,730	100%
Total circular economy (NACE codes)	310,986,249	0.2%
Total circular economy (Text-mining)	22,694,224,333	12.1%
Total circular economy (both)	22,930,031,811	12.2%

Source: Analysis performed using JRC-WIFO database [3] and text analysis techniques to identify projects in circular economy.

Analysing territorial patterns of ERDF circular economy projects

Regions in Eastern European countries and Greece are the ones with the highest values of ERDF per capita spent on circular economy projects (Figure 1) however, this geographical distribution is strongly correlated with the countries/regions allocations of funds coming from ERDF. The share of ERDF projects in circular economy over total ERDF projects (Figure 2) shows the concentration/specialisation of regions in circular economy. In addition to regions in Eastern European countries and Greece, some regions in the Netherlands, Belgium, Finland, Austria, Germany and the United Kingdom report a share of circular economy projects (measured by EU funds) higher than 22%.

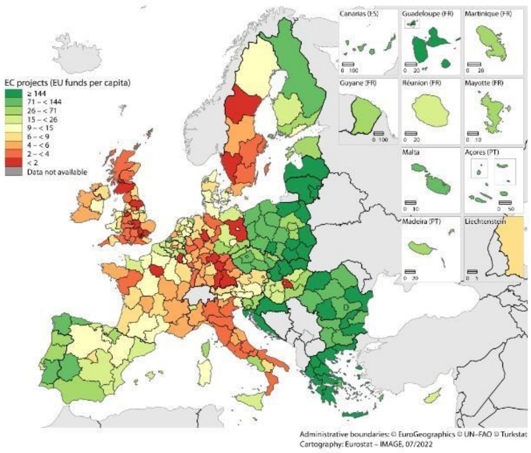


Figure 1. ERDF Circular Economy EU funds (EUR) per capita, 2014-2020

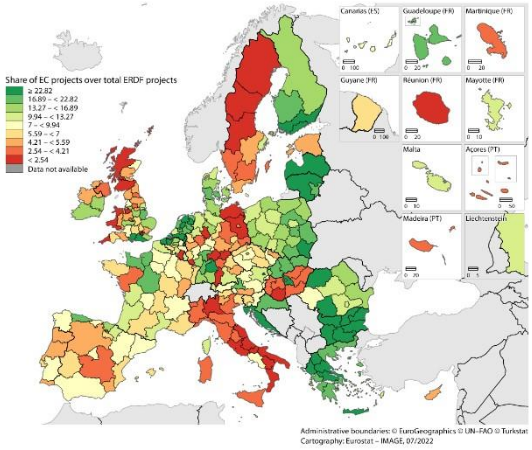


Figure 2. Share of Circular Economy projects, projects over total ERDF projects (% EU funds), 2014-2020

Source: Analysis performed using JRC-WIFO database [3] and text analysis techniques to identify projects in circular economy.

cONCLUSIONS

Policy evaluation is a key element in the policy cycle, especially in the context of a transition to climate neutral economy. It helps to understand policy choice and regional patterns to support more effective policies. The present study provides for the first time a mapping of the circular economy projects funded by ERDF in the programming period 2014-2020. The analysis takes advantage of the novel dataset [3] which includes the description of projects co-funded by the ERDF in 2014-2020.

The analysis is performed combining text-mining analysis technique. Findings show that the use of this methodology can provide a more accurate and full picture of the regional location of ERDF circular economy projects in comparison with NACE code taxonomy.

rEFERENCES

[1] Potting, J.; Hekkert, M.; Worrell, E. and Hanemaaijer, A. (2017). "Circular Economy: Measuring Innovation in the Product Chain", Policy Report, PBL Netherlands Environmental Assessment Agency, The Hague.

[2] Gaikwad, S.V.; Chaugule, A. and Patil, P. (2014). "Text Mining Methods and Techniques", *International Journal of Computer Applications* 85(17):42-45.

[3] Bachtrögler, J., Arnold, E., Doussineau, M., and Reschenhofer, P. (2021). *UPDATE: Dataset of projects co-funded by the ERDF during the multi-annual financial framework 2014-2020*, JRC125008.

[4] Technopolis (2022). *Background paper – Workshop on the development of the ERA circular industrial technologies roadmap for energy-intensive industries, textiles and construction*, 31 May and 1 June 2022, Mimeo.

[5] Kirchherr, J.; Reike, D. and Hekkert, M. (2017). "Conceptualizing the circular economy: An analysis of 114 definitions", *Resources*, *Conservation & Recycling* 127: 221-232. https://doi.org/10.1016/j.resconrec.2017.09.005

Integrating Big Data in the traditional survey of maritime transport statistics: open challenges

Keywords: Big Data, AIS, port calls, tracking ships, maritime transport statistics

INTRODUCTION

The scope of the statistics covered by Regulation (EU) No 1090/2010 amending Directive 2009/42/EC refers to the carriage of goods and passengers by seagoing vessels calling at ports in the territories of the reporting countries. Carriage of goods and passengers by sea means the movement of goods and passengers using seagoing vessels, on voyages which are undertaken wholly or partly at sea. Vessels with gross tonnage greater than 100 are included within the scope of the Regulation when they carry goods and/or passengers only for commercial purposes.

The aim of this work is integrating Big Data with traditional data sources to increase the quality of the maritime traffic statistics and the timeliness of data dissemination. Automatic Identification System (AIS) is an automatic tracking system on ships extensively used in the maritime world for safety and management purposes. AIS provides a Big Data source containing information regarding the identity and characteristics of the ships, as well as details on their location and speed during navigation, landing and anchorage. Eurostat invited countries to transmit quarterly data on vessels calling at ports (set F2 of the maritime statistics legal act) after each reference quarter instead of providing all quarters at the end of the year. Italy has not yet responded to Eurostat's request and provided quarterly dataset F2 after the end of each quarter. In support of the Italian survey of maritime transport, TRAMAR [1], AIS data would allow to anticipate the estimates of ships arriving in Italian ports, as requested by Eurostat.

The use of AIS data to support the official statistics about port visits has been investigated in several existent works. In [2] the *Port Visits Geo-Solution prototype* monitors the ships movement inside the Piraeus Central Port, defined by a polygon, in order of computing the number of arrivals and departures. In the United Nations Global Platform (UNGP) Handbook page [3], several case studies on the use of AIS are described, such as the experimental statistics of the daily number of vessels visiting Danish ports using AIS data published from Statistics Denmark [4]. In [5] two methodologies for generating port visits are compared. The first one, in particular, uses polygons to identify ships inside a port area, as in [2]. As well the aforementioned works, we are interested in calculating the number of visits in the ports of Italian coasts. However, our algorithm returns also the entire travel of a vessel, meant from port of origin to port of arrival, where origin or destination of a travel can be located outside the Italian coasts. Furthermore, similarly to [2] but unlike [4], we distinguish an arrival from a departure from the movement and the speed variation of a ship along the time, instead of using the navigational status in AIS data, that it turned out to be not very reliable when considering port stops of short duration.

In this work we present a methodology to use AIS observations referred to ships calling Italian ports, in a given time period, to reconstruct the entire ships' travels. We use the identified ships' travels to compute statistics about the number of arrivals and departures from each Italian port. We compare the results with Istat survey TRAMAR [1].

In the following sections we are going to explain our algorithm and to show our experimental results.

METHODOLOGY

Data source

AIS data used in this work were supplied by the Task Team on AIS Data of the UN Committee of Experts on Big data and Data Science for Official Statistics (Task Team on AIS Data - UN-CEBD [6]). These data were accessed through the UN Global Platform (UNGP) [7], which holds a global repository of live and historical AIS data (UN-AIS).

UN-AIS dataset contains both live and global archive data from 1st December 2018. Periodical observations of all types of ships are registered. The time interval between two observations of a ship is 10/11 minutes. The available attributes comprise of three information categories: static data (MMSI code, IMO code, vessel name and type), dynamic data (coordinates of ship's location, navigational status, speed and course), and voyage-related data (destination and ship's draft). The UN-AIS dataset makes available also the ships position by the H3 index in multiple resolutions. H3 [8] is a geospatial indexing system that approximates the GPS coordinates through a hexagonal tessellation of the earth's surface. The H3 index identifies the hexagon containing the ship's coordinates and the hexagon size depends on the adopted resolution.

Besides AIS dataset, we used a world ports dataset taken from [9].

AIS-based algorithm

We present an algorithm to compute vessels' travels by exploiting AIS data. A travel consists in: vessel identifiers, date and port of departure, date and port of arrival, where at the most one of the ports can be located outside the Italian coasts. The first crucial step of this algorithm is the identification of the port calls of a vessel. This step could be performed through the navigational status attribute present in the AIS observation for a vessel, as in [4]. Unfortunately, this attribute turned out to be not really accurate for some types of the ships we considered, such as small passenger ships. Thus, in our algorithm we identify a port call if a vessel is located in a port area and has a speed value close to zero (less than 0.2 knots), as we want to analyze only almost stationary ships. We define a port area by using the H3 index [8]. In particular, we choose an index of resolution 7 and we consider both the center and the ring of level 1 (see the example in Figure 1), as this choice resulted the most suitable for the various types of port analyzed in this work. As we map every world's port through an automatic procedure, it may happen that part of the polygons cover an area that is out of scope (i.e. land instead of sea).

The whole procedure consists of the following steps.

1. We create a lookup table made up of pairs <H3 index, port name> for each port of the world.
2. We reduce data by filtering UN-AIS data to select the list of vessels we are interested into. Only AIS observations related to ships passing close to Italian ports, in a given time period T (i.e. an interval [start_date, end_date]), are taken. Distinct MMSI identifiers of the so

collected AIS data are used to filter all the AIS observations in T. Furthermore, AIS observations are filtered by type (passengers, cargo and tanker) and speed close to zero (< 0.2 knots).

3. For each vessel, we identify the port calls, sorted by MMSI and timestamp.
4. We match the different port calls to calculate the vessel travels.
5. We calculate statistics about the number of arrivals and departures in each Italian port in the time period T, in order to compare the results with Istat TRAMAR survey.

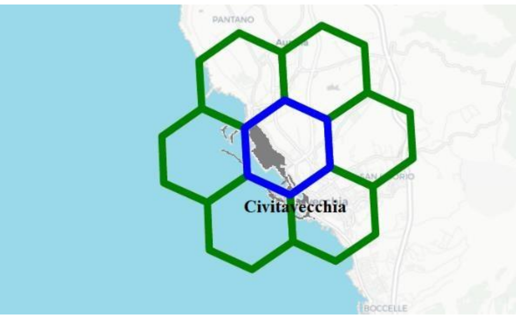


Figure 1. The port area of Civitavecchia (UNLO Code ITCVV) using the H3 index of resolution 7 - central hexagon and the 6 other hexagons that compose the level 1 ring.

EXPERIMENTAL RESULTS

We present experimental results of our AIS-based algorithm for two very different ports, in the month of September 2021. The first one, Carloforte (ITCLF), is a little port located in the small island of San Pietro (Sardinia) and is only used from passenger ferries that connect San Pietro Island to the nearby mainland. Here port calls are frequent and of very short duration. The second one, Civitavecchia (ITCVV), is a big port near Rome and is used both from passenger (including cruises) and commercial lines. Here port calls are less frequent and of long duration. Our algorithm output compared with TRAMAR survey results highlight some differences both in the number of ships and port calls (see Table 1).

Table 1. Comparison between AIS data and TRAMAR survey data: number of port calls for September 2021

Port name	UNLO Code	Month	Ships in TRAMAR	Ships in AIS	Port calls in TRAMAR	Port calls in AIS-based algo
Carloforte	ITCLF	Sept - 21	4	6	1,774	1,842
Civitavecchia	ITCVV	Sept - 21	83	81	551	414

In Carloforte port, our algorithm detects 2 more ships than TRAMAR, for an amount of 230 port calls, one of which is out of scope for the Regulation (EU) No 1090/2010. Otherwise, for the 4 common ships, AIS has 162 fewer port calls (i.e. about 9%). This difference has two reasons: (1) part of AIS messages of the ships have been lost, thus our algorithm does not detect some landings; (2) other landings are lost because port stops take less than 10 minutes and the ships do not appear in the AIS data with speed 0 (we verified that this happens during the early hours of the morning).

The number of ships visiting the port of Civitavecchia found from our algorithm is very similar to the number of ships resulting from TRAMAR survey even if, by comparing the ships' IMO codes, only 70 of them result to be in common. However, these 70 ships make 97% of marine traffic for Civitavecchia. For 7 of these ships, our algorithm returns a lower total number of port calls than TRAMAR. The amount of lost records is 135. In this case the main reason of mismatch is the loss of part of AIS data. Moreover, 13 ships are detected from TRAMAR but not from AIS and this is due to three causes: (1) the ship is anchored outside the port and never lands inside it (see Figure 2), so our algorithm does not link the stop to the port while in TRAMAR survey this port call is registered; (2) the ship never reaches the port but this call is in TRAMAR because of an error in the questionnaire compilation; (3) AIS data are lost during the period of landing in the port. Finally, 11 ships (for an amount of 15 port calls) are detected from our algorithm and not from TRAMAR. In this case, either the ship is out of scope of the survey or we discovered a missing record for the survey.

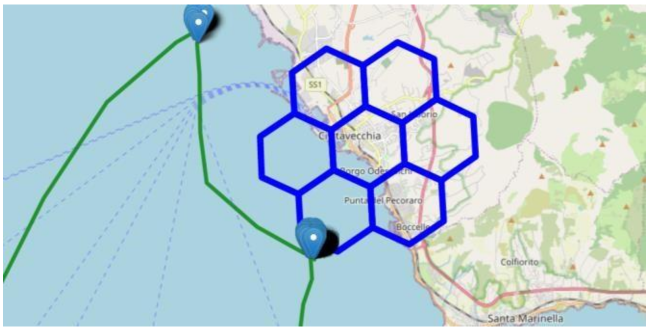


Figure 2 – In this example the ship is anchored out of the Civitavecchia port area, delimited by the 7 blue hexagons, and never lands inside it.

CONCLUSIONS

We presented an algorithm based on AIS data, taken from the UN-AIS dataset, to compute the entire travels of ships calling in Italian ports. We performed experiments by choosing two different kind of ports and we used our results to evaluate the quality of AIS data and of the implemented algorithm. We made an in depth comparison of our output with the Istat survey TRAMAR statistics. We obtained very encouraging results that suggest that integrating AIS data with traditional sources can improve the quality of statistical production. Some open issues remain, the more relevant is the temporary lack of AIS observations. Thus, only facing this problem we could use AIS data to improve the timeliness to generate F2 table for Eurostat. As a future work, we could estimate the port calls missing because of gaps in AIS data through a statistical model to be designed.

REFERENCES

[1] Istat TRAMAR survey - https://indata.istat.it/tramar/index.php

[2] ESSnet Big Data II WPE – Tracking ships – Deliverable E4: Consolidated report on project results (2020-11-19)

[3] United Nations Global Platform (2020). *AIS Handbook Online*. Geneva: United Nations - https://unstats.un.org/wiki/display/AIS/Case+studies

[4] AISDAG: Daily number of vessel (2019) - https://www.statistikbanken.dk/aisdag

[5] Port Visits Using Real-Time Shipping Data – CSO (2022) https://www.cso.ie/en/releasesandpublications/fp/fp-pvrts/portvisitsusingrealtimeshippingdata/datasourcemethodsandquality/

[6] UN AIS Data Task Team - https://unstats.un.org/bigdata/task-teams/ais/index.cshtml

[7] UN Global Platform - https://unstats.un.org/bigdata/un-global-platform.cshtml

[8] H3 (Hexagonal hierarchical geospatial indexing system) - https://h3geo.org/

[9] C. Merrien, Worldwide list of seaports, version 2021 http://dx.doi.org/10.12770/59ab5f6f-79ea-425d-830e-be5ecdb7bdbe

How well-behaved are revisions to quarterly fiscal data in the euro area?

Keywords: Real-time data, data revisions, fiscal policy

Introduction

Most macroeconomic data are revised after the initial release. Revisions originate from various sources with new information becoming available by the time of subsequent releases being the most obvious cause. Conceptual changes to statistical definitions and to compilation and estimation methods constitute another reason. In the case of intra-annual statistics that require seasonal adjustment the revisions may also originate from a re-estimation of seasonal factors. Finally, simple correction of errors and elimination of omissions that take place in the context of a data production process may also lead to data revisions.[87]

Whatever the source of the revisions given their common existence they should be taken as a fact of life. In this context, researchers and policy-makers have no choice rather than understanding them. Only a proper recognition of revisions enables the application of optimal statistical methods that lead to sound analytical conclusions.[88] In the same vein, an acknowledgement of revisions is necessary to place an adequate trust in data available at the time when a policy decision is formed.[89] This paper analyses revisions to quarterly fiscal data in the euro area. Its main objective is to determine how well-behaved fiscal revisions are, especially by contrasting them with macro revisions. To this end, we check to which extent the properties of well-behaved revisions, as outlined by Aruoba (2008), are fulfilled. The criteria are based on the following three characteristics: (1) zero bias, (2) little dispersion and (3) unpredictability given the information available at the time of the initial announcement.

Methods

To answer our research question, we derive a broad set of statistics that allow us to assess all three requirements for well-behaved revisions. To this end, by calculating the mean of revisions we check the degree of a bias across fiscal variables. Moreover, we assess the extent of dispersion in revisions using several indicators. Finally, by running a set of regression models we verify whether revisions are predictable given available information at the time of the initial release. To put the results into perspective, we contrast fiscal revisions with macro revisions, which are significantly better understood in the economic literature. The main contribution of this paper is to deliver a comprehensive analysis of revisions to quarterly fiscal data in the euro

[87] Carson et al. (2004) provides many useful clarifications on statistical revisions, including on typology and terminology.

[88] Multiple studies underline the usefulness of real-time fiscal data for fiscal forecasting, budgetary surveillance or identification of fiscal shocks (see, e.g. Pedregal & Pérez (2010), Asimakopoulos et al. (2020) and Cimadomo (2016)).

[89] Orphanides (2001) in its seminal contribution demonstrates the complexity of policy decision-making in real time. Most notably, the study emphasizes that policy recommendations obtained with real-time data are considerably different from these based on ex-post revised figures.

area. The literature studying revisions to quarterly macroeconomic data has been growing for decades and by now it is very rich (see a literature survey in Croushore (2011)). A large bulk of the literature, like Mankiw & Shapiro (1986), concentrates on the primary indicator of economic activity, which is GDP or GNP. Other papers suggest extensions along various dimensions. Shrestha & Marini (2013), for example, investigate whether the magnitude of revisions to GDP differs during crisis episodes. Also, there are studies analysing revisions to a broader set of economic indicators going beyond the measures of output (see, e.g. Aruoba (2008) for the US, Branchi et al. (2007) for the euro area, Faust et al. (2005) for G7 economies).

According to our best knowledge, no study exists that analyses revisions to the euro area quarterly fiscal data in a comprehensive way. The literature on revisions to fiscal statistics established so far concentrates on annual data often with a view to shedding light on fiscal discipline and budgetary frameworks. De Castro et al. (2013) use real-time vintages of annual budget balance to evaluate the quality of initial data releases, on the basis of which compliance with the fiscal rules is assessed. Maurer & Keweloh (2017) attempt to answer the question whether the quality of annual fiscal data provided in the context of the Excessive deficit procedure (EDP) improved over time in the EU. As far as we are aware, Asimakopoulos et al. (2020) demonstrating usefulness of real-time fiscal data for forecasting purposes, is the only study that provides some limited characteristics of revisions to quarterly fiscal series for the biggest four euro area economies (i.e. Germany, France, Italy and Spain). As concluded in the literature survey on real-time data and fiscal policy analysis in Cimadomo (2016), "more work is needed in this field". With our analysis we try to fill the gap.

Another significant contribution of our study is the creation a real-time fiscal quarterly dataset for the euro area countries. The ability of researchers to conduct real-time analysis depends on real-time datasets, which collect in one place data available at any point in the past. In the US two comprehensive real-time datasets exist next to each other, namely Real-Time Data Set for Macroeconomists by Federal Reserve Bank of Philadelphia (see Croushore & Stark (2001)) and ArchivaL Federal Reserve Economic Data (ALFRED) by the Federal Reserve Bank of St. Louis (see Stierholz (n.d.)). Also, significant efforts have been made to establish a real-time dataset for the euro area (see Giannone et al. (2010)). We contribute to this work by collecting all vintages of Government Finance Statistics for the euro area countries since their publication started in mid-2000s.

Results

Our investigation first concludes that fiscal revisions, like macro revisions, do not satisfy desirable properties expected from well-behaved revisions. This finding is not only relevant for final revisions but it also holds for intermediate revisions. Fiscal variables exhibit a positive bias since most of them grows in annual terms by 0.1-0.3 percentage points more compared to what is published initially. Given the average growth in the sample of around 4% the value of the bias is non-negligible.

Second, the dispersion of fiscal revisions tends to be relatively sizable. Mean absolute revision - our most intuitive summary statistic - amounts to around 1 percentage point for the annual growth in the biggest and most stable categories. It reaches significantly higher values for small and volatile items, most notably government investment. Our analysis also indicates that fiscal revisions became significantly smaller since 2014, which is the moment of the ESA 2010 introduction. While the mean absolute revision for the biggest and most stable categories considerably exceeds 1 percentage point in the first subsample (up to 2014Q2) it is significantly lower than 1 percentage point in the second subsample.

Third, fiscal revisions are in general predictable. While the degree of predictability varies significantly across the variables it is substantial for many of them. The conditional mean with respect to the information available at the time of the initial release is statistically different from zero. As such, revisions do not only reflect new incoming information but also the information known at the time of the initial publication. This feature also speaks in favour of treating fiscal revisions as ’badly’ behaved.

When contrasted with macro revisions, fiscal revisions are quite comparable. Both fiscal and macro revisions are associated with a positive bias of a similar order. At first sight, fiscal revisions appear to be significantly more dispersed than macro revisions, as measured by the mean absolute revision, for instance. We document, however, that since 2014, when the magnitude of fiscal revisions narrowed down considerably, both types are revisions are in the same ballpark. Also, the degree of predictability does not appear to differ between the two types of variables. In this context, we contradict the often heard view that fiscal data in general are subject to particularly large revisions (see, e.g. Cimadomo (2016)).

Conclusions

Our investigation concludes that fiscal revisions are badly-behaved. They fulfil none of the requirements for well-behaved revisions. More specifically, (1) fiscal revisions exhibit a positive bias, (2) they are characterised by a considerable dispersion and (3) they are in general predictable with the information available at the time of the initial release.

While our analysis concludes that fiscal revisions are badly-behaved it is difficult to find support in the data that they are worse than macro revisions presently. Macro revisions are also badly-behaved, which has been already documented in the literature (see, e.g., Faust et al. (2005)). The extent of this ’misbehaviour’ is just similar for the two types of variables. Both macro and fiscal revisions exhibit similar bias and they are subject to a comparable dispersion, most notably since 2014 when fiscal revisions became more contained. Moreover, no major difference emerges in the analysis between the two types of revisions when it comes to predictability. Supplementing the analysis with the intermediate revisions leaves the conclusions unchanged. Notwithstanding this, intermediate revisions do bring additional information to the study. Most notably, they make clear that fiscal variables converge to final values differently from macro variables. While for the former the revisions tend to take place in April and October a more evenly distributed revision pattern is observed for the latter.

References

Aruoba, S. B. (2008), ‘Data revisions are not well behaved’, *Journal of Money, Credit and Banking* 40(2-3), 319–340.

Asimakopoulos, S., Paredes, J. & Warmedinger, T. (2020), ‘Real-time fiscal forecasting using mixed-frequency data’, *The Scandinavian Journal of Economics* 122(1), 369–390.

Branchi, M., Dieden, H. C., Haine, W., Horváth, C., Kanutin, A. & Kezbere, L. (2007), Analysis of revisions to general economic statistics, Occasional Paper Series 74, European Central Bank.

Carson, C. S., Khawaja, S. & Morrison, T. K. (2004), Revisions policy for official statistics: a matter of governance, Working Paper 04/87, IMF.

Cimadomo, J. (2016), ‘Real-time data and fiscal policy analysis: A survey of the literature’, *Journal of Economic Surveys* 30(2), 302–326.

Croushore, D. (2011), ‘Frontiers of real-time data analysis’, *Journal of Economic Literature* 49(1), 72–100.

Croushore, D. & Stark, T. (2001), ‘A real-time data set for macroeconomists’, *Journal of Econometrics* 105(1), 111–130. Forecasting and empirical methods in finance and macroeconomics.

De Castro, F., Pérez, J. J. & Rodríguez-Vives, M. (2013), ‘Fiscal data revisions in europe’, *Journal of Money, Credit and Banking* 45(6), 1187–1209.

Faust, J., Rogers, J. H. & Wright, J. H. (2005), ‘News and noise in g-7 gdp announcements’, *Journal of Money, Credit, and Banking* 37, 403 – 419.

Giannone, D., Henry, J., Lalik, M. & Modugno, M. (2010), ‘An area-wide real-time database for the euro area’, *Review of Economics and Statistics* 94.

Mankiw, N. G. & Shapiro, M. D. (1986), News or noise? an analysis of gnp revisions, Working Paper 1939, National Bureau of Economic Research.

Maurer, H. & Keweloh, S. (2017), ‘Quality enhancements in government finance statistics since the introduction of the euro: Econometric evidence’.

Orphanides, A. (2001), ‘Monetary policy rules based on real-time data’, *American Economic Review* 91(4), 964–985.

Pedregal, D. J. & Pérez, J. J. (2010), ‘Should quarterly government finance statistics be used for fiscal surveillance in europe?’, *International Journal of Forecasting* 26(4), 794 – 807.

Shrestha, M. L. & Marini, M. (2013), Quarterly gdp revisions in g-20 countries: Evidence from the 2008 financial crisis, Working Paper 13/60, IMF.

Stierholz, K. (n.d.), ‘Alfred: Capturing data as it happens’, *Presentation by Federal Reserve Bank of St. Louis (https://alfred.stlouisfed.org/docs/alfred_capturing_data.pdf)* .

An Android app for continuous census management

Keywords: metadata, data security, agile methodology, Android app

iNTRODUCTION

According to Italian Institute of Statistics (Istat) regulation, Population census is based on two yearly sampling survey, respectively referred to as List and Area. While the former uses the individual registry and aims to collect all microdata of interest, the latter, or Areal sampling survey, enumerates individuals who actually inhabit in the territoryIn order to guarantee their goals, the fieldwork is extremely important. In fact, in case of List survey, the enumerator has to work on field to recovery non-respondent people, on the other side, in case of Areal survey, the enumerator has to verify addresses, to identify households and to interview resident population in them.

In order to facilitate the work and to eliminate the use of paper at all, Istat has supplied tablets to municipalities, consequentially IT department developed different technical solutions complaint with the use of the tablets. During the first and second wave of Population census (2018 and 2019) enumerators used the same web application for online and offline operations. The advantage of this approach was that users had to learn and to update only one system: moreover, they did not install any software on tablet because only the browser was mandatory. They could independently choose when going in offline mode or manage the system in offline mode, in case of no net coverage. Despite these advantages, a careful analysis of the problems emerged during the data collection phase showed some issues. Often enumerators were confused by skipping from online to offline. Furthermore, even if a web application was responsive to all types of devices, it did not showed information in the best way on tablet. These issues, linked with the possibility of the development of a mobile app, led Istat to give another possibility for fieldwork; so, in the 2021, it has made available an app called Rilevo that is able to manage every functionalities needed on field. It is an Android mobile app, presented on the Samsung tablets provided by Istat and it makes available features both for verifying the territory and for interviewing families. It works only in offline mode: data locally managed will be synchronized "on demand" with those present on the central server. It has developed using user centricity approach to obtain the best result in terms of usability. [1]

The paper explain how Istat managed the developing, focusing on the possibility of its reuse for other surveys. It is organized as follow: it starts with a focus on the designed mode and functionalities then it goes on a highlight on synchronization of data and metadata and on the modeling of questionnaire.

mETHODS

Design mode and functionalities

Istat considered the Agile methodology the best solution for design: during the initial phase, the stand-up meetings were useful for as-is analysis and then finally arriving at the to-be proposal.

Using Mural tool, all user requirements are collected and all possible use cases are investigated. Code development follows the standards of Agile methodology, providing for regular releases every two weeks and for related acceptance tests. Rilevo is developed with Dart language and built with Flutter framework. Data are locally stored using SQLite, encrypted with the 256-bit AES algorithm via the SQLCipher extension. In addition to the mobile app, the architecture also includes a Java Web Application (jdk 8), called API Gateway, based on the Spring framework and released on Apache Tomcat 8. The Gateway manages secure access and authentication through Shibboleth, the Istat Identity Provider; furthermore, it exposes Restfull services as gateways to micro-services currently used by online applications for survey management and data collection.

Whenever a new version of the app is available on the Play Store, a notification appears on the home page. The enumerator can both decide to update it autonomously or wait for the update to be carried out by the operating system in the next 24 hours.

Only enumerators with assigned units to work can access to the app menu. Rilevo provides a series of different functions both for the management of the operation relating to the Areal survey and for those relating to the List survey. For example, it is possible to check the addresses, identify the households and enter the individuals residing in them. For both Areal and List survey, the questionnaire is available and then the enumerators can conduct interviews.

Data and metadata synchronization

Rilevo app fits any new configuration without the need for intervention on the code. It is a container that starts empty and that specializes through the configurations it receives from the server.

The communication between Rilevo and the server takes place through a synchronization service that provides using specific calls, all the data necessary for the operation of the app itself. This service sends a series of metadata that allow the configuration and the creation of SQLite databases structure. Configuration means also the behavior of functionalities that enumerators can use and the structure and navigation rules concerning the Questionnaire. All this information arrives from two web applications related to census management: the first one is SGI that cooperates all the collection phases in the field, while the second one is the online Questionnaire, for data collection.

The synchronization performs a security check on users who access the app, verifying that they are authorized to carry out operations for the specific survey. An additional security check is the sending of the only enabled data to work. The synchronization phase is bidirectional and the server receives the data coming from the client. In this process, which is carried out at runtime and at the request of the user, the app sends offline processed data to the service; the synchronization processes them and validates each individual record received. If this validation is successful, the data is saved on the server to be available updated both in the online SGI web application and on the app. If validation is not successful, data are discarded. The saving and retrieval of data takes place in a distributed and transactional way on different Oracle Databases, distributing the according to their context (e.g. SGI, Questionnaire and micro-services).

Questionnaire

Rilevo implements a generalized engine for rendering and compilation of electronic questionnaires through Computer-Assisted Personal Interviews (CAPI) technique. This means that the structure of the questionnaire, the partition into sections and pages, the number and the type of questions, the layout, the multilingual texts and the rules guiding the completion of the questionnaire are defined through a set of structured metadata files. These metadata files are provided as input to the Rilevo app, which is able to interpret them and build at run-time the questionnaire to be submitted to the respondent. This approach is the same jet utilized for Computer-assisted web interviewing (CAWI) by the Panda data collection system and has significant advantages in terms of flexibility and system reuse. Moreover, Rilevo and Panda share the same xml metadata files, generated by Panda. In this way, there is a single system dedicated to configurations files generation and the advantage of this approach is measurable in terms of code reusable and errors decreasing. Questionnaire was deeply modified in the last census wave, with respect to the one of previous wave, but it was not necessary to update the app code and it was just required to act on its configuration files.

The first step to use Rilevo is the connection to the central servers and downloads the configuration metadata files to set up questionnaire and rules. In the same way, it acquires all the multilingual texts of the guides and questions.

Data resulting from the interviews are first encrypted and then stored on the device. During data exchange, Rilevo app checks for updated questionnaire metadata files too. When Rilevo detects updated versions of the metadata files, it automatically downloads them and proceeds to update the questionnaire according to new metadata.

Results

Mobile devices have been a major step forward in field surveys, potentially making them simpler and more reliable. However, this required a significant effort on the software development front. While building a web application would seem the cheapest and most obvious choice, on the other hand, many of the features native to mobile apps and necessary for surveying purposes would have to be developed from scratch. To date, even using the tools available for creating web applications, aiming to make one's software responsive, secure, fast, able to run offline and take advantage of GPS means spending time and resources to achieve what mobile apps already have and with arguably inferior results [2].

Rilevo, on the other hand, is a native Android app that can take full advantage of the features that make mobile devices such a suitable tool for field surveys. It allows the enumerators to operate even in mountainous, insular or otherwise areas without network coverage. Geolocation is fast and accurate, data acquisition and synchronization can be done at separate times and, finally, interaction with the GUI is smooth and reliable, as it takes advantage of native touch capabilities.

Given the ever-increasing number of mobile device users, the next goal is to provide respondents with the opportunity to fill out the census questionnaire via mobile app, in addition to the desktop version.

Analyzing results over the years, we can see a very important increment of questionnaires processed after the introduction of Rilevo app. In fact, in the previous wave of the census of Population and Households, when the enumerators could only utilize the web application in offline mode, the numbers were:

- During 2018, on a sample composed by 1607599 families, questionnaires processed offline have been 139200, with a percentage equal to 8.66%. The total number of enumerators who have worked during the period equal to 26174 of which 7126 used Rilevo;
- During 2019, on a sample composed by 1582683 families, questionnaires processed offline have been 121777, with a percentage equal to 7.70 %. The total number of enumerators who have worked during the period equal to 18722 of which 8816 used Rilevo;

During 2021, 2552654 families composed the sample, while the enumerators were 26604 of which 14657 worked with Rilevo: the number of questionnaires processed offline have been 743603 with a percentage equal to 29.13%.

rEFERENCES

[1] Department of Economic and Social Affairs Statistics Division "*Guidelines on the use of electronic data collection technologies in population and housing censuses*", United Nations New York, January 2019

[2] Brian Fling, "*Mobile Design and Development*" (2009), O'Reilly Media, Inc. ISBN: 9780596155445

Bridging independent taxonomies using AI: an application to skills mismatch using PIAAC, ESCO Skills and Online Job Ads

Keywords: Embeddings, Taxonomy alignment, ESCO, PIAAC, Online Job Ads, Skills

1. Introduction

In recent years, the use and design of AI algorithms and frameworks to analyse labour market information for supporting decision-making has grown exponentially. In a context characterised by an increasing number of very diverse data sources in which the information is not classified in a standardised way, the need for accountable and transparent methods to bridge among different classifications is urgent. We develop a framework that combines word embeddings, taxonomy-alignment AI techniques and human validation to provide a crosswalk between the PIAAC background questionnaire and the ESCO Skills Pillar. Our final mapping links, for instance, the skills “Coaching young people” and “Instruct others” provided in ESCO to the PIAAC skill item “Teaching people”. We also present an application that can be performed with the online job ads-enriched PIAAC data on a topic with high relevance but also a high need for better data sources: skill mismatch. We propose a novel measure of the gap between the skills demanded by employers and the skill provision in the workforce. On the demand side, we rely on online job ads (OJA) data from the European Center for the Development of Vocational Training (CEDEFOP) collected in 2019. On the supply side, we use survey data from the first round of PIAAC, which comprises a representative sample of working-age individuals from 40 countries, among which we select 17 European countries. In PIAAC, respondents working in different occupations are asked about their skill use at work in different skill domains: digital, numeracy, literacy, and social skills.

2. Methods

This section describes the global approach to align the taxonomies, data and definitions and measures used in the application section.

2.1. Taxonomy alignment

The first step allows us to train and select the best word embedding model, which is then used in the second step to suggest for each leaf concept possible alignments. The last step consists of validating the suggestions provided to the domain experts to narrow the choices for the alignment that would otherwise be done from scratch. The procedure follows the work done by [1] and apply it to the context of PIAAC and ESCO. Figure 1 graphically presents the process of obtaining the validated aligned taxonomies.

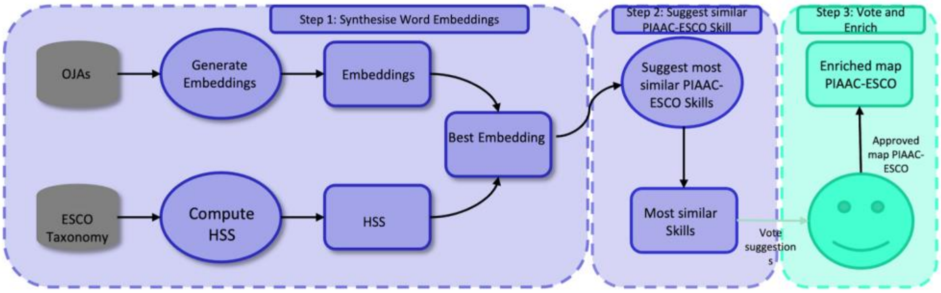


Figure 1: Graphical overview of the process to identify the best taxonomy alignment

The main goal of the first step is to induce a vector representation of taxonomic terms that represent the similarity of words within the taxonomy as much as possible. To accomplish this, we perform three distinct tasks. First, we generate word embeddings through a state-of-the-art method. [1] employ the state-of-the-art method FastText ([2]). This word embedding method considers sub-word information and can deal with out-of-vocabulary words. Following [3], we perform an intrinsic evaluation to select the best embedding model. The authors select the word vectors model with a maximum correlation between their cosine similarity and a benchmark semantic similarity value. In [3], the authors use a handcrafted dataset of pairwise semantic similarity between common words as the gold benchmark. However, those resources usually have low coverage, especially in specific domains like the labour market. For this reason, we resort to a measure of semantic similarity in taxonomies developed by [4], which measures semantic similarity in a taxonomy based on the structure of the hierarchy itself without using any external resource, thus, in a sense, preserving the semantic similarity intrinsic to the taxonomy. The latter method, Hierarchical Semantic Similarity (HSS), has proven to help select embeddings for several applications, like taxonomy enrichment ([5] and [6]) and job-skill mismatch analysis in the labour market ([7]). This method gives a limited number of suggestions to the domain experts to simplify their work of taxonomy alignment that otherwise would be all manual. The last step is validating the suggestions provided to complete the alignment procedure. Figure 2 reports the results of the validation phase, made by involving experts of the partner institutions of the project within which this research has been developed. Each member was asked to vote if and to what extent the ESCO skills suggestions were relevant and consistent with the PIAAC questions, using a Likert scale. Vote scores are concentrated in the upper part of the graph for most skills suggestion with a high level of agreement.

2.2. Data

We use the ESCO (European Skills, Competences, Qualifications and Occupations) Skills Pillar classification, the Survey of Adult Skills (PIAAC) and Online Job Advertisement (OJA) data. Our analysis requires the whole set of skills provided in the classification, Skills Pillar, and the list of selected items of the PIAAC Questionnaire. ESCO provides a multilingual dictionary of occupations and related skills organised as a network; nonetheless, it gives no information on the importance of skills in the considered occupation. PIAAC is a survey conducted by the Organisation for Economic Co-operation and Development (OECD) and contains information on critical cognitive and workplace skills of adults aged 16-65 across OECD countries. The main aim of the PIAAC survey is to assess literacy, numeracy and problem-solving skills in technology-rich environments using tests in each of these domains. We focus on the “skill use at work” items that elicit the frequency of skills used in various domains of job tasks on a Likert scale. We focus on 21 questions, covering the digital, numeracy, literacy, and social skills domains.[90] Online job-advertisements data are obtained from Eurostat and Cedefop as part of the Web Intelligence Hub - Online Job Advertisements (WIH-OJA) project. A representative sample is provided as part of the natural language processing (NLP) dataflow. The sample is made of job ads that include title and description. The sample is designed to provide a balanced coverage of occupation, type of contract, salary, working time, education, economic activity and experience. We select data for UK in 2019 for the embedding, while we cover 17 European countries to analyse skills mismatch.

2.3. Application of the method to skill mismatch: definitions and measures

We match PIAAC and ESCO skills, then using the WIH-OJA data to assign each matched skill the number the number of online job ads. We follow the approach developed by [9] that calculate the RCA using the O*NET dictionary of occupations and skills, which surveys a sample of workers in the US to assess the relevance of a skill for each occupation. We calculate it using online job ads as proposed in [10]. The relevance is computed as the frequency of a skill in the job ads of a specific occupation, relative to the skill's frequency in job ads in all other occupations. Analogously, the RCA of a skill in PIAAC is computed as the frequency of skill use among survey respondents in a given occupation, relative to survey respondents in all other occupations. Finally we develop a measure of skill mismatch based on the importance of each skill for each occupation. Specifically, our measure of skill gap is based on differences in the revealed comparative advantage (RCA) of each skill in each occupation between the OJAs and PIAAC. First, we calculate the RCA of each skill for each occupation in both OJV and PIAAC. Next, for each skill, we can compute the rank of the RCA across all occupations on both the demand and supply side. Thus, differences in the RCA ranks of skills between demand and supply reflect potential mismatches in skill relevance. We define the gap in skill ranks as the percentile rank on the demand side minus the percentile rank on the supply side. Thus, a positive value indicates that the RCA of the demand for a skill is larger than the RCA of the skill supplied in a particular occupation. We refer to this case as skill shortage.

3. Results

The literature often distinguishes between four occupation types: manual routine, manual non-routine, cognitive non-routine, and cognitive routine occupations (Autor, Levy, and Murnane, 2003). These occupations differ in the tasks workers need to perform on the job. For instance, food preparation assistants perform predominantly manual, routine intensive tasks, such as manual assembling and quality checks. On the other hand, teaching professionals perform predominantly cognitive and non-routine tasks, such as using advanced mathematics and teaching people. At the same time, structural transformation and technological change have different impacts on different types of tasks. Automation technologies have particularly

[90] For countries in our sample, PIAAC is only available for 2012 and 2014. However, the earliest job ads data from CEDEFOP stem from 2019. To remedy the temporal misalignment between PIAAC and OJV, we use the skill change in the US, for which the PIAAC survey was conducted in 2012 and 2017, to project skill changes for all other countries. Assuming that changes in the occupational skill content in the US represent changes at the technological frontier (e.g., [8]), these changes can be used to project an upper bound of how skills have evolved in other PIAAC countries.

rendered codifiable routine and manual tasks susceptible to substitution by automation. As the task composition and thus the skill requirements of different occupations are affected differently by technological change, this also renders occupations more or less susceptible to changing skill demands and skill mismatch, which we also refer to as skill gaps. We find that while cognitive non-routine and cognitive routine workers have a skill supply surplus on average (negative skill gap), manual non-routine and manual routine workers exhibit skill supply shortage (positive skill gap). Introducing the regional dimension, we observe that skill shortage for workers in manual-intensive occupations and the skill surplus for those in cognitive-intensive occupations is persistent across EU countries. For almost all countries, cognitive workers show skill supply surplus, while manual workers have skill supply shortage on average. One potential driver of the positive skill gaps of manual workers could be a supply shortage of specific in-demand skills that have gained importance in recent years, such as digital skills. To investigate this, we separate the skill gap for each occupation group by different domains: digital, numeracy, literacy, and social skills. We find similar skill gaps across all skill domains for our four occupation types: cognitive workers show skill supply surplus on average, while manual workers exhibit shortage on average. Further, the supply shortage across all skill domains is largest for manual routine workers, while the surplus is highest among cognitive non-routine workers.

4. Conclusions

We present a novel methodology to support the construction of crosswalk between classifications using AI techniques and human validation. In our application we show the usefulness of the method in enriching an existing survey, PIAAC, with an independently-developed classification of skills, ESCO Skills Pillar, to be able to link online-job-ads from WIH-OJA data and study skill mismatch in Europe. This methodology can be applied to other contexts that share analogous issues. For instance, it could help bridge national classifications and international standards or enrich existing surveys with data sources even when a crosswalk is missing.

References

[1] A. Giabelli, L. Malandri, F. Mercorio and M. Mezzanzanica, Weta: Automatic taxonomy alignment via word embeddings. Computers in Industry (2022) 138, 103626.

[2] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information (2017) Transactions of the Association for Computational Linguistics 5, 135–146.

[3] M. Baroni, G. Dinu and G. Kruszewski, Don’t count, predict! a systematic com-parison of context-counting vs. context-predicting semantic vectors (2014) Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 238–247.

[4] L. Malandri, F. Mercorio, M. Mezzanzanica and N. Nobani, Meet-lm: A method for embeddings evaluation for taxonomic data in the labour market (2021) Computers in Industry 124, 103341.

[5] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica and A. Seveso, Neo: A tool for taxonomy enrichment with new emerging occupations (2020) International Semantic Web Conference, Springer. pp. 568–584.

[6] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica and A. Seveso. Neo: A system for identifying new emerging occupation from job ads (2021) Proceedings of the AAAI Conference on Artificial Intelligence, pp. 16035–16037

[7] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica and A. Seveso. Skills2job: A recommender system that encodes job offer embeddings on graph databases. (2021) Applied Soft Computing 101, 107049

[8] J. Caunedo, E. Keller and Y. Shin, Technology and the task content of jobs across the development spectrum. (2021) NBER WP No. 28681.

[9] A. Alabdulkareem, M. R. Frank, L. Sun, B. Al Shebli, C. Hidalgo and I. Rahwan, Unpacking the polarization of workplace skills (2018) Science advances, (4)7.

[10] A. Giabelli, L. Malandri,, F. Mercorio and M. Mezzanzanica. GraphLMI: A data driven system for exploring labor market information through graph databases (2022) Multimed Tools Appl, 81, 3061–3090.

Use of multilevel grids to release protected grid data from the French 2021 Census

Keywords: grid data, census, suppressive method,

INTRODUCTION

In the context of releasing Census 2021 data on a 1km² grid for France, Insee chose to protect these data with a method of geographical aggregation based on quadtree approach applied on grid data. This method allows to release gridded data on multiple levels and to handle statistical disclosure control at each level, especially the finest one. The R package *gridy* is developed for that purpose. Unfortunately, some disclosure issues are not directly addressed by this method, especially the attribute disclosure problem.

Some problem of geographical differentiation also arise, since the same information is released, at the meantime, on the administrative areas. The R package *diffman*, developed at Insee several years ago, is particularly powerful to handle these issues with a graphbased approach.

Thus, after having introduced the two methods, we present how we articulate them to tackle as many confidentiality issues as possible.

The results presented in this abstract are provisional, because the final data are not yet available.

METHODS

Multi level aggregation or the “gridy” approach

Insee is used to releasing gridded data on multiple level at once [1]. To lower the reidentification disclosure risk, the main confidentiality rule is here not to release proper data with less than 11 households in a cell. Nevertheless (and fortunately), total of population in each cell is disclosed without any protection.

The “quadtree" approach

In the case of grid data, one strategy may be to join contiguous cells into larger polygons (e.g. larger rectangles and cells) so that each polygon meets the threshold. The new polygons can be obtained by aggregation, *i.e.* by grouping the cells until the threshold is reached, or by disaggregation, starting from the largest cell and splitting it until it is no longer possible to cut it without going below the threshold [2].

These methods have the advantage of preserving the additivity but also allow to avoid the creation of "false zeros". In addition, the threshold rule is respected by construction of the algorithm. The areas first obtained by disaggregation are called the "natural" areas and they can be of different sizes depending on when we stop splitting.

The main drawback is that disseminated information is not homogenous : some highresolution tiles are available as well as some large areas, grouping for instance 16 initial tiles. Considering that the tiles are defined to be comparable, it seems to be counterproductive. The worst case is when a highly populated tile is near from another one cell with a low population.

The “gridy” approach

However, it is possible to obtain more accurate information if we continue to divide the resulting cells in smaller cells and if we decide to hide the information from the cells that are below the threshold. Obviously, it is not sufficient to hide only the cells below the threshold since the information can be found by geographical differentiation comparing the finer level of detail to a coarser level of detail. Therefore, we have to hide the information contained in another cell at the same level. This process requires disseminating information on several grids ("composite" grid) corresponding to different levels of detail, which is equivalent to disseminating information on the same grid with cells having different shapes and sizes. To avoid the issues related to differentiation between levels, the suppression process consists of grouping deleted cells and ensuring that each group exceeds the threshold. The process, presented in the Figure 1, is implemented in an R package called “gridy”, not yet published. At last, instead of releasing blanked cells, the total of a group for a given sensitive variable is distributed proportionally among the cells that make it up.

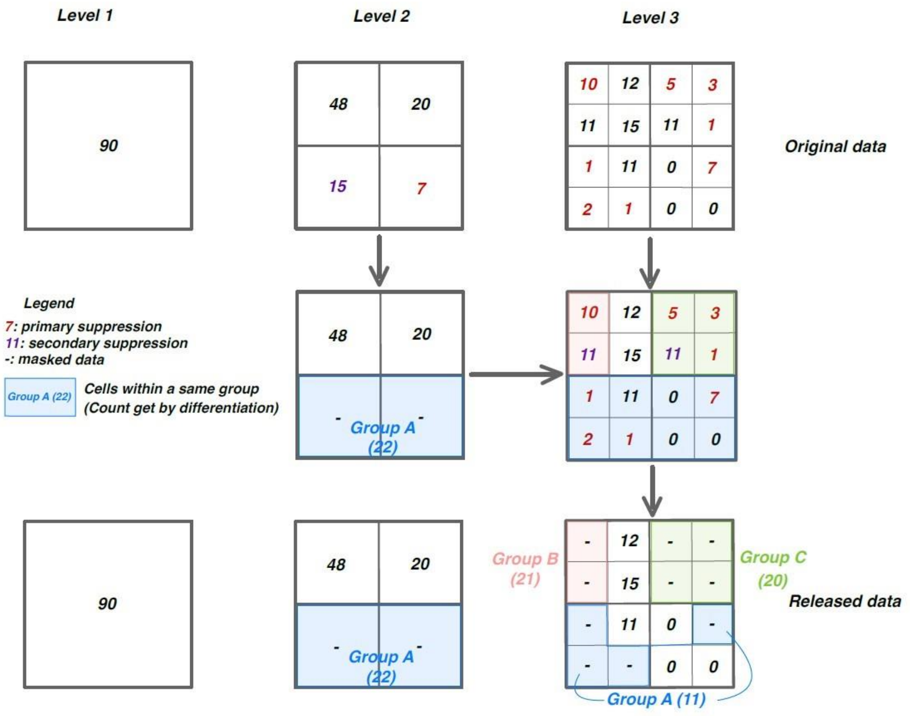


Figure 1: Gridy approach to handle disclosure control in multilevel gridded data

Definition of at risk cells : Geographical differentiation

The new contours introduced with this method, combined with other administrative contours can generate statistical disclosure through geographical differentiation. A package R called *diffman* makes the detection of these potential disclosures easier with a graph-based approach [3].

In the case of a further diffusion to the municipality level, the implemented method allows to search for differentiation problems among all possible combinations of municipalities. The method is based on the observation that the areas at risk of differentiation are only located at the intersection of one municipality and a set of cells (see Figure 2). Therefore, the focus here is only on the intersections between cells and municipalities. The number of municipality combinations to check for differentiation is huge, so we model the problem by a *graph* where the municipalities are the vertices and where two municipalities are connected if a cell overlaps them. This defines a set of connected components composed of municipalities that can be checked for differentiation separately and efficiently using *graph* reduction methods.

The algorithm output are the risk areas corresponding to the geographical difference between a group of cells and a group of municipalities (see Figure 2). These risk areas must be protected by removing one or more cells among those involved in the differentiation operation that created the risk area. The method used to choose the cell to be removed is still being assessed at this stage.

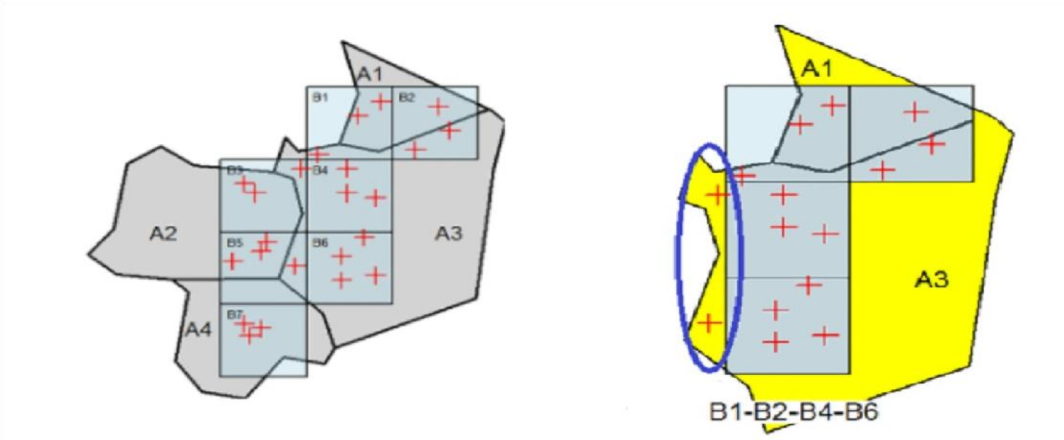


Figure 2: Example of area at risk of differenciation

The overall approach

The “gridy” method tackles the reidentification disclosure risk. The “diffman” approach tackles the differentiation issues. How to articulate them to ensure that both problems are handled? And what about the attribute disclosure risk?

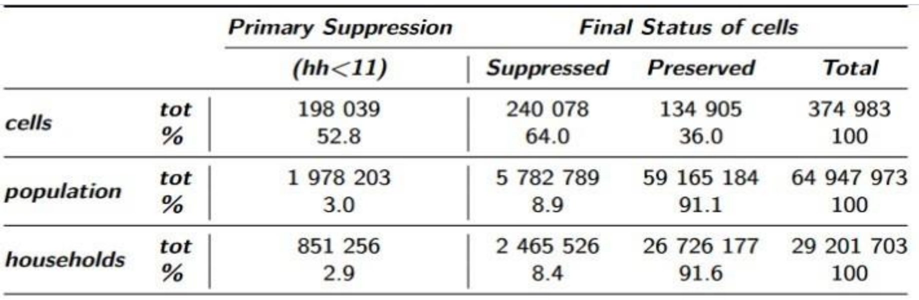
Consider a set of joint data for cells and municipalities.

The overall approach consists in the following steps:

i. Identification of cells that can lead to differentiation issues, thanks to the “diffman” method; ii. Identification of risky cells in terms of attribute disclosure issue: for example, a cell above the threshold but with individuals sharing the same sensitive attributes is considered risky;
iii. Secondary suppression with the “gridy” approach; iv. Filling of the blanks within the groups of suppressed cells.

RESULTS

Table 1 presents the provisional results of the multi-level aggregation algorithm applied on the French 2021 census data starting with the set of cells under the threshold of 11 households.



		Primary Suppression	*Final Status of cells*		
		(hh<11)	*Suppressed*	*Preserved*	*Total*
cells	*tot*	198 039	240 078	134 905	374 983
	%	52.8	64.0	36.0	100
population	*tot*	1 978 203	5 782 789	59 165 184	64 947 973
	%	3.0	8.9	91.1	100
households	*tot*	851 256	2 465 526	26 726 177	29 201 703
	%	2.9	8.4	91.6	100

Table 1: Secret generated by the aggregation algorithm

Among the 375 000 initial cells, 52% are below the 11 households threshold, applying the algorithm increases this ratio to 64%. In the end, only 9% of the population is suppressed.

A *graph* of municipalities has been built from the French 2021 census data. There are 1902 subsets of connected municipalities, the largest connected component contains 18762 municipalities. The search for risk areas here would not have been tractable without the graph reduction techniques implemented in the *diffman* R package.

CONCLUSIONS

Using these two methods, we are able to protect data without hiding any information. At Insee, the multilevel aggregation method was originally implemented to release very sensitive grid data such as income and other data from tax sources and has been in use for several years. The ideal dissemination of such data does not consist in a single 1km² grid, but in a whole set of grids of different tile sizes, which are to be released by Insee. The method of detection of differentiation problems will have to be implemented for the release of the Census 2021.

REFERENCES

[1] M. Branchu, V. Costemalle, *Données carroyées et confidentialité*, 2018, http://www.jms-insee.fr/2018/S23_2_ACTE_COSTEMALLE_JMS2018.pdf

[2] Lanigor, Oller, Martori, *A quadtree approach based on European geographic grids : reconciling data privacy and accuracy*, SORT, n°41, 217, p139-158

[3] Costemalle, *Detecting Geographical Differencing Problems in the Context of Spatial Data Dissemination*, Statistical journal of the IAOS, 2019, p559 – 568

Hedonic House Price Index for Poland based on listings

Keywords: house price indices, listings, transactions.

Introduction

House price indices (HPIs) serve various functions, including informing the general public, banks, the financial sector, and the government about changes in the housing market. Given that housing booms and busts can threaten the financial system's sustainability, HPIs are also crucial for macroprudential supervision. They are mainly built based on the transaction prices indicated in the notarial deeds.

The main advantages of using transaction prices are that they are the most reliable indicator of a property's market value. In addition, this data is readily available in countries with developed real estate monitoring systems. On the other hand, they are often delayed by the need for public institutions to enter information into databases. In addition, it should be noted that the transaction price is set several to several months before the transaction date, which means that it does not reflect the housing market at any given time.

Another source of information, based on which real estate price indexes are built, can be listings prices. The main advantage of this type of data is that their level, in most cases, is suggested by real estate agents, who, knowing the market in question, can correctly determine the value of a given property. In addition, the number of properties offered for sale is far more significant than the number of transactions concluded. The disadvantages of using offer prices for constructing the indexes in question are due to the following circumstances. First, offer prices may differ significantly from the market value of a given property, mainly due to poor qualifications of the intermediary or inadequate knowledge of the owner. Second, properties with particularly low offer prices relative to market value may be transacted relatively quickly. In contrast, properties with high offer prices close to market value may be put up for sale for a long time, and in extreme cases, transactions may only take place at a time. However, the biggest advantage is the availability without almost any delay.

The idea of asking prices as a source of information for computing housing price indexes is not new [1]. Few research articles compare asking and transaction price indexes. A few studies, however, indicate that the offer data are a good reflection of the changes in the property market and offers, which may be an adequate substitute when transaction data are unavailable [2–5].

The problem with lags concerned with transaction prices can be seen in the case of Poland. The lag for the transaction prices delivered by Property Price Register is around six months. The index has been lagging for several months because of data unavailability. In Poland, the National Bank of Poland (NBP) and the Central Statistical Office (CSO) published the index of apartment prices for provincial cities. The NBP was the first to publish hedonic indices in 2010 (data from Q3 2006) for the largest cities in Poland.

The CSO has published the average prices of residential units in provincial cities since 2015. Previous attempts proved unsuccessful due to difficulties with access to data. Only the dynamic construction of the Property Price Register database could overcome this barrier. It should be emphasised that these indicators are based on full property ownership; the cooperative ownership right is omitted (around 30% of the existing housing market). The CSO provided the house price index for Poland starting in 2010 (stratification method).

In this paper, the hedonic house price index for Poland was constructed based on the listings from the 27 biggest cities in 1996 - 2022. The index was then compared with the official published by CSO. Despite both being based on different sources, methods show similar behaviour.

Methods

The study used a unique database of more than 5 million housing offers in 27 cities in Poland from 1996 to 2022. In 2019 in the secondary market (existing buildings), 52% of transactions occurred in these cities. The earlier data were obtained from archival advertisements (various local periodicals from the whole country, such as regional editions of the newspapers) in the form of photocopies, photographs or magazines, which were digitally reproduced and arranged in a database. The data from 2008 were collected from advertising portals several times a quarter with a web-scrapping procedure.

The archival data needed to be digitalised. The scope of available information varied and depended on the time of publication (adverts from the end of the 20th century often did not include the asking price; instead, they provided information on whether there was a phone in the flat) and the publisher's requirements. The information on an apartment’s price, location and size were the easiest to find; however, other information was also available to a lesser extent. The scope of information about offers from ad portals was broader. The most important data include the location (district, housing estate and street), asking price, location in the building (floor), type of ownership, floor size, construction technology, parking facilities and standard of completion. Since this information came from different external sources, it was necessary to adapt the datasets to arrange them into a uniform pattern. The essential task was to design a homogeneous database with a uniform system of recording variables. The main problem concerned identifying location, which is the main factor determining the value of a property. As this factor was defined with a different degree of detail, from the general level (city, district, housing estate, street) to the specific location (street with a house number), properties were grouped and assigned to a proper category.

Initially, more than 5,5 million offers were collected in the database (after preliminary cleaning to remove offers without price, surface area or a specified general or specific location) by both ownership and cooperative ownership of the premises. In addition, repeated offers had to be removed as a result: offering the same apartment in subsequent months (e.g., January, February, March), offered by different intermediaries (the same apartment was offered more than once) and offering the same apartment in a given quarter at different prices. The last offer was left in the dataset. Moreover, the outliers were removed using the Cook distance procedure. After these steps, the dataset accounts for 2.7 million offers.

The hedonic HPI for each city was constructed with the rolling time-dummy method [6]. Since the information about the offers came from two sources, the range of explanatory variables describing each apartment was also different. Variables used in the models:

- listings from 2009 – district/estate, area, age, technology, quality of the apartment
- listings up to 2008 – district/estate, area

The quarterly hedonic HPI for each city is presented in figure 1.

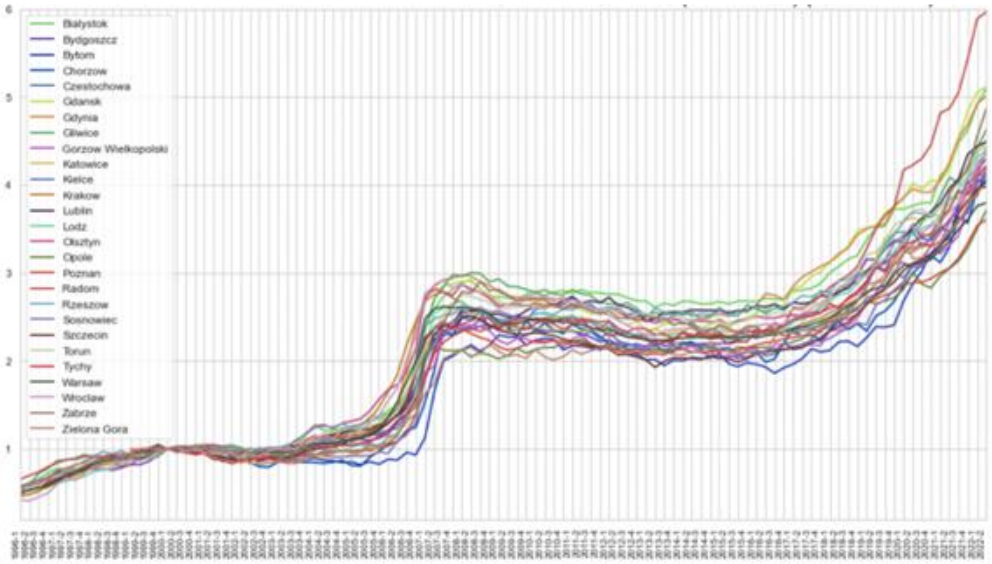


Figure 36. Hedonic HPI for selected cities in Poland Q1 1996- Q2 2022 (Q1 2000 =1)

Results

The index for Poland was then constructed, taking into account the share of transactions in 2019 (data from CSO). The nominal and real hedonic house price index for Poland in 1996-2022 is presented in figure 2.

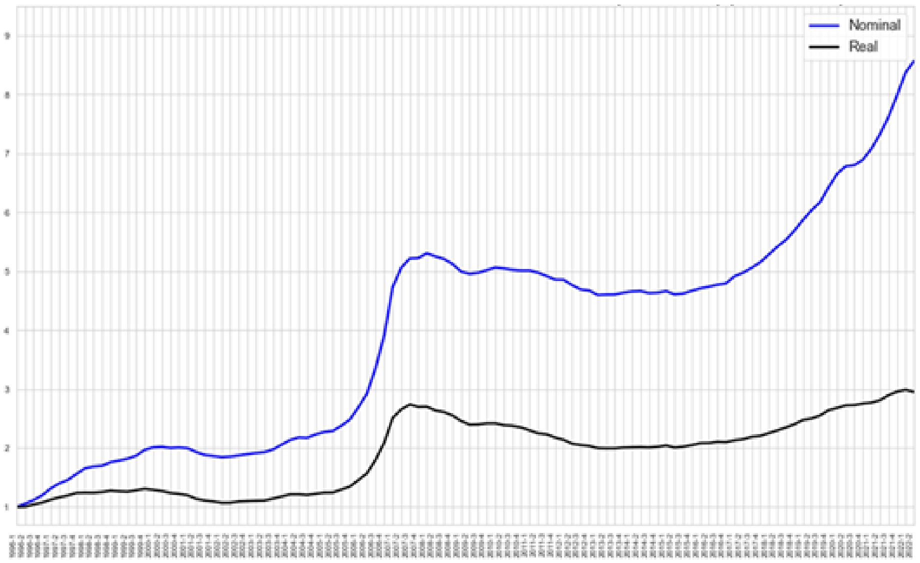


Figure 2. Nominal and real hedonic HPI for Poland Q1 1996-Q2 2022 (Q1 1996 = 1)

In the case of HPI, an important issue, in addition to the greatest possible control over changes of a qualitative and quantitative nature in the states of features between periods, is their timeliness. As previously written, indexes based on transaction prices are published with a lag of up to several months. Indexes of offer prices are published more quickly but are based on data close to transaction prices (they are often subject to negotiation, the differences may depend on the market situation or the buyer/seller's ability to bargain). Given the above, an attempt was made to compare housing price indexes based on offer and transaction data in Poland. In the next analysis step, the listing HPI was compared with the index provided by CSO. The comparison involves 2010 – 2022, the longest possible period for contrast.

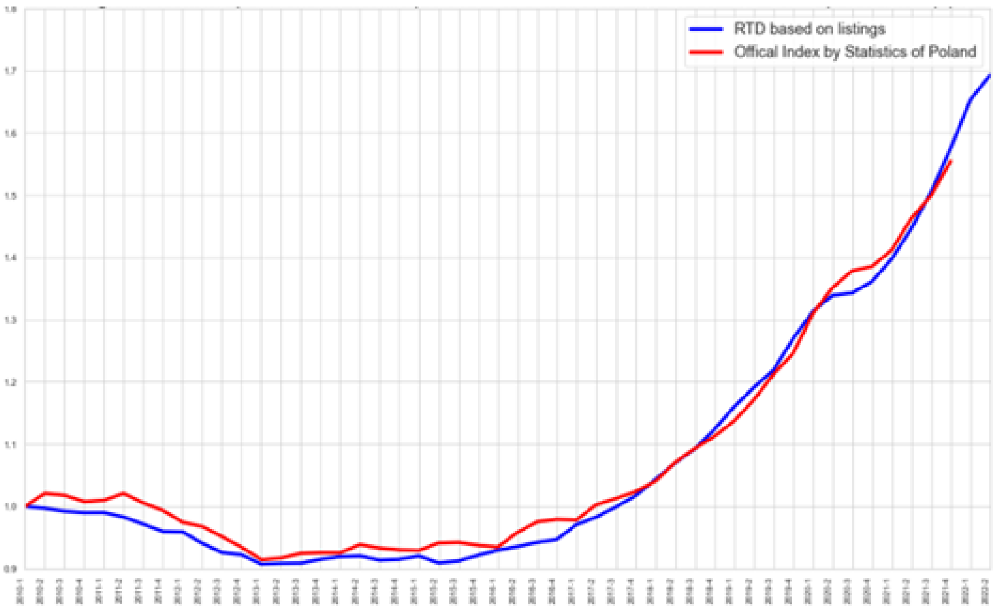


Figure 3. HPI's – RTD listings and CSO based on transactions Q1 2010 – Q2 2022 (Q1 2010 = 1)

Even though the indices were built on different data, using other methods, and different subject coverage (the CSO index does not take into account the cooperative ownership right to the premises), they show very similar price changes in the analysed period. The CSO index lagged by two quarters as the data were unavailable at the comparison time. Some slight differences seem to be not crucial.

Conclusions

Using a micro-level dataset of over 2.7 million listings, the house price index for Poland over the period 1996 to 2022 was constructed. The hedonic indices were determined for each city and then aggregated. The created index has some limitations but provides information on the housing market in Poland for much longer than it officially published. Moreover, the comparison with the officially published index shows very significant similarity. Considering the ease of gathering the data on listings and the possibility city of providing HPI without delay, listings HPI appears to be a good indicator of changes in the housing market.

References

1. Pollakowski, H.O. Data Sources for Measuring House Price Changes. *J. Hous. Res.* **1995**, *6*, 377–387.
2. Anenberg, E.; Laufer, S. A More Timely House Price Index. *Rev. Econ. Stat.* **2017**, *99*, 722–

734, doi:10.1162/REST.

3. Lyons, R.C. Can List Prices Accurately Capture Housing Price Trends? Insights from Extreme Markets Conditions. *Financ. Res. Lett.* **2019**, *30*, 228–232,
4. Shimizu, C.; Nishimura, K.G.; Watanabe, T. House Prices at Different Stages of the Buying/Selling Process. *Reg. Sci. Urban Econ.* **2016**, *59*, 37–53,
5. Kolbe, J.; Schulz, R.; Wersing, M.; Werwatz, A. Real Estate Listings and Their Usefulness for Hedonic Regressions. *Empir. Econ.* **2021**, doi:10.1007/s00181-020-01992-3.
6. Hill, R.J.; Scholz, M.; Shimizu, C.; Steurer, M. An Evaluation of the Methods Used by European Countries to Compute Their Official House Price Indices. *Econ. Stat.* **2018**, *2018*, 221–238, doi:10.24187/ECOSTAT.2018.500T.1953.

Geospatial Data Store - centralization of spatial data in Statistics Poland

Keywords: GIS, GSBPM, geospatial data, statistics;

Introduction

Currently at the Statistics Poland, works are carried out as part of the 'Gates of Statistics' project, which aims to improve the quality, usefulness and availability of statistical information for the largest possible group of recipients. In principle, this project is carried out with the use of modern technologies and adapting the current methodology and organization of research to modern technological possibilities. It also provides for the modernization of MPPS (Model of the Statistical Production Process – Polish implementation of GSBPM) to adapt it to current needs.

Providing reliable, credible and independent statistical data of high quality is the overriding task of official statistics services, serving to meet the information needs of state bodies, public administration, entities of the national economy, and the entire society, including individual citizens. Collecting data, processing it, and finally making it available gives the recipients the opportunity to get acquainted, among others, with the economic, demographic and political situation of the country, the state of the natural environment or changes taking place in them.

As part of the work, it was planned to create many databases (repositories) so that different processes could access them and, as a result, base their activities on the same sets of data without the need of duplication. One of such repositories is the Geospatial Data Store (GDS) - a database containing geometric data of geospatial objects along with their spatial location, with the accuracy of x, y coordinates. Database objects allow for geocoding (linking to a spatial location) both address points and units of the administrative or statistical division of the country.

Building the GDS taking into account the assumptions of the MPPS model, and thus the consistency and standardization of data structures, technical processes and metadata, and, consequently, also business processes taking place as part of the works of the MPPS phase 7 (sharing), will extend the scope and improve the sharing of statistical information, and increasing their usefulness.

The main task of the GDS, which is one of the database structures planned to be implemented under the Gates to Statistics project, is the centralised storage of geospatial data.

Methods

The GDS will be a database structure containing geometric data of geospatial objects along with their spatial location, with an accuracy of x, y coordinates. The base objects will allow for geocoding (linking to a spatial location) both to address points and units of administrative division, statistical division of the country and grids. The basis of the GDS will be a geospatial database created as part of the project "Spatial Statistical Data in the State Information System (PDS)". In the 'Gates of Statistics' project, this database will be expanded to achieve full functionality of the Geospatial Data Store and thus fit into the entire statistical production process.

Results

GDS will provide a centralized place for collecting geospatial data and provide mechanisms for their management. Geospatial data is necessary both to conduct preparatory work for surveys or censuses, to manage and monitor the work of interviewers and enumerators in the field during data collection, and finally, after the completion of a statistical survey - to perform multidimensional spatial analyzes.

Conclusions

The use of geospatial data collected and managed in one place will allow, in the context of the entire statistical production process, to:

- unify the information resource described by geospatial data,
- organize the entire production process in this area,
- more efficient quality control of the shared data,
- centralized management of the full range of geospatial data.

Business statistics (GASP3A.1)

Session Chair: **Søren Schiønning Andersen** *(Statistics Denmark)*

The Italian Social Mood on Economy Index: sub-indices focused on business
Luca Valentino *(National Institute of Statistics–ISTAT)*, Elena Catanese *(National Institute of Statistics–ISTAT)*

Business register improvements: a balance between search, scrape and 3rd party web data
Olav ten Bosch *(Statistics Netherlands-CBS),* Arnout van Delden *(Statistics Netherlands-CBS)*

Innovating Business Data Collection: System-to-System Data Communication applied to an agricultural survey
Ger Snijkers *(Statistics Netherlands-CBS)*

The Italian Social Mood on Economy Index: sub-indices focused on business

Keywords: Sentiment analysis, Twitter, Macroeconomic indicators

Introduction

In October 2018, the Italian National Institute of Statistics released the Social Mood on Economy Index (SMEI) [1], an experimental high-frequency sentiment index based on Twitter data. The index, derived from samples of public tweets (in Italian) captured in real-time, gained a good spread among economic analysts.

Before and during the pandemic, the SMEI showed different trends compared to the main monthly series of macroeconomic indicators, making unclear the economic cyclical signal captured. Therefore, the index has undergone a process of revision, aiming at enhancing its economic interpretability. Since the set of keywords relates to many topics (expenditures, financial situation, inflation, fiscal policies), it is not so evident what the index is measuring. In particular, due to its multivariate nature, it is not clear with which univariate economic time series the index should be consistent.

Our work aimed both at the presentation of SMEI and to understand if it is possible to improve its quality and its economic interpretability. We derived from the SMEI sub-indices more related to the business activities and expectations for the period 2018-2021. We adopted a specific strategy of analysis to improve our first sub-index to produce a final sentiment index on business (B2).

Background

Sentiment analysis is an increasing area of research and different studies presented the main algorithms for text mining and sentiment analysis with supervised and unsupervised learning techniques [2; 3]. Analysis making use of lexicon-based methods mainly refers to the English language [4; 5]. There are few resources available in Italian. Castellucci et al. [6] developed a Sentiment Italian Lexicon (DPL), a dictionary whose lemmas are associated with pre-computed sentiment scores used by Istat in producing the SMEI.

Data and Methods

The SMEI is an experimental daily-frequency index aimed at measuring public sentiment on the economy based on messages of Italian Twitter users. We use Twitter's streaming application programming interface (API) to collect samples of public tweets matching a filter made up of 100 keywords relevant to the study of the general and personal economic dimensions. We compute daily index values using filtered tweets reported in a single day (a mean of around 30K daily tweets from February 2016 to the end of June 2021) as a single block. We pre-process tweets through different phases of text cleaning and normalization: converting every letter of a word into lowercase and tokenizing the text into words [1].

Sentiment scoring is obtained using the DPL [6]. Lexicons are vocabularies whose lemmas are associated with pre-computed sentiment scores. We compute the sentiment per tweet as the weighted mean of the entries that match the lexicon. The daily index value is then derived as an appropriate central tendency measure of the score distribution of all the daily tweets.

However, different biases can affect estimates. The free download from Twitter API did not allow full access to all the Italian active users' tweets and the Twitter users are not a representative sample of the Italian population due to different Twitter penetration rates among various sub-population (e.g., young people). On these rates, there is no official information.

Aimed at developing an index, which summarizes conversations on aspects of business activity, firstly, we selected a subset of words to build a sub-index, the so-called Business index. As the new series were still too correlated to the SMEI, we performed a Topic analysis, with a Latent Dirichlet analysis approach (LDA) on the conversations on aspects of business activity underlined by the sub-index during the first and latest quarters of our period. In this way, we selected a smaller amount of words[91] able to narrow down business conversations and calculated the B2 index.

We analysed how the sub-indices correlate with the SMEI. Aimed at verifying which index was closest to the macroeconomic indicators, we correlated SMEI and sub-indices with the series of main Istat macroeconomic indicators (quarterly series of Household consumption expenditure, GDP, Value added, and Indexes of turnover in other services; monthly LFS employment rates; monthly series of Consumer confidence, Business confidence climate, Industrial turnover index, and Industrial production index). As the period is quite short (2018-2021), we mainly used graphical analysis and correlation analysis, as the standard econometric analysis could face serious hindrances and possibly turn out unsuitable.

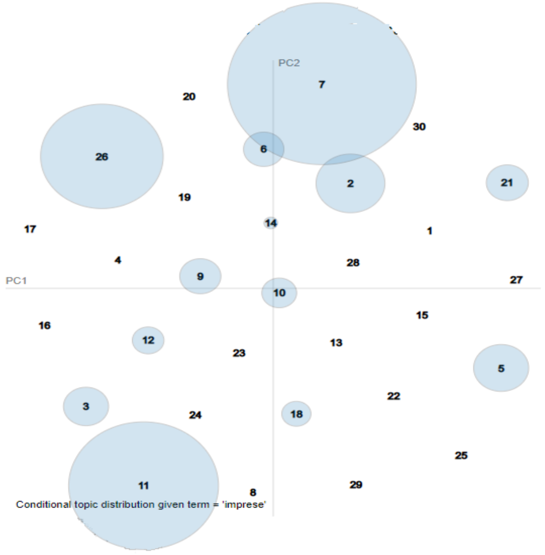
Results

As the correlation between the monthly means of the SMEI and the Business index was 0.93, we analyzed with an LDA the conversation of the first and last quarter of the period 2018-21. For the first quarter of 2018, we identified 30 clusters. We estimated that only 17.6% of tweet conversations focus on enterprises. In figure 1 we show the clusters containing the word "enterprise" and their main theme.

Thus, we focused on words characterizing the core conversation of these clusters to produce a sub-B2 index (enclosing words such as e-commerce, investments, etc.). The correlation between the SMEI and the B2 index diminished, indicating a wider diversification of the meaning of the two indices.

However, the selection of the keywords on core business activities produced a reduction in the volume of tweets used for the three indices: SMEI (daily means 30K tweets), Business index (13K), and B2 index (4K). They reached their maximum volume at the end of March and April 2020, when Italian Government proclaimed the first emergency law for the Covid-19 pandemic.

[91] Filters used refers to words or expressions like balance, sales, investments, financing, sales, debts, credits, etc.



Cluster 7: (3.8% of tweets) Policies in support of business (investements and taxes)

Cluster 11: (3.5% of tweets) Enterprises selling on-line

Cluster 26: (2.7% of tweets) Womens flexibility in enterprises

Cluster 2: (4.7% of tweets) Innovation and sustainability in enterprises

Cluster 21: (2.9% of tweets) Start-up

Figure 37 – Results of the topic analysis on Business index conversations during the first quarter of 2018

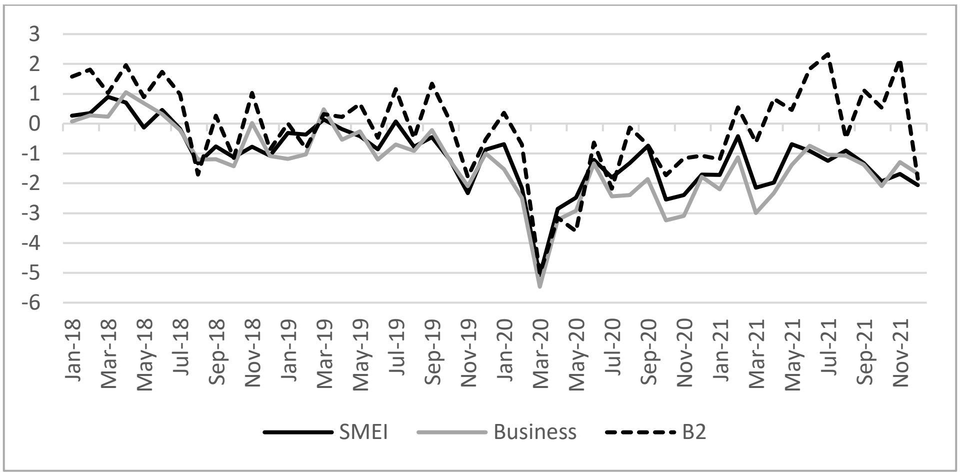


Figure 38 – SMEI and sub-indices – monthly means, Jan 2018-Dec 2021

The monthly means of the SMEI and Business indices (showing no seasonality) are in figure 2. The SMEI series and the sub-indices negatively peaked in March 2020 during the first lockdown. Afterward, the indices decreased again in October and November 2020 (during the second lockdown). The behavior of the sub-indices varies a lot after spring 2021 when the Italian economy recovered. The B2 index was much more sensitive to this economic improvement, reaching its maximum values in July 2021 and November 2021.

The correlations of the B2 index with macroeconomic indicators improved a lot with respect to the SMEI and the Business indices (table 1). The B2 index improved the correlation with all the quarterly/monthly indicators compared to the SMEI.

The improvement is particularly relevant and statistically significant for the indicators directly related to the business cycle: Indexes of turnover in other services (from 0.4 to 0.7), Industrial turnover index and the Business confidence climate (from 0.4 to 0.6), Industrial production index and Household expenditures (from 0.6 to 0.7). The B2 index correlates better than SMEI with the GDP and the Value added (from 0.1-0.2 to 0.4). Its relationship with employment rates remains weak.

Table 1 – Correlations of SMEI and sub-indices with main macroeconomic indicators, Jan. 2018-Dic. 2021

Macroeconomic indicators		**SMEI**	Business index	B2 index
Quarterly indicators	GDP	**0.1**	0.3	0.4
	Value added	**0.2**	0.3	0.4
	Household expenditures	**0.6**	0.7	0.7
	Index of turnover in services (2015=100, Seas.adj.)	**0.4**	0.5	0.7
Monthly indicator	Business confidence climate Seas. adj.	**0.4**	0.5	0.6
	Industrial turnover index (2015=100, Seas. adj.)	**0.4**	0.4	0.6
	Industrial production index (2015=100, Seas. adj.)	**0.6**	0.6	0.7
	Consumer confidence (no seas.)	**0.5**	0.6	0.6
	Employment rate (15-64)	**0.3**	0.4	0.4

Conclusions

Thank to topic modeling techniques, we were able to build the B2 sub-index, which turned out to be more related to the business cycle, thus showing a more effective economic meaning for some series.

We plan to monitor in the next months the performance of the B2 index to enrich the Istat supply of sentiment indicators with one focused on business expectations. We also plan further developments to increase the accuracy of the SMEI and the sub-indices related to the review of the lexicon-based approach and more fine-tuning in defining the samples of the sub-indices.

References

[74] M. Bruno, E. Catanese, R. Iannaccone, A. Righi, M. Scannapieco, P. Testa, L. Valentino, D. Zardetto and D. Zurlo, The Social Mood on Economy Index, Methodological note. Rome: Istat, (2022).

[75] A. Gandomi and M. Haider, Beyond the hype. Big data concepts, methods, and analytics. International Journal of Information management, 35(2), (2015) 137-144.

[76] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics, 37 (2), (2011), 267–307.

[77] G. A. Miller, WordNet: a lexical database for English. Communications of the ACM, 38(11), (1995), 39-41.

[78] A. Esuli and F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, LREC, 6, (2006, May), pp. 417-422.

[79] G. Castellucci, D. Croce and R. Basili, A Language Independent Method for Generating Large Scale Polarity Lexicons. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16), European Language Resources Association (ELRA), Portoroz, Slovenia (2016).

Business register improvements: a balance between search, scrape and 3rd party web data

Keywords: Business registers, web scraping, search engines, machine learning, NACE

INTRODUCTION

These days the web is an integral part of life. We use it in our work, for communication and interaction, to compare, buy and sell products, to plan holidays, to order food and many more things. For enterprises, the web is often much more than a communication channel. Depending on the activities they perform the web may play a role in advertising, in managing their logistics, in customer care, or in recruiting new employees. Hence the web is full of digital traces from economy that may help National Statistical Institutes (NSIs) in making official statistics.

NSIs typically maintain a statistical business register (SBR) of enterprises. It contains the statistical units comprising the enterprise population and variables such as number of employees (size class), location and type of economic activity (NACE code). It also contains the relation between enterprises, other statistical units and administrative units such as legal units. It is used as a frame for drawing samples for surveys and as the backbone for statistics on economy. Using the web as a data source the SBR can be improved with better, more detailed or new information that is difficult to grasp in a more traditional way.

Figure 1 contains a high-level view of this concept. Depending on country characteristics the proportion of legal units for which the website(s) are known may vary heavily. In most cases the missing URLs have to be found in the so-called URL finding phase (A). In the next phase (B) statistical variables can be derived or improved from web data. In both phases websites, 3rd party scraped data or other web data sources can be used. In this paper we list various ways of using web data as examined in the international Web Intelligence Network (WIN) project, WP 3 (new use cases), UC5 (business register quality enhancements). We give some examples, and list the advantages and challenges. The main focus is on phase A and we will dive a bit deeper into a practical case of using 3rd party scraped data.

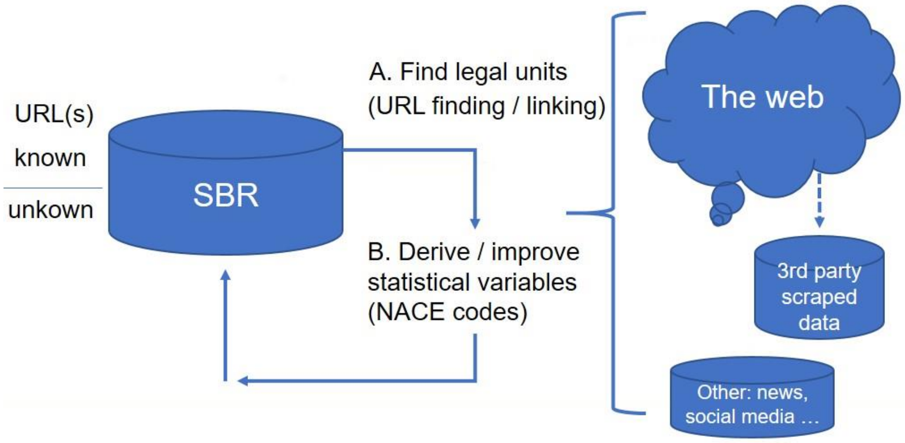


Figure 1: General view on business register improvements from web data sources

sEARCH

Search engines are used to find URLs for legal units or to verify known URLs. A search query is executed containing the name of the unit, possibly augmented with identifying information such as the municipality or a chamber of commerce (COC) number. This needs to be automated, so a (paid) search application programming interface (API) might be a better mechanism than interpreting the human readable search result pages, however both approaches are possible. Search results have to be interpreted to select the best match. The snippet – a short textual extract of the results page – can be used to do this, or alternatively, URLs returned by the search engine can be scraped to select the most relevant hits. This however involves an extra scraping step.

In countries where it is obligatory to mention identification numbers on enterprise websites these can be used for direct exact linkage. When such legislation is absent, machine learning techniques have proven to be useful for selecting relevant search results. Based on a labelled training set of valid and invalid search hits a model can be trained that catches the search engine behaviour for that particular search engine. The set of legal units in the SBR with known URL can serve as a training set. Since search engines evolve over time the model used has to be retrained periodically.

Search engines should be used with care as identifying information contained in the query could be identified in web server logs. This phenomenon is known as *search engine leakage*. One can reduce the risk by carefully designing the queries, spreading them across different search engines or by making a non-disclosure agreement with a search provider. Search engine leakage is to be taken seriously but manageable. A more in-depth discussion of using search engines for URL discovery can be found in [1] and [2].

sCRAPE

Scraping websites that belong to legal units in the SBR can only be done if the URL(s) for that unit is known or discovered by search or linking to other sources. If that is the case scraping is typically done via the *generic scraping* concept, for which unlike *specific scraping*, no prior knowledge on the structure of that site is available. Generic scraping typically starts at the home page and recursively visits deeper pages up to a certain maximum depth. Decisions have to be taken whether to store the complete website, only the texts, or the derived data (or all).

A *focused scraper* does not follow all links but gives priority to those that are expected to contain the most valuable information for the task at hand. For example to detect economic activity, a focussed scraper might give priority to the 'about us' page.

It is important to check that the website visited does match the legal unit at hand. The existence of identifying information on the site is crucial for this task. National legislation might force enterprises to put such information on their website, such as Chamber of Commerce or tax id, but this is not always the case. Special care should be given to manyto-many relations between legal units and websites. Depending on the business activity an enterprise might run many different websites. Contrary, small businesses might not have their own website and use a 'business collection' page to advertise their services.

An example of a more extensive discussion of scraping for statistics can be found in [3].

3RD PARTY WEB DATA

Web data collected by 3rd parties can be a useful additional input. Re-using such data saves resources. On the other hand, a (paid) agreement has to be made and the dependence on the 3rd party has to be managed. This is feasible only if the added value of the data is considerable. An example of a 3rd party collecting web data is the company DataProvider (DP). For over 2 years they provide Statistics Netherlands with a monthly DP dataset with URLs of Dutch businesses and additional variables. This data has been linked to the SBR using contact variables such as COC number, domain, email, zip code and phone numbers. Contact information might be missing on websites or in the SBR. Figure 1 shows the gaps in the DP data (left panel), and the corresponding SBR linkage variables (right panel).

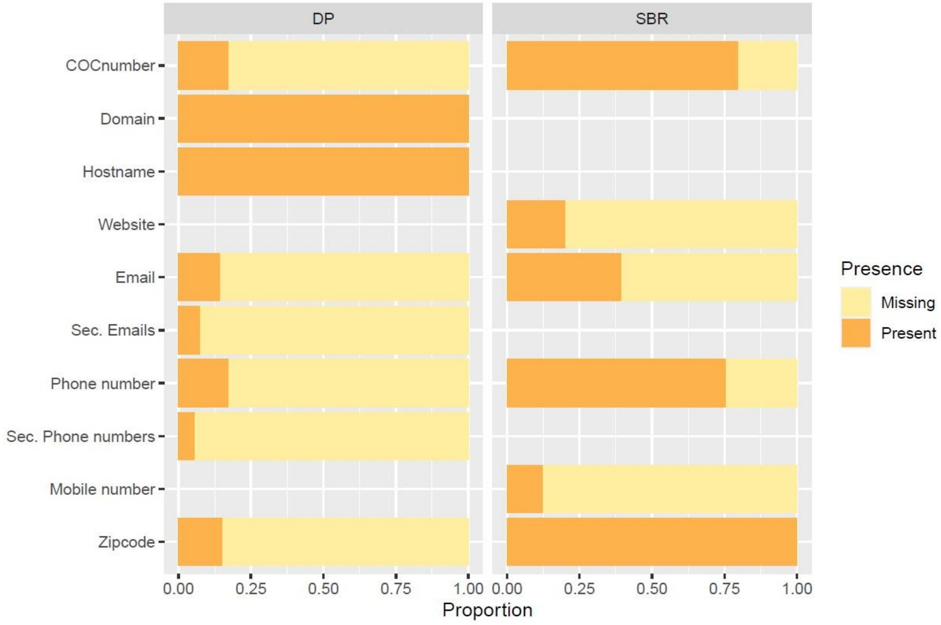


Figure 2: Completeness of linkage variables in DP data and SBR.

Note that the link between DP data and legal units can be one-to-one (1:1), one-to-many (1:n), many-to-one (n:1) or many-to-many (m:n). This makes it a complex task. Table 1 summarizes the results. At 75% linkage probability we have for 528 780 of the 4 630 836 legal units in the SBR (11%) a one-to-one match with the DP data. Moreover for 111 904 + 4 863 legal units (2.5%) there are multiple DP hits that may help improve SBR data. These results can probably be improved if the linking strategy is further refined. The value of the additional variables is still to be examined.

Table 1. Nr. of legal units (LUs) by linkage cardinality at 75% linkage probability

	#URLs in DP			
# LUs	2+ (n)	1	0	Total
2+ (m)	4863	27935	3957354	4630836
1	111904	528780		
0	5057922		X	X

The analysis shows that in this case for about 14.5% of the legal units a URL could be deduced from 3rd party data, which indicates that this approach is valuable. More information on linking all kinds of data (incl. web data) to a SBR can be found in [4].

oTHER WEB DATA SOURCES

One of the other data sources that have been used in this context is the internet *Domain Name System (DNS)*. This register of domain names and IP addresses is present in all countries and could be useful to deduce domain ownership. However, the degree of openness of this data varies per country and domain. Another data source of interest are *news and social media* messages. There might be some descriptive value in a press release or a social media message about the real activities performed by a legal unit. Yet another data source could be *job ads*. Enterprises typically put short descriptions of their main tasks and activities in it and since these texts are created explicitly for this goal by the company itself, they might be a valuable extra information, if the job ad can be linked to a legal unit.

dERIVING STATISTICAL VARIABLES

The scraped data from websites, 3rd party data or additional sources can be used to supplement the SBR or to derive or improve statistical variables. An example of the first is to add email addresses or phone numbers found on websites to the SBR. An example of the second is detecting economic activity (NACE codes) from website texts. The latter requires interpretation of raw texts and usually involves natural language processing (NLP) and machine learning techniques. Other statistical variables that have been derived from web data are, degree of innovativeness, degree of sustainability, operating a web shop or not, or belonging to the platform economy. For a more detailed example on deriving statistical variables from website texts we refer to [5].

wRAP-UP

In this paper we presented a high-level view on business register improvements using web data. URL finding concerns the discovery of URLs of legal units that do not have their website registered in the SBR. Search engines can very well be used for this. The textual paragraphs of the search results (snippets) can be used in combination with machine learning techniques to select valid hits. Alternatively a scrape on the found URLs can be performed. Search engines must be used with care, but the risk of search engine leakage is manageable. Once URLs for legal units are known in the SBR or found via URL finding, statistical variables, such as NACE code, can be derived. This involves scraping and interpreting the results using NLP and machine learning techniques. Third party data can be used as an alternative to scraping by the NSI. Other web data sources such as DNS, news or job ads can also be used. The best approach to improve a business register from web data is probably a mix of web data sources and techniques that best meets the country specific situation in the SBR and in the national web.

rEFERENCES

[1] A. van Delden, D. Windmeijer, O. ten Bosch, *Searching for business websites*, CBS Discussion paper, Dec. 2019, Searching for business websites (cbs.nl)

[2] H. Kühnemann, et. al. *Report: URL finding methodology*, WIN project, 2022-01-31 Report: URL finding methodology (europa.eu)

[3] G. Barcaroli, et. al *Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises*, Qconf, Wien 2014

[4] L. Ryan et. al., *An SBR spine as a new approach to support data integration and firm-level data linking*, IAOS 36 (2020) pp. 767–774, doi/10.3233/SJI-200640

[5] Daas, P.J.H., van der Doef, S. (2020) *Detecting Innovative Companies via their Website*. IAOS 36(4), pp. 1239-1251, doi/10.3233/SJI-200627

Innovating Business Data Collection: System-to-System Data Communication applied to an agricultural survey

Keywords: automated questionnaire completion, APIs, microservice, official surveys

INTRODUCTION

Sample surveys using questionnaires are a primary data collection method. In the 20th century sample surveys have proven to be a cost-efficient method to produce accurate statistics, although they come with substantial costs both for the National Statistical Institutes (NSIs) and businesses, who may experience high response burden [1]. Nowadays in the information age, there are a lot of new digital data sources in smart industries, also called “Industry 4.0” (e.g. [2]), such as smart (or precision) farming (e.g. [3], [4]). Increasingly these data sources provide Application Programming Interfaces (APIs), through which these types of data are available. This API-based communication allows systems to communicate without human intervention and makes System-to-System (S2S) data collection possible ([5], [6]).

Completing (business) survey questionnaires involves a number of steps [1]:

1. Comprehension of a question and the task(s) to answer a question
2. Retrieving the internal data needed to answer the question
3. Computation and evaluation of the answer
4. Reporting of the answer

For business respondents steps 2 and 3 generally need a lot of work. For step 2, this includes collecting the data from internal data sources, which may involve colleagues from other departments. The retrieval process becomes more complicated in case of mismatch between the requested and available data. This is followed in step 3 by calculating the required answer (often totals) based on the retrieved data. Studies show that these activities are time consuming and considered burdensome. Consequently business respondents (like famers) may adopt a satisficing response behaviour, resulting in measurement errors. Farmers asked for a less burdensome and more efficient way of reporting [4]. With modern technology, both the retrieval and computation step can be automated, provided that the data are electronically available and can be accessed using IT technology.

We will discuss an S2S data collection method, combining IT technology and survey methodology, which results in the automated pre-filling of an electronic agricultural questionnaire using an API provided by a smart farming machine manufacturer, John Deere: the MyJohnDeere API. The aim is to automatize and replace the manual completion of questionnaires including the manual retrieval and re-keying of data. This project is part of a Eurostat funded research project.

S2S data collection methods could allow NSIs to:

a) Replace (parts of) surveys, in particular replace manual completion of questionnaires.

b) This would reduce data collection costs of NSIs (in the long-term) and businesses (i.e., response burden).
c) Develop new statistics (including real-time statistics) from the new data sources, and provide useful and more detailed information back to the businesses.

In this paper we will discuss a pilot prototype we developed targeting the first goal. The prototype implements a S2S data collection method using APIs: data available in the MyJohnDeere cloud are automatically collected to pre-fill the Crop Yield Survey questionnaire. This questionnaire is sent yearly to sampled farmers to be completed at the end of each year. An exploratory study that we did prior to developing this prototype showed that the data required to complete this questionnaire are available in MyJohnDeere. These data are generated by sensors in John Deere machines used in precision farming. In this study John Deere is used as a first Proof of Concept.

In section 2 we discuss this new method in detail. While we have tested the pilot prototype with open data (see Section 3), we plan to test the field pilot with real data from farmers. Section 4 concludes this paper.

mETHOD

For the S2S data collection, we have chosen a microservice architecture, including an authentication and data collection microservice, as is shown in Figure 1. The authentication microservice makes sure that the farmer can log-in to his MyJohnDeere data in the cloud. The data collection microservice acts like an intermediary between the electronic questionnaire (eQ) and the John Deere cloud.

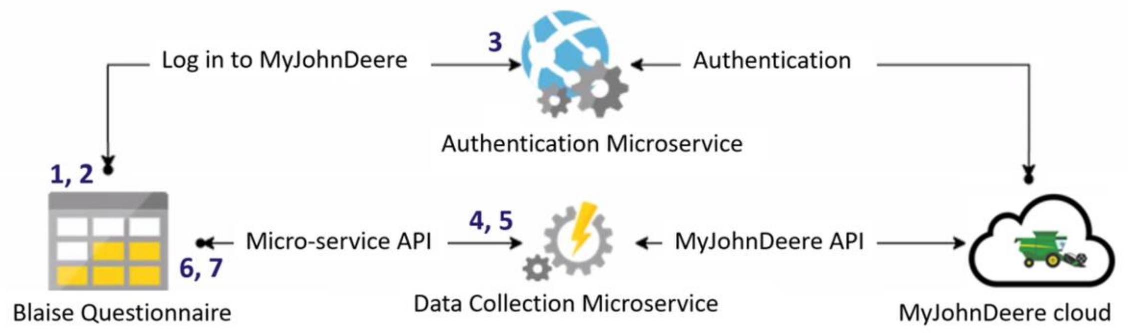


Figure 1. Microservice architecture for S2S data collection based on APIs

The process sequence farmers will follow, starts by logging onto the Blaise eQ:

1) A sampled farmer logs onto the online questionnaire according to the common procedure: sampled units receive an invitation letter with login details, including a web address, user name and password. After having opened the web page and after having keyed in the username and password, they enter the eQ immediately.
2) Instead of starting to complete the regular questions, now the farmer first is asked if they have John Deere smart farming machines, use MyJohnDeere and if they want to use their John Deere data to pre-fill the questionnaire.
3) If the answer is “yes”, the farmer follows an authentication process that gives us temporary and partial access to their data in MyJohnDeere. This is done without sharing the farmer’s credentials. This authentication protocol is based on the standard delegation protocol

OAuth 2.0. Now, the microservice can make API calls to the John Deere cloud (step 4). In case the answer is “no”, the farmer has to complete the questionnaire in the traditional manual way.

4) Via the “Microservice API” the online questionnaire asks for answers to the “Pilot Microservice”. The microservice browses the John Deere cloud looking for the appropriate data. These data are retrieved from the John Deere cloud via the “MyJohnDeere API”, and kept in memory until the answers to the questions are calculated. Right after, the answers are sent to the online questionnaire and imputed.
5) In the context of the Crop Yield questionnaire, in some cases we find ambiguities related to the identification of summer/winter crops per field. For instance, the crop harvested in a specific field is tagged as “wheat” and there is no seeding data. In this case there is not enough information to classify it as “winter wheat” or “summer wheat”. In this case an extra step is introduced where the farmer is asked (through a web form) to select “winter wheat” or “summer wheat” for the ambiguous field. At this point, the online questionnaire makes a second API call, the microservice recomputes the ambiguous crop totals (“winter wheat” or “summer wheat” in this example) and sends the updated answers to the questionnaire.
6) When all answers have been computed, these are are presented to the farmer in the questionnaire. The farmer can check and edit the pre-filled questionnaire. Questions that could not be pre-filled still have to be completed manually (if applicable).
7) After having checked and completed all questions, the farmer decides whether or not to send the answers. They can decide to start a next session at another time. The process ends when the answers to the questions are submitted.

Using this architecture, eQ software functionality can be extended without modifying the questionnaire software itself as long as it supports the following two features:

a) Communication with other computer systems without human intervention via APIs,
b) Electronic handling of data access permissions through an authentication protocol.

Since these features are based on standard software practices, this methodology could be easily adopted by other NSIs.

Most of the S2S data collection process is automated. This methodology automates the completion of the Crop Yield questionnaire but it doesn’t change the questionnaire itself. During login and authentication, human intervention is always required and may also be needed during disambiguation and final editing of the questionnaire.

PROTOTYPE TESTING WITH OPEN DATA AND NEXT STEP

Based on the MyJohnDeere platform, there is an ecosystem of applications developed by third party software companies that provides digital services to John Deere customers. Applications can be run in two different modes: "sandbox" (only for testing purposes and production) and production. Our pilot prototype has been already successfully tested in the "sandbox" mode. In order to do that, we have created a virtual farm in the platform that have been fed with open data provided by John Deere through its GitHub public repository.

The "sandbox" mode allowed us to test the S2S communication in a way that is very close to real conditions. This "sandbox" test showed that technically our system works well: the data collection microservice (as shown in Figure 1) calculated the answers to the questions in the Crop Yield Survey. The next step is a small-scale field pilot, which we are preparing at this very moment. In this field pilot the focus is on user experience (and response burden), costs, and data quality, in order to improve the architecture. At the conference to first results of the pilot will be presented.

cONCLUSION

Even though we still have to assess how this method will work in practice with data from farmers, this “sandbox” prototype demonstrates how data capture and processing can be automated. We concluded that this methods works technically, next we have to make it work in practice.

There are a number criteria that need to be fulfilled for this methodology to be applied in general:

a) Significant overlap between the data source and the questionnaire
b) The data source must have an API for S2S communication.
c) For privacy reasons, electronic handling of data access permissions should be possible.

John Deere and other big farming machine manufacturers that have similar APIs (for instance CNH Industrial, New holland, and Claas) have international presence in markets all over the world. For this reason, our methodology can be applied to arable farming in other countries. In addition, we are studying Farm Management Information Systems (FMIS), the farmer’s crop registration systems, in order to apply this methodology.

With regard to the Crop Yield Survey we focused on field operations data, but the John Deere cloud stores other types of data, like machine, agronomic service providers activity, soil and environmental conditions data. It is a huge potential data source for modernizing agricultural statistics. Other agricultural surveys like the crop protection survey, and surveys for other statistics (for instance, transportation and finance are fields where we can find a lot of APIs) can benefit from this methodology. Nowadays in the information age, there are a lot of new digital data sources in smart industries, like in smart (or precision) farming. In some cases, these data sources provide APIs. One of our next steps is checking business sectors for these conditions.

In addition, for large scale usage, there are still a few challenges that need to be fulfilled, like data harmonization, standardisation of S2S software, stability of the IT architecture in the future, and market penetration of advanced data systems by businesses [4]. We believe that this S2S data communication approach is valuable and promising, and can go beyond the questionnaire, by leaving the questionnaire out of the process, but there is still a long way to go.

rEFERENCES

[1] G. Snijkers, G. Haraldsen, J. Jones and D. K. Willimack, Designing and Conducting Business Surveys, Wiley, Hoboken (2013).

[2] B.R. Haverkort, and A. Zimmermann, Smart Industry: How ICT Will Change the Game! IEEE Internet Computing (2017), 21(1): 8-10.

[3] X. Pham, and M. Stack (2018), How data analytics is transforming agriculture. Business Horizons (2018), 61: 125-133.

[4] G. Snijkers, T. Punt, T., S. De Broe, S., and J. Gómez Pérez, Exploring sensor data for agricultural statistics: The fruit is not hanging as low as we thought. Statistical Journal of the IAOS (2021), 37(4): 1301-1314.

[5] N. Bharosa, R. van Wijk, N. de Winne, and M. Janssen (eds.), Challenging the chain: governing the automated exchange and processing of business information. IOS Press, Amsterdam (2015).

[6] G. Buiten, G. Snijkers, P. Saraiva, J. Erikson, A. G. Erikson, and A. Born, Business data collection: Toward Electronic Data Interchange. Experiences in Portugal, Canada, Sweden, and the Netherlands with EDI. Journal of Official Statistics (2018), 34(2): 419-443 (ICES-5 special issue).

Geospatial statistics (JENK3A.1)

Session Chair: **Marta NAGY-ROTHENGASS** *(Eurostat)*

Towards standardised geospatial statistics
Rina Tammisto *(Statistics Finland*), Jerker Moström (*Statistics Sweden*)

New and improved tools for spatial statistics and centralization of spatial data in Statistics Poland
Mirosław Migacz *(Statistics Poland)*

GSGF Europe – opportunity to re-assess national geospatial and statistical data practices
Igor Kuzma *(Statistical Office of the Republic of Slovenia-SORS)*

Towards standardised geospatial statistics

Keywords: geospatial, statistical, GSGF Europe, GIS

Introduction

Europe is confronted by a mix of societal, economic and environmental challenges such as climate change and its consequences, ageing societies, and political crisis causing economic stress and immigration. Understanding these major cross-border challenges and taking the right decisions requires new insight that we can only derive from new types of data and their combination. Through the adoption by the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) in 2018[1] and the United Nations Statistical Commission (UNSC) in 2019[2], the Global Statistical Geospatial Framework (GSGF) has been recognised as a framework for the world that provides an underlying mechanism to integrate statistical and geospatial information. As GSGF understands it, the integration means real collaboration and knowledge exchange between the geospatial and statistical domains. The goal is that statistics, as well as all other data producers, use the same sources of geospatial information and the same geocoding services – leading to more efficient use of resources and improved integration of data produced by different public organisations and the private sector.

Eurostat has supported a sequence of ESSnet GEOSTAT projects with the goal of building a common methodological foundation for the integration of statistical and geospatial information and to enable a more harmonised production of geospatial statistics in the European Statistical System (ESS). The project series started in 2010 and the last GEOSTAT project ended in spring 2022. The general objective of the GEOSTAT 4 project (2020-2022) was to foster the integration of statistics and geospatial information in the European context by bringing the Global Statistical Geospatial Framework (GSGF) into the European context.

Methods

The project produced a European extension of the GSGF (Global Statistical Geospatial Framework)[3]. The original model was complemented with new perspectives and supported with rich body of implementation guidance. Best practices, new methods and use cases for integrating statistics and geospatial data were described in a number of case studies. A number of requirements and recommendations were elaborated to support a more operational understanding of the different principles of the framework. In addition, the project brought a geospatial data perspective to the ESS quality framework of statistics and developed the first geospatial quality indicators and a quality checklist.

Results

The community behind the GEOSTAT projects is the European Forum for Geography and Statistics. The focus of the community has shifted from harmonisation of output data (gridded population data) to a more fundamental harmonisation of production processes through implementation of the provisions of the Global Statistical Geospatial Framework (GSGF). The implementation of the European extension of the GSGF is now the cornerstone that can be used when the evaluation of the present state needs to be done or actions for the development is the question.

Conclusions

In conclusion, there is now a solid framework in place to guide Member states in their efforts to modernise production processes and enhance the integration of statistical and geospatial data. The presentation highlights the successes of the current state and points of future development needs in order to ensure successful statistical geospatial data integration in the future.

References

[80] UN GGIM (2018). Report of the Session, E/2018/46-E/C.20/2018/19, page 14. https://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/GGIM8-report-e.pdf

[81] UN Statistical Commission (2019). Report on the fiftieth session, E/2019/24-E/CN.3/2019/34, page 29. https://unstats.un.org/unsd/statcom/50th-session/documents/Report-on-the-50th-session-of-the-statistical-commission-E.pdf

[82] GEOSTAT 4 (2022). GSGF Europe GEOSTAT Information Service. https://www.efgs.info/gsgf-europe-geostat-information-service/

New and improved tools for spatial statistics

Keywords: GIS, geostatistics, census, grid, INSPIRE

Introduction

Digitization has become an integral part of modern economies around the world. Innovative digital tools allow you to automate, simplify and speed up many processes, saving resources.

To meet user expectations, following trends in development of innovative products and services, Statistics Poland launched the Geostatistics Portal in a new, overhauled version.

Developed since 2011 and officially launched in 2013, the Geostatistics Portal is the place, where statistical data users can find everything they need to visualize data on thematic maps. Though the system was initially built for disseminating census results, the data scope and tool variety is growing since the launch.

Methods

Up to 2013 statistical data published by Central Statistical Office of Poland (CSO) was limited to official indicators, publications, press releases and tables in data banks. The closest thing to spatially referenced statistics were maps inserted into publications as static images. In 2013 CSO released the Geostatistics Portal – a web platform for spatial visualization of census' and other statistical surveys' results.

Agricultural Census 2010 and Population and Housing Census 2011 were the first censuses in Poland that were carried out without use of paper. That would not be possible without use of geographic information systems. Up to year 2008 vast amounts of geographic data were kept only on paper maps – today spatial data is kept and updated only in numeric form. Currently Central Statistical Office of Poland (CSO) is in possession of numeric data that provides spatial reference for all desirable survey frames.

All statistical data in the 2010-2011 census round was collected with reference to address point coordinates, which allows publication of census results aggregated to any spatial unit, whether it's a statistical, administrative unit or a grid cell. The only restriction is the need to maintain statistical data confidentiality. In order to publish census results CSO prepared the Geostatistics Portal. The portal provides tools for creating all sorts of thematic maps.

In 2013, Local Data Bank – the largest statistics database maintained by Statistics Poland – has been connected to the Geostatistics Portal and thematic maps can be prepared for all data from the Local Data Bank.

In 2016 Statistics Poland completed a project called "Geostatistics Portal – Phase II". Thematic map tool set was expanded to include advanced diagram maps and thematic mapping based on user data. Data from the Population and Housing Census 2011 was published on the European 1km x 1km grid, including: total population, population by sex and biological age groups and femininity ratio. In addition, spatial analyses on microdata were introduced for results of the 2010-2011 census round.

In 2020 Statistics Poland launched a new project called “Spatial Statistical Data in the Information System of the State” Project (PDS) to further develop the Geostatistics Portal.

Results

In the framework of the Geostatistics Portal, following public e-services are available:

- access to geostatistical data and analyses from computer and mobile devices,
- geostatistical exploratory analyses,
- geostatistical modelling,
- automated content enrichment.

All the above-mentioned services use the data possessed by official statistics and the results of geostatistical analyses, crucial for functioning of the state, local governments and local communities. Data can be presented in a convenient graphic form, in the form of maps and charts, which aims at speeding up decision-making processes. Solutions supporting presentation of current statistical surveys’ results are an additional advantage of the Geostatistics Portal.

Geostatistics Portal offers the following set of tools:

- **Resource catalog** for viewing ready-made data visualizations on maps;
- **Map portal - GUS data** for preparing data visualizations on Statistics Poland data, including the Local Data Bank;
- **Geocoding** for attaching coordinates or boundaries of administrative division to tabular data;
- **Analysis studio** for preparing visualizations and analyses and sharing them with other users;
- **Map studio** for preparing cartographic presentations of user data;
- **Composition studio**, for sharing data visualizations via network mapping services;
- **Resource manager** for storing, editing, sharing and managing data.

The Geostatistics Portal offers a broad choice of geovisualization tools. Customizable choropleth maps can be created to visualize statistical indicators, whereas absolute data can be shown using a powerful set of diagram map tools. Single and multiple phenomena can be presented using different kinds of diagram maps. Users can visualize absolute values, reflect a structure of a phenomenon or show its trend across a span of several years. Complex diagram maps allow presenting data in different units and different scales on one single map. Any thematic map created in the Portal can be printed out or saved as a document, while statistical data can be exported as a table.

The PDS project launched in 2020 focused on upgrading two existing e-services – Geostatistics Portal web and mobile versions – and building three completely new e-services for advanced statistical data processing: Exploratory geostatistical data analysis, Geostatistical modelling and Semi-automated user content enrichment. The development of these services results in the possibility of identifying, describing and visualising the spatial distribution of the analysed data, establishing spatial relations, correlations and clusters, examining the spatial heterogeneity or autocorrelation, and building the probabilistic models.

The new functionalities will be available to advanced users, but also to users without specialist knowledge in the field of statistical analyses. The implementation of automatic content analysis mechanisms (i.e. "text mining") and usage of metadata describing geostatistical analyses available in the PDS system, will provide users with the possibility of supplementing their own text with graphic elements.

The new Geostatistics Portal launched in April 2022 replacing the old version of the system.

Conclusions

Statistics Poland has made gigantic progress in the field of geographic information over the last decade. Maintaining a point based statistical survey framework allows georeferencing statistical survey results to point and presenting them on maps in any desirable way.

The overhaul of the Geostatistics Portal was another step forward in providing users with a convenient and intuitive way to access statistical data on maps. With a huge database and a wide range of visualization tools at their disposal, users can design their own dynamic map presentations rather than rely on static maps. New tools for advanced analyses on microdata and exploratory geospatial analyses are further improvements of an already robust set of tools for spatial statistics – all this to bring statistical data closer to the people and authorities and make governance easier on all administrative levels.

GSGF Europe – opportunity to re-assess national geospatial and statistical data integration practices

Keywords: Global Statistical Geospatial Framework (Europe), geospatial and statistical data integration

Abstract

Institutional collaboration between the statistical and geospatial communities has a long tradition in Slovenia. Geospatial information has been an integral part of official statistical production since the 1970s and has gradually resulted in the establishment of the Central Population Register and the Register of Spatial Units. Unique identifiers of addresses, buildings and dwellings equipped with coordinates of their locations were applied to all registers and records that followed, thus enabling the dissemination of various geospatial statistical products of high geographical detail. However, this compilation of georeferenced statistical data sources has never been systematically described or assessed considering statistical reference architectures or models. Introduction of the Global Statistical Geospatial Framework (GSGF) thus offered the opportunity to examine the national practices of geospatial and statistical data integration regarding the five principles and key elements of the framework.

The Register of Spatial Units represents the fundamental geospatial infrastructure for official statistics in Slovenia. It ensures the hierarchical consistency of the administrative boundaries and contains unique identifiers of centroids (coordinates) of buildings with an address for the period since 1995. These identifiers are integrated in core registers applied to the statistical production process (Population Register, Business Register, Register of Agricultural Holdings, Register of Employment, Tax Register, etc.). For geocoding and georeferencing purposes, public registers and records can use only authoritative geospatial information provided and daily updated by the Surveying and Mapping Authority of the Republic of Slovenia. Geospatial data management environment is maintained according to the national legislation and standards. The institutional data entering the statistical production process are primarily already georeferenced and are further managed as point-based statistical data. Thus geospatially enabled statistics can be disseminated in a consistent and harmonised manner. Geospatial statistics in Slovenia are systematically published only by official administrative boundaries from the Register of Spatial Units. In 2008, the national system of hierarchical grid was established that is recognised as standard grid division of the national territory. Additionally, the INSPIRE grid can be applied as well. The smallest administrative unit used for dissemination of geospatial statistics is a settlement and the smallest grid cell size is 100m x 100m. Geospatial statistics can be discovered and accessed by means of a geospatial web service STAGE (WMS, WFS) as open data.

The first review of the national practices on statistical and geospatial data integration considering the GSGF is now being iterated as the GEOSTAT4 project provided the European version of the Framework. The findings of this examination will be presented.

Price data (MANS3A.1)

Session Chair: **Doris Rijnbeek (***Oesterreichische Nationalbank***)**

Evaluating energy prices and costs impacts on households and industry’s costs
Stavros Lazarou *(Eurostat),* Viktor Hauk *(European Commission - DG Ener)*

Automatic classification for price statistics using machine learning methods
Bogdan Oancea *(National Statistical Institute-INSSE and University of Bucharest)*, Marian Necula *(National Statistical Institute-INSSE)*

Price imputation methods based on statistical algorithms
Botir Radjabov *(GOPA Luxembourg)*

Evaluating energy prices and costs impacts on households and industry’s costs

Keywords: electricity, natural gas, prices, costs

Introduction

On the demand side of the electricity market, residential consumption trends are expected to be mostly driven by the increasing number of households, proliferation of electric appliances or the electrification of heating, while energy efficiency measures such as installing LED lightbulbs, more efficient appliances and smart meters may push electricity demand lower. Average temperatures play an important role too, both for heating and for cooling. In the case of businesses, the consumption of electricity is mainly influenced by two similarly countervailing factors: the level of economic activity and energy efficiency measures. Whilst demand is somewhat inelastic with prices, i.e. there is a minimum electricity use necessary for households or businesses, the increases in prices in the last two years are understood to make price an increasingly important driver in reduced demand.

Household electricity prices

EU household electricity retail prices in the DC band grew by 22% between 2010 and 2021 (see Figure 1 [left]); while the energy crisis will lead to further sharp increases in both wholesale and retail prices for 2022. The average annual growth rate of retail prices within the DC band were 2.5%/y from 2010 to 2020; this grew to 8% between 2020-2021. In absolute terms, the average EU27 retail price grew from 215 EUR/MWh to 231 EUR/MWh (+16 EUR/MWh) in the same period of time. This movement between 2019 and 2021 was driven by Energy and supply (+16 EUR/MWh) and Network costs (+3.6 EUR/MWh), whilst taxes and levies declined a little in this period (-3.1 EUR/MWh).

When looking at the cost composition between the cost of the energy, network costs and taxes and levies (see Figure 1 [right]) the taxes and levies category saw its share in the total bill decrease significantly from 40% in 2017 to 37% in 2021 (further analysis of this component is provided in the next section). Meanwhile, energy component of the prices was decreasing between 2010 and 2020, and so was the contribution of energy and supply costs to the total bill. However, in 2021, the contribution of the energy component increased to 37% of the total energy bill in 2021 – the first time since 2014. Most importantly, energy and supply costs in 2021 exceed all recorded values since 2010.

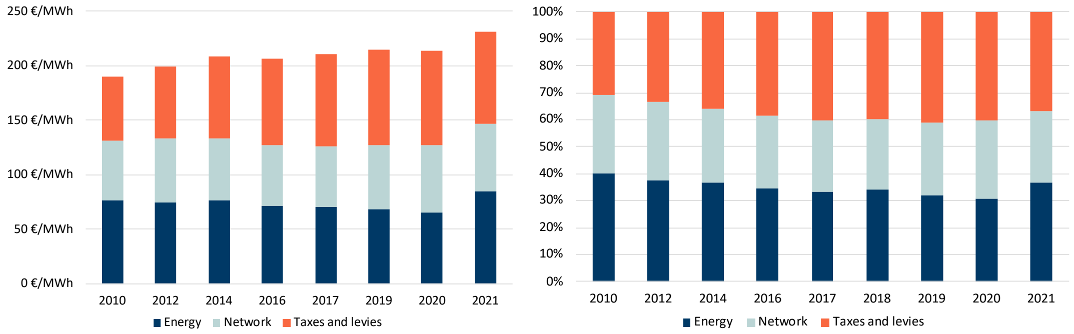


Figure 1 – Evolution and composition (left) and relative composition (right) of the EU household price (DC band)

Source: DG ENER in-house data collection, Eurostat [1]

Composition of taxes, levies, fees and charges

In order to better understand how Member State policies and fiscal instruments impact household retail prices, the taxes, levies, fees and charges category is further broken down into six subcomponents: VAT, renewable taxes, capacity taxes, environmental taxes[92], nuclear taxes and other. Note that only policies and mechanisms that directly impact retail prices are considered, and not all tax subcomponents exist or are applied in all Member States. The following chart displays an evolution of EU27 averages.

Figure 2 shows that taxes and levies associated with policies designed to support renewable energy sources have seen a decline since 2019 from 26 EUR/MWh to 24 EUR/MWh. Meanwhile capacity taxes continue to gradually increase over time. Since 2010, capacity taxes have increased by 288%, although at 3.9 EUR/MWh in 2021, still represent only a minor part of the tax component. While environmental taxes and levies had been following an increasing trend (+49%) since 2010, their contribution has reduced a little after peaking at 19 EUR/MWh in 2019 and were, in 2021, 17.9 EUR/MWh, or 7% less than in 2019. The trends highlight that taxes and levies for renewables and environmental purposes (and the policies they support) can be amongst those most vulnerable when pressures on prices and taxes emerge. The elimination of the EEG surcharge in Germany in 2022 is expected to drive further reductions in the environmental taxes and levies component in the next few years.

[92] This category includes general energy taxes, which are typically classed as having an environmental purpose

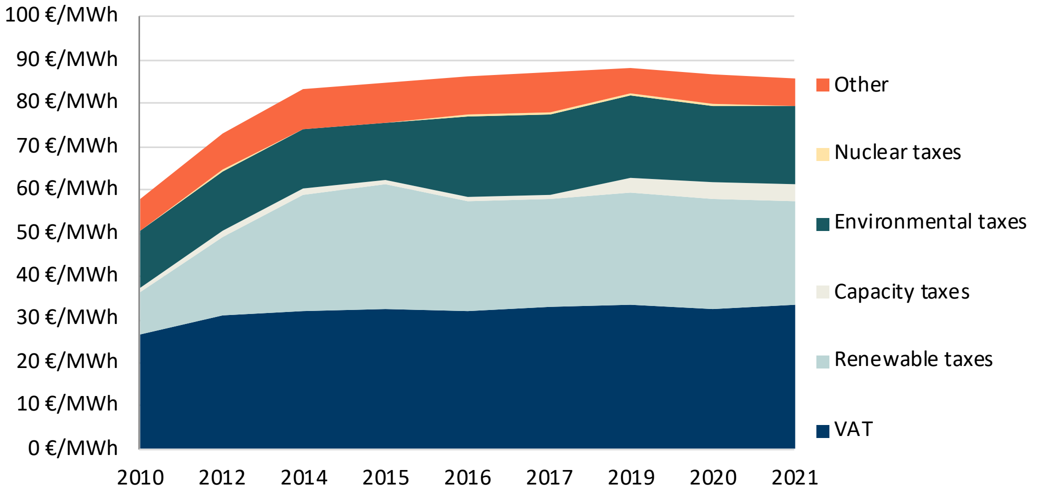


Figure 2 – Evolution of taxes, levies and charges for EU households since 2010 (DC)

Source: DG ENER in-house data collection, Eurostat [1]

The structure of the taxes and levies component between 2019 and 2021 showed minor changes, with the most notable being the small downward shift in the environmental taxes and renewable energies contribution (both -1%), while VAT and capacity tax increased (both +1%) (Figure 3).

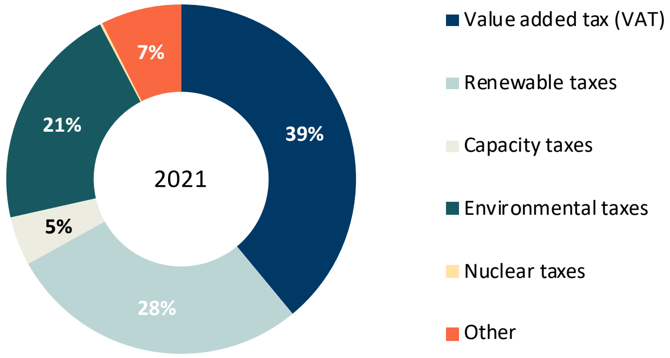


Figure 3 – Composition of the taxes and levies component of household electricity prices in 2021 (DC band)

Source: DG ENER in-house data collection, Eurostat [1]

Non-households electricity prices

Figure 4 shows that industrial electricity prices in the ID band grew at an average annual rate of 2%/y during the last decade, overall showing an increase from 96 EUR/MWh in 2010 to 124 EUR/MWh in 2021. Since 2019, industrial electricity prices increased by 14%, from 108 EUR/MWh to 124 EUR/MWh (+16 EUR/MWh) in absolute terms. This is the highest 2-year increase in prices observed within the past decade.

Due to the exclusion of VAT and other factors related to tariff calculations, industrial electricity prices are more influenced by the energy component compared to households and hence, more driven by developments in the wholesale market. The energy component, despite a small dip in prices in 2020, increased by 24% (+12 EUR/MWh) in 2021 compared to 2019. Network charges also contributed, increasing by 14% (+3 EUR/MWh) since 2019, and these are at their highest level since 2010 at 23.3 EUR/MWh. The lowest increase was observed in levies and taxes which only saw an increase of 1.6% (+0.6 EUR/MWh) between 2019 and 2021. This small increase is in contrast to the small decrease in taxes and levies (-3 EUR/MWh) experienced for household retail prices, see figure 4.

In terms of energy price components, in 2021 the energy component now contributes 51% of the total industrial price for electricity, compared to 47% previously. Meanwhile, network charges' contributions remain unchanged at 19% and a substantial decline was observed for taxes and levies, which in 2021 contributed 30% to the electricity price compared to 34% in 2019.

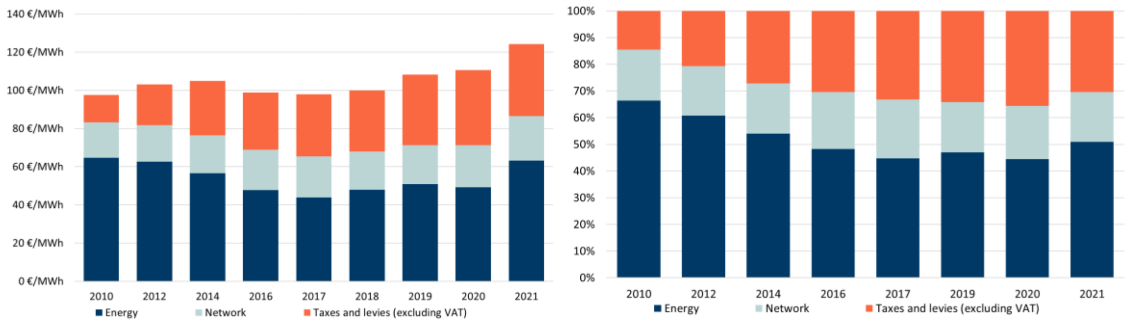


Figure 4 – Evolution and composition of the EUR 27 industrial retail prices (ID band), absolute (left), share (right)

Source: DG ENER in-house data collection, Eurostat [1]

References

[1] Eurostat database (Eurobase): https://ec.europa.eu/eurostat/web/main/data/database

Price imputation methods based on statistical algorithms

Keywords: price statictics, hedonic regression, regression tree, bagging tree, random forest

1. Introduction

Hedonic linear regression is a usual tool to estimate missing prices in price statistics. One of the most significant issues with hedonic linear regression is its functional form, which assumes that products characteristics are perfectly linearly substitutable. This, however, might not correspond to real consumer preferences once they make a substitution from one characteristic to another. Therefore, prediction accuracy of hedonic linear regression might suffer and is chalenged in this paper by prediction accuracy of other statistical algorithms.

2. Methods

2.1 Hedonic linear regression

Hedonic linear regression [1] can be defined as

$$\ln p_{it} = \alpha + \delta D_t + \sum_j \beta_j z_{ijt} + \epsilon_{it}$$

where natural logarithms of prices are regressed on time dummies and on product characteristics. This regression can also be expressed as

$$\ln p_{it} = g(D_t, z_{it}) + \epsilon_{it}$$

or as

$$y_i = g(x_i) + \epsilon_i$$

2.2 Regression tree

The objective of a regression tree [2] is to split observations into several distinct and non – overlapping regions $\{R_1, \ldots, R_M\}$ so that least squares are minimised within each region R_M such that

$$\bar{y}_{R_m} = \underset{\hat{g}(x_i)}{\arg\min} \sum_{y_i \in R_m} (y_i - \hat{g}(x_i))^2$$

The final goal of a regression tree is to figure out the set of regions that minimizes the sum of squared errors such that

$$\{R_1^*, \ldots, R_M^*\} = \underset{\{R_1,\ldots,R_M\}}{\arg\min} \sum_{R_m} \sum_{y_i \in R_m} (y_i - \bar{y}_{R_m})^2$$

2.3 Bagging tree

To improve prediction accuracy of regression trees, a bagging procedure or a bootstrap aggregation procedure has been developed [3]. The main idea of a bagging procedure is to take an average value of predictions on multiple bootstraped samples of the original observations rather than obtaining a single value for each prediction such that

$$\hat{G}(\boldsymbol{x}_i) = \frac{1}{B}\sum_{b} \hat{g}_b(\boldsymbol{x}_i)$$

2.4 Random forest

To account for a possible correlation of predictions, a random forest procedure has been developed [4]. The main idea of random forest procedure is to select different sets of independent variables for each tree with an aim to reduce correlation of predictions.

3. Results

To be able to test prediction accuracy of all of the above methods, the initial dataset of consumer electronics (tablets) has been divided into training dataset (70 percent of observations), which is used to train the respective models, and testing dataset (30 percent of observations), which is used to test the respective models` predictions. The following table illustrates prediction accuracy of all of the above mentioned methods. The predictions, which are within ± 2% range of true values are considered to be correct.

Table 1. Prediction accuracy of the selected methods

Hedonic Linear Regression	Regression Tree	Bagging Tree	Random Forest
80 %	93%	94%	96%

4. Conclusion

As it can be seen from the results of this paper, prediction accuracy might be improved if not hedonic linear regression but tree based algorithms are used. The importance of prediction accuracy is vital for producing valid price indices in case the underlying data experiences a lot of missing products.

References

[1] de Haan, J. (2010). Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and Re-Pricing Methods. Jahrbücher für Nationalökonomie und Statistik, 230(6), 772-791.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Routledge.

[3] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[4] Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Automatic classification for price statistics using machine learning methods

Keywords: machine learning, classification, word embeddings, price statistics

Introduction

The modernization strategy of NSI Romania includes the adoption of new methods for statistical processes and the integration of new data source into the statistical production. Thus, 3 years ago we started a process of collecting data regarding prices through web scraping techniques from the main e-commerce national sites (Oancea and Necula, 2019). Web scraped data were not directly used for the computation of CPI yet, but they complement the traditional data collected by interviewers for comparisons, monitoring and other quality checking. The volume of the web scraped data is very high, around 50.000 records collected every week. In order to be possible to use them, products must be classified according to the categories used for CPI computation. Manually labelling is almost impossible due to the high volume of data, therefore we piloted an automatic classification process, starting with a reduced set of product categories and we chose 15 categories from ECOICOP international classification, from food and household appliances categories. Table 1 shows the categories and the number of products in each category.

Table 9. Number of products in each class

ECOICOP class	Number of products	ECOICOP class	Number of products
01.1.1.2	74	01.1.6.1	17
01.1.1.3	14	01.1.7.3	21
01.1.4.1	45	01.1.7.4	21
01.1.4.2	36	01.1.8.1	33
01.1.4.7	59	05.3.1.1	931
01.1.5.1	45	05.3.1.2	767
01.1.5.3	88	05.3.1.3	660
01.1.5.4	42		

Methods

Firstly, we randomly selected a sample with 2853 records and manually labelled the products according to ECOICOP international classification.

Secondly, in order to be able to use different machine learning techniques for automatic classification, we transformed the product names, which are text data, into numeric vectors (embeddings). We used different techniques to build product name embeddings: Count Vectorization introduced Spark (1972), TF-IDF (Rajaraman and Ullman 2011), Word2Vec (Mikolov et al., 2013, Fasttext (Joulin et al, 2016) and Glove (Pennington et al, 2014). For

Word2Vec we built the embedding of a product name in two different ways: by adding the embeddings of each word composing the product name and by averaging the embedding of each word. When building product names embeddings we limited to 3000 the dimensions of the feature vectors for Count Vectorization and TF-IDF methods and to 50 for Word2Vec, Fasttext and GloVe to keep the running time under acceptable limits for an experiment. For Count Vectorization and TF-IDF methods, we considered not only the words composing the product names, but also n-grams with a maximum of 3 words. For Word2Vec and Fasttext methods we applied both the CBOW and SKIP-Gram approaches. Thus, we built 9 different embeddings for each product.

Thirdly, having the numerical representation of product names, we proceeded with a series of supervised machine learning techniques for automatic classification: logistic regression (Mertler and Vannatta, 2002), Naive Bayes (Xu, 2018), Decision trees with Gini and information gain to perform the split (Wu et al., 2008), Bagged trees (Kotsiantis, et al., 2005), C4.5 and C50 variants of decision trees (Quinlan, 2014), Support Vector Machines with radial and sigmoid kernels (Cortes and Vapnik, 1995), Random forests (Breiman, 2001), K-Nearest Neighbours (Mucherino, 2009), Artificial Neural Networks (McCulloch and Pitts, 1943), and eXtreme Gradient Boosted Trees (Chen and Guestrin, 2016).

After dividing the data set into training and testing subsets, we fitted each method mentioned above on the training data set and then applied the fitted model on the testing data sets. We even implemented a grid search procedure to choose the optimum values for the parameters of the classifiers. However, the grid search is time consuming, and some parallel programming techniques should be used in the future to keep the running time reasonable. For this experiment, we used the grid search only for SVM, XGBoost, KNN and ANN.

Results

We computed the confusion matrix, accuracy and F1 metrics for each classification method. Regarding the F1 score, we computed it in two variants: as a simple average and as a weighted average of F1 for each class. The weights were computed as the inverse of the frequency of each class, coping thus with the pronounced imbalance of the classes.

The results showed an impressive accuracy of classification with values ranging from 0.76 to 0.99. The best classifiers proved to be the multinomial logistic regression, the Naive Bayes, Random Forests, ANNs, and SVM with the radial kernel if one considers the accuracy metric and Random Forests, ANNs, and SVM with the radial kernel if the weighted F1 is to be considered. Regarding the vectorization methods, the best results were obtained with the Count Vectorization, TF-IDF, GLOVE and FASTTEXT, while Word2Vec showed the lowest accuracy for almost all classification methods used in our study. All data processing was performed using the R software system ver. 4.2 on a desktop computer with an Intel Xeon E3-1246 at 3.50GHz and 16 Gb of RAM.

Table 2 shows all the ML methods tested together with the word vectorization technique that gave the highest accuracy while table 3 shows one method (C50 variant of the decision trees) with all word vectorization techniques.

Table 2. Accuracy of classification for different ML techniques

Classification method	Word vectorization	Accuracy	F1	Weighted F1
Logistic regression	Count vectorization	0.976	0.924	0.859
Naïve Bayes	Count vectorization	0.989	0.943	0.877
Decision trees (Gini)	TF-IDF	0.963	0.961	0.882
Decision tress (information)	TF-IDF	0.952	0.937	0.852
Bagged decision trees	TF-IDF	0.983	0.945	0.879
C4.5	Count vectorization	0.984	0.945	0.880
C50	Count Vectorization	0.984	0.946	0.881
Random forests	GLOVE	0.991	0.971	0.904
KNN	TF-IDF	0.987	0.945	0.880
ANN	GLOVE	0.991	0.975	0.906
SVM (radial kernel)	FASTTEXT SKIP GRAM	0.997	0.988	0.922
SVM (sigmoid kernel)	FASTTEXT SKIP GRAM	0.982	0.931	0.861
XGBoost	Count Vectorization	0.986	0.945	0.880

Table 3. Accuracy of classification (with C50) for different word vectorization techniques

Word vectorization	Accuracy	F1	Weighted F1
Count vectorization	0.984	0.946	0.881
TF-IDF	0.983	0.945	0.879
Word2Vec CBOW (ADD)	0.893	0.737	0.678
Word2Vec CBOW (MEAN)	0.883	0.695	0.633
Word2Vec SKIP GRAM (ADD)	0.869	0.678	0.621
Word2Vec SKIP GRAM (MEAN)	0.850	0.679	0.620
FASTTEXT CBOW	0.910	0.741	0.679
FASTTEXT SKIP GRAM	0.944	0.746	0.679
GLOVE	0.932	0.766	0.621

Conclusions

The results obtained are very encouraging, automatic classification methods tested so far showing very good performances. These performances can be explained by the fact that the product names do not vary much from one retailer to another. Once a classification model is fitted on a training data set, the “unseen” data provided largely follow the same rules to build product names and the classification will show good results.

However, there are some issues still to be solved in the future. We only run the classification algorithms on a small dataset and even so, the computing time was high especially for the methods where we used a grid search to choose the optimum values of the parameters. This problem will be more acute when we will try to classify bigger data sets and special parallel programming techniques should be considered. Another problem that we should consider for the future is the out-of-vocabulary (OOV) words. While the FASTTEXT embedding method can cope with such words, the other methods cannot treat OOV words. A strategy to alleviate this problem should be devised too.

References

[83] Oancea, Bogdan and Necula, Marian. “Web Scraping Techniques for Price Statistics – the Romanian Experience”, Statistical Journal of the IAOS, 2019, vol. 35, no. 4, pp. 657-667

[84] Sparck Jones, K., "A statistical interpretation of term specificity and its application in retrieval", 1972, Journal of Documentation, Vol. 28 No. 1, pp. 11-21, https://doi.org/10.1108/eb026526

[85] Rajaraman, Anand and Ullman, Jeffrey (2011). Data Mining. Mining of Massive Datasets. 2011, pp. 1–17 Cambridge University Press, doi:10.1017/CBO9781139058452.002

[86] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean Jeffrey, “Efficient Estimation of Word Representations in Vector Space”, 2013, arXiv, https://doi.org/10.48550/ARXIV.1301.3781

[87] Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas,” Bag of Tricks for Efficient Text Classification”, 2016, arXiv, https://doi.org/10.48550/arxiv.1607.01759

[88] Pennington, Jeffrey and Socher, Richard and Christopher Manning. “GloVe: Global Vectors for Word Representation”, In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[89] Mertler, C. and Vannatta, R. Advanced and multivariate statistical methods (2nd ed.). 2002, Los Angeles, CA: Pyrczak Publishing.

[90] Xu, Shuo, „Bayesian Naïve Bayes classifiers to text classification”, Journal of Information Science, 2018, vol. 44, pp. 48-59.

[91] Wu, X, and Kumar, V and Quinlan, JR and Ghosh, J and Yang, Q and Motoda, H, et al. Top 10 algorithms in data mining. Knowledge and information systems. 2008, 14(1):1–37.

[92] Kotsiantis, S.B., Tsekouras, G.E., Pintelas, P.E. Bagging Model Trees for Classification Problems. In: Bozanis, P., Houstis, E.N. (eds) Advances in Informatics. PCI 2005. 2005, Lecture Notes in Computer Science, vol 3746. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11573036_31

[93] Quinlan J. C4.5: programs for machine learning. Elsevier; 2014.

[94] Cortes, C and Vapnik V. Support-vector networks. Machine learning. 1995, 20(3):273–97.

[95] Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[96] Mucherino, A., Papajorgji, P.J., Pardalos, P.M. k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications,2009, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

[97] McCulloch, W. S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 1943, 5(4), 115–133

[98] Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016 (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Geospatial statistics 2 (JENK3A.2)

Session Chair: **Hannes Reuter** *(Eurostat)*

Wasserstein Distances for Spatial Statistics: The Spatial-KWD library
Stefano Gualandi *(University of Pavia)*, Fabio Ricciato *(Eurostat)*

On the Application of the Fourier-based Distance to Spatial Statistics
Stefano Gualandi *(University of Pavia)*, Andrea Codegoni *(Università degli Studi di Pavia)*, Fabio Ricciato *(Eurostat)*

Exploring Spatial and Demographic Official Statistics on Personal Insolvency
Sven Brocker (University of Duisburg-Essen), Jonas Klingwort *(Statistics Netherlands-CBS)*, Christian Borgs *(IT.NRW)*

Geocoding millions of addresses in a reproducible manner for Big Data Climate Risk analysis
Jörn Franke, Daphné Aurouet, Malgorzata Osiewicz (*European Central Bank*)

Wasserstein Distances for Spatial Statistics: The Spatial-KWD library

Keywords: Kantorovich-Wasserstein Distances, Spatial Statistics, Density Estimation

1. INTRODUCTION

Spatial statistics, geography, and other statistical domains often deal with measurements or estimates of physical or social quantities over a finite geographic region. In several practical applications, the geographic space is discretized, in a regular square grid and the target variable to be measured (or estimated) is non-negative. An important example of such application in the field of demography is the spatial density of the present population published, for example, in the Eurostat GEOSTAT 2018 population grid[93].

In these types of spatial applications, the empirically measured/estimated value of the target quantity is represented by a 2-dimensional histogram. Each grid unit or tessellation region represents a *bin*, and the collection of the variable values across all bins represents a 2-dimensional distribution, or *map*. In this paper, we use the term *map* as synonymous with a 2-dimensional empirical distribution (i.e., 2-dimensional histogram) defined over the geographical (Euclidean) space.

In spatial statistics, the same target quantity can be measured/estimated in multiple alternative ways that may either represent entirely independent approaches or, more often, methodological variants of a single general approach. Hence, the ability to assess quantitatively the goodness of one method against the others is critical not only for selecting the best measurement/estimation method among the set of concurrent candidate options, but also for guiding the development of future improved methods. This is precisely the case for the problem of estimating the spatial density of the present population based on Mobile Network Operator (MNO) data, for which multiple estimation approaches are being actively researched (e.g., see [1] and references therein).

In this work, we present the Spatial-KWD package [2], which implements several features to compare very large spatial maps as defined above. At the core of the package, there is an efficient implementation of a numerical algorithm to compute KantorovichWasserstein distances which is optimized with respect to both runtime and memory consumption. The Spatial-KWD package has features that should facilitate the pipeline analysis of a spatial statistician, such as, the comparison of one-to-one, one-to-many, many-to-many maps; several options to deal with unbalanced empirical measures; the possibility to handle non-convex grid regions representing a country or administrative units. The package is developed in standard C++11 and provides wrappers for R and Python. As a use case, we will discuss in Section 4, the use of the Spatial-KWD package for the estimation of spatial population density based on MNO data, where we compare the solutions obtained with different approaches to identify the best estimation method.

[93] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat

2. cOMPARING mAPS

In several statistical fields, it is quite common to resort to Root Mean Square Error (RMSE) to summarise the spatial distribution error. When the distribution is interpreted probabilistically, statistical distances such as Kullback-Leibler Divergence (KL), CrossEntropy (CE) or Hellinger distance (HD) are occasionally used. All such choices have a common fundamental limitation: they do not account for *geographical proximity* among the bins. In other words, such metrics *treat spatial bins like independent categorical bins*, disregarding their mutual geographic distances. We argue that systematically missing the crucial role of geographic distances sounds like a paradox when the problem is about comparing geographical distributions. Hence, we advocate the use of an alternative distance metric for comparing maps, and specifically the Kantorovich-Wasserstein distance (KWD for short), which takes into account geographical proximity *by design*.

KWD is known by several alternative names in different scientific fields: Earth Mover Distance, Mallow Distance, and Gini-Kantorovich Distance among others (e.g., see [3] and references therein). It has been originally introduced in the field of Optimal Transport, and later adopted in the field of Image Processing, from where it has recently made its way into the field of Machine Learning and Computer Science [4]. With this extended abstract, we aim to raise awareness about the potential use of KWD in the fields of geography and spatial statistics, and at the same time make available the Spatial-KWD package to the statistical community.

On the practical side, the main key issue to be considered when using KWD is the high computational complexity that prevents the computation of the *exact distance value* for very large maps. However, the recent work by Bassetti, Gualandi, and Veneroni [3] has shown that a *close approximation* within a provable bound can be computed in reasonable time on standard off-the-shelf machines, paving the way towards the application of KWD also to large maps of practical interest for spatial statistics.

3. sPATIAL-KWD

The application of KWD to practical problem instances in the field of spatial statistics involves a number of methodological design choices that were carefully considered when implementing the Spatial-KWD package. We describe next two such functionalities, referring the interested reader to the online manual of the package for full details [2].

3.1. Non-convex regions: constrained or unconstrained link cost?

In practical applications, the area of interest has an irregular shape, corresponding for example to a country or some other administrative unit. If the region is convex, all direct segments between any pair of its points (bin centres) by definition lie entirely within the region, and there are no complications as to the definition of the minimum-cost ground distance between any bin pair. If the region is non-convex, depending on the specific statistical application one may need to constrain the minimum-cost path to lay entirely within the region. In Figure 1, we show an example of non-convex region, for which the direct path (red line) crosses the region border. For such bin pairs, the question arises as to whether the link cost should be set equal to the (unconstrained) Euclidean distance or, alternatively, to the (constrained) shortest path that lies entirely within the region of interest. The answer is a matter of methodological choice, and both options may be preferred in different statistical applications, also depending on the "physics" of the phenomenon at hand. The Spatial-KWD package offers the possibility to select both options.

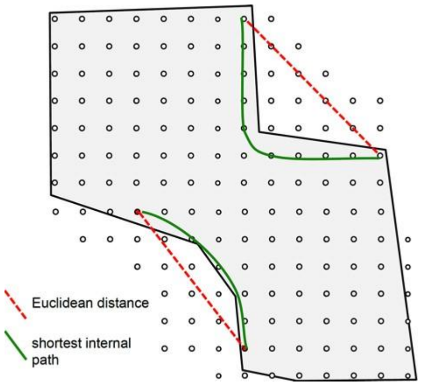


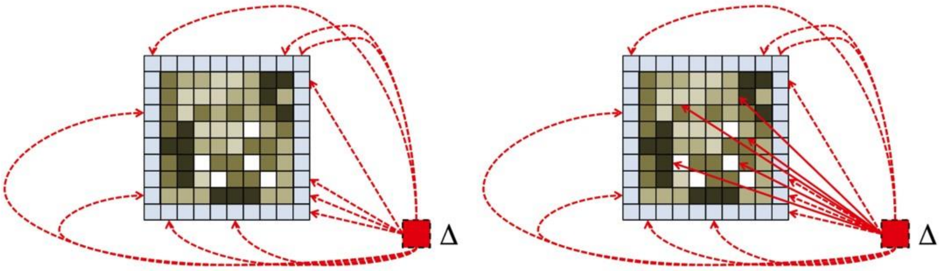
Figure 1 - Irregulars non-convex regions.

3.2. Options for dealing with mismatched mass

A typical case in the field of spatial statistics is encountered when the input map is underestimating or over-estimating the total mass in the considered region. In other words, mass mismatching represents a global under-/over-estimation error. From an algorithmic point of view, there are different ways to handle the extra mass, with different methodological implications. Again, the choice of one particular strategy must be tied to the particular application at hand, and specifically to the physical interpretation of "mass", and to the (known or supposed) root cause for the mass gap. If the mass gap represents a global under-/over-estimation error, we must consider the *expected characteristics of the estimation error* in order to choose a suitable strategy for handling mass mismatching.

The two general categories for handling mismatched mass are: (i) methods that distribute the mass gap to the geographic bins according to some fixed distribution, and then conduct the KWD computation to the case of matched mass; (ii) methods that logically assign the mass gap to an auxiliary "virtual bin", connected with the geographic bins (or a subset thereof) through "virtual links" with pre-defined fixed cost, and from there let the mass flow from/to the geographic bins in the way that minimised the overall transportation cost. Both options are supported in the Spatial-KWD package via dedicated input parameters.

In the second methods of "virtual bin and links", the physical bins connected to the virtual bin represent the (physical) *insertion points* of extra mass. Configuring the set of insertion points represents again a design choice, but it is a "softer" choice rather than the direct assignment of mass to bins. Figure 2 shows two different possible ways to connect the virtual bin with the physical bins of the map.



(a) Border insertion (all border bins serve as insertion points).

(b) Specific gateways points

Figure 2 - Two examples of “virtual bin” approaches to handle mass mismatching.

3.3. Computational Evaluation

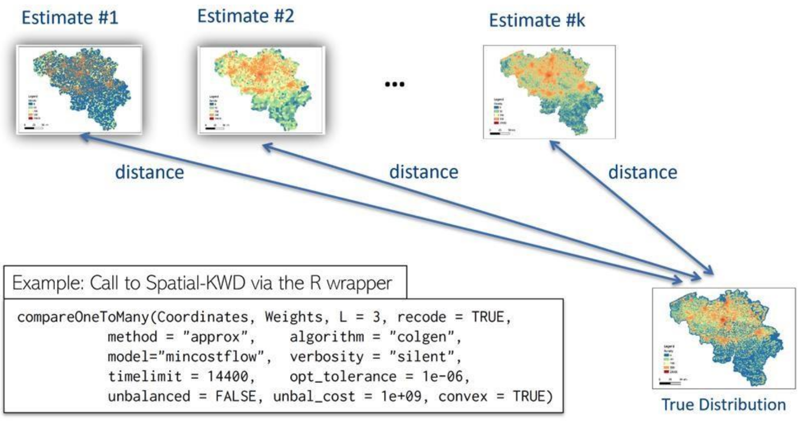
To assess the efficiency and the scalability of our solver we have run a number of experiments using synthetic data derived from the official data by StatBel (from https://statbel.fgov.be). In Table 1, we compare the runtime (in seconds) of the SpatialKWD package on a standard desktop computer for different maps representing the whole territory of Belgium at different bin size resolutions.

Table 1 - Runtime seconds at different grid map resolutions (the leftmost column shows the bin size in meters).

resolution	nodes (tiles)	arcs	iterations	runtime in sec.
1000 × 1000	39 303	64 349	89 227	0.52
500 × 500	156 240	243 231	162 366	1.32
250 × 250	621 868	1 092 105	671 988	5.06
125 × 125	2 482 118	4 425 571	5 962 135	572.81

4. Spatial Density from Mobile Network Operator data

The Spatial-KWD package was successfully used in [5] to estimate the spatial distribution of mobile phones from MN) data. Traditional geolocation solutions rely on Voronoi tessellations and approximate cell footprints by mutually disjoint regions. In [5], the authors test multiple probabilistic approaches. The numerical solutions obtained with different methods were compared to the true distribution (built based on semi-synthetic data) by menas of the Spatial-KWD package. For more details on this work we refer the reader to the online R notebook available: https://r-ramljak.github.io/MNO_mobdensity/. Full documentation of the Spatial-KWD package in R and python is available from [2].



Example: Call to Spatial-KWD via the R wrapper

```
compareOneToMany(Coordinates, Weights, L = 3, recode = TRUE,
           method = "approx",    algorithm = "colgen",
           model="mincostflow",  verbosity = "silent",
           timelimit = 14400,    opt_tolerance = 1e-06,
           unbalanced = FALSE, unbal_cost = 1e+09, convex = TRUE)
```

REFERENCES

[1] F. Ricciato , G. Lanzieri, A. Wirthmann. *Towards a methodological framework for estimating present population density from mobile network operator data*. IUSSP Research Workshop on Digital Demography in the Era of BigData, Seville, 2019 [2] S. Gualandi. *Computing Kantorovich-Wasserstein distances for large spatial maps*. 2021. https://github.com/eurostat/Spatial-KWD

[3] F. Bassetti, S. Gualandi, M. Veneroni, M. *On the Computation of KantorovichWasserstein Distances Between Two-Dimensional Histograms by Uncapacitated Minimum Cost Flows*. SIAM Journal on Optimization, 30(3), pp. 2441-2469, 2020

[4] G. Peyré, M. Cuturi, M. *Computational optimal transport: With applications to data science*. Foundations and Trends® in Machine Learning, 11(5-6), pp. 355-607, 2019

[5] F. Ricciato and A. Coluccia. *On the estimation of spatial density from mobile network operator data*. IEEE Tran. on Mobile Computing. DOI: 10.1109/TMC.2021.3134561

On the Application of the Fourier-based Distance to Spatial Statistics

Keywords: Fourier-based distance, Wasserstein Distance, Spatial Statistics

INTRODUCTION

In this work, we present the Fourier-based metric introduced in [1] as an alternative to the Wasserstein metric [2] for applications in spatial statistics. The main advantage of the Fourier-based metric is that it can be computed very efficiently via the FFT algorithm [3]. It was formally proved in [1] that these two metrics are *equivalent* in mathematical terms. Recall that two metrics d_1 and d_2 defined in X are mathematically equivalent if there exist positive constant c and C such that it holds that $cd_2(\mu,\nu) \leq d_1(\mu,\nu) \leq Cd_2(\mu,\nu)$ for every (μ,ν) in X. Mathematical equivalence is a weaker notion than isometry and does not imply that the two metrics yield the same results for the same input argument. However, motivated by the formal property of mathematical equivalence, we explore *empirically* the relationship between the Wasserstein metric and the Fourier-based metric and found that in practical real-life applications, the two metrics are strongly interlinked. For instance, when applied over estimates of spatial population distributions obtained from mobile network operator data with different estimation methods, we observe empirically a clear a monotonic relationship between Fourier-based metric and the Wasserstein metric. This indicates the possibility of using the Fourier-based metric as a proxy for Wasserstein metrics in real-life applications where the computation load of the latter is critical.

METHODS

Here, we focus on the relationships between the Wasserstein distance of order 1, denoted by W_1, and the Fourier-based metric, denoted by $f_{1,2}$. In [1] the following Theorem was demonstrated:

Theorem. Given μ, ν two discrete probability measures with finite support with cardinality T, the following equivalence relationship holds

$$\frac{2\pi}{T^2} W_1(\mu,\nu) \leq f_{1,2}(\mu,\nu) \leq W_1(\mu,\nu) \tag{1}$$

In general, this relation is rather loose when the size of support T increases, since the lower constant tends quadratically to 0. Therefore, equation (1) does not express a relation of perfect equality (i.e., isometry). The looseness of the lower bound in (1) is due to the need to take into *the worst case*, which is inflated by the periodicity of the Fourier transform. In real-life applications, scenarios close to the worst case may be extremely rare, and the actual relationship between the two distances in *the average case* is considerably tighter, as we show empirically below.

Numerically, the computation of the Fourier metric relies on the following formula

$$f_{1,2}(\mu,\nu) := \frac{2}{|T|} \sum_{k=1}^{\frac{T}{2}} \frac{|\hat{\mu}_k - \hat{\nu}_k|^2}{|k|^2} \tag{2}$$

where the Fourier transforms of μ and ν appear in the numerator. Hence, to calculate $f_{1,2}$ (μ, ν) we must first transform the two probability measures in phase space by Fourier transform, and then apply a weighted distance of type L^2 in phase space. Clearly, the step that requires the higher computational cost is the Fourier transform which has computational complexity $O(T \log(T))$.

The study conducted in [1] on the DOTmark dataset [4] shows that, in practice, the relationship between these two metrics is much tighter than that expressed by the equivalence relationship (1). Figure 1 shows the main results of the experiments reported in [1]. In particular, the coefficient of determination in this experiment is around 0.89, meaning that the distance $f_{1,2}$ (μ, ν) is a good predictor for the distance W_1 (μ, ν) when adequately scaled.

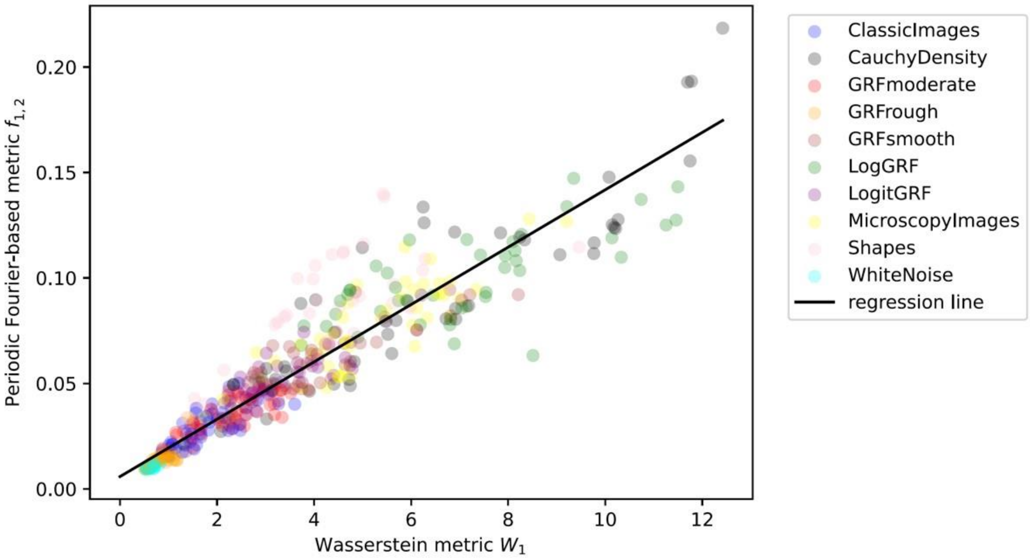


Figure 1 – Comparing the Wasserstein and the Fourier-based metrics: Distances between every pair (per class) of images in the DOTmark dataset and relative regression line.

APPLICATION TO SPATIAL STATISTICS

To confirm that the Fourier-based metric behaves well with respect to the Wasserstein metrics in practical scenarios, we compare it with the results obtained with the Spatial Kantorovich-Wasserstein Distance (Spatial-KWD) library [2] that was used in [5] to compare true distributions and reconstructed distributions of spatial population densities based on Mobile Network Operator (MNO) data.

In [5] the authors used the Spatial-KWD values to comparatively assess the goodness of different estimation approaches. In this work we are interested to assess whether resorting to Fourier-based metric would have led to similar conclusions with respect to relative goodness of the various approaches. To this aim, we will compare the results obtained through the Spatial-KWD library with the results obtained using the Fourier metric $f_{1,2}$ (μ, ν). To perform this comparison, we have at our disposal 9 test instances, i.e., 9 different distributions obtained with different estimation approaches. Each test instance is aggregated 4 different spatial resolutions, resulting in rectangular histograms of different size: 267×228, 534×455, 1068×910 and 2134×1819. These instances were carefully created to test the Spatial-KWD library during the software development phase. Testing on these instances, therefore, gives us reliable results on the behaviour of the Fourier-based metric in relation to the Spatial-KWD library.

As we can see from the example in Figure 2, these densities are very sparse. The SpatialKWD library is optimized to gain advantages from the geometric configuration of these types of problems. For what concern the Fourier-based distance implementation, we do not perform any kind of code optimization at this stage to leverage the sparse geometry of the problem.

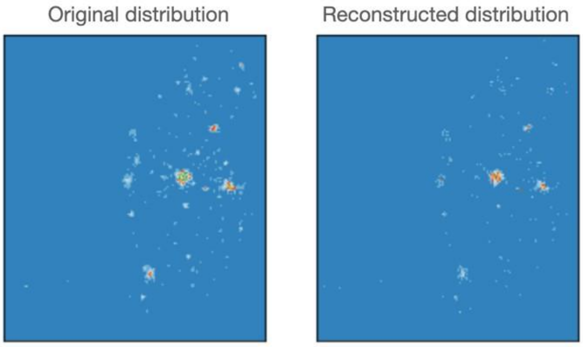


Figure 2 - Original and Reconstructed spatial population density for Belgium (semi-synthetic data derived from real data).

Figure 3 shows a roughly quadratic relationship between the values calculated with the Spatial-KWD library and the values calculated with Fourier-based metric $f_{1,2}$. As such relationship is monotonic, the ranking of different (pairs of) distributions by Fourierbased metric would deliver the same result as the ranking based on Spatial-KWD. In other words, if the goal of computing a distance metric between the estimated and true value is to assess comparatively the goodness of different spatial estimation methods, as done in [5], one would obtain practically the same result by using Spatial-KWD or the Fourier-based metric. The advantage of using Fourier-based metric is a much faster computation, with computation times several orders of magnitude lower than SpatialKWD: for instance, for the higher resolution instances with a total of 3.88 millions of bins (2134×1819) the Fourier-based metric is calculated in 17.5 seconds, while the Spatial-KWD took between 45 minutes and 9 hours depending on the test instance. Conversely, the Spatial-KWD package offers more versatility and flexibility to handle particular situations in real-life applications, e.g., in case of mismatched mass or nonconvex domains (see the Spatial-KWD package documentation [2] for details).

cONCLUSIONS

The Fourier-based metric offers a valid and computationally cheap alternative to the Wasserstein distance that may be of interest for practical applications involving very large problem instances. Despite the mathematical equivalence between the two does not provide a tight bound, we observe *empirically* that the numerical values obtained with Fourier-based metric are linked *monotonically* to corresponding Spatial-KWD values for the same instances, with an approximately quadratic relationship. This indicates the possibility of using the Fourier-based metric as a proxy for Wasserstein metrics in reallife applications where the computation load of the latter is critical.

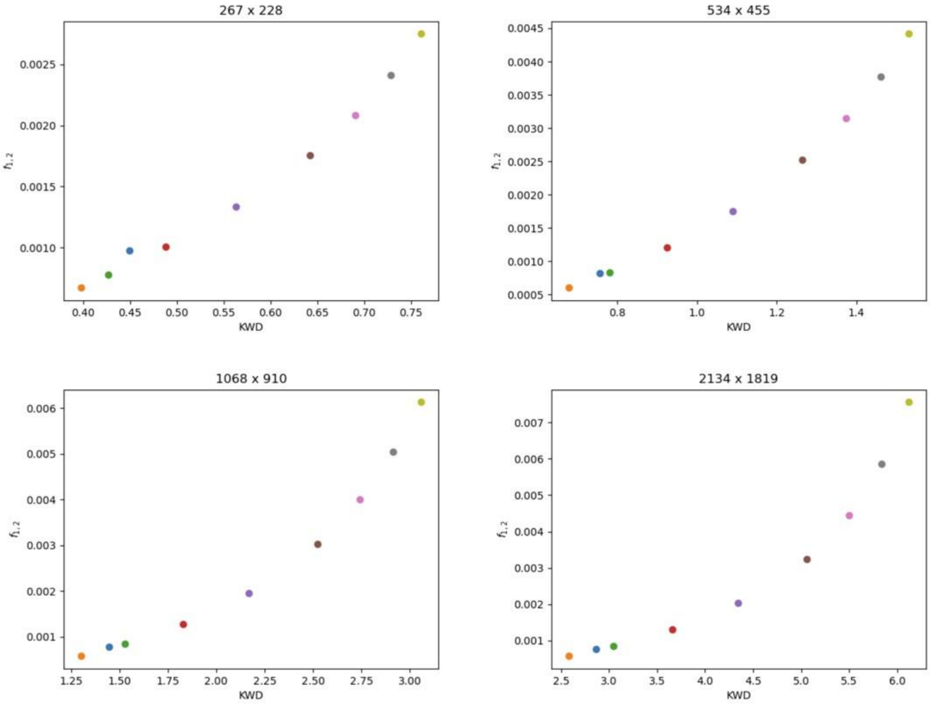


Figure 3 - Correlation between Spatial-KWD and $f_{1,2}$.

REFERENCES

[1] G. Auricchio, A. Codegoni, S. Gualandi, G. Toscani, M. Veneroni. *The equivalence of Fourier-based and Wasserstein metrics on imaging problems*. Rendiconti Lincei, 31(3):627–649, 2020

[2] S. Gualandi. *Computing Kantorovich-Wasserstein distances for large spatial maps*. 2021. https://github.com/eurostat/Spatial-KWD

[3] E.O. Brigham. *The fast Fourier transform and its applications*. 1988, Prentice-Hall

[4] J. Schrieber, D. Schuhmacher, C. Gottschlich. *Dotmark - A benchmark for Discrete Optimal Transport*. IEEE Access, 5:271–282, 2017.

[5] F. Ricciato, A. Coluccia. *On the estimation of spatial density from mobile network operator data*. IEEE Trans. on Mobile Computing, doi: 10.1109/TMC.2021.3134561

Exploring Spatial and Demographic Official Statistics on Personal Insolvency

Keywords: administrative data, record-linkage, spatial data, text mining, geo-referencing, social inequality, experimental statistics, personal bankruptcy, new digital data, spatial autocorrelation

1 Introduction

Current crises, such as the COVID-19 pandemic or the war in Ukraine, have increased economic and financial burdens on individuals and households. Since 2011, the total number of personal insolvencies per year in Germany has constantly decreased from 136.000 (2011) to 56.000 (2020) but increased by 95% to 109.000 in 2021 [1].[94] However, a large amount of this increase is likely caused by a law amendment (personal insolvency processes have been shortened from six to three years) and not caused by the COVID-19 pandemic exclusively [2]. Nevertheless, it remains to be evaluated whether the pandemic might show delayed effects on these numbers. In addition, further increases can probably be expected due to the war in Ukraine. Moreover, flooding in 2021 caused the devastation of entire regions in Western Germany. Here, regional effects of this natural disaster causing personal insolvencies can be expected. Until now, official statistics publish absolute numbers on personal insolvency at the German federal state level. These numbers are based on insolvency announcements from local courts (Amtsgerichte) in Germany. This paper provides novel methodological analyses for this administrative data source to improve current official statistics using data science tools. These newly derived statistics will inform on the demographic and spatial distribution of personal insolvency and provide more accurate information for policymakers. Hence, these statistics can serve as a base to potentially inform the development of tailored social policy programs.

2 Background

Research shows that demographics play an essential role in personal insolvencies since family structure, ethnic background, income, and education are strong predictors [3, 4]. There is also evidence for spatial differences between regions [5]. Spatial differences include, for example, individuals' socioeconomic status (i.e., access to toptier schools or well-paid jobs [6]), the housing market, environmental factors, and local crime rates. These variations reflect social inequalities, but on the other hand, may themselves further reproduce or even exacerbate them [7]. However, there is little research and evidence from Germany (for example, [8]), and analyzing the relationship between personal insolvencies and spatial differences adds to the existing body of literature.

3 Methods

Declarations of insolvencies are to be made public in Germany (§ 9 InsO). They are published in a large administrative database after the respective court decision. Delaying personal insolvency

[94] Personal insolvency is a simplified procedure for handling the insolvency of a natural person (private individual). It is intended to equally satisfy the creditors of an insolvent debtor on a prorata basis.

is a criminal act; thus, the reports are expected to be reported on time, and a complete picture of all insolvencies is shown. The individuals' age and geo-location, the type of declaration, and the dates of the court decisions (results of the latter two are not shown in this paper) were extracted from this database using text mining. The extracted information is used to generate statistics on the age and geospatial distribution of individuals affected by personal insolvency. Therefore, the information was linked to geographical data with different resolutions: federal state level (16 geographical units) and postal code level (8725 geographical units). Moran's I correlation coefficient is used to study whether adjacent and bordering regions are similar (spatial autocorrelation). The results indicate whether there is a clustering of different values (dispersion), no autocorrelation (randomness), or perfect clustering of similar values (opposite of dispersion). After analysis, all personal information was deleted.

4 Data

The period under study is March 2022 untile the end of August 2022. Within this period, 64.285 entries related to personal insolvency proceedings have been selected from the database. It was impossible to determine the age for about 5% of these. For about 0.3%, the geo-location could not be found. This is because it was not specified or is located outside of Germany.

5 Results

The age distribution is shown in Figure 1. There is a steep increase in insolvencies after the legal lower limit of 18 years (German law prevents individuals under 18 from going bankrupt), with a plateau in the early thirties. The number of insolvencies then decreases until the group of people aged 50-60. There is no noticeable increase in insolvencies for retirees (ages 63 to 67, depending on the birth cohort), as insolvencies show a steady decrease with age. On average, people are 44 when they are affected by personal insolvency. The youngest individual is aged 18, and the oldest is 99 years.

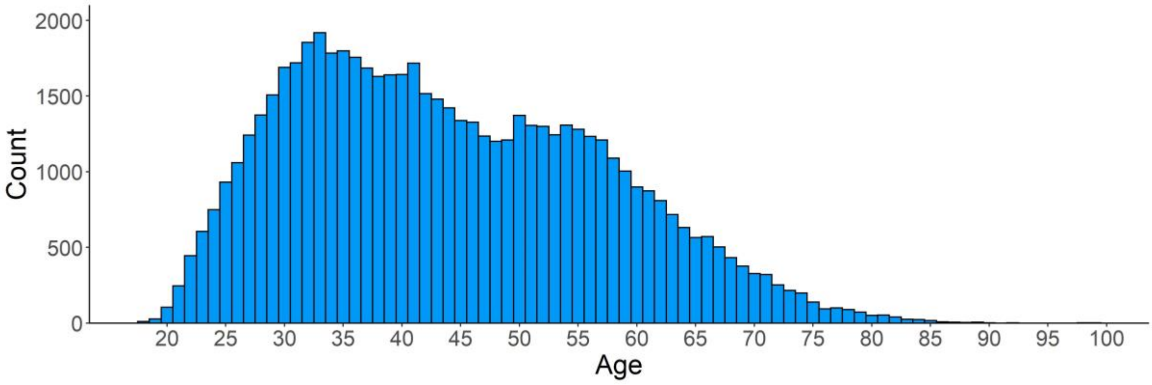


Figure 1: Histogram of personal insolvency in Germany by age (binwidth = 1).

Figure 2 shows the spatial distribution of personal insolvencies on the federal state level. The color gradient indicates the number of personal insolvencies in each federal state per 100.000 inhabitants. A north-south disparity is evident. The maximum of personal insolvencies is found in Hamburg (119 per 100.000 inhabitants) and the minimum in Bavaria (43 per 100.000 inhabitants). Averaged over all federal states, there are 82 personal insolvencies per 100.000 inhabitants. Here, $I = 0.38$, indicating moderate clustering of similar values.

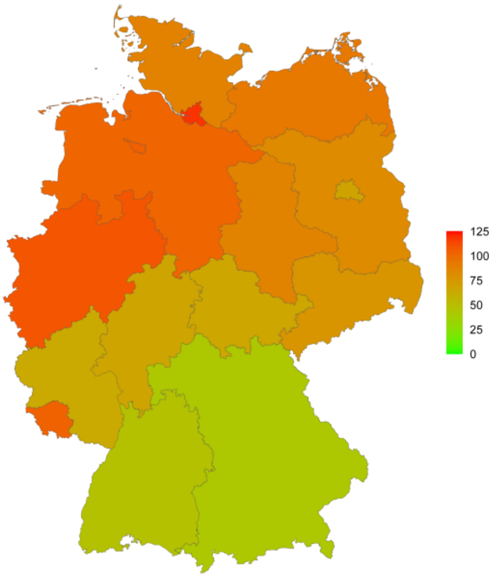
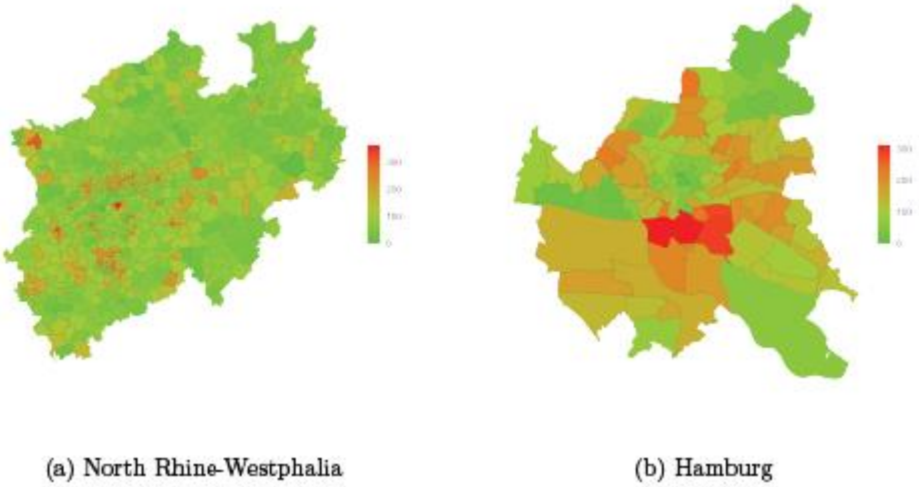


Figure 2: Spatial distribution of personal insolvencies on the federal state level. Color gradient indicates the number in each federal state per 100.000 inhabitants.

Figure 3 shows results (personal insolvencies per 100.000 inhabitants) on the postal code level for the most populated federal state (North Rhine-Westphalia) and for the federal state with the most personal insolvencies per 100.000 inhabitants (Hamburg). Both panels show smaller values for rural and larger for urban areas. For North Rhine-Westphalia, the mean personal insolvencies per 100.000 inhabitants is 108 and the maximum is 363. For Hamburg, For Hamburg, the mean is 119 and the maximum is 312. For North Rhine-Westphalia, the spatial autocorrelation is $I = 0.38$ and $I = 0.44$ for Hamburg. Both values indicate a moderate clustering of similar values.



(a) North Rhine-Westphalia

(b) Hamburg

Figure 3: Spatial distribution of personal insolvencies on the postal code level. Color gradient indicates the number in each federal state per 100.000 inhabitants.

6 Conclusions

This research demonstrated using data science tools to extract previously unused information from an existing administrative database. This new information enables the production of more precise official statistics on personal insolvencies concerning demographic and spatial statistics. The added value of these statistics will become particularly apparent when the factors mentioned above (pandemic, war, and flood) are reflected in regional effects. Furthermore, this database can serve as a source for various socio-political questions concerning social inequality. For example, demographic data and spatial information from a German regional database can be linked using the municipality key as an identifier. This would allow studying spatial correlations between personal insolvency and, for example, education or income and an improved understanding of the reasons for personal insolvency.

Current and future work is dedicated to improving the text mining techniques used and studying the quality of the obtained data. For example, it is intended to extract gender as an additional feature. Moreover, different regional classifications will be used for in-depth studies on spatial autocorrelation.

However, the period considered in this study is small, and the results reported should be considered indicative or experimental statistics. Nevertheless, implementing these novel analyses in producing official statistics could serve as a basis for decision-making and the development of tailored social policy programs in the medium and long term.

References

[1] Destatis. Insolvencies by year. URL destatis.de, 2022.

[2] CRIF. Studie Schuldenbarometer 2021: Fast doppelt so viele private Insolvenzen. URL crif.de, 2022.

[3] L Lefgren and F McIntyre. Explaining the puzzle of cross-state differences in bankruptcy rates. *The Journal of Law and Economics*, 52(2):367–393, 2009.

[4] P Bishop. The spatial distribution of personal insolvencies in England and Wales, 2000–2007. *Regional Studies*, 47(3):419–432, 2013.

[5] G Galster and P Sharkey. Spatial foundations of inequality: a conceptual model and empirical overview. *Rsf: The Russell Sage Foundation Journal of the Social Sciences*, 3(2):1–33, 2017.

[6] J Lindley and S Machin. Spatial changes in labour market inequality. *Journal of Urban Economics*, 79:121–138, 2014.

[7] BI Bryant and P Mohai. *Race and the Incidence of Environmental Hazards: A Time for Discourse*. Routledge, Abingdon, 2019.

[8] A Farwick and W Petrowsky. Überschuldete Privathaushalte. *Jahrbuch StadtRegion*, 5(1), 2008.

Geocoding millions of addresses in a reproducible manner for Big Data climate risk analysis

Keywords: geocoding, big data, matching addresses, climate geospatial data, machine learning

Introduction

Climate risks, especially related to natural hazards, such as droughts, heat waves or floods, are very specific to a location or area. Determining physical risk of a business requires precise address information of an entity and its assets, such as production sites and distribution centres, which are then linked to potential hazard intensities at specific location. Indicators of the physical risk can then be calculated based on the intensity of physical hazard in combination with the properties of an asset (e.g. material and construction year of a building, industrial vs a residential area,) which serve as a proxy for vulnerability of an asset and potential damage.

The first part – obtaining the exact location of an entity, is usually very challenging. Often one has only address information in text format of various quality. However, we need the exact geographical coordinate of this address to be able to match it to risk. This process is known as geocoding. In this paper, we shed light on how such exercise can be done in a scalable manner for millions of addresses.

We start with a simple example of the following address: "ECB, Sonnemannstrasse 22, 60134 Frankfurt am Main, Germany". An exact geographical location of this address would be (50.11038, 8.70148) (EPSG:4326). This exact location can be then matched with climate maps from public sources that describe in which area a phenomena (e.g. floods, heat waves, draughts etc.) is likely to occurs and with which intensity (e.g. water depth in case of flooding).

There are several cloud services, such as Google Maps or Bing Maps, that allow to geocode addresses, but they are 1) very expensive and 2) slow (over 50 days for our addresses according to some internal estimations) 3) the quality is difficult to assess.

Hence, we defined and implemented our own geocoding process suitable for millions of addresses:

- We use the public free Open Street Map (OSM) [1] dataset that contains geospatial objects and their (mostly) standardized address as metadata attached to those objects
- We use 1) heuristics and 2) a machine learning model provided by libpostal to convert the addresses in our economic datasets into a clean standardized address format that can be matched with the metadata in OpenStreetMap
- We do several geospatial operations to get the exact location from the matched data, such as in case of buildings we select the centroid of the building, in case of addresses without street number, we select the middle of the street etc.

Methods

Geocoding

The first step to geocode is to get a reliable inventory of the cities, streets and buildings in Europe. OpenStreetMap (OSM) is such an inventory. One can describe it as “Wikipedia” for map data worldwide. Hundreds of thousands of users contribute to it and update it in a timely fashion. In addition to volunteers, also government institutions and (non-) commercial entities provide their input to the OSM. It is free and open to use. The commercial providers mentioned before do not make their map data freely available or one limited number of entries can be queried (e.g. what is at this address).

Additionally, to a rather full coverage of addresses, it also contains train tracks, highways, bridges and millions of so-called point of interests (POI), such as train restaurants, banks, factories, data centres which could be relevant for economic analysis.

Preparation of the reference data

We follow the following technical steps at a high level:

- Download the files is Geofabrik [7] (Geofabrik: Download server for openstreetmap data, s.d.). We plan in the future to use the torrent protocol to be independent of an individual mirror.
- The OpenStreetMap data is in a format called ProtocolBuffer Binary format (PBF) [8]. We convert this format to another format suitable for parallel big data processing: Apache Parquet [8]. The conversion is done with osm-parquetizer [10]
- The data is further transformed from the OpenStreetMap concepts [11], such as Nodes, Ways, Relations, into geometries suitable for mapping to the climate risk data
- The result is the metadata, such as street names, cities and their associated geometrical objects. The Geometrical Objects are stored in a dedicated column in the Parquet format in Well Known Text Format (WKT) in different projections matching the projections of the climate data. The other columns contain the metadata, e.g. streetname, city etc.
- We join this data based on the geometrical objects with data on communes by Eurostat [12] to cater for missing/incomplete data on city names, e.g. sometimes only a street name and number where associated with a geometrical object as the city, country can be determined from its spatial location. Also, it enabled us to provide the transliteration of certain city names. We want also to explore in the future the administrative boundaries in OpenStreetMap to enrich the data as they are very detailed [13].

Preparation of the address data in the ECB economic datasets

It is very difficult to ensure data quality of addresses in economic datasets. There are many non-standardized conventions to write an address within different European countries. Often the address data is incomplete as it was reported incomplete. Some put the street number before the street. Street names often contains abbreviations. E.g. in Germany “Sonnemannstraße” can be also written “Sonnemannstr.”, “Sonnemannstrasse” etc. Some contain the apartment number in a building. Zip codes can be included before, after a city, sometimes they are also included in the country (e.g. 69314-DE). Addresses are written in Latin or Cyrillic characters.

Until now there was also no tooling to find quality issues in address data in an automated fashion. Additionally, this makes exact matching very difficult to impossible.

As a first objective is to harmonize the addresses: i) identify elements in an address (e.g. street name, house number, country, city, zipcode), and ii) based on those elements convert the address into a standardized canonical form suitable for matching with the OpenStreetMap data. We use for both steps libpostal [14] an open source project sponsored by Mapzen: OpenVenues which support 60 languages and is based on a large machine learning model for those tasks. Harmonizing the addresses ensures a better match with the OSM data because the addresses are coming from different systems where they have been often manually filled with different degree of precision. First, they are normalized by expanding the abbreviations, removing upper cases, punctuations, etc. and then they are parsed (with a Conditional Random Fields model) to detect the roles of the components of the address (house name, house number, road name, city, etc.).

Matching the OpenStreetMap data to the harmonized addresses in the ECB economic datasets

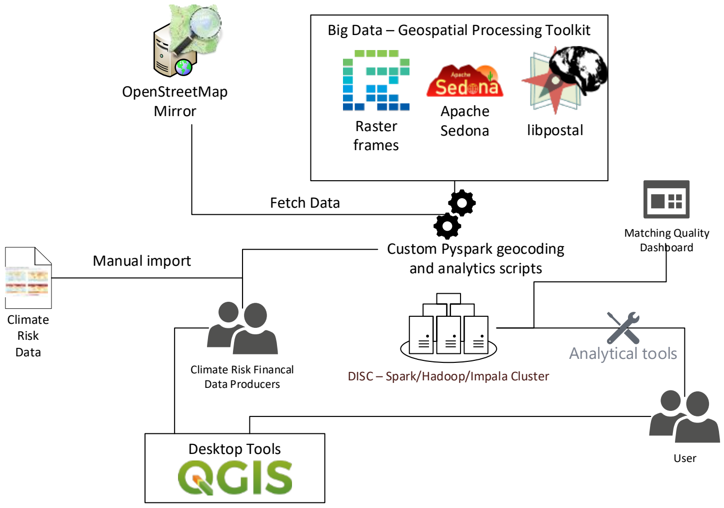
On a high level, we try to match the address data as follows:

- as it was exactly written in the business address data,
- on the full address provided after normalization by libpostal,
- on parts of the address provided after normalization by libpostal.

IT Architecture

This process is run on a powerful Big Data Cluster where we thousands of tasks can be run in parallel. Figure 39 gives an overview of the solution architecture from an Information Technology (IT) perspective.

Figure 39 Solution for Geocoding and Analysing Climate Data


OpenStreetMap Mirror
Big Data – Geospatial Processing Toolkit
Raster frames
Apache Sedona
libpostal
Fetch Data
Matching Quality Dashboard
Manual import
Custom Pyspark geocoding and analytics scripts
Climate Risk Data
Climate Risk Financal Data Producers
DISC – Spark/Hadoop/Impala Cluster
Analytical tools
User
Desktop Tools
QGIS

The core elements are the data, the cluster, and the actions associated to them. The data is stored on-premise in a data lake containing all datasets. The climate risk files are manually uploaded while the data from the OSM initiative is fetched via a manually triggered Python script. To exploit the data sources, custom scripts use PySpark which is the Python interface for Apache Spark [2]. Spark is an Open Source Big Data processing framework. The gains of such framework come from leveraging the resources of the computing cluster and the sound parallel execution of tasks. These scripts implement the geocoding process and the analysis. Furthermore, they include also important geospatial processing libraries, namely Apache Sedona [3] , Rasterframes [4], and libpostal [4]. The dashboard on the address data matching quality is presented in a commercial dashboarding tool but users have access to the open source Desktop Tool QGIS [5] for spatial data visualization and ad-hoc analysis using an interface without need for programming.

The solution is highly cost-efficient in the already existing infrastructure. The rest of the section focuses on the content of the custom scripts which contain the data preparation and the geocoder.

results

By evaluating the proposed geocoding solution, we found heterogeneity in quality across the countries. For 20 of the 25 European countries, we could find a pair of coordinates for 90 percent of entities, and for 12 countries more than half of the matches were done at the street level.

The proposed geocoding solution was evaluated by comparing to the results with a commercial cloud service for 10,000 entries. However, this alternative approach based on commercial solution did not return any result for half of the entities. For the matched results, the median distance between the two approaches is less than one kilometre for 17 out of the 25 countries, while for half of the countries the difference in distance is within 5 km for 95% of percent geolocated entries. It should be highlighted that the distance should not be interpreted as the distance from a true location but rather a consistency between two different approaches. We find the results satisfactory even for the assessment of flooding risk where the precise location plays a bigger role than for other hazards.

The principal sources of errors were: exclusion of overseas territories (e.g., for France, Portugal, and Spain), transliterations from local alphabet[95] to Latin alphabet or to English (Bulgaria, Cyprus, and Greece), missing information in the addresses (in all countries), confusion between homonym places (e.g., towns, street, avenue and square with the same name) (Poland), equivalent or shortened names (e.g., names of historical figures) (e.g., France, Italy), very small settlements not reported, or outdated records. Curiously, we also noticed that the results of the commercial provider returned at times a location based on the matched street name even if the city was different and on other side of a country, while in our approach the priority was given to a city on the expense of more precise matching at street and number level.

[95] Note that for ©OpenStreetMap the best practice is to keep the local alphabet.

Conclusions

Presented analysis underlines the complexity of geocoding and challenges in the processing of address information. The bigger the dataset and countries covered, the higher the number of special cases encountered. While these caveats in mind, overall, we see a convincing case for using data from OpenStreetMap and we believe that further pre-processing applying country-specific transformations would lead to more successful matching. Improvements in the quality and completeness of reported addresses would be the largest contributor to more precise geocoding.

references

[1] OpenStreetMap, “OpenStreetMap.org,” [Online]. Available: https://www.openstreetmap.org. [Accessed 28 07 2022].

[2] Locationtech, “Rasterframes,” [Online]. Available: https://rasterframes.io/.

[3] QGIS Project, “QGIS,” [Online]. Available: https://www.qgis.org/.

[4] “Geofabrik: Download server for openstreetmap data,” GeoFabrik, [Online]. Available: http://download.geofabrik.de/. [Accessed March 2021].

[5] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker and I. Stoica, “Apache Spark: a unified engine for big data processing,” *Communications of the ACM,* vol. 59, no. 11, pp. 56-65, 2016.

[6] A. Barrentine, “libpostal: international street address NLP,” OpenVenues, 2018. [Online]. Available: https://github.com/openvenues/libpostal.

[7] J. Yu, Z. Zhang and M. Sarwat, “Spatial data management in Apache Spark: the GeoSpark perspective and beyond,” *GeoInformatica*, vol. 23, no. 1, pp. 37-78, 2019.

Energy Statistics (MANS3A.2)

Session Chair: **Madelaine Mahosky** *(Eurostat)*

Energy consumption decomposition analysis using European official statistics – methodology and input data
Stavros Lazarou *(Eurostat)*

Y-0 Estimates of supply side of Energy balances and key indicators
Maja Božičević Vrhovčak *(Energy Institute Hrvoje Požar)*, Stavros Lazarou *(Eurostat)*

Import dependency by origin indicator
Cristina Martello (*Eurostat*)

Evaluating energy prices and costs impacts on households and industry’s costs
Stavros Lazarou *(Eurostat),*

Energy consumption decomposition analysis using European official statistics – methodology and input data

Keywords: decomposition, energy consumption, European official statistics

Introduction

For several decades now, the importance of energy efficiency has been well understood, both in the context of clean energy transition, and for its benefits to national economies in general. This has evolved to such an extent that after being labelled for some time as the 'hidden fuel', it is now more commonly referred to as the 'first fuel' [1], thus showing its prime role in the current energy efficiency debate. Another sign that energy efficiency is recognised in the political agenda is its presence in the United Nations' 7th Sustainable Development Goals (SDG7). This states that 'by 2030, double the global rate of improvement in energy efficiency' [2].

In the European Union, energy efficiency improvement was one of three 20% targets that should be reached by 2020 (Energy Efficiency Directive 2012/27/EU). The target has since been updated to 32.5% by 2030, compared to projections of expected energy use (Energy Efficiency Directive 2018/2002).

If these ambitious goals are to be implemented effectively, actions need to be carefully planned and monitored. National administrations, academics, researchers, international organisations, and other stakeholders will need gain a solid understanding of what drives energy consumption. Decomposition analyses have, however, been widely used for many decades. Both ex-ante and ex-post analyses of energy efficiency can be improved because of newly implemented policies and from the penetration of new technologies.

Theory

When attempting to assess energy savings, two approaches are possible: top-down and bottom-up. In the bottom-up approach [3], data from specific energy efficiency improvement measures are collected and then aggregated. Alternatively, with the top-down approach, data for aggregated sectoral energy consumption are used. The top-down approach is therefore easier to calculate because less data is required.

In the initial studies on energy efficiency, the main indicator used at national level was the Total Energy Consumption (TES) per unit of Gross Domestic Product (GDP). This is also referred to as the TES per GDP or energy intensity of the economy. The reasoning behind this was that a decrease in this ratio implied a reduction in energy requirements to generate a unit of national output and therefore, it was considered a proxy for efficiency improvements of the overall economy.

However, studies [4] showed that variations in this indicator do not identify energy savings achieved thanks to specific policy or new technologies, but rather aggregate a variety of other confounding factors, which can include:

- structural changes of the economy, for example shift toward less energy intensive industries;

- behavioural factors, such as increase in the number of appliances used in households;

- price effects, rebound effects;

- macro-economic factors, such as recession, which can cause a sub-optimal use of resources;

- environmental factors, such as colder winters than average;

- actual change in the energy intensity of the sector.

Figure 40 depicts how energy intensity in a sector can decrease even though the energy consumption remains quite constant, due to an increase in the sector's Gross Value Added (GVA). This could be partially imputed to energy efficiency gains made in the industry sector.

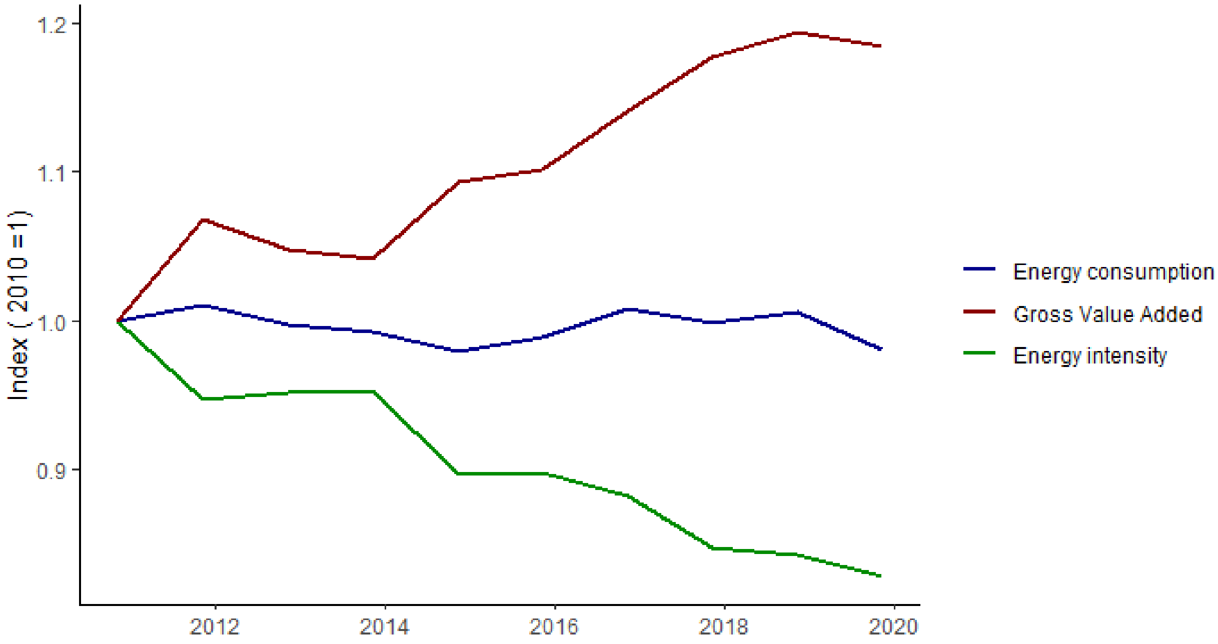


Figure 40: Indexed indicators for Germany's total industry energy consumption, GVA and energy intensity variation (base: 2010)

However, when observed more closely, the value added share of each sub-sector has changed over the period, as seen in Figure 41. Specifically, in this example, the GVA share in the transport equipment sub-sector increased quite significantly, while the construction sector share decreased. This structural change may partially explain the decrease observed in overall energy intensity in the industry sector.

The objective of the decomposition analysis is to separate the respective impact of each of these drivers on the total energy demand.

In the context of energy statistics, energy intensity is defined as the amount of final energy used to generate a unit of desired output. The energy use could be in the form of direct fuel consumption, including both fossil fuels and biofuels, or of electricity and heat consumption. The output can take multiple forms depending on the final purpose of the process being studied. This could be the transport of one person over a given distance, the heating of a predefined dwelling surface, or the production of one ton of steel. A decrease in energy intensity is referred to as energy efficiency improvement.

Separating the effects of the variation of each component on the total energy demand requires some mathematical operations that will allow the quantification of each impact in terms of energy demand. To isolate the change in energy consumption which is caused by variation in the respective shares of each sub-sector[96], mode or end use, disaggregation of the activity and energy consumption data are required. When this decomposition is carried out, data on the sub-sector activity should be in a physical rather than monetary unit, to avoid distortion due to price change, and furthermore, the unit should be the same across each sub-sector. However, this is not always feasible, for example, the iron and steel industry production level may be described by the mass of steel produced, while for other sectors with diverse outputs, the GVA is the only appropriate measure.

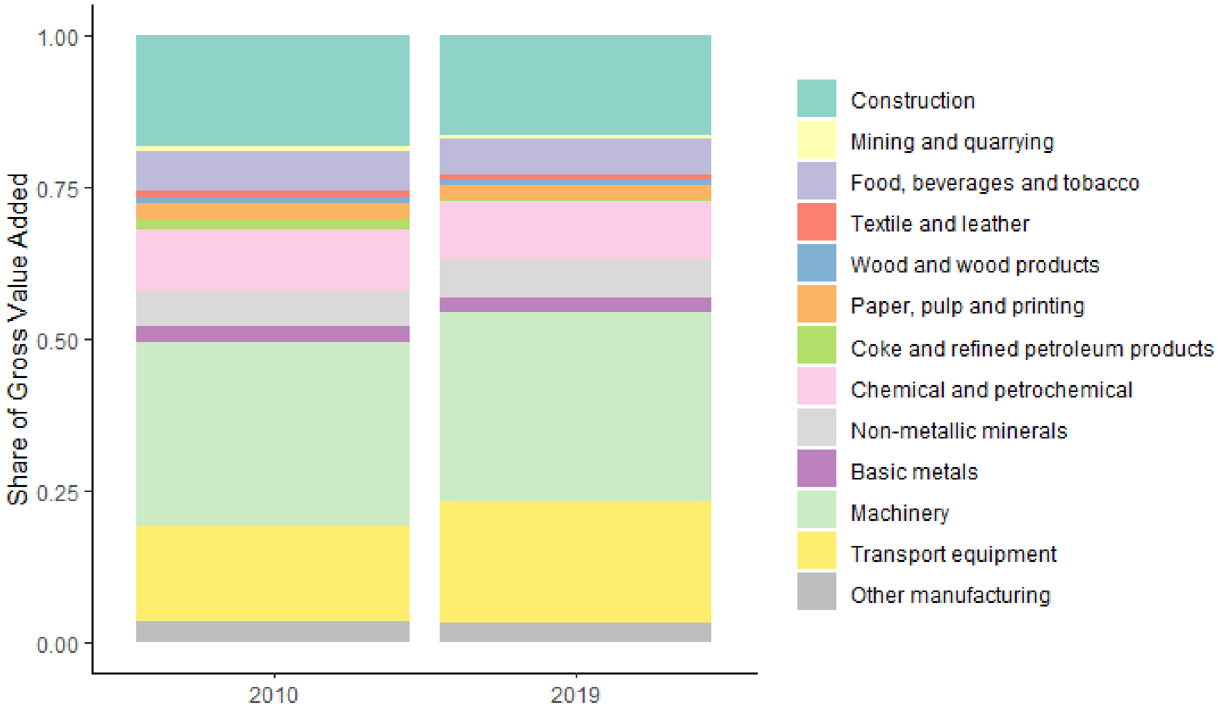


Figure 41: Sectoral share of the industry GVA for Germany

One way to separate the effects of each component measured in different units is to use an index of the activity level, thus giving flexibility in the choice of the unit. This sort of analysis involves the index number theory, which is studied in economics and referred to as Indexed Decomposition Analysis (IDA).

[96] It should be noted here, that in some instances of this report, sub-sectors can refer to the sectoral classification (for example: steel manufacturing is a sub-sector of Industry), to a mode (for example bus and personal car are different modes for passenger transports) or to an energy end use (for example, heating is an end use of the residential sector).

Methods

When attempting To perform the decomposition, the starting point is to establish an identity, where the energy consumption of a sector (E) is expressed as a product of each element whose impact are to be quantified. In practice, the decomposition identity is usually expressed as follows:

$$E = \sum_i E_i = \sum_i A \times \frac{A_i}{A} \times \frac{E_i}{A_i} = A \times \sum_i S_i \times I_i \quad (1)$$

Where:

E_i: Energy consumption of sub-sector i

A_i: Activity level of sub-sector i

$S_i = \frac{A_i}{A}$: Share of energy consumption of sector i

$I_i = \frac{E_i}{A_i}$: Energy intensity of sub-sector i

The IDA allows the respective contribution (measured in terms of energy consumption) of each of these driving factors to be determined. This can be expressed either additively (notation Δ) or multiplicatively (notation D), as is shown below:

$$\Delta E = E_t - E_0 = \Delta E_{ACT} + \Delta E_{STR} + \Delta E_{INT} + \Delta E_{RES} \qquad (2)$$

$$DE = \frac{E_t}{E_0} = DE_{ACT} \times DE_{STR} \times DE_{INT} \times DE_{RES} \qquad (3)$$

The additive form decomposes the difference between two points in time, while the multiplicative form decomposes the ratio of change with respect to the base year.

The activity effect (E_{ACT}) accounts for changes in energy consumption due to the change in the economic activity of the sector: the activity effect is positive (i.e. the energy consumption increases) if the overall activity increases.

The structural effect (E_{STR}) accounts for changes in energy consumption that are due to the change in the relative importance of more or less energy-intensive sectors. The structural effect is positive if the share of energy-intensive sectors grows.

The intensity effect (E_{INT}) is represented by the ratio$\frac{E_i}{A_i}$. It accounts for changes in total energy consumption due to technology advancements, efficiency improvements, policy, and other effects. The intensity effect is negative if there is a drop in energy intensity.

The residual effect (E_{RES}) is an undesirable output from an imperfect decomposition, which occurs with some of the mathematical methods.

Decomposition is commonly calculated on a yearly base. In the case here, this would mean the first year of the period is marked as 0 and the last year as T. However, it could also be conducted on a shorter period if the corresponding data are available. Regarding the actual calculation, two approaches are possible:

Chaining decomposition uses annual time-series data, and decomposition is made on changes between consecutive years. The results for each effect are then 'chained' to generate a time series.

Meanwhile, non-chaining decomposition is conducted using data for only the first and last year of the period, without calculating it for the intermediate years.

Conclusions

The decomposition analysis result consists in a series of coefficients which represent the estimated effect of each component of the initial equation on the overall energy consumption between different periods covered in the dataset. These coefficients can then be used to perform various analysis, and presented in graphical format.

References

[99] IEA. (2019, December 19). Energy efficiency is the first fuel, and demand for it needs to grow. Retrieved from International Energy Agency:

https://www.iea.org/commentaries/energy-efficiency-is-the-first-fuel-and-demand-for-it-needs-to-grow

[100] UN. (2021). SDG7 - Ensure access to affordable, reliable, sustainable and modern energy for all. Retrieved from United Nations: https://sdgs.un.org/goals/goal7

[101] Vreuls, H., Thomas, S., & Broc, J.-S. (2009). General bottom-up data collection, monitoring, and calculation methods.

[102] Liu, N. (2006). Energy efficiency monitoring and index decomposition analysis. National University of Singapore.

Y-0 ESTIMATES OF SUPPLY SIDE OF ENERGY BALANCES AND KEY INDICATORS

Keywords: energy balances, now-casting, ARIMA, Chow-Lin

Eurostat's annual energy statistics and its key output energy balances are published at the end of January of the second year following the reference year. Since 2018, Eurostat is complementing the release of regular energy balances with the release of early estimates of energy balances in June – 6 months after the end of reference year. However, some users need data even faster. This work investigates if additional improvements in timeliness can be achieved, by estimating the data relevant for the on-going year at the end of the year – the so called Y-0 estimate.

In terms of applied methodology, the estimation consists of the following sequential steps. In the first step, the aim is to estimate as many yearly flows as possible using corresponding monthly indicators. As the indicators need to be forecasted using an algorithmic approach, we believe a univariate time-series approach is preferable to a multivariate one because multivariate time-series models are done on a case-by-case basis which is not feasible for this analysis. Furthermore, since monthly indicators usually exhibit seasonal patterns, a seasonal ARIMA approach [Kocenda, Evzen, i Alexandr Cerný. Elements of Time Series Econometrics: An Applied Approach. Karolinum Press, Charles University, 2014.] is considered appropriate to forecast the missing monthly data points until December of the year which is being estimated.

Subsequently, as the monthly flows are not necessarily fully consistent with yearly flows in levels, to predict a low frequency series by a high frequency indicator series we use the Chow Lin maximum log likelihood method [Eurostat. Energy balance guide: Methodology guide for the construction of energy balances. Eurostat, 2019.; Sax, Christoph, and Peter Steiner. “Temporal Disaggregation of Time Series.” The R Journal 5, no. 2 (2013); Eurostat (2018): ESS guidelines on temporal disaggregation, benchmarking and reconciliation, Manuals and guidelines, Publications Office of the European Union, Luxembourg] to disaggregate the yearly flows to a monthly frequency. While other temporal disaggregation methods could be used as well, it is important to note that monthly indicators may not be directly related to the yearly energy flows for all series to be estimated. Since when using Chow-Lin the trend component follows the trajectory of the annual data, this method may perform better in cases when the trajectories of the indicators temporarily deviate from the trajectory of the yearly series [Marini, M. (2016): Nowcasting Annual National Accounts with Quarterly Indicators: An Assessment of Widely Used Benchmarking Methods, IMF Working Papers, Washington]. Then, as the consistency between the low and high frequency series is achieved, an estimate of the yearly flow is simply the sum of the monthly series estimated by the Chow-Lin procedure.

As only a part of the yearly energy balance flows have corresponding monthly indicators available, additional methods are needed to estimate the yearly energy flows for which we do not have corresponding indicators. These methods include forecasts based on regressions using time trend, extrapolation formulas based on technologically founded ratios (e.g. shares of other energy inputs used in oil refining), extrapolation formulas based on previous year’s shares and the estimation of the aggregates using formulas based on energy balance identities.

In this paper, a methodological approach to the estimation of the current yearly energy balance flow values has been proposed. In addition, this research provides a brief analysis of the characteristics of monthly and yearly energy balance flow data which are used as inputs in the estimation. Finally, it provides an analysis of the forecasting performance of the proposed methodology, as well as possible reasons for the lower forecast performance for a part of the estimations.

Import dependency by origin indicator

Keywords: energy trade, import dependency, oil

Introduction

The indicator tracks the dependency linked to the imports of crude oil and finished petroleum products made from crude of a specific origin. This is to trace the dependency by origin connected to the purchase of primary commodities by other countries. The assumption that must be accepted as a necessary simplification is that the exports of a country which produces finished products, are made from indigenous crude and from imported crude proportionally (the same for domestic consumption of such finished products).

In other words, the imported finished products to country C from country J (which produce them in its refineries) are assumed to be in proportion to the imported crude and the indigenous crude of country J.

The formula corrects for the different methodology in reporting trade of crude vs trade of products which asks countries to report finished products by last consignment rather than ultimate origin. To achieve this, the imported finished products from country J to country C are also assumed to be in proportion to imported oil to country J from other countries.

The indicator measures only the first layer of secondary dependency. It is possible for example, that crude oil extracted in P is imported by the reporting country D, then refined into finished products, which are exported to reporting country E, which are then re-exported to reporting country C (country of interest); the second level of secondary dependency related to these amounts would not be picked up by the indicator. This was a conscious choice because adding further complexity to the formula to track additional levels of secondary dependency was deemed unjustified as the subsequent corrections would have been much smaller than the first level correction.

The indicator captures the secondary dependency only if the countries involved report to Eurostat. It is possible that crude from P for example is imported to the non-reporting country X and refined into finished products, which are then exported to EU country C (country of interest); the ultimate origin in this case would not be picked up in the calculation.

The indicator tracks the combined dependency by origin of both crude oil and refined products imports. This is to accomplish the calculations explained above. It also avoids being misleading, as separating the indicator into two, one for crude oil and one for refined products, could lead to misunderstandings as the components can have very different absolute amounts and therefore relative weights. For example, a 90% dependency for crude and 10% dependency for products might sound as a relevant dependency overall if it is not noticed for example that crude oil is only a minor fraction of the total.

The indicator refers to the supply side of the market. It shows the dependency of a country with regards to the origin of the imports and indigenous production. The aim is not showing the dependency related to the demand side or to the domestic energy consumption. For example, the use of oil products in international maritime bunkers or in international aviation or for non-energy purposes is not deducted. One should be careful to consider these amounts in the evaluation of the complete picture of energy dependency of a country.

FORMULA

If $\mathrm{I}_{\mathrm{Tot,C}} \leq \mathrm{E}_{\mathrm{Tot,C}}$ then we define $\mathrm{UOIO}_{\mathrm{C,P}}$= :M (indicator is not calculated for net exporters)

If $\mathrm{I}_{\mathrm{Tot,C}} > \mathrm{E}_{\mathrm{Tot,C}}$ then

$$\mathbf{UOIO_{P,C}} = \left(\mathbf{I_{P,C}} + \Big(\sum_{\mathbf{j \in allRC}} \mathrm{I_{j,C}} \times \frac{\mathbf{I_{P,j}}}{\mathbf{PROD_j + I_{Tot,j}}} \Big) - \Big(\mathbf{E_{Tot,C}} \times \frac{\mathbf{I_{P,C} + SIGMAc}}{\mathbf{PROD_C + I_{Tot,C}}} \Big) \right) / (\mathbf{PROD_C} + \boldsymbol{net}\mathbf{I_{Tot,C}})$$

Exception: United Kingdom for 2015 to 2019 and Norway: these are origins which report or reported data to Eurostat and therefore for which we can further adjust the formula correcting for the declared imports and indigenous production:

$$\mathrm{UOIO_{P,C}} = \left(\mathrm{I_{P,C}} \times \frac{\mathbf{PROD_P}}{\mathbf{PROD_P + I_{Tot,P}}} + \Big(\sum_{\mathbf{j \in allRC}} \mathrm{I_{j,C}} \times \frac{\mathrm{I_{P,j}}}{\mathbf{PROD_j + I_{Tot,j}}} \Big) - \Big(\mathrm{E_{Tot,C}} \times \frac{\Big(\mathrm{I_{P,C}} \times \frac{\mathbf{PROD_P}}{\mathbf{PROD_P + I_{Tot,P}}}\Big) + \mathbf{SIGMAc}}{\mathbf{PROD_C + I_{Tot,C}}} \Big) \right) / (\mathbf{PROD_C} + \boldsymbol{net}\mathbf{I_{Tot,C}})$$

$\mathrm{UOIO}_{\mathrm{C,P}}$ is the indicator for dependency by origin of imports of oil for country C from origin country P. It is expressed as a percentage of total net imports and indigenous production of oil for C

C is the country of interest (i.g. EU Member States).

allRC refers to all reporting countries, which is all 27 EU Member States plus Norway, Iceland, EU candidate countries, EU potential candidates and Energy Community Contracting Parties

P are Russia, United States, Norway, Saudi Arabia, United Kingdom, Kazakhstan, Nigeria, Iraq, Azerbaijan, Algeria, Libya that are the top 11 non-EU origins of oil (This calculation excludes the category “Not specified” which in 2020 was only around 1% of the total)

$\mathrm{I}_{\mathrm{P,C}}$ is the imports of oil from country P to country C. In other words, imports declared by country C as originating from country P

$\mathrm{I}_{\mathrm{P,j}}$ is the imports of oil from country P to country j. In other words, imports declared by country j as originating from country P

$\mathrm{I}_{\mathrm{j,C}}$ is the imports of oil from country j to country C. In other words, imports declared by country C from trade partner j

$\mathrm{E}_{\mathrm{Tot,C}}$ is the total exports of oil of country C

$\mathrm{I}_{\mathrm{Tot,C}}$ is the total imports of oil of country C

$\mathrm{I}_{\mathrm{Tot,j}}$ is the total imports of oil of country j

SIGMA$=\sum_{\mathrm{j \in allRC}} \mathrm{I_{j,C}} \times \frac{\mathrm{I_{P,j}}}{\mathrm{PROD_j + I_{Tot,j}}}$

$PROD_j$ is domestic (indigenous) production of oil in country j

$PROD_C$ is domestic (indigenous) production of oil in country C

$\mathit{net}I_{Tot,C}$ is the total net imports (total imports – total exports) for country C

:M represents Eurostat's database convention for dissemination of "missing value – data cannot exist"

Application

$$UOIO_{P,C} = \left(I_{P,C} + \left(\sum_{j \in allRC} I_{j,C} \times \frac{I_{P,j}}{PROD_j + I_{Tot,j}} \right) - \left(E_{Tot,C} \times \frac{I_{P,C} + SIGMAc}{PROD_C + I_{Tot,C}} \right) \right) / \left(PROD_C + \mathit{net}I_{Tot,C} \right)$$

The import dependency from Russia for Country C is calculated as:

The direct imports of crude oil and petroleum products of Country C from Russia

+

The secondary dependency: The indirect imports assumed coming proportionally from Russia via other reporting countries (Petroleum products produced in other reporting countries using Russian crude oil and/or petroleum products produced in Russia but imported and re-exported by other reporting countries)

-

Country C's exports corrected for the proportion of crude oil and petroleum products re-exported but assumed to be originating from Russia

/

Total net imports and indigenous production of oil for country C (to express the indicator as a percentage)

References

[1] Eurostat database: https://ec.europa.eu/eurostat/web/main/data/database

Evaluating energy prices and costs impacts on households and industry’s costs

Keywords: electricity, natural gas, prices, costs

Introduction

On the demand side of the electricity market, residential consumption trends are expected to be mostly driven by the increasing number of households, proliferation of electric appliances or the electrification of heating, while energy efficiency measures such as installing LED lightbulbs, more efficient appliances and smart meters may push electricity demand lower. Average temperatures play an important role too, both for heating and for cooling. In the case of businesses, the consumption of electricity is mainly influenced by two similarly countervailing factors: the level of economic activity and energy efficiency measures. Whilst demand is somewhat inelastic with prices, i.e. there is a minimum electricity use necessary for households or businesses, the increases in prices in the last two years are understood to make price an increasingly important driver in reduced demand.

Household electricity prices

EU household electricity retail prices in the DC band grew by 22% between 2010 and 2021 (see Figure 1 [left]); while the energy crisis will lead to further sharp increases in both wholesale and retail prices for 2022. The average annual growth rate of retail prices within the DC band were 2.5%/y from 2010 to 2020; this grew to 8% between 2020-2021. In absolute terms, the average EU27 retail price grew from 215 EUR/MWh to 231 EUR/MWh (+16 EUR/MWh) in the same period of time. This movement between 2019 and 2021 was driven by Energy and supply (+16 EUR/MWh) and Network costs (+3.6 EUR/MWh), whilst taxes and levies declined a little in this period (-3.1 EUR/MWh).

When looking at the cost composition between the cost of the energy, network costs and taxes and levies (see Figure 1 [right]) the taxes and levies category saw its share in the total bill decrease significantly from 40% in 2017 to 37% in 2021 (further analysis of this component is provided in the next section). Meanwhile, energy component of the prices was decreasing between 2010 and 2020, and so was the contribution of energy and supply costs to the total bill. However, in 2021, the contribution of the energy component increased to 37% of the total energy bill in 2021 – the first time since 2014. Most importantly, energy and supply costs in 2021 exceed all recorded values since 2010.

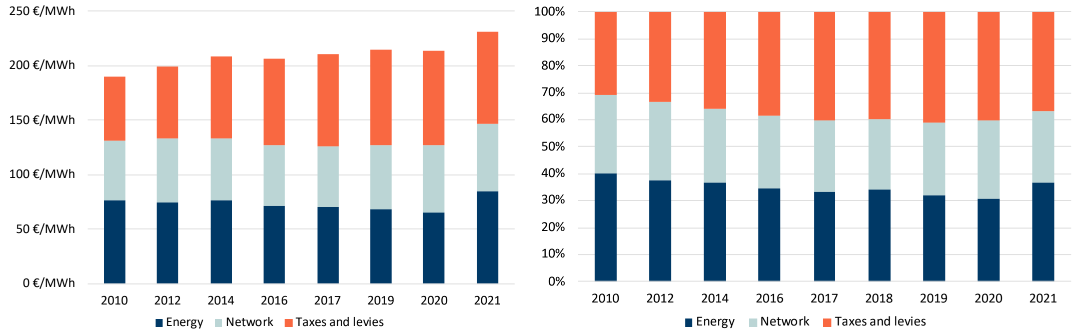


Figure 1 – Evolution and composition (left) and relative composition (right) of the EU household price (DC band)

Source: DG ENER in-house data collection, Eurostat [1]

Composition of taxes, levies, fees and charges

In order to better understand how Member State policies and fiscal instruments impact household retail prices, the taxes, levies, fees and charges category is further broken down into six subcomponents: VAT, renewable taxes, capacity taxes, environmental taxes[97], nuclear taxes and other. Note that only policies and mechanisms that directly impact retail prices are considered, and not all tax subcomponents exist or are applied in all Member States. The following chart displays an evolution of EU27 averages.

Figure 2 shows that taxes and levies associated with policies designed to support renewable energy sources have seen a decline since 2019 from 26 EUR/MWh to 24 EUR/MWh. Meanwhile capacity taxes continue to gradually increase over time. Since 2010, capacity taxes have increased by 288%, although at 3.9 EUR/MWh in 2021, still represent only a minor part of the tax component. While environmental taxes and levies had been following an increasing trend (+49%) since 2010, their contribution has reduced a little after peaking at 19 EUR/MWh in 2019 and were, in 2021, 17.9 EUR/MWh, or 7% less than in 2019. The trends highlight that taxes and levies for renewables and environmental purposes (and the policies they support) can be amongst those most vulnerable when pressures on prices and taxes emerge. The elimination of the EEG surcharge in Germany in 2022 is expected to drive further reductions in the environmental taxes and levies component in the next few years.

[97] This category includes general energy taxes, which are typically classed as having an environmental purpose

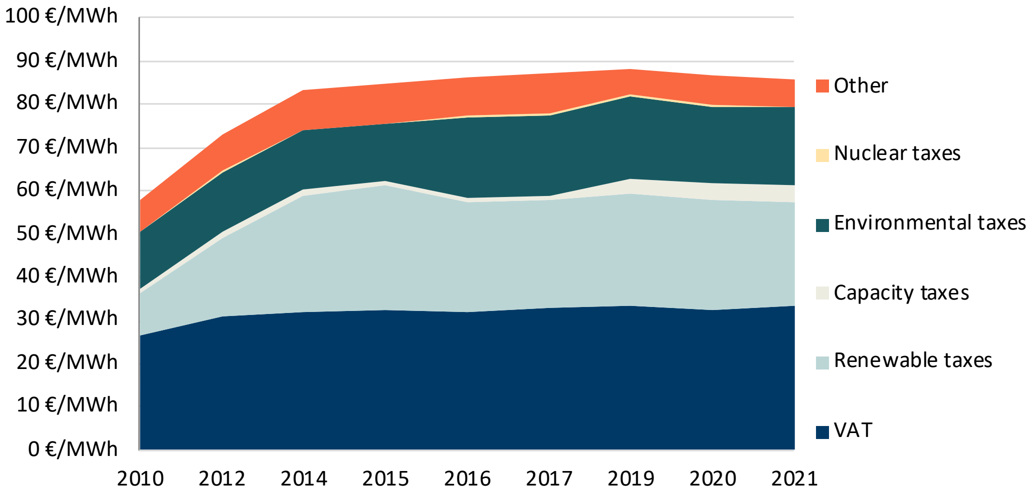


Figure 2 – Evolution of taxes, levies and charges for EU households since 2010 (DC)

Source: DG ENER in-house data collection, Eurostat [1]

The structure of the taxes and levies component between 2019 and 2021 showed minor changes, with the most notable being the small downward shift in the environmental taxes and renewable energies contribution (both -1%), while VAT and capacity tax increased (both +1%) (Figure 3).

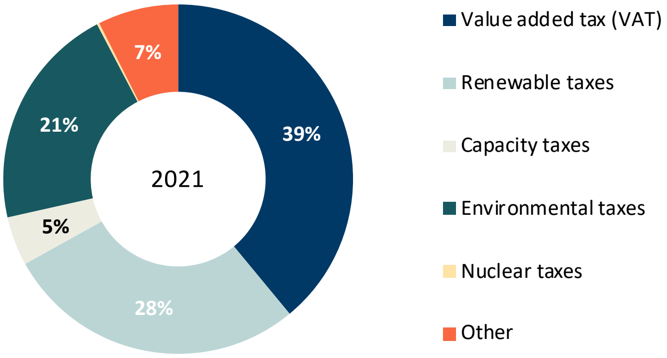


Figure 3 – Composition of the taxes and levies component of household electricity prices in 2021 (DC band)

Source: DG ENER in-house data collection, Eurostat [1]

Non-households electricity prices

Figure 4 shows that industrial electricity prices in the ID band grew at an average annual rate of 2%/y during the last decade, overall showing an increase from 96 EUR/MWh in 2010 to 124 EUR/MWh in 2021. Since 2019, industrial electricity prices increased by 14%, from 108 EUR/MWh to 124 EUR/MWh (+16 EUR/MWh) in absolute terms. This is the highest 2-year increase in prices observed within the past decade.

Due to the exclusion of VAT and other factors related to tariff calculations, industrial electricity prices are more influenced by the energy component compared to households and hence, more driven by developments in the wholesale market. The energy component, despite a small dip in prices in 2020, increased by 24% (+12 EUR/MWh) in 2021 compared to 2019. Network charges also contributed, increasing by 14% (+3 EUR/MWh) since 2019, and these are at their highest level since 2010 at 23.3 EUR/MWh. The lowest increase was observed in levies and taxes which only saw an increase of 1.6% (+0.6 EUR/MWh) between 2019 and 2021. This small increase is in contrast to the small decrease in taxes and levies (-3 EUR/MWh) experienced for household retail prices, see figure 4.

In terms of energy price components, in 2021 the energy component now contributes 51% of the total industrial price for electricity, compared to 47% previously. Meanwhile, network charges' contributions remain unchanged at 19% and a substantial decline was observed for taxes and levies, which in 2021 contributed 30% to the electricity price compared to 34% in 2019.

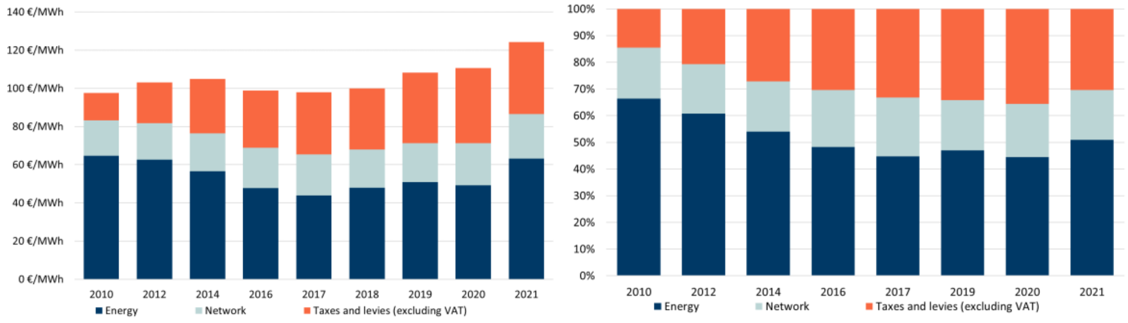


Figure 4 – Evolution and composition of the EUR 27 industrial retail prices (ID band), absolute (left), share (right)

Source: DG ENER in-house data collection, Eurostat [1]

References

[1] Eurostat database (Eurobase): https://ec.europa.eu/eurostat/web/main/data/database

On line Job Advertisment (GASP3A.3)

Session Chair: **Fernando Reis** *(Eurostat)*

Using online job adverts data to understand labour market demand for skills
Nora Condon *(SLMRU (SOLAS)*

Towards statistics on skills: considerations on using WIH OJA data
Vladimir Kvetan *(Cedefop)*

Changes of skill requirements across jobs and labour markets: evidence from web vacancies
Emilio Colombo *(Universita' Cattolica del Sacro Cuore)*

Demand for digital skills: experimenting with information extraction from online job advertisements
Joanna Napierala *(Cedefop)*

Online Job Advertisements: the Italian Case Study
Annalisa Lucarelli, Elena Catanese, Francesco Amato, Francesca Inglese, Giuseppina Ruocco
(National Institute of Statistics–ISTAT)

Towards statistics on skills: considerations on using WIH OJA data.

Keywords: Online job advertisements, Skills, Skills statistics, Skills intelligence.

5. Introduction

In the last two decades, "skills[98]" became a central element of European labour market policy [1]. In the currently rapid changing labour markets, information on employers' skills needs is key for shaping the education and training sector. Therefore, skills intelligence is becoming a crucial element for strengthening Europe's competitiveness, improving people's life chances, and minimising skills mismatches. Although skills intelligence is defined as "the outcome of an expert-driven process of identifying, analysing, synthesising, and presenting quantitative and/or qualitative skills and labour market information," [2], it is mostly based on proxies as occupations or qualifications. Therefore, more granular, specific, and timely information on skills is in high demand.

5.1. Online job advertisements as a source of data on skills requirement

Gathering data on skills requirements directly from employers via a survey is highly challenging and resource intensive [3]. Moreover, it can provide still only limited information [4] and is subject to severe time lags between the publishing date and reference period. Therefore, methods providing more granular and timely data are necessary.

Since 2015, the European centre for the development of vocational training (Cedefop) and Eurostat were involved in activities to understand the richness of information gathered from the Web, including online job portals. Early analysis done throughout pilot studies confirmed that information contained in an online job advertisement can provide reliable data on skills requirements. Therefore, after successfully setting up a data production system (DPS), Cedefop [5] has moved it into the Web intelligence hub (WIH) [6] to make it a core of new and experimental sources of statistics.

Currently, the DPS operated by Eurostat's WIH is providing regular data feed on online job advertisements in Europe. This article identifies some challenges in extracting skills from OJAs and suggest possible ways to overcome them. In addition, it also presents ways in which processed data could be used to produce official statistics on skills and so to provide better information to policymakers. In the following part, the methods used for skills extraction in DPS, and possible alternatives will be described. The section on results will focus on discussing the richness of skills extraction by various methods. Conclusions will build on the richness of the results and develop ideas on how to overcome the weaknesses of the individual method with the strengths of the others. The possibilities for types of statistical indicators from OJA will be also described.

[98] For the purpose of this article and for simplicity we use term "skills" as a common word for skills, knowledge and competencies.

6. Methods

The DPS consists of various modules developed to ingest and process OJAs. The pre-set sources (websites and web portals), assessed and selected by experts enter the "data ingestion phase", which involves identifying OJAs and downloading their content. During the next 'pre-processing' phase, the downloaded data are cleaned of irrelevant content and de-duplicated. After that, the 'information extraction', process can follow, to translate the relevant content of OJAs into a database organised by taxonomies and classifications for each variable (more information is available in [5].

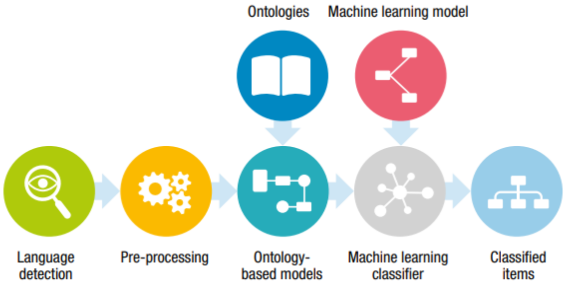


Figure 1: Information extraction process
Source [5]

6.1. Ontology based extraction of skills

Each OJA is processed in its original language within its individual language production pipeline. The multilingual classification of European skills competences and occupations (ESCO)[99] is used as a basis for building the ontologies to extract occupations and skills. In the skills extraction process, the algorithm first tries to classify the OJA using text matching and/or similarity with the ontology terms. If no result is obtained, a machine learning algorithm decides on the classification. The machine learning algorithm uses statistical techniques to give computers the ability to 'learn' (progressively improve performance on a specific task) without being explicitly programmed.

6.2. Data-driven approach to skills extraction

In the highly dynamic labour markets, where occupations and skills are subject to fast changes, an ontology-based system might not capture all skills required by employers. Therefore, a more flexible, data-driven approach has been tested with the purpose to process OJAs and identify new terms that are not yet included in the skills ontology. The system was tested in the Italian, Spanish, French, and Romanian languages. It allowed identifying those terms that might represent a new emerging skill. These terms were subsequently tested by language experts along with the most similar ESCO skills and the occupation code in which the terms appear.

The system is based on word embeddings. The key idea behind word embeddings is that words with similar frequency distributions tend to have similar meanings. Words are represented by semantic vectors, which are usually induced from a large corpus using co-occurrence statistics

[99] https://esco.ec.europa.eu/en

or neural network training. Word embeddings [7] can capture the context of a word in a document, semantic and syntactic similarity between words and other linguistics patterns.

7. Results

7.1. ESCO powered extraction

For this exercise, ESCO version 1.1 (the most recent one) was used. It contains 13,890 concepts, organized within skills, knowledge and attitudes and values pillars. A separate pillar describes language skills and knowledge[100]. All these terms are available across all 24 official languages of the EU. Under ideal circumstances, the WIH DPS system should capture all these almost 14 thousand terms across the countries it covers; However only small percentage of terms is successfully classified (Figure 2).

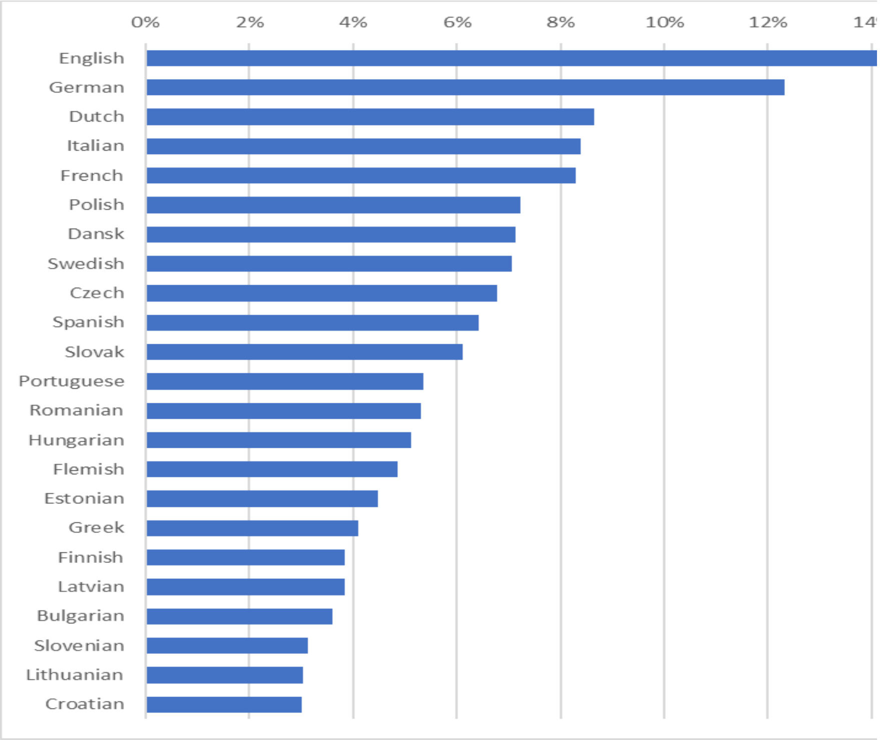


Figure 2: Percentage of skills identified in OJAs, by language.
Source: Own calculation based on WIH-OJA data.

Even the most successful language pipeline, English, can match the terms found in OJA to only 17% of the ESCO taxonomy, and most other language pipelines succeed only in a small fraction of that. Our analysis points out to several reasons for this.

1. ***Difference in quality of language versions*** of ESCO across the EU Member States due to translation errors. The translated terms used by ESCO may not be preferred by employers who draft vacancy notices.

2. ***Lack of alternative labels.*** Language difficulty and existence of many different word endings poses a challenge for the classifier, which is not able to correctly match the ESCO term to a term used by the employer, for example because of different declension of a noun or an adjective.

[100] https://esco.ec.europa.eu/en/classification/skill_main

3. The comprehensiveness, scale, and a long update process of ESCO leads to ***late introduction of terms*** associated with new and emerging skills.

4. ***Many skills are implicit***, for example a "computer use" skill may not be specifically asked for a software developer, as the employer considers it as obvious.

5. Employers may use ***different forms of signalling***, such as formal qualification, to indicate their requirements, instead of providing the long list of skills that stem from such qualification.

7.2. Data-driven approach

To understand better the data and especially skills content which was not classified, we have examined 5 million OJAs in four languages: French, Italian, Spanish and Romanian. The outcomes (terms identified) generated by AI were examined by human experts. In the end, the data exploration has discovered 831 new skill terms, not covered by ESCO. In relation to ESCO, these terms fell into one of 4 categories: generalisation of an existing term, specification of an existing term, a synonym of an existing term (an alternate label), or a truly novel term. Overall, the experts have rated the validity of the proposed terms highly: from 60% validity in French to almost 80% validity in Italian.

8. Conclusions

In the current dynamic labour markets, the information on skills requirements of employers is key for education and training, labour market and other economic policies. The WIH is working towards the development of statistics on skills. Currently there are three ways to do it – descriptive statistics (what skills are required in what occupations); skills specific statistics (skill capturing one of the skills dimension – e.g. green, digital, transversal) or contextualised (building maps how occupations are close to each other with respect of the skills mix, occupations with more dynamics skills needs). However, for each of this type of potential statistics, the sound and consistent skills extraction is a must.

Although the ESCO provides an existing solution for a project that requires work with OJAs written in national languages, the use is faced with various challenges. At the same time, improvements must be considered and introduced, to increase the information yield from OJAs. As a way forward more bottom up approach seems to be the right way forward to mitigate for large differences between amount of information gathered in different countries. However, for this exercise, the terms generated by AI need to be validated by human experts. Therefore for further development of the project it is necessary to strengthen this element.

References

[103] European commission, Communication on a European skills agenda for sustainable competitiveness, social fairness and resilience (June 2020), https://ec.europa.eu/social/BlobServlet?docId=22832&langId=en

[104] Cedefop, Crafting skills intelligence, 2019. Available at: https://www.cedefop.europa.eu/en/blog-articles/crafting-skills-intelligence. Accessed on 10.10.2022.

[105] Cedefop, User guide to developing an employer survey on skills needs, Research paper No.35, Luxembourg: Publications Office of the European Union, 2013, https://www.cedefop.europa.eu/files/5535_en.pdf

[106] Cedefop, Piloting a European employer skills survey, Indicative findings, Research paper No.36, Luxembourg: Publications Office of the European Union, 2013, https://www.cedefop.europa.eu/files/5536_en.pdf

[107] Cedefop, Online job vacancies and skills analysis: a Cedefop pan-European approach. Luxembourg: Publications Office, 2019. http://data.europa.eu/doi/10.2801/097022

[108] Descy, Pascaline et al. 'Towards a Shared Infrastructure for Online Job Advertisement Data'. 1 Jan. 2019 : 669 – 675.Descy, Pascaline et al. 'Towards a Shared Infrastructure for Online Job Advertisement Data'. 1 Jan. 2019 : 669 – 675. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190547

[109] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Changes of skill requirements across jobs and labour markets: evidence from web vacancies

October2022

1 Motivation

In the past few decades technical progress, globalization and the re-organization of the production process with outsourcing and offshoring have radically altered the demand for certain skills.[101]

Albeit the overall effect is extremely complex we can roughly identify two main dimensions of it. One that we can call the *extensive margin* pertains to the creation of new jobs and the destruction of existing ones. This is probably the most debated issue at the center of the policy debate.

The second dimension operates along the *intensivemargin*. In addition to create and destroy jobs, technology is changing profoundly existing jobs, in particular the tasks and their skill requirements modifying considerably the skill-mix employers require and placing greater emphasis on soft skills such as problem solving, ability to work in team, communication abilities etc.

In terms of overall effect the intensive margin is likely to be more important than the extensive one as it potentially affects the entire stock of the labor force. However to analyze it, is necessary to develop tools that are able to measure the characteristics of the jobs, their skill requirements and how these change over time and across occupations.

In this paper we develop a set of innovative tools to online job advertisements (OJAs) on the Italian and UK labor market. Our approach allows to calculate, for each occupation, the different types of skills required. Those skills are mapped into a standard classification system. We subsequently develop measures of the relevance of soft, hard and digital skills within occupations. We develop measures of the *variation* over time in the skill intensity, identifying the change in the skill content of occupation.

More in detail we will use the distinction made above between extensive and intensive margin to decompose the variation of the skill content of occupation into.

- Changes in the **Skill set** (Extensive margin). The same occupation requires new skills which were not required before (e.g. Scala programming language for computer scientists)

- Changes in the **Core set** (Intensive margin). Change in the intensity of use of skills within occupation (e.g. Python is still required for computer scientists but with increasing intensity)

[101] See Acemoglu (1998, 2002); Acemoglu and Restrepo (2019); Autor, Levy, and Murnane (2003), Card and DiNardo (2002)

2 Dataandmethods

2.1 Data

The source of the data is the project that Eurostat and Cedefop are currently carrying to collect online vacancies in Europe from job-portals since 2018. The sources are the major portals that advertise vacancies and include newspaper websites, job boards and employment agencies. The data is stored in Eurostat's Web Intelligence Hub. The current analysis is performed on Italy and UK data of 2019 and 2021. The dataset contains information about the occupation (ISCO), sector (NACE), educationrequirement(ISCED),region(NUTS),skillsclassifiedfollowingESCOtaxonomy.

2.2 Methods 2.3Measuring skill intensity

To measure the intensity of a given skill (or category of skills) within occupations, we use the concept of revealed comparative advantage (RCA) developed by Alabdulkareem et al. (2018).

Given a set of occupations $\bar{O}$ = {o_k,k = 1,...,m}, a set of skills $\bar{S}$ = {s_j,j = 1,...,p}, the measure *rca* for o_i and s_l is defined as:

$$rca(o_i, s_l) = \frac{sf(o_i, s_l) / \sum_{j=1}^{p} sf(o_i, s_j)}{\sum_{k=1}^{m} sf(o_k, s_l) / \sum_{k=1}^{m} \sum_{j=1}^{p} sf(o_k, s_j)} \qquad (1)$$

which ranges between [0,+∞).

Basically RCA is a variant of the TF-IDF where skills are considered as terms and occupations as documents. The idea is to assign a value to the relationship between a pair of skill and occupation, taking into account both (i) the frequency with which the skill is mentioned in a vacancy for the occupation (proportional), and (ii) the frequency with which the skill appears in all the job vacancies (inversely proportional). The aim of the RCA is to give more importance to skills which are specific to the occupation.

2.4 Change of skill requirements

We decompose the change in skill requirement within occupation in two components

Change in the **coreset**, measured by computing the weighted Jaccard distance

$$\Delta\_core\_set_i = WJ_i \cdot \frac{|S_{19i} \cup S_{21i}|}{\max_i(|S_{19i} \cup S_{21i}|)} \cdot 100$$

where

$$WJ_i = 1 - \frac{\sum_{j \in |S_{19i} \cap S_{21i}|} \min(rca_{19ij}, rca_{21ij})}{\sum_{j \in |S_{19i} \cap S_{21i}|} \max(rca_{19ij}, rca_{21ij})}$$

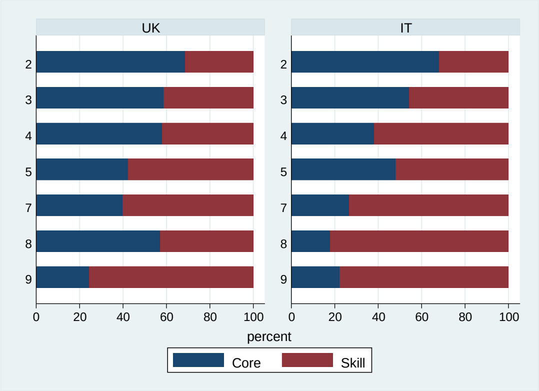
Change in the **skillset**

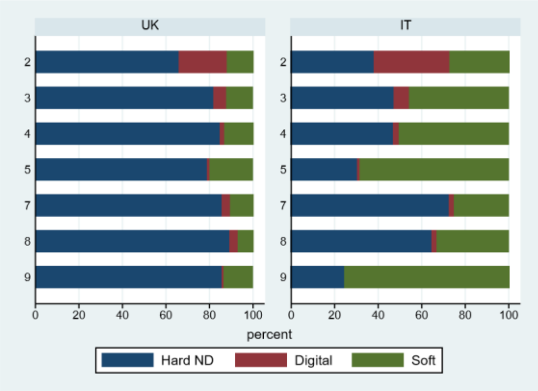
$$\Delta\_skill\_set_i = \frac{|S_{21i} - S_{19i}|}{|S_{21i} + S_{19i}|} \cdot 100$$

From the RCA it is possible to compute the *effective use* of the skill for the occupation. This is a binary measure of the importance of a skill s_l for an occupation o_i defined as follows:

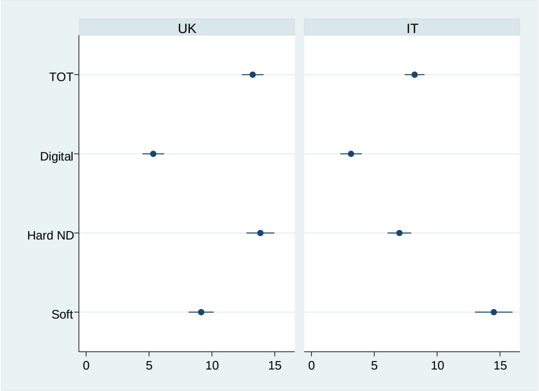
$$e(o_i, s_l) = \begin{cases} 1 & rca(o_i, s_l) > 1 \\ 0 & otherwise \end{cases} \quad (2)$$

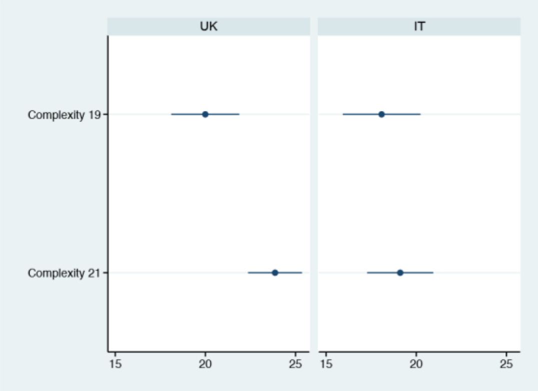
The *effective use e* is a binary metric which allows comparing the relative importance of a skill s_j to an occupation o_i (the numerator in equation rca) to the expected relative importance of that skill on aggregate (the denominator in rca). Intuitively when $rca(o_i,s_j) > 1$ means that the occupation o_i relies on skill s_j more then expected on aggregate.





(b) Change in the core set





We use the effective use to define **job complexity**. More complex jobs require more skills with higher intensity with respect to the average

3 Results

(a) Change of Extensive vs intensive margin

(c) Change in the core set

(d) Change in job complexity

Figure 1: Change in core set and extensive vs intensive margin, within occupation 4d, 2019-2021

3.1 Skill convergence?

Focusing on job complexity we regress the change in job complexity over average complexity for each occupation in 2019 and a bunch of covariates. The negative sign shows that jobs are becoming more complex with those less complex in 2019 displaying the fastest growth

Table 1: Convergence in complexity. Dep var. Δ complexity by occupation

Complexity2019	-0.266***
	(0.016)
Dit	-3.521***
	(0.351)
Admin, accounting	1.245**
	(0.596)
Delivery services	-1.879***
	(0.533)
Design RD	0.920
	(0.598)
ICT	2.635***
	(0.764)
Marketing	1.432**
	(0.638)
r2	0.560
N	288

References

Acemoglu, Daron. 1998. "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality." *The Quarterly Journal of Economics* 113 (4):1055–1089.

———. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature* 40 (1):7–72.

Acemoglu, Daron and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33 (2):3–30.

Alabdulkareem, Ahmad, Morgan R. Frank, Lijun Sun, Bedoor AlShebli, Csar Hidalgo, and Iyad Rahwan. 2018. "Unpacking the polarization of workplace skills." *Science Advances* 4 (7).

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118 (4):1279–1333.

Card, David and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *JournalofLaborEconomics* 20 (4):733–783.

Demand for digital skills: experimenting with information extraction from online job advertisements

Keywords: digital skills, skills extraction, online job advertisements, big data, skills intelligence.

The rapid technology changes have revolutionised the ways we work and impacted the skill demand. The need of equipping the workforce with digital skills to allow them to fully benefit from these new technologies at work is well acknowledged by many EU policy documents. Yet, the monitoring of demand for the skills workers will need is still missing, making the informed policy making decisions for education sector challenging. The approach presented in this paper is an example of a successful attempt made towards improving the labour market statistics providing information not only about number of vacancies, but also about skills demanded for IT roles.

Introduction

The interest to monitor changes in demand for skills related to the twin (green and digital) transition is growing across EU Member States.[102] These two big trends: transition toward green economies and the rapid changes related to digitalisation, automation and robotization, are changing and shaping the situation on the labour markets at the same time. These advances call for a better understanding of changes in demand for workers' skills. In order to understand which, the most demanded skills are brought about by the recent changes related to this twin transition or help in identification of emerging skills, Cedefop together with Eurostat is developing a WIH-OJA database. This database is based on the content extracted from online job advertisements (OJAs).[103] The analysis of such non-traditional data characterised by higher granularity and timeliness, when compared to traditional sources of knowledge about vacancies (e.g. European Job Vacancy Survey), might allow to address such pending questions. It may provide all stakeholders, including labour market and education system policy makers, and training providers, with valuable support in adaptation to the changes by bringing better understanding about directions of these expected changes.

Before this kind of support based on the skills intelligence will be achieved, the challenges related to the information extraction from OJAs must be addressed. It has been shown ([1], [2]) that information extraction solutions based on structured and fully semantic ontological approaches or taxonomies work. They allow for better extraction of meaningful information. However, such high-quality training datasets are rarely available. Moreover, the quality of extractions will be always related to the frequency of taxonomy revisions, which are known for becoming outdated rather fast. This become problematic in particular for analysis of the digital

[102] e.g. Towards a green, digital and resilient economy: our European Growth Model, 2nd March 2022

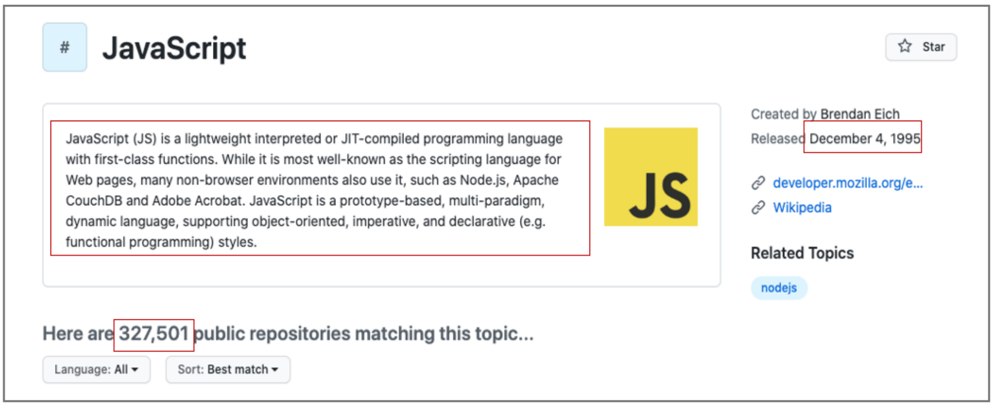
[103] More information about the project can be found on the European Web Intelligence Hub website.

and green skills emerging on the labour market which very often are requested by employers in relation to implementation of innovations or new technologies (e.g. new software, patents).

The extraction of information about job titles and skills for WIH-OJA database is now based on the application of the European Skills, Competences, and Occupations classification (ESCO). As the initial list of digital skills available from ESCO counted only 21 skill terms[104], an experimental project was launched to test if an alternative taxonomy would allow for the improvements in the extraction of information about digital skill terms. In this project, a potential of using tags extracted from stackoverflow.com website was explored to build list of latest IT technologies which were used to compare with the content of OJAs. The method, presented in this paper, seems to successfully allow for identification of the new skill terms for IT roles.

Description of method to extract digital skills

The Stack Overflow is a community-based platform[105] that anyone can access and use to either pose a question, find an answer or contribute with a solution to the technical challenges. The popularity of this public platform is confirmed with a high number of users (100 mln each month)[106] who collectively develop the knowledge. Each single question or answer is categorised by receiving a tag that allows to link and group similar topics of discussion (see Fig 1.).



JavaScript (JS) is a lightweight interpreted or JIT-compiled programming language with first-class functions. While it is most well-known as the scripting language for Web pages, many non-browser environments also use it, such as Node.js, Apache CouchDB and Adobe Acrobat. JavaScript is a prototype-based, multi-paradigm, dynamic language, supporting object-oriented, imperative, and declarative (e.g. functional programming) styles.

Created by Brendan Eich
Released December 4, 1995

Here are 327,501 public repositories matching this topic...

Figure 42. Illustration of the information about tag “Javascript” obtained from GitHub page

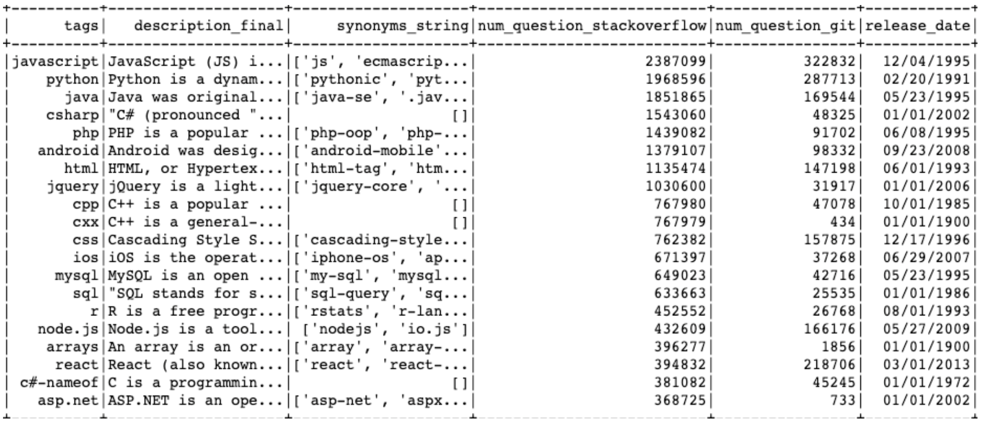
The introduction of tags was made to increase the functionality of the platform by improving the easiness of navigation. The most popular tags have more than 2mln recorded questions. The platform offers also the possibility to create a list of synonyms or alternative labels for the same tag. For our project we found it useful to use this information to create a list of IT technologies together with the information about the number of times this tag was used by users as a proxy of its popularity. The first extraction gave us more than 65ths tags that were used at least once

[104] After launch of this project the ESCO published a classification list, with a proposal to group 1200 skill terms as digital.

[105] https://stackoverflow.com

[106] Information extracted from the main website on 6/10/2022.

by any of the users of this platform since its opening in 2008. In order to narrow down this number, and exclude technologies of minor relevance to the current situation, the information from Stack Overflow was further complemented. The information on description of the technology and the number of questions asked on this specific tag coming from the GitHuB platform was used.[107] The 80% of most frequently asked questions on GitHub included 580 tags, while in Stack Overflow this number was equalled to 1300 and the later was used for annotation of OJAs. This final list after cleaning of terms with similar meaning included only 134 matched terms out of 1200 existing ones in the ESCO classification of digital skills.



```
+----------+--------------------+--------------------+--------------------------+----------------+------------+
|      tags|   description_final|     synonyms_string|num_question_stackoverflow|num_question_git|release_date|
+----------+--------------------+--------------------+--------------------------+----------------+------------+
|javascript|JavaScript (JS) i...|['js', 'ecmascrip...|                   2387099|          322832|  12/04/1995|
|    python|Python is a dynam...|['pythonic', 'pyt...|                   1968596|          287713|  02/20/1991|
|      java|Java was original...|['java-se', '.jav...|                   1851865|          169544|  05/23/1995|
|    csharp|"C# (pronounced "...|                  []|                   1543060|           48325|  01/01/2002|
|       php|PHP is a popular ...|['php-oop', 'php-...|                   1439082|           91702|  06/08/1995|
|   android|Android was desig...|['android-mobile'...|                   1379107|           98332|  09/23/2008|
|      html|HTML, or Hypertex...|['html-tag', 'htm...|                   1135474|          147198|  06/01/1993|
|    jquery|jQuery is a light...|['jquery-core', '...|                   1030600|           31917|  01/01/2006|
|       cpp|C++ is a popular ...|                  []|                    767980|           47078|  10/01/1985|
|       cxx|C++ is a general-...|                  []|                    767979|             434|  01/01/1900|
|       css|Cascading Style S...|['cascading-style...|                    762382|          157875|  12/17/1996|
|       ios|iOS is the operat...|['iphone-os', 'ap...|                    671397|           37268|  06/29/2007|
|     mysql|MySQL is an open ...|['my-sql', 'mysql...|                    649023|           42716|  05/23/1995|
|       sql|"SQL stands for s...|['sql-query', 'sq...|                    633663|           25535|  01/01/1986|
|         r|R is a free progr...|['rstats', 'r-lan...|                    452552|           26768|  08/01/1993|
|   node.js|Node.js is a tool...| ['nodejs', 'io.js']|                    432609|          166176|  05/27/2009|
|    arrays|An array is an or...|['array', 'array-...|                    396277|            1856|  01/01/1900|
|     react|React (also known...|['react', 'react-...|                    394832|          218706|  03/01/2013|
| c#-nameof|C is a programmin...|                  []|                    381082|           45245|  01/01/1972|
|   asp.net|ASP.NET is an ope...|['asp-net', 'aspx...|                    368725|             733|  01/01/2002|
+----------+--------------------+--------------------+--------------------------+----------------+------------+
```

Figure 43. The excerpt from the final list of IT technologies

Quality of data extraction

Application of this list of 1300 tags obtained from Stack Overflow platform to annotate content of OJAs helped us to identify around 144 terms which were found at least once in WIH-OJA database, out of which there were 118 new terms and only 26 terms that overlapped with existing terms in ESCO classification. However, the absence of some skill terms in OJAs could be related to the fact that possession of these skills indicates the knowledge of other IT technologies. For example, as “Drupal” is a free and open-source web content management system written in “php”, someone can assume that the knowledge of “php” suffice to be mentioned in OJAs to recruit a person who will be capable to use Drupal. Yet, by looking at tags which were included in ESCO classification as relevant digital skill terms, but not found in the content of OJAs, could be used as a sort of indication on which of the IT technologies became obsolete. For example, the knowledge of programming languages like Pascal, designed in 1969, or Prolog developed in 1972, which in the meantime became replaced by other programming languages may have disappeared from OJAs for such reason. Instead, looking at the list of programming languages requested by employers and not included yet in ESCO classification we could propose new relevant skill terms. For example, Golang (sometimes termed as Go Programming Language) would be one of them.

The application of taxonomy-based approach comes also with some challenges. The new terms need to be evaluated to understand if they are not prone to errors when used for annotation. It is quite common that the names of IT technologies are one-word terms and for that reason they

[107] GitHub.com is another tool used by milions of software developers to exchange knowledge and collaborate.

might be confused by artificial intelligence algorithms with other non-significant terms. For example, the high detection of demand for “Lua” in Romanian OJAs compared to overall demand for knowledge of this software might indicate a problem with this term due to the fact that this term means in Romanian “to contact” or “to get”. This needs to be addressed during the data cleaning in the corresponding language pipeline. Furthermore, the cross checks of the data made at the occupational level may reveal other problems related to a double meaning of some terms. For example, a term “Rails” used for tagging question related to a web application framework written in Ruby,[108] may lead to a wrong identification of IT skills for occupation of plasterers. In the OJAs recruiting for roles as plasterers we may find word “rails” which does not refer to digital skills (e.g. “we search for a person who knows how to install: rails, smooth walls etc.”). On top of terms causing problems in certain languages or occupations, we have encountered also one-digit words used to describe IT technologies which will be impossible to extract as these are by default cleaned by tools that remove stop words, sparse terms, and similar particular words before the content of OJAs is being processed. One of examples will be the “R” or “R-cran” which is a free programming language that is very often used by researchers, statisticians, or data scientists.

This thorough cross check of terms across all language pipelines and occupation levels, together with analysis of text mining results applied to the description of the tags will allow for building up a list of cleaning rules. Having such cleaning rules in place will eventually allow for making an informed selection of terms valid for annotation aiming at extraction of digital skill terms from content of OJAs.

Conclusions

Despite the high importance of digital skills in the EU policy agenda, the current information about vacancies made available by Eurostat does not provide policy makers with detailed information about demand for specific IT occupations and required skills. The potential of using information obtained from content of OJAs is in this area indisputable, as information about occupations is coded at 4-digit ISCO levels, and it also includes information about the economy sector and demanded skills. Yet, the taxonomy-based information extraction (such as the ESCO driven approach), becomes obsolete quite fast as new technologies emerge, which is especially true for the IT domain. The exploration of non-standard approaches in information extraction is needed to keep the skills intelligence up to date.

The method presented in this paper proves successful in identification of new digital skills terms. Nevertheless, as in any other method, a careful evaluation of extracted terms by labour market experts is a prerequisite before these skills can be added to the well-established taxonomies. The first results of annotation with extended taxonomy that included new terms allows for better understanding of skills required across various IT roles. This information could help not only to better understand what skills are required for certain occupations, but also to deliver feedback to experts working on the developments of competence frameworks (e.g. Digcomp).

[108] It helps in simplifying the building of complex websites.

It is worth mentioning that our experiment was carried out using tags described in English. This decision was made as very often the IT related OJAs are published in English. However, the matter if the machine translated tags would allow for further increase in the number of terms extracted from OJAs content could be further investigated.

References

[110] International Labour Organization. (2020). The feasibility of using big data in anticipating and matching skills needs (978-92-2-032855-2).

[111] Sadro, F., & Klenk, H. (2021). Using Labour Market Data to Support Adults to Plan for their Future Career: Experience from the CareerTech Challenge.

Online Job Advertisements: the Italian Case Study

Keywords: Online job advertisements, Job vacancies, Occupation

INTRODUCTION

The growing use of online job advertisement (OJA) portals has great potential for job market and skills analysis. OJAs do not replace other types of information on the labor market but their potential lies precisely in the combination with other traditional sources. In this perspective, they can provide further detailed and timely insights into labor market trends, that are difficult to gather with traditional sources, and can be used to produce supplementary indicators, enriching the current official statistical production.

The information acquired from OJAs, thanks to advances in web crawling technologies, machine learning and big data techniques, constitutes the source of data for the analysis of online job vacancies.

Currently, a pan-European system for collecting and analyzing data in OJAs [1] has been developed by the European Center for the Development of Vocational Training (Cedefop). Cedefop actively collaborates with Eurostat's Big Data Task Force and the European Statistical Systems Network (ESSnet), with the purpose of exploring the use of online job advertisements as a source of data for producing official statistics, as well as the challenges related to quality assurance.

Advances and improvements on OJAs domain constitute one of the main tasks of the ESSnet Web Intelligence Network project (WIN) which aims to implement a modular European Web Intelligence Hub (WIH) platform, providing a set of tools and services to collect web data for statistical purposes. WIN project starts from the results of the previous projects, Essnet Big Data I and Big Data II, which targeted multi-purpose statistics based on an array of non-traditional data sources. The Italian National Institute of Statistics (ISTAT) is a partner of the WIN project formed by a consortium of 17 organizations from 14 European countries.

The work we present in this short paper is placed in this context and has the aim to provide a contribution to the challenges that OJA data pose in relation to the assessment of the accuracy of the statistical output. The main goal of the paper is to present the approach used for analyzing the quality of the OJA data, especially in terms of coverage and representativeness. Potential bias in online job advertisements as vacancy measures is determinated by different causes, such as over and under-coverage errors, selection error [2], while the lack of representativeness is due to the non-probabilistic nature of the OJA sample [3].

In general in the assessment of representativeness OJAs comparison with other official statistics (JVS - Job vacancy survey, Labor force survey) carries out a main role, but also other job opening-related variables, derived from either administrative sources or business surveys.

In the next session, first, we describe the approach used to analyze the online job advertisements, then we discuss some preliminary results obtained by benchmarking OJA aggregations with other statistical index. Some findings about quality are discussed as well as further actions to improve it.

THE APPROACH USED FOR ANALYSING ONLINE JOB VACANCIES

OJAs refer to advertisements published on job portal revealing an employer's interest in recruiting workers with certain characteristics for performing certain work. These advertisements normally include a big information richness on the characteristics of the job (e.g. occupation, location, type of contract, working time and salary), characteristics of the employer (e.g. economic activity sector) and job requirements (e.g. education, skill and experience) and also on the advertisement itself (e. g. job portal and publishing and the expiring date of the ads). Part of this information is available only as natural language textual data. Therefore, this type of big data requires specific methodologies in terms of processing, classifying and analysis.

OJA data at European level are collected on a Data Lab. From there, it is possible execute queries returning the OJA data needed to work with. OJA data are made available on the Data Lab following all phases of collecting, processing, cleaning, standardising, classifying. Furthermore, in order to check that the data does not display implausible values, a data validation process has been set up for OJA data. The validation rules currently applied to the OJA dataset are both consistency and plausibility rules, which covers a variety of statistical properties of the data. For example, consistency with Eurostat's official code lists and within hierarchical classifications are required; the distribution of ads within categories of a classification is required to be reasonably stable over time and across data releases; for some variables, there should be no missing data, etc. Besides these rules (and other minor ones) operating at the level of the data record or of the distribution, there are some others that check the consistency of the whole database with basic IT and numeric requirements. These rules (also called structural validation rules) include correct naming of datasets and variables, absence of empty fields, etc.

OJAs data are disseminated on the Data Lab on a quarterly basis – even if they are available at daily level – and a release notes and all information about the various update and upgrade steps on the data are posted on a dedicated blog. At the moment, the last updating of the data regards the second quarter 2022, while data are available starting from the third quarter 2018.

2.1. Exploratory analyses

For the analyses described in this section, the Italian data available on OJA Data Lab were used.

The Italian OJA data were analysed on a monthly and quarterly basis - from Q3 2018 to Q2 2022 - taking into account the stock of OJA still active at the end of each month and at the end of the last month of each quarter. This is due to the fact that we compared the OJA data with those of the Italian Quarterly Job Vacancy Survey (conducted by Istat), which takes the last day of the quarter as the reference date for the stock of vacancies. For a similar reason, the General Industrial Activity sections (GIAs) included in the analysis cover the NACE Rev. 2 economic activity sections B to S, which are the scope of Istat's quarterly job vacancy survey.

The level of JVS derived from Istat official survey was taken as a benchmark to assess the problems of OJA coverage and the limits of representativeness. An initial exploratory analysis therefore involved comparing the levels of the OJAs with those of the JVSs, separately for the economic activity sections considered. This analysis revealed the degree of over/under coverage and the economic activity sectors where it is most relevant.

Furthermore, to investigate the causes of over/under coverage and the dynamics of OJAs, the set of job portals contributing to the total number of OJAs in each observation period (month or quarter) was analysed. While the presence of a different set of job portals in the different observation periods, with an inflow and outflow of job portals, points to a real phenomenon (e.g. new job portals might respond to an increasing demand for advertisements to be published), it also constitutes a source of instability in the data. Hence, there is a need for a stabilisation process of the OJAs data, with a freezing of job portals to be considered during the period under analysis.

The dynamics of the OJAs, for specific sectors and sections of economic activity, were compared not only with the JVS but also with those of other official macroeconomic indicators released by Istat. In particular, the OJAs year-on-year variations – obtained by comparing OJAs in each quarter with the OJAs of the same quarter of the previous year -were analysed with the same changes in the raw index of industrial production, in particular with the index of production in construction and the turnover in services which are prominent in the Italian economy.

Finally, the exploratory analysis also included a preliminary assessment of the quality of a variable of particular economic interest, namely occupation (according to the ISCO-08 classification). The consistency between the economic activity section and the major ISCO-08 occupation groups was focused on considering the OJAs percentage composition by occupations in each activity section separately.

MAIN RESULTS AND CONCLUSIONS

The potential of web data sources to complement and enrich traditional surveys is well known. More in detail, the information extracted from OJAs can highlight the dynamics of the labour demand, providing details about vacancies, without increasing the respondent burden for the enterprises.

However our preliminary analysis pointed out several inconsistencies when benchmarking with traditional official statistics both in term of level and dynamics. Thanks to these analysis we were able to identify some quality issues about OJA data coverage. Amongst we can mention::i) the number of web-scraping sources (portals) may not be stable over time: ii) the same OJA may be present in different portals thus rendering it difficult to evaluate the presence of duplicates; iii) each portal may have a different policy removal of the ad, as for instance it is not clear why an ad keeps remain present in the portal because the vacancy is still open or because they simply do not remove it.

In order to improve the quality of our OJA estimates, some major restriction should be carried out.

First we plan to exclude those OJA that fall in the last class of duration because we are not sure that those really correspond to still open vacancies, while we are sure that if the portal removes the ad in a shorter time it really corresponds to a vacancy which has been removed. Then, we plan to restrict our analysis to portals considered as “reliable”, i.e. those who on average remove their vacancies within a short period. Hopefully these operations will improve the stability of the dynamics and may reduce the so far coverage problems observed. In the long term, this could allow to improve estimates of the labor trends disaggregated in terms of major groups of occupation. In addition, further macro-analysis will be carried out, in order to reduce the observed incoherencies between the economic activity and occupation classification.

REFERENCES

[1] Cedefop, “Online job vacancies and skills analysis: a Cedefop pan-european approach”. 2019

[2] M. Beręsewicz and R. Pater, “Inferring job vacancies from online job advertisements”, *Statistical Working Papers*, Eurostat. 2021.

[3] J. Branka, V. Kvetan, J. Napierala “From the online job advertisements to official statistics – the aspects of quality assurance”. *Q2022*, Vilnius. 2022.

GISCO (JENK3A.3)

Session Chair: **Hannes Reuter** *(Eurostat)*

New spatial snow statistics in Finland using Earth observation data and modelling
Sari Metsämäki *(Finnish Environment Institute)*

Innovative in situ and other Earth Observation platforms for air quality official statistics: the GAUSS project
Orestis Speyer *(National Observatory of Athens)*

Web cartography for gridded statistics – the Gridviz library
Julien Gaffuri *(Eurostat)*

The Polish Use Case in Project GAUSS – extent and quality of green areas Smart Statistics based on EO data
Ewa Panek *(Institute of Geodesy and Cartography)*,Orestis Speyer *(National Observatory of Athens)*, Phillip Harwood *(Evenflow)*

New spatial snow statistics in Finland using Earth observation and modelling

Keywords: Snow, snow water equivalent, hydrological model, snow load, snow melt

Introduction

The Finnish Environment Institute SYKE is the national authority for monitoring and forecasting of water discharge and water level, evidently associated to snow conditions. The amount of water stored as snow and information on snow covered area, are variables with spatiotemporal changes and therefore demand constant monitoring. This is carried out through in-situ measurements, modelling and via satellite Earth Observation (EO).

Statistics derived from daily information gives a user a possibility to identify anomalies and trends and are therefore important for planning of activities in different application fields. These are for instance risk assessment (e.g., flooding due to snow melt and snow load on roofs), transport and logistics and hydropower planning. SYKE provides the information to regional entities taking care of water management and flood prevention as well as to several hydropower companies. So far, snow information is provided to Statistics Finland as an average (inside Finland) snow water equivalent on March 15th each year. In ESA-funded project GAUSS (Generate Advanced Update of Smart Statistics) more statistics are generated, with the aim of improving the input to Statistics Finland and also to evolve the snow-related studies at SYKE.

Methods

Within this use case, we aim at delivering (1) Average, minimum and maximum snow load on a monthly basis for provinces (to be defined), (2) Anomaly from past twenty years situation, and (3) possible trends using EO data. In addition, we provide the (4) yearly (2001-2022 at the moment) information on the first snow-free day (referred to as Melt-off day, MoD) derived from EO-based Fractional snow cover maps. In future, together with information on snow accumulation, it is possible to estimate the (5) length of the snow period for each of these years and calculate possible trends.

Snow course data

An important part of SYKE's hydrological monitoring is the network of appr. 150 active snow courses, where observers make monthly or bimonthly observations on snow depth, snow water equivalent (SWE) and fractional snow cover (FSC). These are processed to give land use weighed mean SWE of the measurement day for each snow course. For the days between the monthly measurements the daily SWE value at the snow course location is estimated using the *Snow Course Model*, which uses snow course observations, weather data from FMI and topography as input. The model is forced to follow the measured SWE values. These daily SWE values at the snow course locations are then further processed to cover the whole Finland in 10 km grid. This gridded daily SWE data is the input for the monthly estimation on max, min and average snow load at County level.

EO-based SWE and snowmelt-off day

EO SWE products used in SYKE are provided by FMI. The data is based on combination of satellite-based microwave radiometer and ground-based snow depth observations. The spatial resolution of daily SWE product is 0.25 degrees and it covers all non-mountainous areas in Northern hemisphere.

Melt-off day (MoD) is calculated from FSC time series with a dedicated algorithm described in [1]. The FSC-estimates are fetched from Copernicus CryoLand portal; the estimates are based on the *SCAmod* algorithm developed at SYKE [2]. Besides the information in the original 500m resolution, melf-off statistics is also provided for 19 Finnish counties. The country level statistics is provided as an average of all 500m MoD-pixels within the polygon defining each county.

The same *SCAmod* algorithm with slightly different parameterization is applied in the provision of average FSC for 3rd division sub-basins. The calculation is implemented to computing platform of the Arctic Space Centre of FMI and is maintained in collaboration of SYKE and FMI. Average FSC-values are provided to SYKE hydrologist as a text file.

Watershed simulation and forecasting system

SWE is very important part of the hydrological model of the Watershed simulation and Forecasting System (WSFS), which is used in flood forecasting. The WSFS simulates areal precipitation, snow water equivalent, snow depth and density, soil moisture, groundwater storage and water levels and discharges in rivers and lakes in Finland including transboundary catchments. The model is calibrated against the water level, discharge observations and snow course observations.

In this project the aim is to compare the WSFS simulated SWE with *Snow Course Model* and EO SWE-product. Traditionally the WSFS is calibrated and updated against the snow course observations, but also the EO SWE could be used in the model calibration and updating in the areas lacking the reliable snow course data. Also EO FSC is used in the model for calibration of the snow-off dates, which improves the model simulations in the areas with coarse or inaccurate discharge observations.

Results

This study presents the results of EO SWE (from FMI) and FSC (from SYKE), snow course model and hydrological model of WSFS. Snow simulation will combine hydrological model, snow course observations and EO-data. The statistical data includes snow load, snow melt-off day and snow water equivalent anomaly maps. Statistics are available for whole Finland and in 19 Finnish counties.

Snow load statistics

The statistics on snow load can be either direct statistical values of the whole dataset such as mean or maximum snow load at County area, or it can be defined so that it better defines the effect of snow load on local conditions, such as days in month when snow load exceeds certain thresholds. Figure 1 presents examples of monthly statistics of snow load in Finland in January and March of 2017.

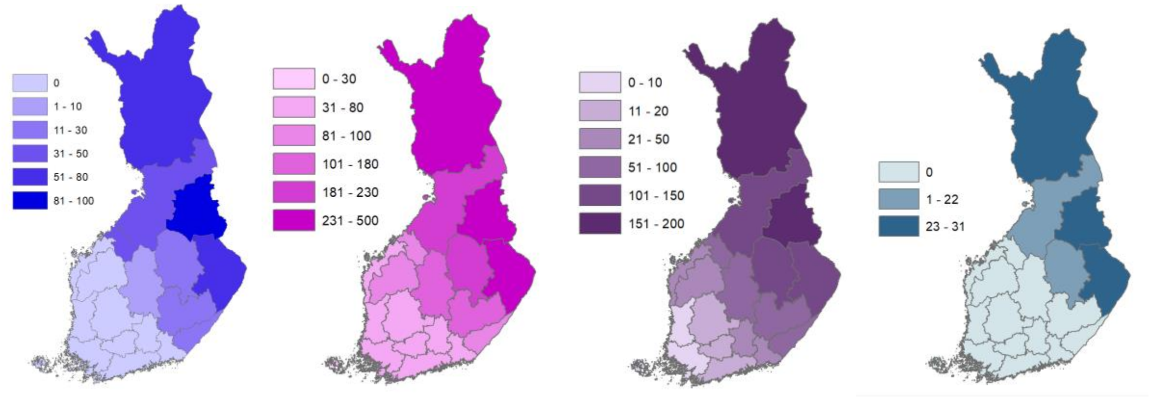


Figure 1. Snow load (kg/m^3) in Finland in March and January 2017. a) Minimum, b) maximum, c) mean snow load at counties in March 2017, and d) number of days with snow load exceeding 100 kg/m3 in counties in January 2017.

Snow melt-off dates

A couple of examples on yearly snow melt-off maps is presented in Figure 2. The country level trend for part of the 19 counties are presented in figure 3

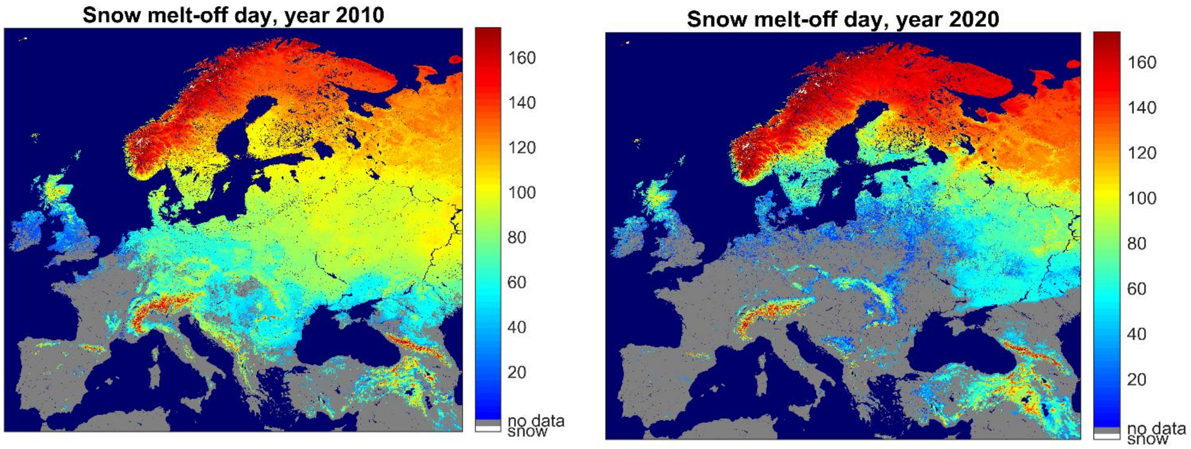


Figure 2. Melt-off days maps for years 2010 and 2020 (years 2001-2022 available)

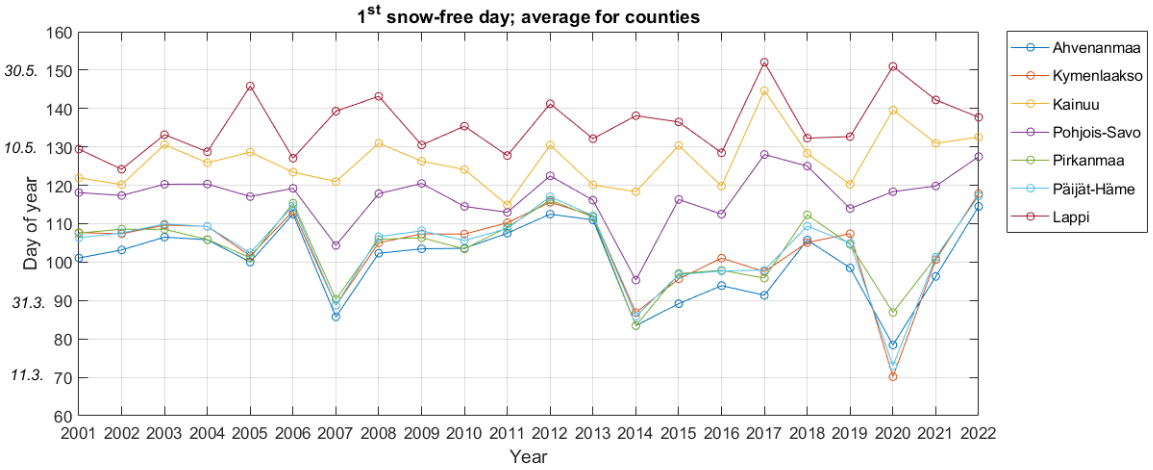


Figure 3. Yearly Melt-off days, averaged for counties

Snow water equivalent anomalies

The mean monthly anomaly maps of SWE are calculated using the simulated SWE of WSFS and EO SWE. The control period used in the anomaly map is the last 20 years period. The resolution of the map is based on the sub-basin division of Finland and the average area of the sub-basins is approximately 50 km^2. In the watershed scale the WSFS anomaly map is well in line with EO SWE in the accumulation phase of the snow season. During the melting period the accuracy of the EO SWE is reduced, when it is compared to WSFS, snow course model and EO SCA data.

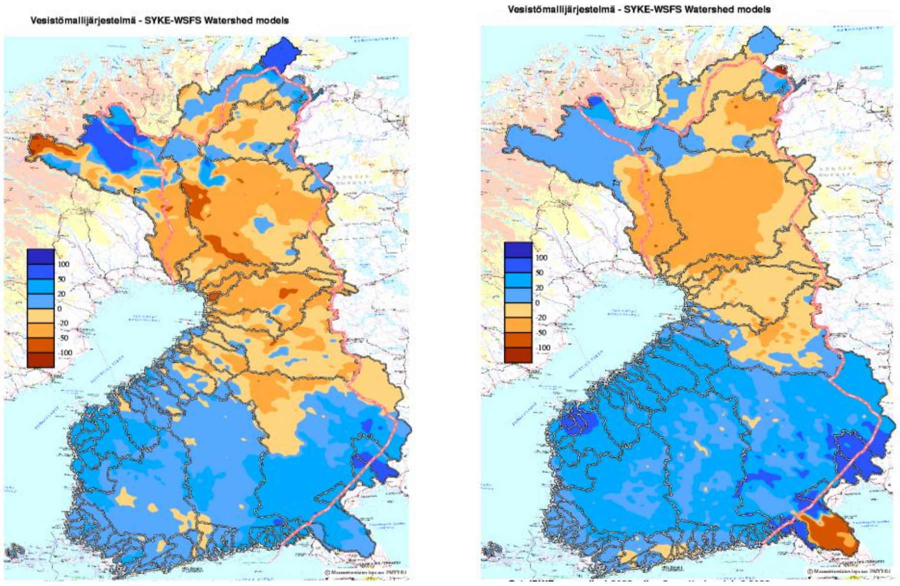

Vesistömallijärjestelmä - SYKE-WSFS Watershed models
100
50
20
0
-20
-50
-100
Vesistömallijärjestelmä - SYKE-WSFS Watershed models
100
50
20
0
-20
-50
-100

Figure 4: Mean montly anomaly maps (mm) of January 2022 for simulated SWE (left) and EO SWE (right)

Conclusions

This study presented new spatial statistical parameters as examples with improved spatial and temporal coverage. Combining data from Earth observations, field measurements and model outputs have proven to be potential for impactful and meaningful parameters for national statistics on seasonal snow cover.

References

[1] Metsämäki, S., Böttcher, K., Pulliainen, J., Luojus, K., Cohen, J., Takala, M., Mattila, O.-P., Schwaizer, G., Derksen, C., & Koponen, S.. The accuracy of snow melt-off day derived from optical and microwave radiometer data — A study for Europe. *Remote Sensing of Environment (2018), 211: 1-12.*

[2] Metsämäki, S., Mattila, O.-P., Pulliainen, J., Niemi, K., Luojus, K., Böttcher, K. . An optical reflectance model-based method for fractional snow cover mapping applicable to continental scale. *Remote Sensing of Environment (2012), 123: 508-521*

Innovative in situ and other Earth Observation platforms for air quality official statistics: the GAUSS project

Keywords: Air quality, low-cost sensors, Earth Observation, IPR reporting

Introduction

A growing convergence is lately experienced between the domains of Earth Observation (EO) and statistics. The 47th meeting of the European Statistical System Committee in 2021 was entitled “Earth observation (EO) for official statistics'' and culminated in the so called Warsaw Memorandum [1], which underlines the increasing importance of exploiting EO data for statistical purposes, while the CROS ESSNet Big Data II project explored this interface through several case studies [2]. The implementing regulation for the 2021 population and housing census [3] calls for the geocoding of the census at a 1km^2 grid, and EO and statistics are considered on par regarding their status as High Value Datasets according to the open data Directive [4]. The SDG framework has offered a prime opportunity for some tangible examples where EO can support National Statistical Institutes (NSIs) [5], and similar efforts are carried forth by the GEO’s EO4SDG initiative [6].

Air quality remains mostly unaddressed perhaps due to the fact that the relevant Indicator (11.6.2) is of Tier 1 classification (see Shaddick et al. [7]). In Europe, the reporting entities are mostly the national air quality management authorities that report to the designated repositories and the European Environment Agency (EEA), according to the Implementing Provisions on Reporting (IPR) framework (Implementing Decision 2011/850/EU). This use case, implemented in collaboration with the Greek NSI (ELSTAT) as part of the ESA funding project GAUSS (https://eo4smartstats.com/), aims at exploring whether exploiting a - wider than the current regulatory National Air Pollution Monitoring Network (NAPMN) - array of EO platforms (i.e. sensors, satellite, model) could extend the scope of existing IPR methods. This is pursued by enhancing the spatial disaggregation of reporting (up to Local Administrative Unit), capturing the intra-urban variability of regulatory pollutants and including smaller cities with no official AQ monitoring.

Methods

The overall approach aims at the statistical product No 114 “ENV- AIRQ_0 Air Quality Directive IPR Report” of the ELSTAT Hellenic Statistical Programme 2020-2022, and the focus is on four pollutants ($PM_{2.5}$, PM_{10}, NO_2 and O_3). While the CAMS Regional Ensemble Reanalysis model will form the basis of the geospatial enhancement, it is the NAPMN of Greece and a nation-wide network of low cost sensors that will be exploited to nudge the model towards more realistic values across the entire domain of interest.

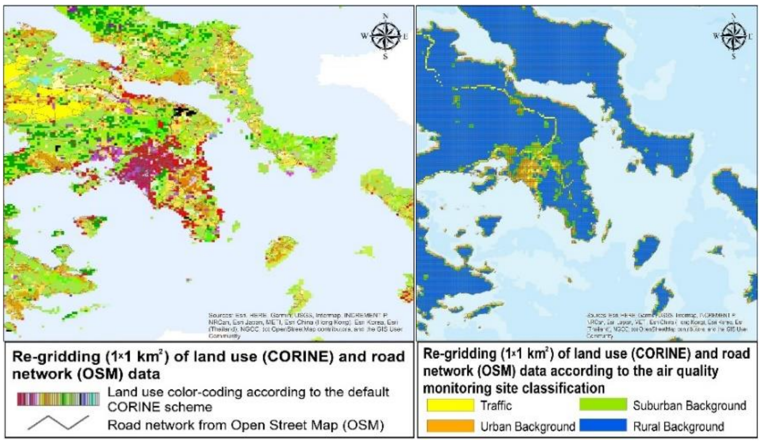
CAMS Regional Ensemble Reanalysis model and calculation grid

Hourly air pollution data from CAMS for the year 2020 (approx. 10km resolution) are attributed to a 1km^2 calculation grid, performing spatial averaging in case of cells overlap. The hourly data resolution was selected since it allows for taking into account the short-term concentration variability, extraordinary modifications in the emissions (e.g. Covid-lockdown), and is also a requirement in IPR reporting. GIS processing is done under a Mollweide projection, while also attempting to align with the Global Human Settlement Layer (GHSL) grid to facilitate future implementations of human exposure (such as the 11.6.2 Indicator) and compatibility with the upcoming census grid.

Re-classification of grid cells through Corine and OpenStreetMap (OSM)

In order to utilise the in situ data, all cells in the fine 1km^2 grid were re-classified in 4 categories, relevant to major monitoring site types of the 2008/50/EC directive, namely traffic, urban background, suburban background and rural background. The driving dataset is Copernicus CORINE LULC 2018 (cell size = 100m) which is aggregated into the aforementioned 1km^2 grid by attributing the dominant by area LULC type. The aggregated CORINE data were rebadged to the 3 background categories depending on their type. Specifically, cells corresponding to “continuous urban fabric” were classified as “urban background”, “discontinuous urban fabric” was linked to the “suburban background category” and the remaining non-urban types were associated with “rural background”. “Green urban areas, sports/leisure facilities and airports” were either classified as urban or suburban background depending on the prevalent type in the neighbouring cells. Cells containing monitoring stations are classified outright according to the station type.

To assign “traffic characterization” to 1km^2 cells expected to be more influenced by vehicular emissions, GIS data from the OpenStreetMap (OSM) database were used. The area covered by roads in a cell was calculated based on the length of the roads and assumptions regarding their width. The area (as a ratio of cell area) for all cells containing monitoring stations of the NAPMN was examined, in order to identify a threshold ratio threshold above which “traffic stations” dominated. It was found that for a major road coverage more than 10% in cells, >80% of station-containing cells were accurately classified as traffic. The re-classification of 1km^2 cells (displayed in Figure 1 for the region around the greater area of Athens) is deemed important since adjustment factors of CAMS data based on in-situ data will be calculated separately for the four site categories.



Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

Re-gridding (1×1 km^2) of land use (CORINE) and road network (OSM) data

Land use color-coding according to the default CORINE scheme
Road network from Open Street Map (OSM)

Re-gridding (1×1 km^2) of land use (CORINE) and road network (OSM) data according to the air quality monitoring site classification

Figure 44. Regridding of LULC and road network GIS data (left) according to the air quality monitoring site classifications (right)

Compilation of in situ database

The major part of in situ data derive from the NAPMN and refer to NO_2, O_3, $PM_{2.5}$ and PM_{10}. Over 40 NAPMN sites spread over Greece are included. The second data source, exclusive for $PM_{2.5}$ data, is the air quality monitoring network of the National Research Infrastructure PANACEA that numbers over 100 sites with the emphasis being on urban areas [8]. The site type characterization is a critical parameter for the methodology as it showcases the need for densification of monitoring networks, and also provides a good example for the importance of emerging environmental sensing technologies. In the case of Greece, the spatial coverage of $PM_{2.5}$ monitoring by the NAPMN is rather sparse, and the sensor network, provided a five-fold increase of monitoring locations covering also urban areas otherwise unrepresented.

Satellite data retrieval and inference of surface NO_2 observations

Through TROPOMI (Sentinel-5 Precursor (S5p) mission), a level-2 product is available that provides daily measurements of the NO_2 vertical tropospheric column (one overpass at approx. 11:00 UTC). The approach entails inferring ground-level NO_2 concentrations at $1km^2$ spatial resolution from this product with the help of NO_2 vertical data from the CAMS Ensemble Model, and data from ECMWF's ERA5 weather reanalysis model (namely the Boundary Layer dataset and the Pressure and the Geopotential datasets). The S5p L2 products are clipped and regridded to a $1km^2$ grid, using an area-weighted oversampling approach. A vertical spline interpolation of CAMS NO_2 data from the 8 sampling altitudes to the TM5-MP vertical levels is performed, i.e., the levels of the TROPOMI product averaging kernel. To obtain the CAMS model simulated satellite NO_2 column, the CAMS model partial column profile is multiplied with the tropospheric averaging kernel [9]. The surface concentration is finally calculated [10] using a simulated surface-to-column conversion factor from a chemical transport model that also corrects for biases related to vertical mixing assumptions.

Results

The final scope of the use case is to aggregate the $1km^2$ adjusted hourly values to the officially reporting zones in an equivalent manner to current IPR reporting, and check its validity while providing the high geospatial resolution required. In a further step, concentrations and statistics relevant to the air quality standards (e.g. limit values, exceedances) will be also compiled at the municipality level (LAU). The reporting zones can be seen in Figure 2 with un-nudged CAMS values.

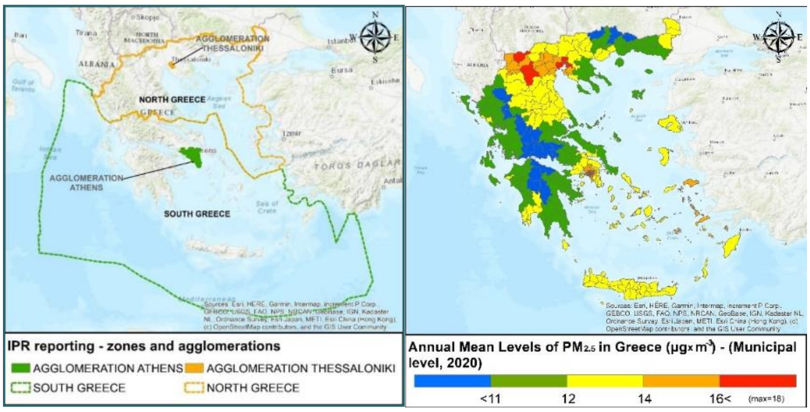

AGGLOMERATION THESSALONIKI
NORTH GREECE
AGGLOMERATION ATHENS
SOUTH GREECE
IPR reporting - zones and agglomerations
AGGLOMERATION ATHENS
AGGLOMERATION THESSALONIKI
SOUTH GREECE
NORTH GREECE
Annual Mean Levels of PM2.5 in Greece (μgx m³) - (Municipal level, 2020)
<11
12
14
16<
(max=18)

Figure 2. IPR reporting zones (left) and finer spatial disaggregation (LAU) that will be implemented in this study, here for CAMS original values (right)

The critical step in the integration of the CAMS, GIS and in-situ monitoring data is the calculation of adjustment factors (ratios). First, zones of influence are calculated around station-containing cells, and measured concentrations are propagated. The area of each zone of influence will depend upon the cell type (i.e. 2008/50/EC directive site representativeness , urban - few km^2; suburban - tens of km^2; and rural - hundreds of km^2. Initial values of 5km^2, 25km^2 and 200 km^2 were set for further validation. Traffic cells were not assigned a zone of influence as they are representative of very limited areas. Factors are calculated only for cells containing monitoring sites, and adjustments are applied by site type over the whole domain. The dependence of the correction factors upon region and season will be also explored. For cells not containing a station or not within a zone of influence, their concentrations is determined based on the cell type classification In this way, hourly concentration data will be assigned to all cells of the extended domain.

The inferred NO_2 surface concentration values from S5P will firstly undergo an evaluation against the NAPMN measurements. These values will not be utilized for nudging the CAMS model identification but instead they will help identify case studies where the nudged model cannot inherently (due to agnostic emission inventories) perform well, as is the case of Covid-lockdown or of a forest fire, and for further checks on the efficient capturing of country-wide spatial distribution of NO_2.

Conclusions

The expected outcome of this ongoing work within the frame of the GAUSS project, presents an alternative approach to provide air quality statistics that are meaningful in spatial scales smaller than the ones typically used for official statistics. Such results can motivate or feed supplementary solutions for spatial characterization of air quality and population exposure (including synergies with spatial, atmospheric and exposure modelling). While the study is focused in Greece, the datasets and methodologies are designed from the get-go to be completely scalable across Europe once the added value to current modus operandi is showcased to the competent entities, with novel insights on needed emissions accounting.

References

[112] Warsaw Memorandum: https://dgins2021.stat.gov.pl/warsaw-memorandum, access:12 October 2022

[113] Mleczko et al., WPH Milestones and deliverables: https://ec.europa.eu/eurostat/cros/system/files/wph_h3_final_technical_report_24_03_2021.pdf, last access: 12 October 2022, 2021

[114] Legislation - Population and demography - Eurostat: https://ec.europa.eu/eurostat/web/population-demography/population-housing-censuses/legislation, last access: 12 October 2022

[115] European legislation on open data | Shaping Europe's digital future: https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data, last access: 12 October 2022

[116] IAEG-SDGs Indicators |https://ggim.un.org/meetings/GGIM-committee/11th-Session/documents/The_Geospatial_SDGs_Roadmap_WGGI_IAEG_SDGs_20210804.pdf, last access: 12 October 202

[117] Kavvada, A., Metternicht, G., Kerblat, F., Mudau, N., Haldorson, M., Laldaparsad, S., Friedl, L., Held, A., and Chuvieco, E.: Towards delivering on the sustainable development goals using earth observations, Remote Sensing of Environment, 247, 111930, https://doi.org/10.1016/j.rse.2020.111930, 2020

[118] Shaddick, G., Thomas, M. L., Jobling, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H., Cohen, A., Van Dingenen, R., Dora, C., Gumy, S., Liu, Y., Martin, R., Waller, L. A., West, J., Zidek, J. V., and Prüss-Ustün, A.: Data Integration Model for Air Quality: A Hierarchical Approach to the Global Estimation of Exposures to Ambient Air Pollution, https://doi.org/10.48550/arXiv.1609.00141, 26 September 2016

[119] Stavroulas, I., Bougiatioti, A., Grivas, G., Paraskevopoulou, D., Tsagkaraki, M., Zarmpas, P., Liakakou, E., Gerasopoulos, E., and Mihalopoulos, N.: Sources and processes that control the submicron organic aerosol composition in an urban Mediterranean environment (Athens): a high temporal-resolution chemical composition measurement study, Atmospheric Chemistry and Physics, 19, 901–919, https://doi.org/https://doi.org/10.5194/acp-19-901-2019, 2019

[120] Eskes, H., van Geffen, J., Boersma, F., Eichmann, K.-U., Apituley, A., Pedergnana, M., Sneep, M., Veefkind, J. P., and Loyola, D.:675 Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Nitrogendioxide, Tech. Rep. S5P-KNMI-L2- 0021-MA, Koninklijk Nederlands Meteorologisch Instituut (KNMI), https://sentinels.copernicus.eu/documents/247904/2474726/Sentinel-5P-Level-2-Product-User-Manual-Nitrogen-Dioxide, CI-7570-PUM, issue 4.0.1, 6 July 2021

[121] Cooper, Matthew J., et al. "Inferring ground-level nitrogen dioxide concentrations at fine spatial resolution applied to the TROPOMI satellite instrument." Environmental Research Letters 15.10 (2020): 104013

Web cartography for gridded statistics – the Gridviz library

Keywords: Geospatial statistics, grids, web cartography, visualisation.

Introduction

More and more National Statistical Institutes (NSIs) produce gridded statistics. Such statistics have well-known advantages compared to statistics geographically referenced on irregular statistical units such as administrative units: They allow depicting more objectively the spatial variation of statistical variables by removing the bias introduced by the irregularity of size and shape of other statistical units. Gridded statistics also represent an opportunity to produce statistics at more finer geographical resolutions based on GIS and earth observation data and technologies. Those "geospatial statistics" are expected to become more and more available.

The potential of gridded statistics is however not fully exploited – existing and future gridded statistics such as the European census 2021 gridded statistics require specific analysis tools. This abstract presents the Gridviz tool [1] proposed to support the exploration and analysis of gridded statistics on the web with improved cartographic techniques.

Web cartography for gridded statistics

Most existing gridded statistical datasets produced by NSIs are available on their website as raw tabular data, based on standard encoding formats. Overviews of the content are sometimes provided through web mapping applications, which usually show some selected statistical variables as coloured squares, but unfortunately do not allow an efficient and in-depth exploration of these datasets and their dimensions. The two following opportunities could be considered:

Methodological opportunity - Cartography

Cartography is the science of representing geographical information – it provides a set of formal methodologies to represent complex information related to the geographical space in an efficient (and unbiased) manner. The role of cartography is not only to show information in an appealing and colourful manner, but to support the understanding, exploration and analysis of complex information with efficient visual mechanisms.

Existing cartographic methodologies show the possibilities offered to support the analysis of complex spatial information [2, 3]. Jacques Bertin proposed in [2] an analysis of the properties of the six visual variables (figure 1) used to represent information graphically. The "visual grammar" he proposed should be further adopted in statistical cartography to go beyond traditional choropleth maps [4, 5], especially for gridded multi-variate statistics.

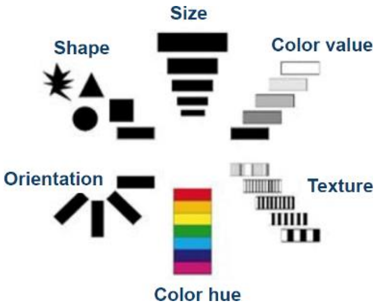


Figure 45. The six visual variables according to [2]

Technical opportunity – Web technologies

Cartographic technologies, especially web-mapping technologies, are evolving quickly and offer new possibilities. The most direct way to show gridded statistics on the web is to use traditional image-based web mapping tools, which allow showing the statistics as static choropleth images only. A more advanced emerging approach consists in transferring raw data to the user and display it on-the-fly as proposed in [6]. This allow more interactivity and enable efficient exploration of the data at different scales, along different dimensions, as well as multi-variate analyses.

This approach relies on the utilisation of specific data formats (for multi-resolution tiled data), and specific rendering technologies such as HTML5 Canvas and WebGL which have now become widespread web standards.

The Gridviz library

To take benefits of these opportunities, Eurostat started developing the Gridviz [1] library, whose version 2 was released in September 2022. Gridviz is an open source JavaScript library specially designed for the visualisation of gridded statistics on the web. It can be used as a base software component to develop web applications showing any type of gridded statistical dataset on a website. It allows exploring these datasets in a web interface by zooming and panning, as common web mapping interfaces. Gridded statistics can be shown using various pre-defined and customisable cartographic styles such as:

- Shape/Color/Size Style: This style shows grid cells with changeable shape, colour and size depending on different statistical dimensions. These visual variables can remain the same for all cells, or change together in combination allowing showing multi-variate information.

- Composition style: This style shows the composition of each grid cell as different types of charts: Pie chart, bar chart, radar, ring, age pyramid, etc. A colour is assigned to each category to show. The size of the chart can also be set based on a total value.

- Segment style: This style shows each cell as a segment with changeable orientation, colour, length and width. It is suitable to show multivariate information. Orientation is particularly suitable to show a variation compared to a previous period.

- Side style: This style shows differences between adjacent cells as a side segment with changeable colour and width. It is particularly suitable to depict strong spatial discontinuities in the data.

- Pillar style: This style shows grid cells in a perspective view, as vertical pillars. The height and colour of the pillar can be set based on different statistical variables.

- Kernel smoothing style: This allows applying a Gaussian kernel smoothing to an input variable to depict main variation trends over space.

- Other styles are available and documented on the library website [1]. An overview of some of these styles is shown on figure 2.

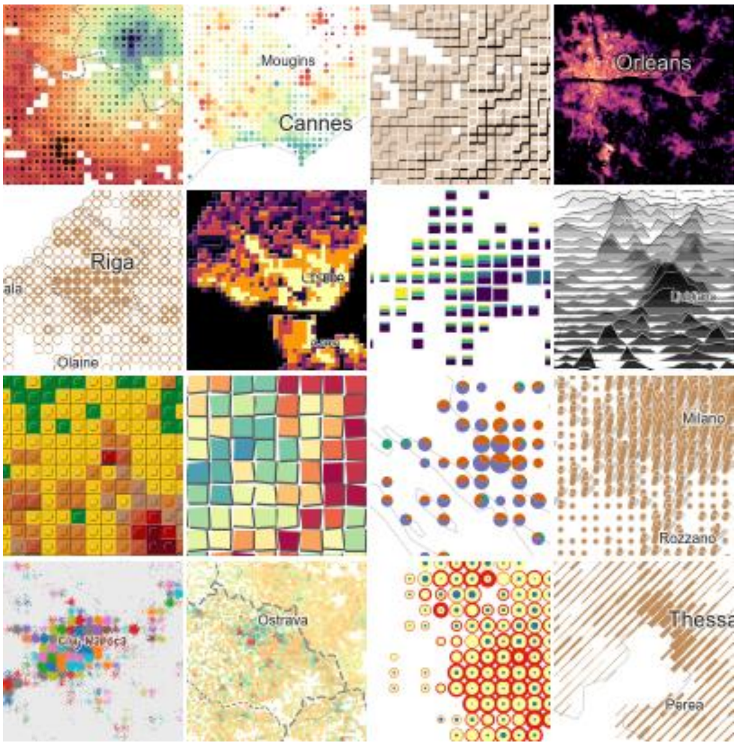


Figure 46. Overview of some Gridviz cartographic style for gridded data

Input data are usual tabular data, as CSV format – possibilities exist to show multi-resolution data depending on the zoom level. Input data can also be decomposed into “tiles” for better efficiency especially for large datasets. Contextual information can be provided such as a map background and place names (such as city names).

Examples

Advanced Gridviz utilisation examples have been developed from existing gridded statistical datasets and can be seen online from the library website [1]:

- 1km grid of Europe: https://eurostat.github.io/gridviz/examples/EUR.html

- 1km grid of Croatia: https://eurostat.github.io/gridviz/examples/HR.html

- 200m grid of France: https://eurostat.github.io/gridviz/examples/FR.html

These first examples illustrate the possibilities offered by Gridviz tool to expose the richness of these gridded statistics with advanced multi-scale cartographic techniques.

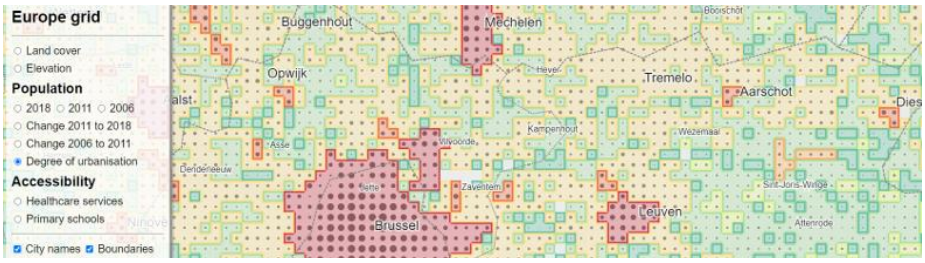


Figure 47. Europe 1 km grid example

Conclusions and perspectives

Gridviz tool improves the exploration of gridded statistical datasets on the web with cartographic techniques beyond the traditional pixel colouring style. It helps promoting the utilisation of gridded statistics and may encourage a further development of this type of statistics, especially based on GIS data and technologies.

The high genericity level of the tool and its source code openness enables future reuses and evolutions: Contributors are welcome to improve and enrich the tool on the current collaborative development platform on Github (https://github.com/eurostat/gridviz). Possible contributions could be on:

- New cartographic styles, for more specific types of gridded statistics. New perspective styles, multi-variate styles (value-by-alpha, Chernoff faces, etc.). Some new styles could be developed for temporal data.

- New input data formats (Cloud Optimized GeoTIFF, Arrow, protobuf, etc.) for improved efficiency.

NSIs willing to improve the visibility of their existing gridded datasets are welcome to reuse Gridviz on their website, either for a simple overview or for a more advanced exploration tool. Gridviz plans to be used in several Eurostat digital publications, and for the dissemination of the incoming Census 2021 data.

References

[122] Gridviz visualisation tool for gridded statistics. https://github.com/eurostat/gridviz

[123] Bertin, J. (1968). Semiology of Graphics: Diagrams, Networks, Maps, Esri Press books, ISBN: 9781589482616.

[124] Field, K. (2018). Cartography: a compendium of design thinking for mapmakers, International Cartographic Association, Esri Press books, ISBN: 9781589484399.

[125] Pieniazek, M., Zych, M. (2020). Statistical maps - Data visualisation methods, Statistical research papers, Statistics Poland.

[126] Gaffuri, J. (2021). Statistical cartography, EMOS 2021. https://emos2020events.ec.unipi.it/statistical-cartography/

[127] Gaffuri, J. (2012). Toward Web Mapping with Vector Data, International Conference on Geographic Information Science GIScience 2012, LNISA, volume 7478, pp 87-101, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33024-7_7

The Polish Use Case in Project GAUSS – extent and quality of green areas Smart Statistics based on EO data

Keywords: EO data, Smart Statistics, Green space, Green cities

Introduction

The Polish Use Case Study in Project GAUSS (Generating Advanced Usage of Earth Observation for Smart Statistics) [1] submitted in response to ESA AO/1-10632/21/I-BG "Earth Observation for Smart Statistics" [2] employs the application of Earth Observation and in-situ data to assess the extent and quality of green areas for deriving information on natural capital statistics and human well-being conditions for the Polish chief government executive agency (Statistics Poland) [3].

Statistics Poland is interested in improving the quality and timeliness of regional well-being statistics. According to OECD, Poland is within the bottom two deciles in terms of well-being indicators related to environmental quality [4]. In particular, Statistics Poland is interested in gathering information on the extent and quality of vegetation at a commune level (LAU). At the national level, this information is currently inhomogeneous, outdated and often with gaps, because it is produced individually by local governments, and with large intervals. This makes it difficult to combine the quality-of-life indicators such as health status with information about the impact of green areas.

Methods

The data from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-5 climate reanalyses [7] and Sentinel-2 high-resolution satellite data from the Copernicus Land Monitoring Service [5] accessed by WEkEO EU Copernicus Data and Information Access Services (DIAS) [6] were used to generate smart statistics for green areas. In addition, the Dynamic World product as a result of collaboration between Google and the World Resources Institute [9] and also the Polish Database of Topographic Objects (BDOT10k) [8] were used to determine different types of green areas, especially in urban agglomerations. The annual statistics generated on green areas are also supported by the raster preview for the entire territory of Poland. Thanks to this, the End-User can monitor the green areas created by the JupyterLab [10] system and modify the procedure of delimiting the areas at every stage of the data processing.

Results

Two main outputs are planned to be prepared for administrative units (LAU) in Poland. The first one is the information on the extent of green areas within LAU obtained based on high-resolution satellite data (Sentinel-2), Copernicus Land Monitoring Service and Dynamic World and validated by the Polish Database of Topographic Objects (BDOT10k). The second one is the information about the quality of green areas (vegetation status) at the LAU level obtained from vegetation indices based on optical sensors, derived from high-resolution satellite data (Sentinel-2), together with additional climate indicators. The project's most innovative result is the opportunity for the Statistics Poland to observe green areas by using local resources and the interface with the Google Earth Engine application [11]. Thanks to this, Statistics Poland End-User will have full control on the generation of statistics from individual municipalities with 10 m resolution.

Conclusions

The proposed indicators address the issues that are of a great interest for Statistics Poland and for other statistical offices, because they are crucial to fulfil the organization of the Public Services Monitoring System [12]. The purpose of this System is to provide local government entities, businesses, and the public with information necessary to comprehensive evaluation services provided at local level. The actual data of green areas at the LAU level, extents and vegetation conditions are crucial for the Smart Statistics and will be used when only they are ready.

References

[1] https://eo4smartstats.com/index.php/author/admin3586. Home. GAUSS. Published January 26, 2022. Accessed October 16, 2022. https://eo4smartstats.com/

[2] Earth Observation for smart statistics - ESA Commercialisation Gateway. ESA Commercialisation Gateway. Published April 7, 2021. Accessed October 16, 2022. https://commercialisation.esa.int/opportunities/earth-observation-for-smart-statistics/

[3] GUS. Statistics Poland. Stat.gov.pl. Published 2020. Accessed October 16, 2022. https://stat.gov.pl/en/

[4] OECD Environmental Review of Poland - OECD. Oecd.org. Published 2022. Accessed October 16, 2022. https://www.oecd.org/environment/country-reviews/oecdenvironmentalreviewofpoland.htm

[5] Copernicus Land Monitoring Service. Copernicus.eu. Published 2018. Accessed October 16, 2022. https://land.copernicus.eu/

[6] Copernicus and Sentinel data at your fingertips. Wekeo.eu. Published 2022. Accessed October 16, 2022. https://www.wekeo.eu/

[7] ECMWF. ECMWF. Published 2022. Accessed October 16, 2022. https://www.ecmwf.int/

[8] Geoportal.gov.pl. Published 2022. Accessed October 16, 2022. https://bdot10k.geoportal.gov.pl/

[9] Dynamic World - 10m global land cover dataset in Google Earth Engine. Dynamicworld.app. Published 2021. Accessed October 16, 2022. https://dynamicworld.app/

[10] Project Jupyter. Jupyter.org. Published 2022. Accessed October 16, 2022. https://jupyter.org/

[11] Google Earth Engine. Google.com. Published 2022. Accessed October 16, 2022. https://earthengine.google.com/

[12] SMUP. Smup.gov.pl. Published 2022. Accessed October 16, 2022. https://smup.gov.pl/

Time series analysis (MANS3A.3)

Session Chair: **Jean Palate** *(National Bank of Belgium)*

Online job ads' time series: A novel approach
Gabriele Marconi *(Sogeti)*, Alexandros Bitoulas *(Sogeti)*, Anca Maria Kiss *(Sogeti), Fernando Reis (Eurostat)*

Fixing the model for the seasonal component: a new revision policy
Maria Novás Filgueira *(National Statistical Institute-INE)*, Carlos Sáez Calvo *(National Statistical Institute-INE)*, David Salgado *(National Statistical Institute-INE)*, Luis Sanguiao-Sande *(National Statistical Institute-INE)*

The complexity of the turning points detection in 2022
Gian-Luigi Mazzi *(Senior Consultant),* Rosa Ruggeri Cannata *(Eurostat)*, Piotr Ronkowski *(Eurostat)*

Testing for cointegration in STAR models: A simulation-based study
Gülşah Sedefoğlu *(Istanbul Commerce University)*

JDemetra+ 3.0: New (R) tools for (high-frequency) time series analysis
Anna Smyk *(National Institute of Statistics–INSEE)*

Online job ads’ time series: A novel approach

Keywords: Big data, online job ads, scraping, time series, chaining, survival analysis

Introduction

Aim and data

We present a new methodology to calculate time series for the total number of online job ads (OJAs) found in Europe between 2019 and 2022, broken down by country and occupation (ISCO 2-digit categories). To achieve this, we discuss the general problems with the time series of data scraped irregularly and from multiple sources; we describe the new methodology; and we compare the main results obtained through the application of this methodology to results obtained with other methods.

The dataset used in this paper is the Web Intelligence Hub's Online Job Advertisement (OJA) database, developed by Cedefop and Eurostat. This dataset covers over 100 million ads posted in EU countries and the UK since July 2018, and collected from several hundred web sources including job search engines and public employment services’ websites. We focus our analysis to the period January 2019 – June 2022, for which the data collection is considered to have been more stable [1].

Problems with OJA time series

OJA data ingestion happens through scraping or through API download and its regularity is less than ideal. Scrapers can fail for reasons like changes in website structure or temporary overload; APIs may change structure or stop updating, leading to gaps in the data ingestion process. In addition, the market for online ads changes, leading to variations in the list of data sources from which data needs to be collected to maintain coverage of the OJA population. This leads to the following three problems.

Missing data. If a delay occurs between two consecutive ingestion dates, some ads may not be collected, for example because the advertised position is quickly filled and the ad removed. This has two implications: first, the total number of ads is underestimated. Second, if the ads are not missing at random, the distribution of ads across different categories (e.g. occupations) is altered.

0-to-N peaks. Data ingestion delays induce an irregular data pattern on the affected source, with 0 ads recorded for each day without data ingestion, and a potentially large number (N) of ads collected in a single day when ingestion is restored. In the most extreme case, 8 ads were collected from a source on 5 October 2018, just before a long delay in the data ingestion (with 0 records in each date), which ended on 20 February 2020 (869 ads found). This increases the noise of the time series.

Incomparable source sets. Data from some sources are not collected at all before or after a certain date, leading to a changing source population over time. This limits the time comparability of OJA data, especially over the medium-long period.

Three approaches to OJA time series analysis

Raw time series are produced by the simplest method: summing up the number of ads collected each day (possibly by country or other variable). The results obtained with this method need to be interpreted cautiously, given the strong limitations [1].

Stable-source time series are produced by summing up only ads collected from “stable” sources, meaning that their number of ads has a relatively low variability over time [2]. This method can lead to a large data loss, because not all sources are “stable” from the beginning to the end of a data collection. In the case of the OJA dataset, this is partly addressed by assessing source stability over shorter, 15-month time periods. This still results in losing 40% of the data, in addition to reduce comparability over time (because the same source may be included for some periods and not for others).

Using probability and statistical models it is possible to deal with missing data and generate reliable time series. We apply tools from survival analysis [3] and chaining [4] to build time series accounting for missing data, 0-to-N peaks and incomparable source sets, while avoiding the large data loss characterizing the stable-source method.

Methods

Applying probability models to address missing data and 0-to-N peaks

We address the above-mentioned problem of missing data through inverse probability-of-censoring weights (IPW) [3]. In short, we frame ads as units that appear and disappear from the web over time (Figure 48), and weight them based on the probability that they were missed by the data collection, estimated through a survival analysis model. The main challenge is that while the posting and removal intervals are known, the exact dates are not. This is called interval censoring [5], and it substantially complicates estimation. We found no ready-made software packages for the OJA case of double interval censoring. Therefore, we programmed a maximum-likelihood solution in R for a constant-hazard-rate survival model, which we apply by data source and occupation.

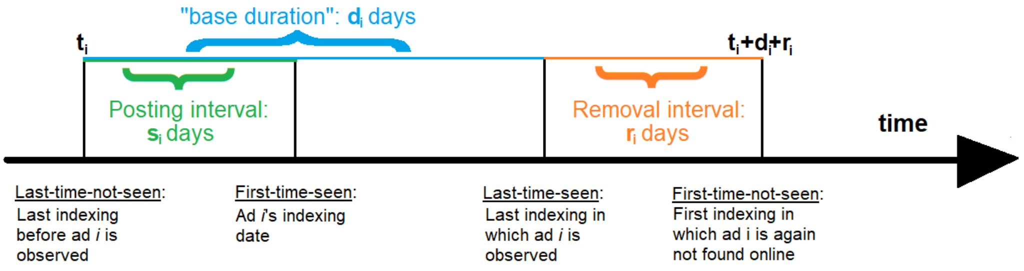

"base duration": d_i days
t_i
t_i+d_i+r_i
Posting interval: s_i days
Removal interval: r_i days
time
Last-time-not-seen: Last indexing before ad *i* is observed
First-time-seen: Ad *i*'s indexing date
Last-time-seen: Last indexing in which ad *i* is observed
First-time-not-seen: First indexing in which ad i is again not found online

Figure 48 The time structure of web scraping

Probability models can also address the problem of 0-to-N peaks, by allocating the N ads collected on the ingestion date across all the days that passed without ingestion. For the sake of simplicity, we assume that ads are equally likely to appear on any day of the posting interval, so that the best estimate of the number of ads found in each day is N (weighted by its IPW) divided by the number of days.

Applying chaining to solve the problem of changing source sets

The method of chaining is widely used in statistics, particularly in the domain of price indexes. Chaining is designed to aggregate multiple time series (e.g. of product prices) into a single one, even in the context of a "changing universe of products" [4]. This makes this method suitable to OJA data, with data sources replacing products and the number of ads from each source replacing the price of each product. The rate of change is derived as the ratio between the sum of ads across sources in two points in time. This is a Dutot index, an elementary index used for chaining with acceptable basic properties [4]. Since the absolute value of the index has no meaning in itself, we calibrate it so that the estimated number of ads for 2022Q2 matches the actual number found in that quarter.

Chaining implies that a long time series is derived through the aggregation of rates of change calculated for shorter periods – in our case, daily. The daily rates of change are calculated based on the sum of ads collected for two consecutive days from a fixed set of sources. Fixing the set of sources ensures data comparability for each calculated rate. A single regularity criterion determines which sources are selected: that data has been collected from each source at least twice in the proximity of the date (i.e., the month surrounding it) for which the rate is computed. Importantly, the set of sources involved in the calculation of two different daily rates does not need to be the same, in the same way that the set of products involved in the calculation of a price index changes over time.

An important advantage of relying on the established framework of chaining is that we can draw from the underlying knowledge and experience. To deal with OJA's interval date censoring, we calculate each source ads as a 14- and 30-day moving average, mirroring the recommendation to use product-level moving averages in case of uncertain transaction times [4]. In addition, we draw on recommendations to deal with outliers, i.e. products with an oversized impact on the price index [4]. We identify outliers as sources with an above-threshold impact on a country's daily rate of change, and bound their values to match this threshold. Given that the interquartile range of the source impact's distribution is quite narrow, we fix this threshold at 2000 or 10000 interquartile ranges (the median is not meaningfully different from 0).

Results

We compare aggregate EU + UK trends in the number of ads based on raw time series, stable-source time series, and three different specifications for chained series (Table 10). The first important result is that chaining allows producing time series with a very limited loss of data. In addition, important similarities are observed between the chained series of the 14- and 30-day moving averages (which also use different thresholds for outliers). This is promising in terms of identifying a stable specification of the chaining model. Only the "Covid drop" is substantially different across the two series, but this is probably because the 30-day moving average spreads the January 2020 peak into February (Figure 49), resulting in a lower estimate for the following decrease.

Table 10. Time series methods: summary description and aggregate EU+UK results

Time series method	*Description*	*Data loss*	*Coefficient of variation*	*Covid drop*	*Overall increase*
Chaining (30-day moving	Based on the 30-days moving averages of the number of ads found in each data source. Only considers outliers with impact over plus/minus	7%	0.26	-50%	1.49

average)	10000 times the interquartile range (i.e. ± 14%). This does not impact EU + UK estimates, because such extreme impacts are observed only in national series				
Chaining (14-day moving average)	Based on the 14-days moving averages, considers outliers with impact over plus/minus 2000 times the interquartile range (i.e. ± 5%)	5%	0.24	-64%	1.59
Chaining (gross)	Chained series without moving averages and outlier treatment	3%	0.56	-79%	7.15
Raw	Time series generated without a specific data treatment, based on the data collection date	0%	0.29	-66%	1.31
Stable-source	Time series based on discarding ads that do not come from sources that are not "stable" for a country and 15-months period	40%	0.39	-67%	2.18

Note: "Data loss" is the proportion of ads that needs to be discarded to produce the estimate. The "coefficient of variation" is the standard deviation of the number of ads across months, divided by the average. The "Covid drop" is the drop of ads between January 2020 and April 2020. The "overall increase" is the factor by which the aggregated quarterly number of ads increased between the beginning and the end of the time period (2019Q1 to 2022Q2).

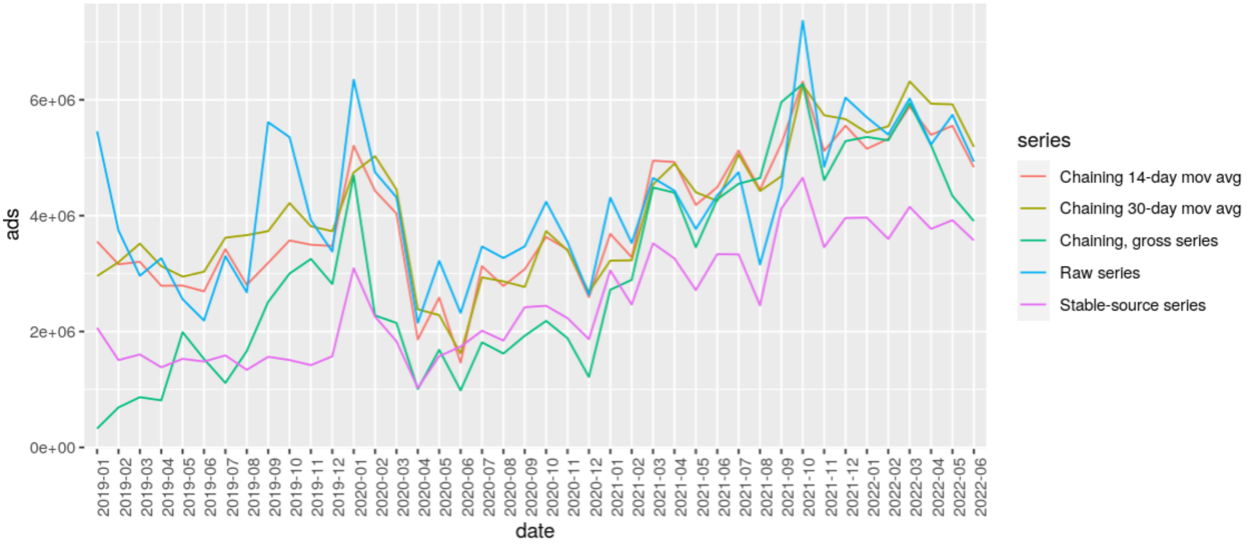


Figure 49 Evolution of the total number of ads in the EU + UK (January 2019-June 2022), by statistical method

Conclusions

Our next steps are as follows. Firstly, to analyse time series at lower levels of aggregations, i.e. by country, occupation and even for some country-occupation combinations. Secondly, we already started applying machine learning models (Ridge and Boosted GLM regression, bagged CART and MARS models) to test the predicting power of each time series method against labour market developments (in particular job vacancies and new jobs). Finally, while the methodology presented in this paper applies to the flow of published OJAs over time, we would like to extend it to the stock of open OJAs.

References (only essential ones to meet the 4-page requirement)

[128] A. Ascheri, A. Kiss-Nagy, G. Marconi, M. Meszaros, R. Paulino, and F. Reis, Competition in urban hiring markets: evidence from online job advertisements, Eurostat Statistical Working Papers, 2021

[129] E. Colombo, M. Fontana, A. Gatti, F. Mercorio, M. Pelucchi, A. Scrivanti and A. Vaccarino, Real-time Labour Market Information on Skill Requirements: Setting up the EU System for Online Vacancy Analysis, Chapter 9 (Unpublished version), 2021

[130] S. Seaman and I. White, Review of inverse probability weighting for dealing with missing data, Statistical Methods in Medical Research, 2013, 278-295

[131] IMF, ILO, Eurostat, OECD and WB. Consumer Price Index Manual. Concepts and Methods 2020, International Monetary Fund, 2020

[132] Z. Zhang and J. Sun, Interval censoring, Statistical methods in medical research, 2010, 53–70

Fixing the model for the seasonal component: INE's revision policy

Keywords: seasonal adjustment, revision policy, partial concurrent adjustment, ARIMA parameters, canonical decomposition

Introduction

One of the multiple decisions that statisticians must face on the release of seasonal and calendar adjusted series, is the revision policy when new data are available. The National Statistics Institute of Spain (INE) used to apply the policy of “Partial Concurrent Adjustment: ARIMA Parameters” in JDemetra+, but huge revisions from the beginning of the series were occasionally observed. Analysing this issue deeply, we realized that revisions were due to two main reasons: model changes because of lack of admissible decomposition, and especially, changes in the autoregressive roots assignment.

We present the revision policy applied at the INE, which may be considered a compromise between the “Partial Concurrent Adjustment: ARIMA Parameters” policy and the “Partial Concurrent Adjustment: Fixed Model”, both implemented in JDemetra+. This policy avoids model changes by: (i) fixing the last estimated model with admissible decomposition when a model change is triggered and (ii) adjusting root assignment parameters to make sure autoregressive roots remain in the same component. In doing so, we improve the estimation of the model parameters with the new data, while avoiding big revisions.

Model-based seasonal adjustment

There are two widely used methodologies for seasonal adjustment: the model-based approach and the fixed filters approach. The methods discussed here refer only to the first method.

The model-based approach to seasonal adjustment has two steps. In the first step, a regARIMA model that best fit the series is adjusted. In the second step, the ARIMA model is decomposed in several ARIMA models, one for each of the unobserved components (trend-cycle, seasonal, transitory and irregular components). This decomposition is determined by assigning each root of the autoregressive polynomial of the ARIMA model for the series to each component. However, it is not always possible to obtain an admissible decomposition. If there is some admissible decomposition, in order to identify the components, as much noise as possible will be extracted from each component, except for the irregular one where all the extracted noise is added. The decomposition so achieved is unique and it is called canonical. Each of these unobserved components are then estimated using the Wiener-Kolmogorov filter, which is the minimum mean squared error estimator. In practice, this filter is approximated by a finite 2-sided filter with a window of width between 3 and 5 years. Assuming the length of the series is much bigger than the width of the window, the estimator can be considered final for the central observations of the series. However, at the beginning and the end of the series the filter cannot be realized, and we must use the preliminary estimator, which is obtained by applying the WK filter to the series extended with forecasts and backcasts. Since forecasts and backcasts are actually a function of the observed time series, we are using a non-symmetric filter instead of the optimal WK filter. As new data arrives, it replaces the forecasts, and the filter gives new estimates of the unobserved components. These new estimates are called revisions and can be substantial.

The described approach for seasonal adjustment is implemented in JDemetra+ using the TRAMO-SEATS method. TRAMO is a program for estimation and forecasting of ARIMA models with regression variables, missing observations and outlier detection. From the ARIMA model obtained by TRAMO, SEATS derives appropriate models for the unobserved components (Trend-cycle, Seasonal, Transitory and Irregular, see [1]).

Revision policies

The revisions of seasonal adjusted data take place for two main reasons. First, due to the availability of new observations in the observed time series, and second because of a better estimate/identification of the seasonal pattern. The former is unavoidable, but only causes revisions near the time points where the data set changes, often near the end of the time series. Regarding the latter, a different estimate of the seasonal pattern produces small revisions from the beginning of the series, because of the small changes in the estimated coefficients in the Wiener-Kolmogorov filter. However, a different model for the seasonal pattern results in a completely different Wiener-Kolmogorov filter, even if the model for the observed series is preserved, leading to big differences from the beginning of the series. The challenge is to find a balance between the need for the best possible seasonally adjusted data, specially at the end of the series, and the need to avoid revisions that may later be reversed.

There are different revision policies in use, which differ in when the model, filters, outliers and regression parameters are re-identified and re-estimated. In current adjustment, they are re-identified and re-estimated at appropriately set review periods, while in concurrent adjustment they are re-identified and re-estimated every time new data become available. Both strategies have some drawbacks. Under current adjustment, we are keeping the model until the next review even though according to the new evidence perhaps it is no longer acceptable. On the other hand, concurrent adjustment does completely review the model each time we have a new observation of the series. While this ensures that we have a well specified model at each step, it might also result in major revisions from the beginning of the series after each new observation becomes available, because re-identification might lead to a model change, thus changing the Wiener-Kolmogorov filter used to extract the seasonal component ([2]).

In practice, the revision policies applied are “Partial concurrent adjustment”, where some of the model, filters, outliers and calendar regressors are re-identified once a year and the respective parameters and factors are re-estimated every time new data become available and “Controlled current adjustment”, where current adjustment is applied, checking the results with the ones obtained by using “Partial concurrent adjustment”, which is preferred if a significant difference exists. These revision policies are implemented in JDemetra+.

Methods

The INE used to apply the “Partial concurrent adjustment: ARIMA Parameters” policy, which consists in re-estimating all the regARIMA parameters while leaving the model fixed, because this policy has the advantage of the incorporation of the new information when it becomes available and fulfils the requirement of presenting comprehensible data to the users. This last advantage disappeared when we observed the effect of this policy on the huge revisions from the beginning of the series.

The INE started to analyse a new policy, which may be considered a compromise between the "Partial Concurrent Adjustment: ARIMA Parameters" policy and the "Partial Concurrent Adjustment: Fixed Model".

It can be summarized in the following principles:

1. Identify the proper model (ARIMA model, calendar regressors and outliers) once a year (in general, with data up to December of the previous year).
2. Re-estimate all coefficients every time new data become available (monthly or quarterly period) avoiding model changes, on the unobserved components, by:

    (1) Fixing the last estimation of the model with admissible decomposition when a model change is triggered.

    (2) Adjusting root assignment parameters to make sure autoregressive roots remain in the same component.

In this way, we still improve the estimation of the model parameters with the new data, while avoiding big revisions.

Results

We present a comparison of the "Partial concurrent adjustment: ARIMA Parameters" policy and the proposed revision policy using the Service Sector Turnover Index.

The series starts in January 2000 and ends in March 2016. The model selected at the beginning of 2016 is (2,1,0)*(0,1,1), with 4 calendar regressors (that account for the effects of 17 working days with holidays, Easter working days, Easter holidays and leap year) and one additive outlier in August 2012.

In February 2016, the seasonal and calendar adjusted series was published applying the "Partial Concurrent Adjustment: ARIMA Parameters" policy. In March 2016, huge revisions were noticed from the beginning of the series, as shown in Table 1. The second column of the table shows the monthly growth rates of the seasonally adjusted series with data until February 2016, applying Partial concurrent adjustment: ARIMA Parameters. The third column shows the rates with data until March 2016, same policy, which are very different. The fourth column shows the rates obtained applying our new policy, which are much more similar to the second column. The rest of the columns show the same data for the end of the series.

The reason for the large discrepancy between the February and March growth rates is the different assignment of the complex AR roots to the unobserved components, which yields a different theoretical model for the components. This leads to a change in the corresponding Wiener-Kolmogorov filter, with similar consequences to those of a model change.

Another case with a possible change in the theoretical model for the seasonal component, applying "Partial Concurrent Adjustment: ARIMA Parameters" policy happens when SEATS changes the model selected at the beginning of the year for the series because of a non-admissible decomposition for a specific period. Our revision policy implemented checks if SEATS has changed the model, and if that is the case, the estimation of the coefficients of the model obtained in the more recent period (in a year) with admissible decomposition is used instead of changing the model in the middle of the year. This is the same as applying the "Partial Concurrent Adjustment: Fixed Model" policy for a specific period.

Table 11. Comparison of monthly growth rates in initial and final periods

Period	Final date February	Final date March	New policy March	Period	Final date February	Final date March	New policy March
2000-02	-0.1	-0.4	-0.1	2015-02	-0.4	-1.0	-0.6
2000-03	-1.3	-3.6	-1.2	2015-03	0.8	-0.1	0.1
2000-04	2.8	6.3	2.6	2015-04	-0.5	0.6	-0.2
2000-05	-0.6	-1.8	-0.5	2015-05	0.1	-0.1	0.2
2000-06	-1.2	-2.5	-1.2	2015-06	-0.4	-1.1	-0.3
2000-07	1.2	3.0	1.1	2015-07	2.0	3.7	2.0
2000-08	0.3	-1.0	0.3	2015-08	0.6	-0.8	0.7
2000-09	2.1	3.5	2.0	2015-09	1.4	2.1	1.6
2000-10	-0.2	-1.5	-0.1	2015-10	0.7	1.9	0.8
2000-11	0.3	0.2	0.3	2015-11	-1.3	-3.0	-1.2
2000-12	0.9	3.2	0.9	2015-12	0.9	2.0	0.7
2001-01	-1.9	-4.7	-1.9	2016-01	0.4	0.8	0.9
2001-02	0.5	1.6	0.5	2016-02	0.8	1.6	1.3
2001-03	1.0	2.8	1.0	2016-03		2.9	1.7

Conclusions

Although we know that revisions are necessary in Seasonal Adjustment when new data become available, i.e., the adjustment reflects the new information, it is important to verify that these are reasonable. If revisions in the final period, (and also in the two previous years), appear, this may be considered as reasonable. If, however, we observe huge revisions from the beginning of the series, this should always raise a red flag.

So, the revision policy applied at the INE, which can be considered a compromise between the "Partial Concurrent Adjustment: ARIMA Parameters" policy and the "Partial Concurrent Adjustment: Fixed Model", both implemented in JDemetra+, avoids model changes by: (i) fixing the last estimation of the model with admissible decomposition when a model change is triggered and (ii) adjusting root assignment parameters to make sure autoregressive roots remain in the same component.

In this way, we still improve the estimation of the model parameters using the new data, while avoiding big revisions.

References

[133] A. Maravall, G. Caporello, D. Pérez and R. López, New Features and Modifications in Tramo-Seats, Banco de España. (2014)

[134] S.C. Hillmer and G.C. Tiao, An ARIMA-Model-Based Approach to Seasonal Adjustment, Journal of the American Statistical Association 77 (1982), 63–70.

The complexity of the turning points detection in 2022

Keywords: Business cycle analysis, dating algorithms, turning points detection indicators, data analysis

Introduction

In the last years, Eurostat has developed a comprehensive online tool for a statistical analysis (Business Cycle Clock (europa.eu)) of the business cycle situation in the Euro Area and its member countries. This tool, described in detail in [1] and [2], is mainly composed of two well distinct parts: one providing an assessment of observed turning points occurred in the far past, past and recent past providing a statistical dating chronology; the second based on a set of turning points coincident indicators providing quasi real-time information on the possible occurrence of turning points. Both the dating and the detecting parts are designed to provide information on the three main economic cycles: business, growth and acceleration cycle following the so-called αABβCD sequence. Having quite different aim and scope, the two part of the tool show some significant methodological and empirical differences, which can be synthetized as follow:

- While the dating part is based on a non-parametric dating algorithm, the detecting one is designed as a fully parametric and probabilistic system;
- The main reference variable for the dating exercise is the quarterly GDP in volume and for the detecting exercise is the Industrial production index;
- Timing and timeliness of the two exercises appear to be substantially different taking also into account the release calendar of both reference variables mentioned above.

Also considering the differences described above, it is possible that some discrepancies in the assessment of the recent cyclical situation can emerge between the two exercises. Based on our experience, past discrepancies have always been very transitory covering a few months and not that relevant. The observed discrepancies between the dating and the detecting part emerged when the economy was positioned in two different, even if quite close, cyclical phases, as for instance being in an acceleration or in a decelerating phase or in a deceleration phase instead of a slowdown one.

By contrast, since March 2022, we have observed a much relevant disagreement between the dating and the detecting part with the former signalling an expansionary, even if decelerating situation, and the latter indicating an ongoing recession. This disagreement has been particular evident for the Euro Area, Germany and Italy. Obviously, this anomalous situation has requested some deep investigations to understand causes and factors generating it. This paper aims at describing the analysis carried out, the actions performed and the solutions found.

Methods

In this session, we are briefly describing the methodology used in the dating and detecting exercises as well as the situation observed in 2022 and some data issues emerged.

Methodology

As already mentioned, the methodology used in the dating exercise is quite simple. It is based on a non-parametric dating rule (similar to the one proposed by Hardin and Pagan) applied to the three cycles considered where the growth cycle is obtained by applying to original data two kind of HP filter, a law pass and a high pas one, to approximate and ideal band pas filter.

Concerning the detecting part, this one is based on three probabilistic indicator, one for the acceleration cycle (available for the Euroarea only) obtained by fitting a univariate MS model to the sentiment indicator, and two indicators for the growth and the business cycle respectively simultaneously derived by a multivariate MS-VAR model [3]. All these indicators have been recently revised to better account for the pandemic effects, also by introducing some interventions variables whenever needed [4] and have proven to be very well performing also during the pandemic. MS-VAR models have been developed for the Euroarea and all member countries using the most appropriate specification and set of input variables.

Some data issues

Figure 1 below is showing the above-mentioned disagreement observed in 2022.

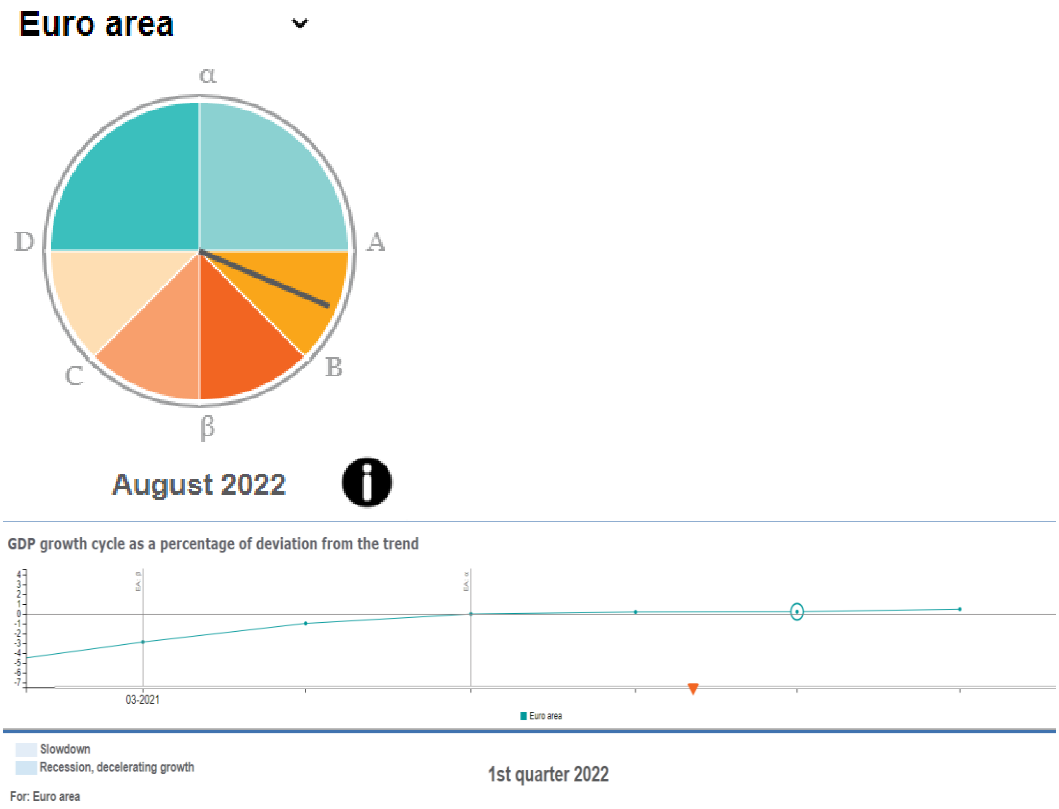


Figure 50. Disagreement observed in 2022.

Since the model specification has been recently revised showing very good results, we have decided to concentrate our attention on some inconsistencies observed among data involved in such models namely:

- *the divergent pattern between GDP and IPI,*
- *the excessive volatility of the IPI,*
- *the post pandemic driving role of services and*
- *the negative results several opinion surveys indicators contradicting the evolution of corresponding hard data.*

(Some examples to be provided)

Results

In this session we are presenting some results of our simulations done by keeping the model unchanged a modifying the set of input data used in order to reduce or cancel the observed discrepancy between the dating and the detecting exercise observed in 2022.

(Section to be completed.)

Conclusions

(Section to be completed.)

References

[135] Anas J., Mazzi G.L., Ruggeri Cannata R., Rieser D., Paracho C. "An overview of Existing Business Cycle Clock Applications and Some Proposals for Further Improvement" in UNSD and Eurostat Handbook on cyclical composite indicators – Chapter 20" Ed. By G.L. Mazzi and A. Ozyildirim – Eurostat 2017

[136] Ruggeri Cannata R. (2021), The Eurostat Business Cycle Clock: a complete overview of the tool, Statistical Journal of the IAOS vol. 37, no. 1, pp. 309–323, 2021.

[137] Anas J., Billio M., Carati L., Ferrara L., Mazzi G.L.: Chapter 14 of the UNSD and Eurostat Handbook of cyclical composite indicators ED by G.L. Mazzi and Ataman Ozyldirim – Eurostat 2017

[138] Billio M., Carati L., Mazzi G.L., Ruggeri-Cannata R., Ronkowski P., Vlachou H. - Mitigating the pandemic effects in the Eurostat system for turning points detection – paper presented at the 2nd OECD and TSACE workshop on Time series methods for official statistics – Paris 2022.

Testing for cointegration in STAR models: A simulation-based study

Keywords: cointegration, Fourier approach, STAR model, structural changes

INTRODUCTION

In time series analysis, economic variables are subject to structural breaks of unknown numbers and forms. In the literature, dummy variables are mostly used to catch the breaks. However, prior knowledge is needed for the date, number, and form of the breaks with the usage of dummy variables although the actual nature of the breaks is generally unknown and major breaks also sometimes do not exhibit their impacts immediately as noted in [1] and [2]. In this respect, the Fourier approach is suggested as an alternative way to overcome the problems seen in the usage of dummy variables (see the papers [3], [4], [5], [6], [1], [7], and [8]). The flexible Fourier approach successfully catches the breaks regardless of the date and number of breaks. The approach captures not only gradual breaks but also works well for sharp breaks in the data as pointed out in the paper [9]. In addition to modeling structural breaks, taking the nonlinear structure into account is another important factor in the cointegration analysis. The information obtained from the linear models may be insufficient for economic inferences and variables may be conditional on the nonlinear dynamics of the models. Power problems may occur in the tests when the nonlinear nature of the variables is ignored in the analysis. Thus, results tend to be biased which makes it more difficult to find a cointegration relation between variables.

In this paper, different from the paper [10] in which nonlinearity is modeled through the exponential smooth transition autoregressive (ESTAR) model and structural breaks are modeled through the Fourier approach, we first practice with the decimal number Fourier frequencies to catch the breaks regardless of the date, number, and form of the breaks in the Fourier ESTAR test. Second, we extend the Fourier ESTAR test by computing the simulation results of the logistic smooth transition autoregressive (LSTAR) cointegration test with the Fourier function. To the best of our knowledge, there is no Fourier LSTAR cointegration test in the literature.

METHODS

In this section, we show the main equations that we consider in the data generating process. Following the papers [2] and [9], we can write the models including the deterministic term d_t as follows:

$$y_t = d_t + \beta' x_t + u_t, t = 1,2, \dots T \tag{1}$$

$$d_{2t} = \mu + \alpha(t), \tag{2}$$

$$d_{3t} = \mu + \vartheta t + \alpha(t), \tag{3}$$

$$\alpha(t) = \alpha_1 sin\left(\frac{2k\pi t}{T}\right) + \alpha_2 cos\left(\frac{2k\pi t}{T}\right). \tag{4}$$

In equation (4), k is the Fourier frequency; T is the number of observations; t is a trend, and $\pi \cong$ 3.1416. In the literature, Fourier frequency is defined in the range of [1,5] since findings prove that working with higher numbers of frequencies may not enhance the power of the test. Furthermore, it is easier to catch the breaks when the frequency k = 1 (see [1], [6], [7], and [8]). However, in this paper, we practice with the frequencies 1.5, 2.5, 3.5, and 4.5 to see how the power properties change when we use decimal numbers in the Fourier function. Nevertheless, the Fourier function $\alpha(t)$ is not added to the deterministic term in ESTAR and LSTAR cointegration tests when the structural changes are ignored in the testing process.

We estimate the following equation after obtaining residuals from equation (1) through Ordinary Least Squares (OLS):

$$\Delta\hat{u}_t = \rho\hat{u}_{t-1}G(.) + \sum_{j=1}^{p} \psi_j \Delta\hat{u}_{t-j} + \epsilon_t, \tag{5}$$

where ϵ_t is the stationary error with zero mean and G(.) is the transition function formulated as follows for the ESTAR and LSTAR models, respectively:

$$G(.) = 1 - exp\{-\gamma u_{t-1}^2\}, \tag{6}$$

and

$$G(.) = (1 + exp\{-\gamma(u_{t-1})\})^{-1}, \tag{7}$$

where the parameter γ specifies the smoothness of the function and is assumed to be greater than zero. Based on equation (5), the null hypothesis of no cointegration relation is tested against the alternative of the presence of cointegration relation with a STAR adjustment as follows:

$$H_o: \rho = 0, \tag{8}$$

$$H_1: \rho < 0.$$

The grid search approach is applied to overcome the nuisance parameter problem that arises because the transition parameter γ is defined only under the alternative hypothesis. The advantage of this approach is that the infimum type *t* statistic is obtained easily.

RESULTS

The data generating process starts with estimating the critical values of the cointegration test for the different number of independent variables and sample sizes. Following, the size and power properties are computed with the obtained critical values. We use R programming for the simulation process.

The simulation results indicate that the size of the proposed tests is not affected by the change in the variance and the autocorrelation coefficient. Size properties are also close to the nominal level. The power of the test increases as we enhance the sample size. Results of the comparison of the Fourier ESTAR, Fourier LSTAR, ESTAR, and LSTAR tests show that the power of the Fourier ESTAR and Fourier LSTAR tests is higher than ESTAR and LSTAR tests for all sample sizes and parameters. Thus, the probability of rejecting the incorrect null hypothesis is easier when the structural breaks are modeled by the Fourier function. However, the power of the LSTAR test is the lowest compared to other tests.

cONCLUSIONS

The simulation results indicate that the Fourier ESTAR test with decimal frequencies maintains good size and power properties. It is important because the optimal frequency, which is chosen from the model that gives the minimum sum of squared residuals among the estimated models with different frequencies, can fit the model best with the decimal numbers and meet the needs in the empirical examples. The proposed Fourier LSTAR cointegration test also maintains the good size and power properties. The test performance is higher when the structural breaks are taken into account through the Fourier function and hence we are more likely to reject the null hypothesis correctly. In particular, the probability of rejecting the null hypothesis correctly is higher when we define the exponential transition function in the STAR model instead of the logistic function.

rEFERENCES

[1] R. Becker, W. Enders, and S. Hurn, A stationarity test in the presence of an unknown number of breaks, Journal of Time Series Analysis 27 (2006), 381–409.

[2] C. Tsong, C. Lee, L. Tsai, and T. Hu, The Fourier approximation and testing for the null of cointegration, Empirical Economics 51 (2016), 1085-1113.

[3] A. R. Gallant, On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Flexible Fourier Form, Journal of Econometrics 15 (1981), 211–245.

[4] A. R. Gallant, The Fourier Flexible Form, American Journal of Agricultural Economics 66 (1984), 204-208.

[5] A. R. Gallant, and G. Souza, On the Asymptotic Normality of Fourier Flexible Form Estimates, Journal of Econometrics 50 (1991), 329-353.

[6] R. Becker, W. Enders, and S. Hurn, A general test for time dependence in parameters, Journal of Applied Econometrics 19 (2004), 899–906.

[7] W. Enders, and J. Lee., A Unit Root Test Using a Fourier Series to Approximate Smooth Breaks, Oxford Bulletin of Economics and Statistics 74 (2012a), 574-599.

[8] W. Enders, and J. Lee, The Flexible Fourier Form and Dickey-Fuller Type Unit Root Tests, Economics Letters 117 (2012b), 196-199.

[9] P. Banerjee, V. Arčabić, and H. Lee, Fourier ADL cointegration test to approximate smooth breaks with new evidence from Crude Oil Market, Economic Modelling 67 (2017), 114-124.

[10] B. Güriş, and G. Sedefoğlu, A Proposal of Nonlinear Cointegration Test with the Flexible Fourier Approach, Communications in Statistics - Simulation and Computation (2022), https://www.tandfonline.com/doi/abs/10.1080/03610918.2022.2067874?journalCode=lssp20

[11] D. Maki, An alternative procedure to test for cointegration in STAR models, Mathematics and Computers in Simulation 80 (2010), 999-1006.

JDemetra+ 3.0: New (R) tools for (high-frequency) time series analysis

Keywords: *seasonal adjustment, trend estimation, high-frequency data, structural models, tradingday correction*

1 Introduction

The latest version of JDemetra+ [109], 3.0 to be released in December 2022, fills several critical gaps in the tool box of a time series analyst providing extended features for seasonal adjustment and trend estimation, including high frequency data and production tools. A state-space framework gives access to Basic Structural Models, offering a seasonal adjustment procedure with explicit decomposition and time varying trading day correction.

This open source software [1] is unique in its combination of very fast java routines, a graphical user interface and a family of R packages. The graphical interface offers a structured and visual feedback, suitable for refined analysis and training, whereas R tools allow the user to mix the capabilities of JDemetra+ with the versatility of the R world, be it for mathematical functions or data wrangling.

The first package allowing to use JDemetra+ core algorithms {RJDemetra} R was created in 2018 [2] and since then the software has been rapidly expanding: algorithms have been extended and tools upgraded. In version 3.0 {RJDemetra}R will be replaced with a family of packages covering each a very specific perimeter. This modular structure makes the functions more readable, fosters evolution and encourages the user to enhance the functions according to its own needs.

The tedious task of generating calendar regressors for quarterly, monthly but also high frequency data is now possible using simple R functions of the {rjd3modelling} R package, for the first time in R. This kind of feature is of high importance for seasonal adjustment practitioners and even more important is the possibility to refresh estimation when new data is available, which is covered in greater detail in the methods section.

Our paper aims at describing and illustrating the new capabilities of JDemetra+ 3.0 as well as the R packages allowing to access them. In the remainder of this abstract we highlight three categories of features which seem to be much sought-after by users: seasonal adjustment of high frequency data, trend estimation and tools for building seasonal adjustment production chains entirely in R.

[109] JDemetra+ is an open source software for time series analysis. It has been officially recommended by Eurostat to the European Statistical System members since 2015.

2 Methods

2.1 Seasonal adjustment of High-frequency data

Infra-monthly economic time series have become increasingly popular in official statistics in recent years, more and more users ask for timely weekly and even daily indicators of economic developments. Many of those indicators display seasonal behavior and, thus, are in need for seasonal adjustment. JDemetra+ seasonal adjustment algorithms have been augmented to meet this need, offering an enhanced reg-ARIMA pre-treatment model and extended versions of the ARIMA model-based, STL and X-11 seasonal adjustment approaches able to deal with multiple and non-integer periodicities common in high frequency data, as described in Smyk and Webel (2022) [3]. These extensions are accessible through the {rjd3highfreq} R [4] and {rjd3stl} R packages as well as through the graphical user-interface. Some key features like extended X-11 to any (fractional) periodicity or fractional airline model are unique to JD+. Fractional periodicities are tackled using a Talyor approximation for the backshift operator $B^{s+\alpha} \tilde{=} (1-\alpha)B^{s} + \alpha B^{s+1}$, where α is the decimal part of the periodicity. A short illustration is provided in the Results section.

2.2 Trend estimation

Trend estimation algorithms have been upgraded in the extended x-11 algorithm available in {rjd3highfreq}R and as a stand alone feature of the {rjdfilters}R package [5]. The refined and final trend-cycle extraction filters of the genuine X-11 method are essentially a set of pre-specified weights for symmetric m-term Henderson filters with $m \in \{3,\dots,101\} \cap \mathbb{N}_{odd}$ and their asymmetric Musgrave surrogates. The extended X-11 approach is founded on applying local polynomial regressions to the input series, as derived in Proietti and Luati (2008) [6].

More algorithms are available via {rjdfilters}R such as Reproducing Kernel Hilbert Space (RKHS) filters of Dagum and Bianconcini (2008) [7] with same kernels, FST filters derived from Grun-Rehomme, Guggemos, and Ladiray (2018) [8] and DFA filters derived from the AST approach of Wildi and McElroy (2019) [9]. This package also provides new functions for building moving average based filters and plotting their properties, which could be used for training purposes.

2.3 Mass production of seasonally adjusted data in R

R has become ubiquitous in official statistics and the demand for its use in production of seasonally adjusted data is growing fast. JDemetra+ offers the speed and the pre-specified refresh policies recommended by Eurostat Guidelines on SA [10]. A wide range of "partial-concurrent adjustment" options, in which parameters and reestimated and/ or re-identified gradually have been long available in JDemetra+. But, until now, in the versions 2 family, these options were linked to updating a workspace (specific data structure) via the graphical user interface or more probably via the cruncher (a batch module). It was quite a liability for full production in R as explained in Smyk and Tchang (2021) [11]. Revisions policies are now even more customizable if implemented in R, as time spans on which options are applied can be chosen by the user. Before version 3.0, the user could chose between re-identifying outliers on the whole series span or on the last year of the data (this is the widely applied "partial concurrent last outliers policy"), now the period is customizable, which really makes sense when progressively remodelling the impact of the covid crisis.

3 Results

3.1 Seasonal adjustment of French daily births series

We consider the series of daily french births from 1968 to 2020. Spectral analysis shows that two periodicities $p_1 = 7$ and $p_2 = 365.25$ are present. The series is first linearized: outliers are detected and calendar effects removed with the following fractional airline model: $(1-B)(1-B^7)(1-B^{365.25})(Y_t - {}^{\mathrm{P}}\alpha_i X_{it}) = (1-\theta_1 B)(1-\theta 2B7)(1-\theta 3B365.25)\epsilon t,...\epsilon t \sim$ NID $(0, \sigma_\epsilon^2)$ with $1-B365.25 = (1-0.75B365 - 0.25B366)$ Then a decomposition is performed with extended X-11, using modified filters with the Taylor approximation, which avoids imputing data.

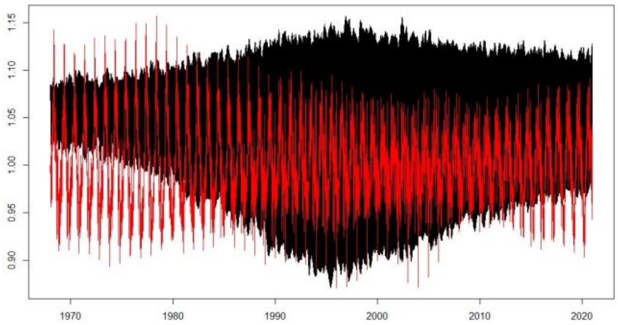


Figure 1: French daily births: estimated seasonal factors , p=7 (black) and p=365.25 (red)

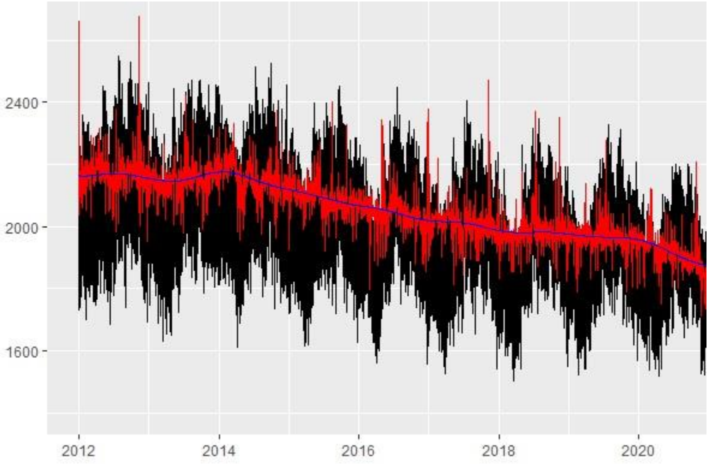


Figure 2: French daily births: raw (black), trend (blue), seasonally adjusted (red)

4 Conclusions

This abstract is meant to give a hint on the new developments of an open source time series software, already widely used in official statistics throughout Europe and the world. JDemetra+ provides new solutions to some critical issues, which could benefit many time series analysts. Our forthcoming paper will describe in greater detail those innovations, providing detailed explanations on algorithm's extensions as well as R code snippets.

Further developments are under way in several domains, among them a better calibration of filters and more adapted tests to high-frequency data.

References

[1] Sylwia Grudkowska et al. *JD+ documentation*. url: https://jdemetradocumentation.github.io. (2019).

[2] Jean Palate Alain Quartier-la-Tente Anna Michalek and Raf Baeyens. *RJDemetra: Interface to 'JDemetra+' Seasonal Adjustment Software. R package version 0.1.6.* url: https://CRAN.R-project.org/package=RJDemetra. (2018).

[3] A Smyk and K Webel. "Towards seasonal adjustment of infra-monthly time series for JDemetra+". In: *2nd Workshop on Time Series Methods for Official Statistics OECD* (2022). url: https://community.oecd.org/docs/DOC-218994.

[4] Jean Palate. *rjd3highfreq: R package for seasonal adjustment of high frequency data*. url: https://github.com/palatej/rjd3highfreq.

[5] Jean Palate Alain Quartier-la-Tente. *rjdfilters: R package for linear filters*. url: https://github.com/palatej/rjdfilters.

[6] Tommaso Proietti and Alessandra Luati. "Real time estimation in local polynomial regression, with application to trend-cycle analysis". In: *The Annals of applied statistics* 2.4 (2008), pp. 1523–1553.

[7] Estela Bee Dagum and Silvia Bianconcini. "The Henderson Smoother in Reproducing Kernel Hilbert Space". In: *Journal of Business & Economic Statistics* 26 (2008), pp. 536–545. url: https://ideas.repec.org/a/bes/jnlbes/v26y2008p536545.html.

[8] Michel Grun-Rehomme, Fabien Guggemos, and Dominique Ladiray. "Asymmetric Moving Averages Minimizing Phase Shift". In: *Handbook on Seasonal Adjustment* (2018). url: ec.europa.eu/eurostat/web/products-manuals-andguidelines/-/KS-GQ-18-001.

[9] Marc Wildi and Tucker McElroy. "The trilemma between accuracy, timeliness and smoothness in real-time signal extraction". In: *International Journal of Forecasting* 35.3 (2019), pp. 1072–1084. url: https://EconPapers.repec.org/RePEc:eee:intfor:v:35:y:2019:i:3:p:1072-1084.

[10] Eurostat. *ESS Guidelines on Seasonal Adjustment*. Tech. rep. Eurostat Methodologies and Working Papers, European Commission, 2015. doi: 10.2785/317290. url: http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines//KS-GQ-15-001.

[11] A Smyk and A Tchang. *R Tools for Jdemetra+, Seasonal adjustment made easier*. Tech. rep. Institut National de la Statistique et des Etudes Economiques, 2021. url: https://www.insee.fr/en/statistiques/5019812.