EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

ABSTRACTS

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

Welcome to Q2024!

Dear colleagues,

Statistics Portugal and Eurostat are pleased to invite you to the **11th European Conference on Quality in Official Statistics (Q2024)** in Estoril (the Greater Lisbon region) from **4 to 7 June 2024**.

Following a longstanding quality-driven tradition in the European statistical community, this conference will gather statisticians, academics, and external stakeholders to **foster the dissemination of knowledge and research on emerging issues related to quality in official statistics**. By bringing diverse stakeholders together, Q2024 aims to enhance and improve cooperation between official statistics and the scientific community — in which students play a significant role — and to promote closer dialogue between users and producers.

Q2024 will focus on **the role of official statistics as a pillar of democracy** and on some of the most challenging issues currently faced by the official statistics community. The conference topics aim to be broad enough to cover the diversity of challenges faced by the statistical community while also addressing the most relevant dimensions within the field of quality. Q2024 will explore how **institutional frameworks** and the role of **innovation and research** impact on quality in official statistics, and it will also focus on the power of data and **NSIs' capabilities to harness opportunities** in a challenging environment.

Bearing in mind the importance of capacity-building in previous editions of the conference, a variety of engaging and informative **one-day training courses will be held on 4 June 2024**. These courses will be organised around four main pillars: quality management, innovation/data science/AI, dissemination/communication, and the integration of administrative and privately-held data.

The Q2024 website has been designed to keep you updated with the most relevant information about this keystone event. Should you have any questions, please contact the organising team at q2024@ine.pt for technical matters, or at q2024@leading.pt for logistical issues.

We invite you to share your insights and projects with this inspiring community, and we are very much looking forward to welcoming you to Estoril!

Francisco Lima, PhD
President, Statistics Portugal

Mariana Kotzeva, PhD
Director General, Eurostat

Scientific Committee

Mariana Kotzeva, *Co-Chair (Eurostat)*

Francisco Lima, *Co-Chair (Statistics Portugal)*

Athanasios C. Thanopoulos, *ELSTAT – Greece*

Jūratė Petrauskienė, *Statistics Lithuania*

Jean-Pierre Poncelet, *Eurostat*

Aurel Schubert, *ESGAB (Former Chairperson)*

Roxane Silberman, *ESAC*

Silke Stapel-Weber, *ECB*

Pádraig Dalton, *CSO-Ireland; CES*

Dominik Rozkrut, *Statistics Poland; IAOS*

Paula Brito, *FEP.UP; CLAD – Portugal*

Pedro Magalhães, *ICS-ULisboa – Portugal*

Maria do Rosário Oliveira Martins, *IHMT-UNL - Portugal*

Programme Committee

Claudia Junker, *Co-Chair (Eurostat)*

Maria João Zilhão, *Co-Chair (Statistics Portugal)*

Avis Benes, *Eurostat*

Veronique Van Der Zande, *Eurostat*

Monika Wozowczyk, *Eurostat*

Pedro Campos, *Statistics Portugal; FEP-UP*

Magda Ribeiro, *Statistics Portugal*

Monika Bieniek, *Statistics Poland*

Giorgia Simeoni, *ISTAT – Italy*

Thomas Burg, *Statistics Austria*

Kornélia Mag, *Quality Expert – Hungary*

Silvan Zammit, *National Statistics Office – Malta*

Patrícia Ávila, *CIES-ISCTE – Portugal*

Pedro Raposo, *Católica-Lisbon – Portugal*

Remi Prual, Quality Expert - Estonia

Programme

Wednesday, June 5, 2024

	Auditorium	Room F5-F7	Room C5-C7	Room B	Room E	Big Hall
9:00 - 10:30	**Opening Session: "Official Statistics as a Pillar for Democracy"**					
10:30 - 11:00	Morning Coffee Break - 05 JUNE					
11:00 - 12:30	Session 1 - Special Session: The Third Round of Peer Reviews in the European Statistical System – Its Implementation, Results and Lessons Learnt	Session 2 - Confidentiality and data protection I	Session 3 - Geostatistics I	Session 4 - Quality management	Session 5 - Combining and integrating Data sources	
12:30 - 14:00	Lunch - 05JUNE					
12:45 - 13:30	Speed Talk Session 1 - Quality frameworks	Speed Talk Session 2 - Smart Survey Implementation	Speed Talk Session 3 - Improving household surveys	Speed Talk Session 4 - Data processing	Speed Talk Session 5 - Statistical leadership/HR management	
14:00 - 15:30	Session 6 - Governance and Quality	Session 7 - Quality of administrative data	Session 8 - Quality of web-based data	Session 9 - Experimental statistics	Session 10 - Metadata Quality I	
15:30 - 15:45	Afternoon Coffee Break - 05 JUNE					
15:45 - 16:30						Poster Session 1
16:30 - 18:00	Session 11 - Quality assessment and review I	Session 12 - Online job advertisement	Session 13 - Innovative methods & machine learning	Session 14 - Communication and Statistical literacy	Session 15 - Smart Surveys Implementation	
18:30 - 20:00	Welcome Reception					

Thursday, June 6, 2024

	Auditorium	Room F5-F7	Room C5-C7	Room B	Room E	Big Hall
9:00 - 10:30	Session 16 - Census & Multisource registers	Session 17 - Assuring quality in statistics: academia & staff development	Session 18 - Special Session: Implementation of Quality Management Frameworks/Systems in the Enlargement and European Neighbourhood Policy (ENP)-East Countries	Session 19 - Statistics and decision making I	Session 20 - Big Data	
10:30 - 11:00	Morning Coffee Break - 06 JUNE					
11:00 - 12:30	Session 21 - Special Session: The ESS Innovation Agenda	Session 22 - User needs & expectations	Session 23 - Machine learning I	Session 24 - Quality of registers	Session 25 - Metadata Quality II	
12:30 - 14:00	Lunch - 06 JUNE					
12:45 - 13:30	Speed Talk Session 6 - Experimental analysis & sources	Speed Talk Session 7 - Quality & Data management	Speed Talk Session 8 - Statistics and decision-making / user engagement	Speed Talk Session 9 - Coordination	Speed Talk Session 10 - Improving business statistics	
14:00 - 15:30	Session 26 - Coordination	Session 27 - Geostatistics II	Session 28 - Cooperation with academia	Session 29 - Data integration	Session 30 - Emerging innovations	
15:30 - 15:45	Afternoon Coffee Break - 06 JUNE					
15:45 - 16:30						Poster Session 2
16:30 - 18:00	Plenary Session: "The challenge of artificial intelligence and its application to official statistics"					
19:30	Official Conference Dinner					

Friday, June 7, 2024

	Auditorium	Room F5-F7	Room C5-C7	Room B	Room E
9:00 - 10:30	Session 31 - Special Session: Communicating Quality	Session 32 - Statistics and decision making II	Session 33 - Data Visualisation	Session 34 - Privately-held data	Session 35 - Quality assessment and review II
10:30 - 11:00	Morning Coffee Break - 07 JUNE				
11:00 - 12:30	Session 36 - Confidentiality and data protection II	Session 37 - Machine learning II	Session 38 - Microdata level perspective	Session 39 - MNO data	Session 40 - Special Session: Best Practices in Quality Management in the Southern European Neighborhood Policy countries
12:30 - 12:40	Intermission				
12:40 - 14:00	**Closing Plenary Session: "Communicating official statistics in the present data ecosystem"**				
14:00 - 15:00	Lunch - 07 JUNE				

Índice

Session 1 - Special Session: The Third Round of Peer Reviews in the European Statistical System – Its Implementation, Results and Lessons Learnt, June 5, 2024, 11:00-12:30

Peer reviews – impact of their recommendations and first results

Mr Vidar Lund[1]

[1]Statistics Norway, , Norway

A peer review can be a powerful driving force for developing the national statistical system. In addition to increasing awareness of the quality framework for European statistics, the process itself is instrumental in bringing forward improvement ideas from various levels in the national statistical institute (NSI) and the other national authorities responsible for European statistics (ONAs).

The second round of ESS peer reviews 2013-2015 had a significant impact on the Norwegian statistical system, contributing to processes resulting in an updated statistical law adopted in 2019, the establishment of an advisory council for Statistics Norway, a multi-year national programme for official statistics and a national coordinating committee for official statistics. Against that background, Statistics Norway entered the third round of peer reviews 2021-2023 with a clear understanding of the potential for change embedded in the process.

The new peer review would be the first evaluation of whether the recent changes in the legal and institutional framework were in line with European standards. In addition, there was an expectation that this round would focus on future-oriented innovations in the statistical system. In consequence, it was decided to follow a decentralised approach involving all relevant parts of the organisation in Statistics Norway and the ONAs, inviting subject matter experts to describe the current situation in their area in the self-assessment phase. The central coordination team in Statistics Norway acted as advisers and editors for the input provided by the contributors across the organisations.

The main goal was set out by the coordination team at the start of the process: To describe the current situation as accurately as possible, without exaggerating positively or negatively, to facilitate receiving recommendations that are useful for the further improvement of the Norwegian statistical system. The observations made by the peer review team and the recommendations received at the end of their visit were in line with these expectations and thus perceived as fair, relevant and helpful for the continued development in the areas concerned.

The same decentralised approach was used when defining the improvement actions to be implemented in the coming years based on the recommendations. In the view of Statistics Norway, the close involvement of experts at various levels in the NSI and ONAs was a key success factor for the recent peer review of the Norwegian statistical system, ensuring a common understanding of the challenges in each area and the most appropriate way to meet them.

Peer reviews – internal and external coordination, cooperation and communication for implementing the peer review

Ms. Monika Bieniek[1]

[1]Statistics Poland, Warsaw, Poland

The 3rd round of peer reviews in the European Statistical System has been a challenge for all its participants, hence the NSIs as well. In a sense, this exercise was not only a test of their maturity as the national institutions coordinating the statistical activities of all other national authorities that develop, produce and disseminate European statistics in a country, but also of their organizational capacity to prepare for peer review. In this context, from the NSI's perspective, the three main factors had the important impact on the whole peer review process in a country and its full implementation. These were the following: well-planned and optimised internal and external coordination within the country, full cooperation between all actors involved in the process (including the NSI's willingness to provide them factual and organizational support), and an active, broad-based information campaign about the peer review process and the value of official statistics. The presentation will be focused on Statistics Poland's practices in the aforementioned three key areas: national coordination, cooperation and communication activities during the 3rd round of peer reviews. In this context, the main steps undertaken by Statistics Poland during the preparatory phase, the peer review visit and the wrap-up phase will be presented.

Peer reviews – challenges in implementing peer reviews from the expert and national perspectives

Maria João Zilhão[1]

[1]Statistics Portugal, Member of the Board , Lisboa, Portugal

Peer reviewing the adherence of the European Statistics Code of Practice (ES CoP) in the European Statitstical System is a very demanding exercise, both for the member states, National Statistical Institutes (NSI) and Other National Authorities (ONA), and the expert teams. But it also constitutes a great opportunity for improvement with respect to the Implementation of the ES CoP: a time to stop and reflect on the principles and indicators of best practices and how organizations are putting in place their own practices to show evidence on that respect to the external expert teams. The presentation will explore the challenges faced from both, member states and external reviewers, and how to efficiently deal with all phases of the process of reviewing, from preparing and studying the most appropriate documentation, to prepare the visit itself, conduct the meetings and conclude in the best possible manner the main findings of the exercise with recommendations, the identification of best practices and also innovative ones. Communication aspects and facilitating an adequate methodology for all parts have been the main key success factors, but considering national specificities of members states are also relevant. All challenges taken into account it was a very rich and refreshing experience for all parties!

The third round of ESS peer reviews – a strategic view on their implementation, main results and lessons learnt.

Mateusz Kaleta[1]

[1]Eurostat, , Luxembourg

The third round of European Statistical System (ESS) peer reviews will be finalised with the public report to the European Parliament and the Council being available by the time of Q2024 Conference. The intervention, building on the messages from the published report, would present a strategic view on peer review implementation, main results and lessons learnt. It would gather information from 31 Peer Review reports and available Improvement Actions Plans of the ESS Member States, final workshops on Peer Review and relevant ESS bodies meetings.

The presentation would take stock of the challenges in implementing the recent round, results, and their possible impact on the future development of the national statistical systems (NSS) of the ESS members as well as draw some lessons from its implementation. Third peer review round strengths, good and innovative practices, progress identified during the peer review process as well as lessons learnt for the future would be shown using qualitive and quantitative summary information and analyses.

The author would present a general overview from Eurostat perspective which was that of a coordinator of the entire process and reflect on the realisation of 3rd Peer Review round objectives, namely reviewing compliance and alignment with the European Statistics Code of Practice by the NSS of the ESS members and at identifying future-oriented, forward-looking recommendations.

Session 2 - Confidentiality and data protection, June 5, 2024, 11:00-12:30

A step-by-step process to deal with the protection of a set of tabular data

Clara Baudry[1], Julien Jamme

[1]Insee, Montrouge, France

Before releasing a set of tabular data to the public, NSIs have to minimize the risk of disclosure of confidential information. The people responsible for protecting sets of tabular data using a suppressive method are confronted with many issues that require a certain level of expertise.

However they struggle to reach this level of expertise since they protect tabular data only once a year.

Insee's department of statistical methods developed a step-by-step process to harmonize the methodology and has been implementing tools (such as the rtauargus package) to reduce the level of expertise required to protect a set of tables.

Even though tools such as Tau-Argus protect tables efficiently, they do not eliminate all the practical difficulties. In fact, one of many challenges for producers is to understand that the set of tables they want to release is not the same as the set of tables they need to protect. For example, a producer might want to release a table that breaks the population down by region and another table that breaks the same population down by municipality. In this case, the two tables must be merged into a single one in order to take into account the hierarchical structure present between municipalities and regions. The step-by-step process is inspired by the roadmap suggested by Hundepool et al. in the Handbook on Statistical Disclosure Control1, but specialized in the tabular data protection.

The first section presents Insee's experience in the protection process for tabular data, which reveals the methodological and implementation challenges faced by our colleagues. In the second section, the enhanced step-by-step process is presented as a series of simple questions to be answered by the user, alongside the description of the provided tools, such as the R package. In the third section, the actions taken to disseminate this new methodological framework throughout the institute are listed, as well as some features under development to foster adoption.

Integrated Risk Management in Quality and Information Security Systems

Joaquim Machado[2], Magda Ribeiro[1]

[1]Statistics Portugal, Lisbon, Portugal, [2]Statistics Portugal, Lisbon, Portugal

In today's complex and dynamic business environment, organizations face multifaceted challenges related to quality assurance, information security, and overall risk management. National Statistical Institutes are examples of organizations that are facing up to these challenges. Our study explores the interconnectedness of these critical domains and emphasizes the need for an integrated approach to address the evolving landscape of risks.

Quality systems play a pivotal role in ensuring that products and services meet or exceed customer expectations. However, the pursuit of quality must be accompanied by a comprehensive risk management strategy to identify, assess, and mitigate potential threats. Integration of risk management into quality systems not only ensures the integrity of products and processes but also improves the overall organizational resilience.

Simultaneously, information security has become paramount in the digital age, where organizations rely heavily on interconnected technologies and data-driven processes. The paper underscores the inseparable link between information security and risk management, emphasizing the importance of a proactive and adaptive security framework. Such a framework must address the evolving nature of cyber threats, privacy concerns, and regulatory requirements to protect sensitive information.

Furthermore, the paper delves into the synergies between quality systems and information security, highlighting how a harmonized approach can amplify the effectiveness of risk management efforts. The convergence of quality assurance and information security fosters a culture of continuous improvement and resilience. It enables organizations to identify vulnerabilities across processes and information systems, implement preventive measures, and respond promptly to emerging risks.

In this study, we will present the approach taken by Statistics Portugal to build an integrated information management, quality and security system, namely by adopting a common risk analysis and management methodology and tool.

The paper concludes by emphasizing the need for a holistic risk management strategy that integrates quality systems and information security. By doing so, organizations can create a robust framework that not only ensures the delivery of high-quality products and services but also strengthens the confidentiality, integrity, and availability of critical information assets. This integrated approach positions organizations to navigate uncertainties with agility, maintain stakeholder trust, and sustain long-term success in an ever-evolving business landscape including national statistical offices.

Synthetic data in official statistics

Mr Simon Xi Ning Simon Kolb[1], Yannik Garcia Ritz[1], Hanna Brenzel[1], Sarah Giessing[1]

[1]Federal Statistical Office of Germany, ,

While the impact of digitization is still growing at an unprecedented pace, the amount and importance of data has already reached unforeseen dimensions. Data driven decision making has become the paradigm of both the private and public sector alike. As the importance and dependence on high quality data becomes more obvious every day, custodians of official data find themselves in a challenging spot when it comes to making data publicly available.

The German Federal Statistical Office faces the risk of violating data privacy whenever data is made accessible to the public. It is nonetheless obliged to provide information to the general public and the independent scientific community. The standard scenario consists of data gathering followed by the publication of aggregates produced from the underlying microdata. It is widely known that aggregation does not offer sufficient protection depending on data granularity and type. Hence, a wide range of so-called post tabular SDC methods have been developed, which modify the aggregated statistics in a way that aims to reduce disclosure risk. These changes will invariably result in reduced data usefulness while leaving the micro data unchanged. However, the federal office of statistics does not confine data publication solely to aggregates. The goal of the Research Data Centre of the Statistical Offices of the Federation and the Federal States is to provide access to microdata for research purposes. This process is strictly governed by the Federal Statistics Law and requires strong data anonymity. So far, achieving thorough microdata anonymization while maintaining statistical usefulness has posed a significant challenge. The rapid advancement in computing power and data availability emphasizes the urgent need for improvements of traditional pre-tabular SDC techniques.

In this work, we explore synthetic data as a means to overcome the limitations of traditional SDC methods for both microdata and aggregates. By synthesizing the microdata, post tabular methods can be avoided, as all contributions to a table are synthetic values. Maintaining analytical utility in the synthesized data and reducing disclosure risk is however no trivial task. We investigate the potential of a synthetic data approach with two separate use cases.

Privacy protection, data validation and secure machine learning with new statistical methods while preserving transparency

Dr Violeta Calian[1]

[1]Statistics Iceland, Reykjavik, Iceland

The goal of the present study is to show that the standard methodology of official statistics may be enriched and its quality may be improved by applying the most up-to-date developments in theoretical and applied statistics. The progress of computational tools and resources in recent years make it possible to apply, in a routinely and timely manner, advanced tests, models and algorithms which have been previously restricted to academic or industrial purposes.

We focus on several stages of the statistical production and illustrate these ideas with concrete applications at Statistics Iceland regarding:

i. Bayesian rule extraction for data validation processes. Data application: comparing domain expert knowledge-based rules and rule mining results

ii. protecting (statistical disclosure control of) tabular data using Bayesian modelling and cryptography inspired methods. Data application: publishing detailed census grid data.

iii. quantifying uncertainty of machine learning classification algorithms while adding interpretability features for transparent and complete communication of results. Data application: random forests and neural networks for demography applications.

One of the main conditions for employing any of these methods is to be able to quantify their performance and to report on the uncertainty in results associated to: data variability, model fit/ complexity, distributional differences between training and predicting data, measurement/register errors or interactions of all these types of errors. We show that these tasks can be achieved for the most complex models and exemplify with results for the cases described above.

Assessment of disclosure risk on financial bases for individuals

Valentina Wolff Lirio[2], Rita Sousa[1], PhD Susana Faria[2]

[1]Banco De Portugal, Porto, Portugal, [2]Universidade do Minho, Braga, Portugal

The Microdata Research Laboratory (BPLIM) has the mission of supporting the production of research projects and studies on the Portuguese economy. For this purpose, BPLIM provides researchers with access to micro databases. Most micro databases, particularly financial databases, contain highly confidential information. The availability of databases with individual information, even if anonymized, requires a prior assessment of the risk of identifying individuals.

In this study, we intend to explore individual and global identification risk assessment methodologies in financial microdata bases at the individual level, with application in a real database.

When security meets privacy we enhance quality

Mr Jose Jabier Zurikarai, Jesús Nieto

[1]Eustat. Basque Statistical Institute, Vitoria-Gasteiz, Spain

Statistical confidentiality has always been a core value of official statistics. However, the big increase of data handling from both public and private sector has raisen the awareness of potential misuse of data; data belonging to individuals, companies or institutions.

Besides, connections between diferent data storage systems have made interconectivity possible but with a higher cost of risk. All the statistical proccess is carried out with connected systems and it opens the door to intrusion. Any public institution is nowadays under the menace of cyber attack. We must ensure that the potential risk is under control as it should be in any other public agency.

In statistics, we have a very high responsibility with our data and our systems. In fact our data don't really belong to us, but to the respondants and that's why we must keep them completely safe. And our data collecting systems must be available 24/7.

In order to ensure this, it is not enough any formal statement telling confidentiality is our core value. We must keep the condidentiality of the data and the proper functioning of our data collection systmes, but we also have to prove that we are doing that properly.

In Spain, the Public Electronic Services law sets the requirements for the Security Management of public institutions and as a result there is the National Security Scheme ruled by a Royal Decree. It contains among ohers, the Security Policy and Minimun Requirements, Systems auditing proccesses and the Categorization of the systems. Aplying the National Security Scheme means that any public organization must audit their Systems and check the level of compliance with the policies and requirements as well as the definition of its criticality level. After all the process an external authorized auditor will certify the level of compliance.

There are different methodologies to aply the National Security Scheme. Some of then are derived from international standards. For instance, the ISO/IEC 27000 family provides a set of standards for the Information Security Management (maybe it is the world best known).

The aim of this presentation will be how to aply Security Management to the Statistical Process, regardless of the metholodology or the standards aplied and focusing on wich are the real issues to certify a system as we have already done in EUSTAT

Session 3 - Geostatistics I, June 5, 2024, 11:00-12:30

Anonymization for integrated and georeferenced Data (AnigeD)

Dr Jannek Muehlhan[1]

[1]Destatis, Wiesbaden, Germany

Data-based information plays a central role in politics, business, science and public life. With digitization and the exponential growth of stored data, as well as new analytical methods such as machine learning, the possibilities for evidence-based decision making have expanded and evolved significantly.

A key challenge in integrating disparate data sets from different data custodians is the protection of personal privacy and trade secrets within organizations. This currently hinders both the wider use of data as a product and the use of integrated data in policy advice and scientific research. Methods for anonymization and statistical confidentiality face the challenge of finding a compromise. On the one hand, they need to protect the information of the data subjects, while on the other hand, the chosen methods should still offer sufficient analysis and information potential for the anonymized data.

Anonymization and confidentiality of individual data leads to information reduction.

In the past it has been shown that common anonymization strategies for individual data in economic statistics led to de facto or absolutely anonymized data sets, which were severely limited for scientific analyses due to the reduced or even distorted information potential. Anonymization and pseudonymization of data, which limits the risk of detection to an acceptable level while preserving sufficient analytical potential, is therefore essential for wider use and value creation.

The AnigeD competence cluster is part of the “Research Network Anonymization for Secure Data Use” of the German Federal Ministry of Education and Research (BMBF) within the framework of the Federal Government’s IT security research program “Digital. Secure. Sovereign”. The thematic focus, which is supported by various research strands, is the further and new development of strategies for the protection of personal and company-related data when using complex integrated data sets. Not only the integration of different data via direct identifiers or probabilities is relevant, but also the integration and linking of data via regional information in the form of georeferencing.

The talk will introduce the challenges in the anonymization process and its possible impact on the quality of statistical products and publications. It will further describe the main research strands within the AnigeD project: 1) Formalization of substantive criteria for the success of anonymization provided by the legal system, 2) Anonymization through synthetic data, 3) Anonymization of georeferenced data, 4) Evaluation of anonymized data according to formal criteria, and 5) Open software tools for anonymization.

A Flexible Approach for Quality Assurance in Geospatial Data: Mapping European Union Agricultural Census Data 2020

Nicolas Lampach[1], Jon O. Skøien, Helena Ramos, Rudolf Seljak, Renate Koeble, Marijn Van der Velde

[1]Eurostat E.1, , Luxembourg

The combination of geographical information and official statistics is elementary to support evidence-based policies and improve effective monitoring in agriculture. Yet, public access to these fully coherent datasets at European Union (EU) level has been impeded owing to lack of appropriate methods assuring data quality. We develop a new flexible approach to rule out the disclosure of confidential and unreliable information from 9.1 million agricultural holdings when mapping the 2020 data of the EU agricultural census. Our work discusses several aspects of the quality assurance framework ranging from coherence, consistency over time to accessibility of the new datasets. We present the first EU multi-resolution grid maps with confidentiality and reliability treatment of key agricultural variables (e.g share of organic farming, livestock density distribution, economic viability) that are based on the statistical data and geographical information from the 2020 EU agricultural census. Additionally, the open-source FSS R package is introduced to make the methodology readily available to the public, the research community, and national statistical institutes

Geospatial Enhancements in Statistical Production at Statistics Portugal

Rossano Figueiredo[1], Ana Santos[1], Magda Ribeiro[1], Francisco Sardinha[1]

[1]Statistics Portugal, , Portugal

Better statistical-geospatial data integration is a strategic priority in the future vision of the European Statistical System (ESS) to support evidence-based decision-making and policy lifecycle and forecasting. In this regard, the consistent production of standardised geospatial statistics is a key line of action in modernising official statistics in the European context through the mission to implement the Global Statistical Geospatial Framework attending the regional statistical-geospatial operating environment (GSGF Europe). Throughout the implementation journey of this statistical-geospatial framework, quality is one of the topics that needs to be more deeply addressed and studied by the statistical community, especially when handling geospatial data, processes and services raises confidentiality and privacy concerns.

Since geospatial aspects are not specifically mentioned in the common quality framework for the ESS, future enhancements on this topic are required to tackle this gap and further appropriately embody it within all four levels of quality assurance, from statistical regulations to methods and tools. To achieve this goal, National Statistical Institutes (NSIs) need to monitor, align and streamline geospatial processes and services into the statistical business processes and services that handle geospatial content, namely by adopting a quality management approach underpinned by quality feedback procedures and geospatial quality reporting guidelines, methods and tools. Thus, geospatial quality management needs to be formally embodied in the statistical business production model as an overarching process and high-level corporate activity to produce high-quality geospatial statistics.

Statistics Portugal (SP) currently documents its statistical production process through its Statistical Production Process Manual (MPPE), which is fully aligned with the Generic Statistical Business Process Model (GSBPM, UNECE), version 5.1. In this manual, SP has also added a third, more detailed level of the process, going beyond the phases and sub-processes, as it describes the tasks and responsibilities of the process. Furthermore, SP acknowledged the review of its MPPE as an opportunity to introduce geospatial enhancements by more extensively documenting geospatial- related activities, services, tasks and respective guiding descriptions and drafting a quality management framework for statistical-geospatial data integration. These added geospatial enhancements will ensure the quality management, assessment and improvement of input and output geospatial data, processes and services considering standardisation, harmonisation and other key aspects outlined in GSGF Europe. This work also intends to guide the experts regarding the geospatial dimension in statistical business production, promote a common understanding and build synergies on tasks and responsibilities to produce geospatial statistics in a systematic and consistent manner.

Enhancing housing surveys in Overseas Departments using Deep Learning and Satellite Data:

Mr Clément GUILLO[1]

[1]Insee, Baie-Mahault, France

The Localized Buildings Directory (RIL) comprises a list of housing units pinpointed by their x, y coordinates. While this directory is comprehensive in metropolitan France, it's not the case in the Caribbean, where there is a lack of reliable sources for this registry. An annual cartographic survey is conducted in the overseas departments (DOM), where surveyors identify certain housing areas on the ground to augment the RIL. The accuracy of the RIL is crucial as it directly influences the quality of the population census and subsequent demographic estimates.

Calibrating the necessary workload in each area presents challenges, especially with rapid new housing developments, including in informal zones. These changes can quickly overwhelm the workload allocated to surveyors. This challenge underscores the importance of using earth observation data. Through satellite imagery, labeled using INSEE's own data, deep learning segmentation algorithms are developed to detect housing. By comparing algorithm outputs at different times, areas requiring more intensive surveying efforts in a given year can be identified. The application of these methods is particularly pertinent in light of the current situation in Mayotte, where conducting a complete census is not feasible this year due to risk factors. Similarly, it is beneficial for newly established institutes such as the National Institute of Saint-Martin, which lack a housing database essential for assessing urban expansion. In both cases, population estimates for each area could be derived from the number of pixels identified as housing on satellite images.

The project utilizes images data from Pleiades (1m resolution) and Sentinel 2 (10m resolution) satellites. A primary concern is the availability of these data, particularly their acquisition frequency. Image labeling before algorithm training is a critical step, though it faces challenges due to the imperfections of current data sources. Historical RIL provides precise geolocations but does not reflect the actual area occupied by housing. On the other hand, IGN's (National Institute of Geographic and Forest Information) topographic database (BDTOPO) offers housing polygons that align with the concept of surface area but are difficult to date accurately.

A method for practical implementation is in the process of development. This method could be integrated prior to the cartographic survey, thereby enhancing the survey's organization and efficiency.

Practical approach in developing the integration of statistics and geospatial information through value chains and data quality

MS Rina Tammisto[1], Godfrey Lowndes, Mervi Haakana

[1]Statistics Finland, Helsinki, Finland

One challenge for the smooth integration and interoperability of statistical and geospatial information lies in ensuring a seamless flow of data between administrative sectors. Achieving interoperability necessitates cooperation, both between and inside of organisations, and a shared understanding of specific development measures and their benefits and agreed-upon implementation.

In the joint GSGF in Finland EU Grant project (2023 to 2025), Statistics Finland, the Land Survey of Finland, and the Finnish Environment Institute aim to lay ground for the future development of integration of statistics and geospatial information in Finland. The project seeks to establish a national adaptation of the GSGF model (GSGF Finland), aligning it with the Global Statistical Geospatial Framework (GSGF) and its European counterpart (GSGF Europe).

Adapting the frameworks to the national level requires more practical tools; as part of our approach, the project is piloting value chains and utilising the quality framework. This work started by conducting a current state analysis, focusing on a case study on geocoding process for Statistic's Finlands Business Register. The evaluation encompassed the statistical production process from data collection to dissemination of statistical products. The evaluation was grounded in the concept of data value chains, tracing information flow of geospatial data from one actor to another and across different production phases. Notably, the innovation emerged from recognizing the significance of quality as the value chain unfolds.

In this pilot, the national data quality framework served as the quality model. Its eleven perspectives, ie. correctness, accuracy, consistency, currentness, completeness, traceability, understandability, compliance, portability, user rights and punctuality, were used in analysing the quality changes in the piloted process.

In the presentation, we will shed more light on the evaluation of the geospatial-statistical production process from a quality perspective, the adapted value chain concept behind it, and initial thoughts on how the results can be utilised as a basis for development roadmap.

Session 4 - Quality management, June 5, 2024, 11:00-12:30

Management of quality in a changing data ecosystem: the case of FAO

Mr Ngarsaim Espoir Beram[1], Ms. Valerie Bizier[1], Aida Khalil Clara

[1]Food and Agriculture Organization of the United Nations (FAO), ROME, Italy

The data ecosystem is being increasingly affected by the emergence of new data sources, reshaping statistical business processes, and data and statistics quality management. Indeed, these dynamics involve opportunities for statistical advancement (increased timeliness, granularity, strengthened decision-making) as well as concomitant risks (data privacy, data integrity and statistics quality, data access and sustainability of data sources) that need to be addressed.

To ensure that the FAO remains relevant in this rapidly changing environment, a paradigm shift towards much closer integration of data and statistics has taken place, in line with the Strategy for the Modernization of Statistics. This led to the establishment of a Data Coordination Group (DCG) with the mandate to ensure a strategic and technical coordination mechanism with recognized authority to make decisions of corporate relevance for data for statistics (e.g., big data) and statistics. The commitment for data and statistics integration is also embedded in the new Data and Statistics Quality Assurance Framework (SDQAF), which ensures a proper use of alternative data sources to fill data gaps and adherence to the highest quality standards of the resulting statistical products.

An additional corporate Standard on the Acquisition and use of non-statistical data sources for statistical purposes, has been developed to address quality aspects specific to different types of input data (big data classes), providing tailored recommendations and guidelines to statistical units. While focused on statistical purposes, this standard can be used to guide experimental statistics analyses, thus encouraging science and research. It also includes provisions on cases where FAO acquires data through a third party or partnerships, thus compensating for the lack of internal resources and promoting knowledge sharing. The Standards on Imputation and Quality indicators have also been revised to reflect the paradigm shift and include implications for the definition and computation of coverage errors and accuracy.

Lastly, the revised Quality Assessment and Planning Survey (QAPS) aiming to assess the compliance of FAO statistical activities with SDQAF principles was implemented in 2023. It provides insights about progress achieved on the quality at both process and output levels and highlights areas of improvement. Also, the systematic review of reference metadata against the recommendations of the revised Standard on Metadata dissemination contributes to improve the accessibility and clarity of FAO statistics.

This paper presents the tools and mechanisms put in place in FAO to ensure the production of high-quality statistical outputs while integrating non-traditional data sources into statistical processes.

Adapting Quality Assurance Frameworks to the Fast-Evolving World of Official Statistics

Mr. Martin Beaulieu[1], Ms Roxanne Gagnon

[1]Statistics Canada, , Canada

The world of official statistics has evolved rapidly in recent years. The need for more detailed and more timely data was already fueling this evolution, and this need has become even more acute with the COVID-19 pandemic. The proliferation of data sources and the development of new technological tools offer great opportunities for National Statistical Offices (NSO). These opportunities, however, bring with them new challenges that are not, or only partially, covered by National Quality Assurance Frameworks.

Although the latest edition of Statistics Canada's Quality Assurance Framework (QAF) is relatively recent (2017) and includes some elements related to the acquisition and use of administrative data, the QAF is primarily written with surveys in mind. For this reason, the QAF is currently under review, to take into account the issues and responsibilities that come with the use of alternative data sources and technologies such as artificial intelligence and machine learning. Assessing the quality of input data, evaluating and communicating the quality of statistics produced, the ethical issues that come with data acquisition and data responsible use, and the data stewardship role of NSOs are just some of the requirements highlighted in the current context. NSOs must adapt their practices to these issues in order to maintain public trust. Several NSOs, including Statistics Canada, have developed tools to define principles and processes to meet these requirements. The integration of these into the Quality Assurance Framework, the highest-level governance tool for quality management, will clearly demonstrate to Statistics Canada employees, data providers and users, and the public that the Agency is committed to meeting these requirements.

This talk will discuss how these issues will be addressed in the new version of the framework, with the aim not only of bringing it up to date, but also of ensuring that it remains relevant in the long term.

The French process-based quality approach, a key role for risk management at INSEE

MRS Pierrette Briant[1], Elodie Kranklader[1], Mahmoud Jlassi[1]

[1]Insee, Paris, France

Within INSEE, a variety of risk management systems exist, some mandatory due to our institutional framework, encompassing risk collection related to working conditions and IT security.

INSEE-specific systems, such as process analysis, play a key role in the Quality policy of the French official statistical system, aimed to « integrate quality into processes, enhancing their safety and efficiency ».

In 2020, INSEE's General Management prioritized securing 21 essential processes, ensuring uninterrupted service. Process quality analyses are considered to be the most unifying framework for dealing with the various aspects of process safety. The last peer review endorsed continuing these analyses, particularly for processes deemed 'essential.'

This article introduces the quality assurance and process safety framework established by the Quality unit. It involves collaborative efforts between process teams and Quality experts to:

- Define processes through GSBPM, fostering a comprehensive understanding.
- Identify and prioritize potential weaknesses or malfunctions based on their frequency and impact on INSEE.
- Develop preventive or corrective actions to address causes or consequences of risks (weaknesses or malfunctions) linked to the process.

The impact of this framework on process security and its role in INSEE's risk management will be assessed, including enhancing risk awareness and supporting security measures at operational levels. However, limitations exist, requiring certain risks to be addressed at higher levels beyond processes and other limitations due to resource constraints, limiting coverage of all 'essential' processes.

Recent examples of process analysis, such as calculating the quarterly unemployment rate from the European labour force survey and monthly turnover indices from administrative sources (process called « ICA »), along with the associated publishing system, will illustrate these points.

The French labour force survey has faced heightened vulnerability following the shift to multi-mode data collection, incorporating Internet-based methods. The process ICA is notably impacted by risks stemming from the corporate tax source, characterized by its intricate and constantly evolving regulations. Concurrently, challenges common to both processes arise from the dispersion of teams across different geographies, spanning from centralized general directorates to regional divisions.

Moreover, the COVID-19 health crisis has exacerbated tensions within teams operating with severely limited staff numbers.

I can see clearly now, outliers are gone - How to improve data quality in official statistics?

Sónia Mota[1], Susana Santos[1]

[1]Banco De Portugal, Lisbon, Portugal

The power of information resides not only in the data itself, but also in how fast you can obtain and transform the data, and how flexible the data are. This means relying on very granular data that is readily available. However, is there a risk that we are compromising data quality for timeliness?

To assess the quality of our statistics, we present a set of indicators that cover different dimensions of data quality, such as information consistency and identification of anomalous values. These indicators have been implemented to continuously assess data quality in the different stages of statistical compilation: they are used to evaluate consistency and correctness in the data that we receive, and in the statistics that are produced. They also check for consistency between different statistical outputs.

This paper not only explores different data management and assessment techniques but aims at inviting the reader to reflect on which standards we want to set for the next era of information.

Management of strategic, integrity and operational risks: the experience of the Brazilian Institute of Geography and Statistics

Mrs. Ana Cristina Martins Bruno[1], Mrs Paula Leite da Cunha Melo[1]

[1]IBGE - Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil

The Brazilian Institute of Geography and Statistics (IBGE) has been working with risk management since 2018. Using a methodology based on ISO 31000, IBGE works in annual management cycles, considering three categories of risks: strategic, integrity and operational risks. The risks management objects are the macroprocesses and processes of the institutional value chain or the strategic projects of the strategic plan (relevance), the recommendations of internal audit or external control bodies (criticality) and the institutional work plan (materiality). In line with IBGE's Risk Management Policy, the work involves the stages of identifying, analysing, evaluating, and treating risks. The result is recorded in the system and generates a report, validated by the Governance, Risks and Controls Committee (CGOV) and approved by the Board of Directors. The status of planned measures to address and mitigate risks is monitored quarterly. In line with the three-line model proposed by The Institute of Internal Auditors – IIA, managers (first line) execute internal controls and the risk management agenda, in compliance with their institutional duties. The second line, responsible for advising and methodological support for risk management, is carried out mainly by the Risk and Processes Management (linked to the Planning and Management Coordination of the Executive Board) and by the Integrity Management (linked to the IBGE Presidency). For risks related to information security, the Information Technology Department also plays a second-line role. Finally, the third line (Internal Audit), as part of the Internal Control System of the Federal Executive Branch, is responsible for evaluating the controls established and exercised in the various activities of the Institution, in collaboration with the General Inspectorate of the Union, Central Body of the Internal Control System. As a way of training employees to implement risk management in their organizational units, Risk Management Methodology Workshop classes are offered annually, remotely, lasting 40 hours. From 2018 to 2022, 19 objects were prioritized, 42 risks were identified, and 225 treatment measures were planned. In 2023, was identified the risk of having the institutional image damaged, considering as object the process regarding Management of External Communication and Institutional Image. Three risk response measures were planned. In relation to risks to integrity, the risks of moral and sexual harassment at work were assessed, taking as a reference the incidents recorded during the demographic census operation. Twenty mitigating measures were planned, which will be implemented in 2024 and monitored.

Change and innovation in a small National Statistical Office

Anton Karlsson[1], Hrafnhildur Arnkelsdóttir[1]

[1]Statistics Iceland, Reykjavík, Iceland

Statistics Iceland is a small statistical office with much of the same responsibilities as larger offices in the European Statistical System with regards to collecting, processing, and disseminating official statistics. This puts a strain on its resources and has bred a distinct type of culture within the office where often there is a one-to-one association between a statistician and subject matter. The main problem has been for a long time that this setup is very challenging when needs for new statistics arise or during times of staff turnover. In short, the setup of the institution was such that agility and flexibility was severely lacking, creating a difficult working environment during times of changes.

In 2023 Statistics Iceland was reorganized along functional divisions instead of dividing the organization by subject matter. This means that as before there was a specific department dealing with statistics within particular domains, now there is a special department working on data, another one with analysis and the third working with dissemination and communications. The fourth department works with finance. Three support units were also founded, IT, Innovation and Human Resources. The reorganization took note of both the Generic Statistical Business Process Model (GSBPM) as well as GAMSO (General Activity Model for Statistical Offices) with the aim of modernizing the statistical production process at Statistics Iceland and creating more agility and flexibility, for example in the case of being able to produce new statistics.

In the presentation we will provide an overview of the changes made and the challenges that Statistics Iceland has phased and are currently phasing. There will be three main points of emphasis:

1) How can innovation be effectively managed by a small statistical office; 2) When doing innovation (and in fact large scale changes like Statistics Iceland has gone through) how do you measure the impact and how can you assess quantitatively the effectiveness of the changes/innovation work; 3) What are the most effective way of breaking up stovepipes and encouraging teamwork, cooperation and the distribution of knowledge within the organization.

Session 5 - Combining and integrating Data source, June 5, 2024, 11:00-12:30

Enhancing statistical registers: innovative integration methods for Buildings and Population Registers

Dott. Damiano Damiano[1], Dr Enrico Orsini[1], Dr Andrea Pagano[1], Dr Armando D'Aniello[1], Dr Stefania Lucchetti[1]

[1]Istat, Rome, Italy

At present, the INS face the challenge of generating official statistics through administrative archives. Specifically, at ISTAT, the development of statistical registers has recently commenced. These registers aim to characterize and provide a snapshot of the socio-economic landscape of the country. One fundamental aspect in the construction of registers is to devise processes to create an integrated architecture among different statistical units. In this work, we describe the use of new innovative methodologies for integrating data from the building and dwellings register with data from the resident population register, with the aim of uniquely placing a family within a dwelling.

These registers are fed with data from the cadastral archive of real estate and buildings, as well as municipal population registers, and both contain information related to addresses with their respective geo-coding. The process begins with geo-coded addresses through geographic coordinates and involves considering two different deterministic methodologies.

The first methodology is based on determining the ownership of a single dwelling. This is done by comparing the proximity between the residential address and the dwelling address at varying levels of geographic precision. The second methodology is applied to all families that do not own a dwelling or own multiple dwellings. For these families, a matrix of resident families at an address is essentially constructed for all available properties at that address.

The methodology for to build a unique family-dwelling association from this matrix has been implemented by calculating a weight that measures the quality of the association. This weight is calculated based on some variables of both geo-spatial and non-geo-spatial nature. The application of this weight allows the use of the harmonized combinatorial optimization method that solves the assignment problem known as the Hungarian algorithm in polynomial time.

The algorithm solves the unique assignment problem applied to families and dwellings by maximizing the sum of the calculated weights. In practice, all possible family-property pairs are encoded through a bipartite graph, where vertices are the elements to be associated, and edges represent possible pair choices, and for each edge, we have the calculated weight.

Applying these methodologies has uniquely allocated the entire resident population to dwellings in the most effective manner possible. The result of this integration enables the calculation of various statistical indicators that were previously obtainable only through surveys.

Taking opportunities to plan quality - the development of a new Secondary data model for CSO Ireland.

Mr Ken Moore[1]

[1]Central Statistics Office, , Ireland

A new CSO Secondary Data Model for CSO, Ireland has been developed to streamline, standardise and where possible automate the process of acquiring, receipting, ingesting and accessing data sources used across the Office. The model was developed in response to the increasing volumes of secondary data sources being received by the Office from the national statistical system, the introduction of a new data hub to manage this data and the strategic move to a "Secondary data first" principle. The model initially focuses on how we manage, govern and reuse administrative data, privately held data and publicly available data within the CSO.

As part of the development of the data model, Quality division placed a key role in ensuring that we took to opportunity to build and plan quality into each step of the new model from data acquisition to data access. This paper will outline the improved quality management elements introduced as the model moves from development to implementation including:

- Developing a quality assessment protocol for examining the quality of inbound secondary data from the statistical system
- Providing for improved engagement with our data providers so we have greater clarity on the data quality checks they carry out and the metadata they associate with the data prior to transmission to CSO
- The introduction of standardised quality checks on the data once received in CSO for ingestion into the Data Hub
- Increased feedback communications with data providers so that common data quality issues are discussed and where possible resolved
- The provision of centralised data services so that commonly used data flows are managed consistently at an initial data processing phase – centralised edits/data matching and linkages and improved metadata standards.

The paper will also outline progress made to date, the challenges being experienced with introducing these changes and the next steps planned for the implementation of the new model.

An application of integrated statistical registers to produce new and systemic indicators on small territorial units

Dr Andrea de Panizza[1], Dr Stefano De Santis[1]

[1]Istat (statistics Italy), Rome, Italy

This paper proposes an application of integrated statistical registers to produce new and systemic indicators portraying small territorial units, which can hardly be addressed by means of surveying. Indeed, statistical registers derived from administrative sources provide a full (census-like) coverage on the population under scrutiny, allowing to get (longitudinal) information on all the individuals and to define à la carte domains. This comes at the price of major limitations in the information available in terms of underlying variables and their robustness, not to mention constraints on dissemination that may result from confidentiality rules. Also, the development of registers is still limited in many European Countries.

Like many other European NSOs, Statistics Italy (Istat) started developing a business register in the early Nineties. In the following decades, through a not always easy dialogue with other public administrations, and borrowing from the experience of the Nordic countries, it managed to establish an ecosystem of registers addressing enterprises, a Base register on Individuals, a LEED for non-farm business employment, and some extensions to registers from additional administrative sources, with the objective to set up and maintain an Integrated Registers System (SIR).

Until now, registers had mostly an infrastructural aim, to serve as basis for statistical production and analysis, and – to a limited extent – to produce register-based statistics and indicators drawing on one or another specific register. In the recent past, two releases of multi-register indicators provided some territorial information (up to the municipality level) on enterprises and the workforce.

The two case studies proposed here show how Official statistics can provide insights even at very fine-grain territorial level and support policies. The first, focuses on the provision of a multi- dimensional perspective on Labour market areas (LMAs), joining information on individuals (employment, education, age, gender, household characteristics) and on productive units. An array of indicators is presented, showing their mutual relationships, and how they can define patterns of LMAs, which add to previous classifications made by Istat. Also, an example of sub-communal level

information is provided for the three Italian FUAs of Rome, Milan, and Naples, considering a proxy for NEETs jointly to household characteristics.

These exercises are meant to serve as a basis to discuss the possibilities and limitations of this approach. Results are to be considered experimental, as well as an anticipation of future work, that could also be performed in collaboration with other NSOs within the ESS.

Strengthening Modernization and Innovation – a holistic approach

Mr. Paulo Saraiva[1], Prof. Francisco Lima, Mr. Jorge Magalhães, Mrs. Sofia Rodrigues, Mr. Almiro Moreira

[1]Statistics Portugal, Lisboa, Portugal

To face the challenges of the new ecosystem in which national statistical offices are now operating, Statistics Portugal has been developing several initiatives to spur innovation and efficiency. The organizational, technological, methodological, communication and human resources initiatives are important on their own, but they are much more impactful when worked together. In this paper, Statistics Portugal will present the main initiatives and challenges it has been working on.

From 2018 onwards, the development of National Data Infrastructure strengthened the appropriation and use of administrative data and other sources and created a single point access to several types of data and made it available to multiple purposes and users, while ensuring the protection and integrity of data. In 2019, Statistics Portugal adjusted its internal organization to strengthen the capacity for data management and analysis. The Statistical Production Process Handbook was updated in 2020 to, among other reasons, introduce the National Data Infrastructure in statistical processes and, in that same year, created a dedicated unit – Administrative Data Unit – centralizing the quality procedures in the treatment of administrative and other data.

Investments in IT and knowledge strengthened the technological infrastructure, the acquisition of new technologies, new skills and techniques, and the introduction of innovative processes. Specific training also plays a crucial role, and several contracts and collaborations with academia are put in place. Resident researchers help the introduction of new developments, namely in the processing of administrative data.

Communication is also a key factor. In what concerns working together with data holders, Statistics Portugal developed a new approach with the personal engagement of the Board and the involvement of a multidisciplinary team in the negotiation of the acquisition of data. The relationship with data providers has long been a focus of Statistics Portugal. Feedback reports for administrative data providers, like those that Statistics Portugal has implemented for business survey respondents since 2013, are under development.

Finally, worth mentioning the creation of StatsLab, an area available on Statistics Portugal's website with new statistical products, presented under development, aiming at taking advantage of new data sources and new methodologies.

Integration of income administrative data into the Portuguese household income distribution: a first national experience using employees' income tax data

Mr David Leite[2], Mrs Eduarda Góis[1], PhD Carlos Farinha Rodrigues[3], Mr Daniel Gomes[1]

[1]Instituto Nacional de Estatística (Portugal), Lisboa, Portugal, [2]Paris School of Economics, Paris, France, [3]Lisbon School of Economics and Management, Lisboa, Portugal

The Portuguese results on poverty and economic inequality are based on a 4-year longitudinal sampling survey of households and their members, carried out every year, which is part of the EU-SILC programme since 2004. The survey collects data on qualitative aspects (e.g. health, housing and material and social deprivation) where alternative sources are difficult to find, and on quantitative aspects related to monetary income where, on the contrary, alternative sources are known, namely tax and social security data.

Until 2021, monetary income data, including employees' income, was obtained exclusively through direct collection from the selected households and individuals, with proxy responses being accepted in situations of individual temporary absence or in incapacity, and frequently without consulting information organised for tax purposes, even though the questionnaire includes the possibility of responding by transcribing data from the annual tax return: that, combined with sampling weighting, increases the possibility of deviations vs. nearly exhaustive administrative data.

From 2022 onwards, considering that the integration between personal income tax data (IRS) and Survey data can be ensured for most incomes, even though not covering for taxpayers exempt from submitting the annual return, and taking advantage of previous studies regarding other data collections about wages and salaries, the Survey started to integrate administrative data on employees' income collected by IRS Model 3, Annex A, in order to improve the consistency and quality of information before deduction of taxes and social contributions.

Overall, compared to data relying exclusively on the original Survey data, the new income distribution that includes imputation of administrative data is more homogeneous, with both a lower average employees' income and the corresponding standard deviation. The adjustment also has a significant impact on employees' income deciles, with a significant change in the first six deciles and on the inequality of the employees' income distribution.

Integration of administrative and survey data in a Short-Term Business Statistics with statistical learning algorithms

Dr Sandra Barragán Andrés[1], David Salgado[1], Sergio Pardina[1], Ester Puerto[1]

[1]Statistics Spain (INE), , Spain

The use of administrative data (and digital data sources) is a must not only for the modernization of the production of official statistics but also for keeping relevance in the new data and AI international ecosystem. These new data sources must be integrated with survey data. However, as it is widely known, this incorporation of new data sources does not come without quality challenges. By and large, the direct substitution, use, or aggregation of administrative data cannot be undertaken since errors both in the representation and measurement lines arise even when formerly they were under control using only survey data.

Representation errors (especially regarding coverage) arise because of unit misclassification errors and other factors. Validity, measurement, and process errors easily occur because of the administrative (non-statistical) purposes of these data sources. Overall, the fact that the data generation mechanism lies outside the control of the statistical process revives both non-sampling errors (validity error, for example) and inferential challenges (non-ignorability, for instance).

We present a proposed end-to-end statistical production process integrating administrative data with survey data in a probability sample. Synthetic values produced from a tax source are computed using a statistical learning model so that validity and measurement errors can be a priori identified and kept under control. The statistical learning algorithm learns from past and present survey and administrative data producing high-quality values for non-influential units, which paves the way to reduce response burden. Influential units are still integrated using survey data.

We share a proof of concept on the monthly Services Sector Activity Indicators using VAT data. We discuss challenges regarding the quality of both the sampling design, the statistical model, and the training data.

Session 6 - Governance and Quality, June 5, 2024, 14:00-15:30

An Approach to Support Data Stewardship through the Implementation of Data Standards across the Irish Data System

Ms Laura Reddin[1]

[1]Central Statistics Office, , Ireland

Data stewardship ensures the ethical and responsible creation, collection, management and use of data, which will improve the quality of data. In addition, public trust will be strengthened across the Irish State. Robust, well-structured data puts decision makers in the best position to make well informed decisions, thereby serving our society. The development of data standards supports data being collected and published in a consistent, harmonised manner. Harmonised data standards provide guidance on best practices around collecting and producing quality data. It also ensures the availability of quality information across the Irish civil and public service.

In order to advance its strategic priority of advancing data stewardship, The Central Statistics Office (CSO), Ireland has begun work on developing and implementing a broad suite of data standards.

This paper will describe the steps taken by the CSO in order to successfully develop and implement the use of data standards across the Irish Statistical System. This involves the research around individual concepts and consultation with key stakeholders in the early development stages of each standard concept being developed. The paper will continue to outline the promotion of established data standards, along with the governance and monitoring of the use of the standards.

It will also outline the approach undertaken by the CSO in the implementation of data standards and it will conclude with detailing the obstacles encountered and the proposed plan to overcome expected barriers.

Risk management through internal audit as a tool to ensure the quality of official statistics - The Statistical Head's role

Dr Stamatios Theocharis[1]

[1]Ministry Of Interior - Greece, Athens, Greece

The production of official statistics according to the European Statistics Code of Practice adopted by the ELSTAT through the Greek National Law, must be based, inter alia, on the guarantee of it's quality. Procedures followed by the organization in order to produce statistics are usually complicated and varied, depending on the type of raw data and their purpose. Towards the achievement of this goal, risk assessment and analysis through the internal audit system adopted by the organization are necessary and useful tools.

Risk management is a rapidly developing field that concerns various areas of activity of organizations as well as the statistics production, and there are many and varied opinions on the content, the way of conducting it, and also on it's specific purpose. According to the raw data collected on a case-by-case basis, the expert must use his judgment, experience and the opinions of those involved with the subject, because each project is separate and contains its own particularities.

In this work we examine the potential of the adoption of an internal audit system and risk management procedures in order to ensure quality by producing official statistics. Basic partners of this effort are the organization statistical head and the relevant internal audit unit. The purpose of this work is : a)to investigate the existing governance structure of the organization, b)to map the procedures related to the production of official statistics by the Ministry of Interior, c)to identify any risks that threaten the effective production of official statistics through the development of specific questionnaire addressed to the producers of the statistics and d)to make proposals that will help to ensure the improvement of the quality and the efficient production of statistics..

In the context of the implementation of the provisions for the production of official statistics, the improvement of the quality of statistics and their certification as official statistics in the Ministry of Interior, a Statistical Head has been appointed as well as a relevant Working Group. This structure also seeks to develop a network of two-way communication among the parts of the system as well as an horizontal and uniform application of directives and observations, regarding the application of the principles of the Code of Good Practice for European Statistics. In this work we analyse how this structure works, and also the potential risks and opportunities associated with improving the quality of statistics.

Independence of official statistics – declarative abstraction versus political reality

Mr Rudi Seljak[1], Bojan Nastav

[1]Private Expert In Official Statistics, Ljubljana, Slovenia

Mounting challenges in modern world, comprising conflicts, migration, environmental disaster, with increasing impact on human lives and sustainable development, require standardized, robust, and reliable evidence base to inform policies tackling the issues, and measure their effectiveness. In modern state systems, the independence of official statistics is thus one of the driving foundations of the democratic political system. Several different tools have been put in place to establish and maintain this independency. Fundamental Principles of Official Statistics by the United Nations and European Statistical Code of Practice (ESCoP) are only two examples of such tools at international or regional levels. Statistical laws or similar legal acts, established by many countries, explicitly or at least implicitly address this independency. Moreover, the issue has been frequently addressed in multitude of academic articles and professional papers in recent decades. While on a declarative level, these tools offer a strong defence against the arbitrariness of political power centres, it is also necessary to ask ourselves, what the real power of these tools is when faced with a concrete situation of political interference with, or influence on official statistics.

In this paper we target a very important stone in the mosaic of professional independence of national statistical organisations (NSO): the issue of dismissal and/or appointment of the head of NSO. This issues is directly addressed in Principle 1 of ESCoP. However, facing political reality, thus-defended professional independence of NSOs turns out to be more abstract than envisaged in the Principles or the Code.

In the paper we use example of dismissal of Director General of Statistical Office of Republic of Slovenia in May 2020 to discuss the power(lessness) of the declarative acts which are supposed to represent a shield against the arbitrariness of political decision-makers. Dismissal, itself having attracted the interest from the general and professional public, was also discussed at two national court instances. The (national) judgments at both instances appear not to share the declarative and intentional institutional independence of NSO; and may have reinforced the movement set in motion back in 2020, potentially damaging the reputation of and trust in official statistics in Slovenia. While recent Peer Review of Slovenian national statistical system identifies compliance issue with CoP, its impact remains to take effect. We aim to show how it is necessary to direct more efforts to address this immeasurably important question.

The evolution of the supervisory reporting framework for the EU banking sector

Mr Paolo Poloni[1]

[1]European Central Bank, , Germany

Supervisory data are typically not conceived for statistical purposes or considered as “official statistics”, but they are disclosed to the public, either directly by supervised institutions or indirectly by the competent authorities. This is because Pillar 3 of the Basel framework on banking supervision aims to promote market discipline, whereby market participants monitor the risks and financial positions of banks and take action to guide, limit, and price banks’ risk-taking, to safeguard financial stability. The disclosure of supervisory data is therefore a public good. In addition, supervisory data can be a reliable source for official statistics such as financial accounts. On the other hand, the nature of supervisory data is different, and its quality is subject to a robust assessment framework, which has distinct peculiarities compared to standard official statistics.

The aim of this paper is to analyse the EU supervisory reporting framework from an institutional and policy perspective, in view of its potential and desirable evolution over time, including its potential integration with the statistical framework.

The paper will be articulated into three main parts. Firstly, it will describe the current EU institutional settings, including the role of the EBA reporting framework and the SSM role focusing on the data quality assessment framework and the publication of supervisory statistics. The current shortcomings will also be analysed.

Next, the paper will describe the possible future evolution in the future, triggered mainly by two elements. The first one concern the recommendations of the EBA feasibility study on integrated reporting (common data dictionary, joint governance, central data collection point). The second element is represented by the European Commission’s strategy on supervisory data in EU financial services.

Finally, the paper will propose several policy principles that should inspire such evolution under certain constraints, including the application of BCBS 239 principles to supervisory reporting, the interlinkages between supervisory reporting and Pillar 3 disclosure, the compliance with Basel core principle 10 on supervisory reporting, and the management of ad-hoc reporting requirements.

Studying the Relation Between Data Quality and Trust in Official Statistics

Ms Joanna F. Lineback[1], Dr. Benjamin Reist

[1]United States Census Bureau, Washington, United States

In this presentation, I discuss the relation between survey participation and trust in official statistics, sharing findings from a recent study on factors influencing survey participation. Response rates, a necessary component of data quality, continue to fall for government surveys. Recent studies provide insight into this phenomenon. In a nationally representative survey that tracked public awareness of the 2020 U.S. Census, respondents were asked about their plans to participate in the census as well as a variety of other questions we thought might be associated with survey participation. Topics included the state of current political and economic environments and civic engagement. Additionally, there were two attitudinal questions, one about trust in federal statistics and one about the fear of the government's misuse of survey responses. One notable finding was that, even after controlling for key demographics, trust in government statistics was negatively associated with survey participation. In the first part of this presentation, I discuss this and related findings. In the remainder of the presentation, I discuss the implications of these findings on response rates as a necessary component of data quality. Common perception is that trust in government is the driving factor of survey nonresponse, but here we have empirical evidence that trust in official statistics is an important predictor of survey nonresponse. These are different concepts, and each has different implications for survey participation.

Session 7 - Quality of administrative data, June 5, 2024, 14:00-15:30

The National Data Quality Framework for Public Sector Data

Ms Kirsti Pohjanpää[1]

[1]Statistics Finland, Helsinki, Finland

Statistics Finland led the development in project on Opening Up and Using Public Data (set up by the Ministry of Finance 2020-2023)

The amount of data is ever-increasing, and so are the data opened and governed by government institutions. It is essential to recognise and describe the quality of the public sector data uniformly. Moreover, the more we wish to reuse the data for some other purposes than the original one, the more important it is to verify the quality of the data.

Questions about reliability, quality and usability of data are crucial for decision-making process, too. The more there is information, the more important it is to assess quality of the data. As the question about the quality of the data is fundamental, widely shared rules and criteria are needed.

The data quality criteria and the indicators together with models and tools supporting their implementation and management form together the National Data Quality Framework.

The data quality criteria and indicators were created in cooperation with a substantial number of stakeholders. They were also tested in several pilots, and they represent the present understanding about data quality. What does data quality mean?

We developed eleven criteria for answering that question. They are divided for three groups. In the first group there are criteria which answer the question: how well information describes reality.

Second, how has the information been described, and criteria in the last group described how can you use the information.

Implementation of data quality criteria requires actions at three different levels at the same time. Actions are needed as well as at national level, organisations, and datasets.

Improving the knowledge of the agritourism sector through integration between survey and administrative data

Dr Roberto Gismondi[1], Maria Grazia Magliocchi, Francesco Truglia, Filippo Oropallo

[1]Istat, , Italy

In Italy, agritourism is a particular secondary activity carried out by about 25 thousand agricultural holdings. Between 2010 and 2020, the number of agritourist farms increased of about 20%, while the overall number of farms decreased more than 30%.

ISTAT (Italian National Statistical Institute) carries out a yearly survey on agritourism, which collects some structural data as agricultural surface, number of beds, number of place settings, services offered to costumers. Available data do not include any economic indicators, as incomes, costs, investments, added valued and profits, as well as any information on employment

The estimation of economic results was possible through the integration between the survey microdata with administrative microdata derived from various sources. ISTAT uses various administrative sources in order to update the Farm Register, which includes all the active farms existing in Italy (more than one million). The farm register contains data as the main economic activity, the technical-economic orientation, the standard production, the main crops cultivated, livestock, the size of the farms and the location. Moreover, additional economic and employment indicators have been linked, at the microdata level, through integration of the following administrative sources: national social security, declarations relating to self-employed agricultural workers and agricultural labour, tax declarations and VAT returns, foreign trade data. This exercise was carried out for the years 2019, 2020 and 2021.

On the basis of data matching, several analyses concerning the agritourism economic performance were possible, at a very detailed territorial level. In particular, main results showed how, on average, agritourisms have higher productivity compared to the other agricultural holdings and play an important role as regards the reduction of the historical gap between Northern and Southern Italy, since they are widespread in every Italian region. Through this integrated analysis, users can be provided with much more data than those available according to the statistical survey.

Census Data and Administrative Data in Portugal: Results and Challenges

Joana Araújo[1], Cecília Cardoso[1], Claúdia Pina[1], Paula Paulino[1], Prof Pedro Raposo[1]

[1]Ine, , Portugal

In the context of democratic governance, census data serves as an indispensable foundation, shaping policies, planning, funding of the municipalities and resource allocation. The creation of the Resident Population Database (BPR) is the central bone of the administrative census. This is crucial and challenging project for Portugal in a particular context related to the nonexistence of a population register and the lack of a unique identifier to link the various administrative sources.

The BPR and the Administrative Census project falls within the scope of the National Data Infrastructure (IND), which embodies Statistics Portugal's strategy of integration and creation of value for society from different data sources.

Drawing from a unique and valuable setting created by the results of Censos 2021 (based on a full field enumeration process), in this paper we evaluate the quality of the administrative data incorporated in the BPR framework and we highlight the strengths and limitations.

The findings reveal that the information gathered in BPR is deemed plausible and comprehensive, underscoring the success of the transformation from "tradicional" census data to administrative census data in Portugal. Among the positive results, we observe an high level of convergence between the two databases in several dimensions. For instance, it is possible to identify similar age profiles by gender and region. Furthermore, when microdata was linked between the two databases, there was a match of 92% of individuals and equal completion rates above 97% for demographic variables (Sex, Age, Marital status, Birthplace and Citizenship).

The paper also unveils challenges, particularly in terms of population coverage. In this regard, we conclude that the BPR underestimates some the population sub-groups (e.g. foreign nationals and some specific age groups).

An innovative approach to improve the quality of the families and nucleus types reconstruction in Italy

DR Rosa Maria Lipsi[1], Anna Pezone[1]

[1]Italian National Institute of Statistics (ISTAT), Rome, Italy

Since 2018, the Italian National Institute of Statistics (ISTAT), as other European countries, moved from the traditional ten-year “door-to-door” census to a yearly “register-based” system (the Permanent Population and Housing Census) in order to produce annual detailed statistics, to enrich the supply & quality of statistical information, to reduce the statistical burden for respondents and the costs by the community. This transition represents a great innovation. However, every ten years, according to European regulations, EU Member States must send to Eurostat information on the main characteristics of their resident population and their social and economic conditions at national, regional and small areas levels, regardless of how they collected them.

A multisource approach, based on a combination of administrative data, registers (as RBI – Based Register of Individuals, RSBL – Statistical Base Register of Territorial Entities) and surveys data, has been used to provides information on Italian population and housing census for the 2021, as required by the EU regulation 2017/712.

The number of families and their characteristics is one of the mandatory information, but also one of the most complex aggregates to detect, validate and disseminate. The main problem to solve is the correct identification of families, as well as Nuclei and Family types. The reconstruction of the family in its internal composition is possible through the correction of individual variables as the kinship, the age, the sex, the marital status, the year of marriage or civil union, analysed in relation to those of the other family members. In 2021, the most important set of the above variables becomes from ANPR (The National Registry of the Resident Population) that contribute to improve the quality of the RBI aiming to produce, at macro-micro level, official statistics on households. In order to obtain these statistics an efficient and efficacy strategy has been planned involving innovative generalized solution of E&I system and specific adaptations, for census demand, of the “Family procedure” for the reconstruction of the Nuclei and Family types, usually used for social surveys.

Our goal is to describe the whole process to produce statistics on the families and their characteristics by using RBI information, ANPR and survey data in order to highlight the main advantages of the innovative integration process and the quality of the data, by suggesting, for the future, how optimizing the process in terms of time and performance too.

Measuring the quality of administrative sources: at macro level with novel indicators and micro level with distributions comparison

Alicia Nieto[1], David Salgado[1], Dr Sandra Barragán Andrés[1], Soledad Saldaña[1], Alba Rodríguez[1]

[1]S.G. for Methodology and Sampling Design, Statistics Spain, , Spain

In the production of official statistics there are three main data sources: surveys, administrative registers, and (privately held) digital sources. The use of administrative sources is lately increasing, however there is a lack of control in the quality of these new sources. The administrative data can be used in different ways, the most challenging is to use them as primary source of data, directly or indirectly to compute the target aggregates.

The advantages of administrative data as a primary source are widely known improving different quality dimensions (reduction of response burden, cost savings, increase of granularity, etc.), but the disadvantages must be considered carefully. In terms of the representation and measurement lines in the Total Survey Error paradigm and the Two-Phase Life-Cycle model by L.-C. Zhang, errors both related to units and to variables are present. Coverage errors arise when identifying units in the target population and validity errors proliferate because of the differences between concepts for statistical and administrative purposes. Therefore, a need to measure the quality of input data emerges as a consequence of the data generation process lying out of the control of NSIs.

In official statistics several quality and performance indicators are used but the focus is on measuring the quality of the output, and most of them have been designed to be used when using survey data as input. So, there is a need to broaden the list of quality indicators to provide room for those quality measures of multisource statistics and even more in the case of statistics based only on administrative data.

At Statistics Spain we are carrying out an exercise to measure the quality of the administrative data used in several short-term statistics of different domains/characteristics. In this work we present the proposal to measure the quality of the input with some indicators for the administrative data source. Moreover, we take advantage of the access to both administrative and survey variables for a part of the sample to directly compare the distributions of the target variable under study.

The ultimate goal is to provide objective measures of the direct use of administrative values without further treatment to gain some knowledge about the quality of the final estimates in comparison with fully survey-based traditional results. This analysis may help us decide regarding the need of treatment of administrative sources to keep under control their disadvantages and ensure the quality of admin-based final outputs.

Administrative Data Quality challenges through the Lens of E-Invoice

Mr João Poças[1], Mr Bruno Lima[1], Mr Salvador Gil[1], Ms Paula Cruz[1], Ms Sofia Rodrigues[1], Mr António Portugal[1]

[1]Statistics Portugal, , Portugal

In Portugal, the Tax Administration has instituted a compulsory electronic invoicing system for reporting all invoices associated with commercial transactions. Statistics Portugal receives this data monthly, providing aggregated taxable amounts per fiscal number for both the issuer and the acquirer. In certain cases, particularly in specific months, the data might be incomplete or contain errors. This leads to the detection of numerous missing or incorrect values that require imputation, presenting a complex and challenging task due to the substantial volume of data.

Despite these challenges, the received data is intended to be accessible to internal users. Before being made available, e-invoice data needs to be processed and validated to ensure its quality, reliability, consistency and completeness. In making this data structure available on a monthly basis, it is necessary to use automated procedures that guarantee the reproducibility of an immediate processing such as the identification and imputation of anomalies.

Standardized procedures have been introduced at the loading stage to verify the data structure, including checks on the number of records, validation of fiscal identification numbers, pseudo-encryption of identifiers, and normalization of attributes. Additionally, there are processes in place to ensure the integrity and consistency of the data when compared to other datasets (either administrative or survey data).

However, these processes are insufficient to support the production of quality statistics. To fulfill statistical requirements and achieve usability, it is necessary to enhance the substantial monthly dataset, consisting of over 90 million records, in terms of quality within a timeframe of less than 12 working hours. This demands extensive coordination and collaborative efforts among various units. Following a predetermined workflow of interrelated tasks, it is essential for Statistics Portugal technicians to work together seamlessly to ensure the smooth running of operations.

The centralised and comprehensive implementation of these procedures, applied to different data sources, promotes a significant improvement in the statistical quality of administrative data. These efforts not only reduce the statistical burden but also contribute to a more coherent and valuable National Data Infrastructure.

Session 8 - Quality of web-based data, June 5, 2024, 14:00-15:30

Web data vs. traditional data sources on real estate – augmenting official statistics

Dr Klaudia Peszat[1], Emilia Murawska[1]

[1]Statistics Poland, , Poland

This paper discusses the potential of web data from real estate online sales and rental offers to augment official statistics. The research is part of experimental studies carried out within the ESSnet Web Intelligence Network project, whose aim is to explore the possibility of producing new and complementing existing statistics with web data.

Online sales and rental offers enable an ongoing observation of the real estate market. The information derived from the web may serve as the basis for flash estimates of many indicators, e.g. price indices. It may also be used for the production of new indicators, such as the standard of dwellings, availability of parking spaces, security and additional amenities, or the observation of the phenomena which are not covered by traditional data sources. In Poland, for example, there is no high-quality reference source on the rental market due to the lack of official registers. In addition, tax data covers only the aggregated information on the rental income, which does not allow identifying any specific information about the apartments the tax corresponds to. Meanwhile, as the outbreak of the war in Ukraine in February 2022 and the inflow of a large number of immigrants to Poland, looking for short-term accommodation demonstrated, more specific and less aggregated data on the real estate rental market have been increasingly needed.

It appears that web data could fill this information gap. However, the use of web data in official statistics brings also a number of methodological and quality-related challenges, such as a difference between offer and transaction prices, lack of coverage of the entire population, duplication of offers within a single portal and across different portals, missing values. Moreover, it requires addressing horizontal issues related to the stability of a data source and changes in the structure of websites.

The paper presents the comparison between web data and official statistics based on surveys and administrative registers for real estate sales market in Poland. The results demonstrate a high degree of convergence in price trends, making web data a promising data source for real estate statistics to generate flash estimates and complement existing statistical information.

Improving quality of web-based data: Human Role in production of Consistent Labour Market intelligence

Mr. Vladimir Vladimir[1], Jiri Branka[1]

[1]Cedefop, Thessaloniki, Greece

In today's dynamically changing labour markets, real-time skills intelligence supported by official statistics is pivotal for shaping effective employment and education policies. Tapping up on web data creates as a powerful avenue to swiftly understand employer demands. Collaborative effort by Cedefop and Eurostat throughout the Web Intelligence Hub (WIH) focuses on utilising web-derived insights to develop official skills statistics. However, the vast linguistic diversity challenges the consistency and cross-country comparability of this data. This highlights the vital need for human input in this process.

This article draws on WIH's work to critically examine abilities of automatised tools to extract coherent and cross country comparable information on skills from web based data. The main departing point of this work is that while the European multilingual classification of Skills, Competences and Occupations (ESCO) provides a comprehensive skill framework, diverse linguistic representations cause challenges to accurate and comparable skills extraction. The article presents the ways how human intelligence was used to refine and standardize skill extraction.

Building on WIH activities to introduce consistent and cross-language comparable extraction of skills terms, this piece advocates for the pivotal role of human judgment and expertise in enhancing the precision and uniformity of data extracted from diverse linguistic sources. It presents validation mechanisms and employing human judgment within automated frameworks and envisions collaborative paradigms integrating human expertise with technological algorithms.

Statistical scraping: informed plough begets finer crops

Dr. Olav ten Bosch[1], Alexander Kowarik[2], Sónia Quaresma[3], David Salgado[4], Arnout van Delden[1]

[1]Statistics Netherlands, The Hague, Netherlands, [2]Statistics Austria, , Austria, [3]Statistics Portugal, , Portugal, [4]Statistics Spain, , Spain

The use of web data for official statistics has been studied extensively in recent years. It is widely recognised that combining such data with traditional inputs improves or speeds-up statistics or opens up possibilities for new indicators that couldn't otherwise be measured. There are successful examples in price statistics scraping web shops, enterprise statistics scraping business websites and social statistics using social media. However, there are also challenges: web data are volatile, rich of biases and of unknown quality, to name a few. But the biggest problem is methodological: how to link, map or cluster the web data not designed for statistics, which uses messy real-world language, into statistical units or aggregates needed for official statistics.

Traditional approaches often involve collecting raw data from various online sources exposing information related to the statistical concept of interest. Over time, new data sources are added resulting in a bulk scraping approach. In contrast, statistical scraping starts at the existing knowledge that statistical offices already have. The web is queried with an identifier, name, category, or statistical definition, so that the result can always be linked back to the statistical context. It's like performing an automated survey on the huge web with messy but linkable results. Or like a farmer that knows from experience what fertilizer and harvesting strategy makes the finest crops. This strategy helps noticeably to cope with representation errors.

A notable example of statistical scraping is in business register enhancement, a subject explored in the ESSnet WIN project. Starting from information in the business register the web is searched to identify digital traces associated with certain statistical units. These traces are then employed to enhance administrative or statistical variables such as NACE codes. Another example can be found in price statistics. Statistical scraping in this context implies a search for well-defined products from the basket to collect high quality price observations for those products. An untapped area with regard to statistical scraping, but where it could yield valuable insights is job market statistics.

In this presentation, we sketch the concept of statistical scraping, a methodology that may complement or in some cases replace bulk scraping methods. We explore its applications, strengths, and limitations, evaluating its impact on data quality. Finally, we ponder the potential implications on ongoing and future projects that utilize web data sources for official statistics, ensuring the preservation of the high-quality standards expected for official data.

Innovation and research to foster quality in official statistics

Loredana De Gaetano[1], Gabriella Fazzi[1], Massimiliano Amarone[1], Serena Liani[1], Samanta Pietropaoli[1]

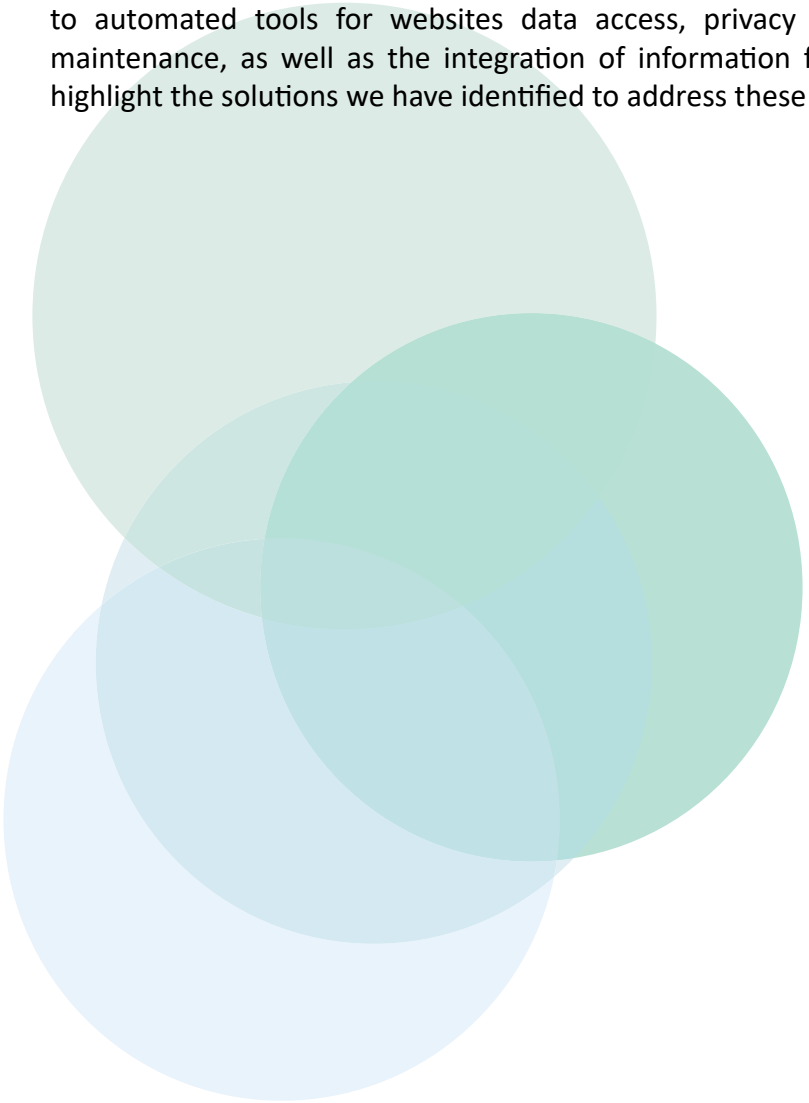
[1]Istat, Rome, Italy

Web scraping is a technique to automatically extract data from websites using specialized software. It is increasingly used in official statistics to replace traditional techniques of data collection because it allows reducing costs, expanding coverage, and improving the quality of data collected. Since 2014, the Italian National Institute of Statistics has been working on collecting websites data for consumer prices statistics, particularly those on the prices of electronics (currently no longer used), transport by train and airfares, electricity on the free market, town gas and food delivery.

This paper aims to explore the possibilities and constraints associate with web scraping for gathering data on prices of package vacations provided by both national and international tour operators.

Given the variety of data sources requiring consultation, customizing programming languages become challenging due to the multitude of websites used for information collection. Additionally, ensuring their ongoing maintenance can be very highly time-consuming task.

We aim to provide a comprehensive overview of the methodological, technical, and legal challenges associated with web scraping for official statistics. Specifically, our focus is on issues pertaining to automated tools for websites data access, privacy concerns, programming, and software maintenance, as well as the integration of information from different sources. Throughout, we highlight the solutions we have identified to address these challenges.



Integrating Social Media and Administrative Data for the Real-Time Prediction of the Consumer Confidence Indicator

Ms Akvilė Vitkauskaitė[1,2], Andrius Čiginas[1,2]

[1]State Data Agency (Statistics Lithuania), Vilnius, Lithuania, [2]Vilnius University, Vilnius, Lithuania

The main objective of this study is to nowcast and forecast the Consumer Confidence Index (CCI). We aim to estimate the current month's CCI values faster than those obtained using the traditional survey methodology, which usually provides results at the end of the month. We achieve this by combining key economic indicators with historical CCI values. We examine the relationship between traditional survey-based indicators and consumer sentiment expressed on social media platforms.

We analyze social media expressions, specifically from X (Twitter), through its official API to develop a Social Media Indicator (SMI). The sentiment analysis of tweets has enabled us to create an SMI that offers a distinct advantage in our predictive models. In addition, we are exploring the possibility of integrating key economic indicators from administrative data, such as inflation rate, income statistics, and unemployment, to increase the accuracy of CCI forecasting. In general, obtaining data for research from popular social platforms such as Facebook and Instagram is not possible due to stringent privacy policies and data protection regulations. Nevertheless, data are easily and legally available from X, but this platform is not so popular in Lithuania. Therefore, the representativeness of X data raises special issues. Taking everything into account, this research aims to advance the forecasting of the Consumer Confidence Index (CCI) with greater rapidity and accuracy. By combining traditional economic indicators with advanced sentiment analysis from X, the study seeks to deliver prompt CCI predictions ahead of standard survey timelines.

From Web Content to Quality Data: Rules, Roles, and Reliability in the Web Intelligence Hub

Fernando Reis[1], Raquel Paulino[2], Vladimir Kvetan[3]

[1]Eurostat, Luxembourg, Luxembourg, [2]Sogeti, Luxembourg, Luxembourg, [3]Cedefop, Thessaloniki, Greece

The Web Intelligence Hub (WIH) serves as a pivotal platform for the retrieval and transformation of web content into high-quality data suitable for official statistics. Ensuring the quality of this data is paramount, not only for its immediate statistical applications but also for its broader public utility. This paper aims to dissect the quality-related aspects of the WIH rules and conditions, aligning them with the Quality Assurance Framework of the European Statistical System. Key principles such as wide availability, equal treatment of users, and data confidentiality are examined to assess how they contribute to data relevance, accuracy, and ethical use. The paper also delves into the types of data and their respective domains of access, highlighting how these classifications serve as quality measures to ensure data reliability and ethical considerations. Additionally, the paper explores the roles and responsibilities defined for data access, emphasizing their importance in maintaining data accuracy and reliability. Suggestions for further quality assurance, such as data auditing, user feedback mechanisms, and automated quality checks, are also discussed. The paper concludes by affirming the critical role of quality assurance in enhancing the utility and credibility of data produced by the WIH, thereby serving the public interest effectively.

Session 9 - Experimental statistics, June 5, 2024, 14:00-15:30

Evolution of the Experimental Statistics project at the Brazilian Institute of Geography and Statistics (IBGE)

Ms Andréa Borges Paim[1], Marcos Paulo Freitas, Raquel Correia

[1]Brazilian Institute of Geography and Statistics (IBGE), , Brazil

This article describes the evolution of the Experimental Statistics Project at the Brazilian Institute of Geography and Statistics (IBGE), from the publication of an internal document in 2019 to its refinement, driven by the demand from the survey departments for user involvement in developing new statistical operations or indicators resulting from new themes or methods. The project was well- received, leading to the first experimental statistical operation in November 2020, during the COVID- 19 pandemic. In the following year, two experimental surveys related to the pandemic were launched—one estimating people with symptoms of influenza-like symptoms and monitoring the impacts on the labor market, and another assessing the pandemic's economic impacts on non-financial firms in the areas of Industry, Construction, Commerce, and Services. Since then, there has been a growing demand, totaling 18 experimental statistical occurrences in 2023.

In response to this demand in 2021, a guide with instructions for producing experimental statistics at IBGE was developed, covering criteria, dissemination procedures, user consultation, and the conditions for removing the experimental status. After two years of maturation of concepts and procedures, based on observations from the initial release of experimental statistical operations, a new version of general guidelines for the preparation and dissemination of Experimental Statistics was crafted. This includes improvements in the dissemination process and the removal of the experimental label.

Moving from experimental to official statistics: increasing the scope of statistics on earnings based on new administrative data sources.

Daniela Ramos[1], Célio Oliveira[1], Ricardo Cotrim[1]

[1]Statistics Portugal, , Portugal

In the context of statistical production, there are well-known advantages in the use of administrative data to provide statistics that are relevant, timely and cost-effective. One of the main advantages relies on the fact that administrative data can broaden the scope of statistics. Specifically, more detailed and disaggregated data can be provided, and the creation of linked datasets can enable not only a more efficient data infrastructure but also shed light on new and emergent phenomena.

Administrative data can also replace and/or supplement sample surveys, which are often costly, increase coverage and data reliability and reduce the response burden on respondents.

Against this background, Statistics Portugal has increased and broadened the scope of the statistics published on earnings to the extent that disseminating consistent and regular data on earnings is an important dimension of analysis of the labour market and its evolution.

This paper aims to present the various statistical production processes implemented, which have made it possible to guarantee the quality of the information released quarterly on earnings and to show the wide variety of statistics released so far.

Initially part of Statistics in Development (StatsLab), as of September 2021, Statistics Portugal releases quarterly official data on gross monthly earnings per employee (per job), calculated based on information at the enterprise level received from the Monthly Statement of Earnings from Social Security, and the Contributory Relation of Caixa Geral de Aposentações. These statistics include various breakdowns by earnings components and enterprise characteristics (economic activity, company size, institutional sector). An important component of the process of moving from experimental to official statistics has been the data flow and data treatment development, including of non-responses, which combines specific pre-defined criteria and a supervised Machine Learning algorithm.

The consolidation of this information also allowed to discontinue the survey data collection on wages for the Labour Cost Index (LCI), with the main advantage being that now the wage component of the LCI considers the universe of enterprises.

More recently, Statistics Portugal has started to receive similar information from the Tax Authority, but at the individual (worker) level, which, within the framework of the StatsLab, has made it possible to extend the scope of analysis through data linkages with other databases available as part of Statistics Portugal's wider National Data Infrastructure, thus allowing additional information to be provided on several sociodemographic characteristics (until now, sex, age and education level).

From experimental statistics to official statistics: state of the art and prospects in Istat

Ms Arianna Carciotto[1], Orietta Luzi[1]

[1]Istat - Italian National Statistical Institute, Rome, Italy

The Italian National Statistical Institute (Istat) has been producing Experimental Statistics (ES) since 2018, following the initiatives carried out by Eurostat and other NSIs in this field. These statistics are defined “experimental” as they use new data sources and/or new methods and/or new tools to meet users’ needs in a more effective and/or timely way. Often, ES are considered not full “mature” since they need to be further refined and validated in terms of compliance with quality requirements and harmonisation rules. In fact, quality assessment is one of the main challenge with ES, especially when transition to official statistics is to be evaluated. However, developing a quality framework for ES is a complex task as they are characterised by a wide heterogeneity. Among others, relevance seems the quality dimension which deserves the greatest attention. Users’ information need is one of the main driver to produce ES while their feedback is deemed as essential both to further improve experimental outputs and to potentially shift them to official statistics.

In any case, ES can always be seen as the result of research studies and analyses aimed at improving efficiency and boosting official statistics.

Comparing some NSIs within the European Statistical System that produce ES, neither a common definition nor shared rules to produce, manage and disseminate experimental products have been agreed upon. However, these kind of statistics originate from the same need to innovate both products and processes, trying to meet stakeholders’ needs in a more timely or thoroughly way. In all situations, stakeholders have an essential role to define the experimental output’s relevance and functionality.

In the paper, some examples of ES particularly appreciated by external users, and others that have shifted to official statistics will be reported, along with an analysis of their strengths in terms of relevance and usability for users.

Some general considerations, at a European Statistical System level, about challenges on the production of ES, possible shared rules, standardisation and exchange of best practices, will also be outlined.

From Experimental to European Statistics: Elevating the Short-term Rentals Project

Simon Bley[1], Christophe Demunter[1]

[1]Eurostat, Luxembourg, Luxembourg

Building on the foundations presented at the Q2022 conference in Vilnius, the short-term rentals project (CETOUR), led by Eurostat, has made significant strides in providing experimental data on short-term rentals sourced from four collaborative economy platforms (Airbnb, Booking, Expedia and Tripadvisor). This successful usage of privately held data for improving tourism statistics serves as a compelling example of how an innovative product can evolve from experimental beginnings to an official statistical product. As such, this use case provides a blueprint for a quality framework applicable to non-traditional statistical production processes.

Since 2021, this project has evolved from its experimental roots, with data releases now following a structured quarterly schedule. These releases meet the high standards of Eurostat's official statistical publications, characterized by regular quarterly releases and seamless integration into Eurobase. This includes the delivery of standardised outputs, such as Statistics Explained articles, news releases, and robust social media engagement.

CETOUR data has earned widespread recognition and application across various user groups, encompassing National Statistical Institutes (NSIs), DG GROW, academia, media outlets, tourism organizations, and more. Notably, users often perceive short-term rentals data as no different from other non-experimental statistics published by Eurostat, underlining its significance in filling a critical data gap concerning short-term rentals.

In evaluating the project's progress against the 16 principles of the European Statistics Code of Practice, this paper sets forth a compelling case for elevating CETOUR from "experimental statistics" to "European statistics." It proposes the project's full integration into Eurostat's 2026 Annual Work Programme.

Experimental Statistics in Finland: A Review of the First Five Years

MR Kristian Taskinen[1]

[1]Statistics Finland, Helsinki, Finland

Experimental statistics play a key role in advancing statistical methodologies and generating novel insights not available through regular high-quality statistics. This article focuses on the first five years of Finland's journey into experimental statistics, providing summary on produced experimental statistics. In addition, some innovations of experimental statistics are also presented, and lessons learned evaluated.

In 2024, Statistics Finland produces around twenty different experimental statistics. Article provides a summary on these experimental statistics including insights on subjects, types, frequencies, quality, and usage of these statistics measured with web page visitors.

Additionally, the article showcases some innovative approaches provided by experimental statistics, such as 1) linking micro and macro data to produce indicators on global value chains 2) using the national income register to produce monthly information on population's economic activity and employment 3) linking statistical data to geographic data and location data to produce information on national traffic network coverage.

As this review coincides with the fifth anniversary of Finland's journey into experimental statistics, the article also evaluates lessons learned so far. Key considerations include 1) the success of the life cycle management of experimental statistics, 2) the alignment of developmental with user needs, and 3) quality of produced experimental statistics.

The article provides valuable reflections on these aspects, contributing to the ongoing evolution of experimental statistical in Finland and in other European countries.

Session 10 - Metadata Quality I, June 5, 2024, 14:00-15:30

IT solution to enhance KAS reference metadata system

Mr Bekim Canolli[1]

[1]Bekim Canolli, Kosovo Agency of Statistics, Prishtine, Kosovo, [2]Burim Limolli, Kosovo Agency of Statistics, Prishtine, Kosovo

The Kosovo Agency of Statistics (KAS) has started producing and using standardized reference and quality related metadata ten years ago. In the past decades a lot of this information has been collected, stored and to a certain extent also disseminated, making it accessible to various (internal and external) users. The quality reports were written mainly in MS Word and were published in KAS web page and sent to Eurostat via Metadata Handler. To overcome this problem, we started to develop a new application that would enable easier and more effective usage of reference metadata produced through the statistical process and would support the evaluation phase of our statistical business model. The application was developed and came to its production life in 2023. The main functionalities of the application are:

- All the collected reference and quality related metadata are stored in one, central database and are ready to be used for different purposes.
- The application enables automated creation of Quality Reports based on the information stored in the central database.
- The application generates html and xml files that can be used for sending metadata via Metadata Handler.

The paper describes the integrated architecture of the application, some further details of its functionalities, and points out main challenges that we met during its development.

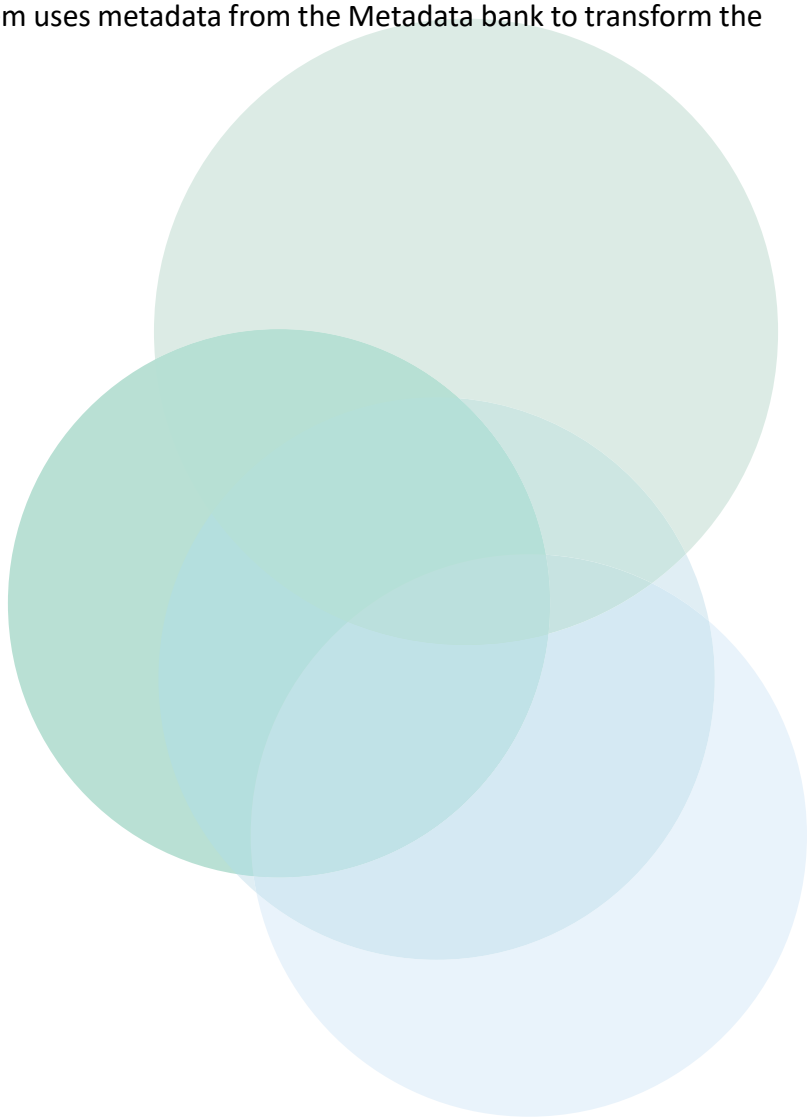
When theory meets practice: GSIM, Metadata and the Danish National Accounts source data system

Ms Anette Morgils Hertz[1], Mr Søren Søren Kristensen[1]

[1]Statistics Denmark, ,

Statistics Denmark has as one of its goals to ensure that the production of official statistics is metadata driven. To support this we have developed a documentation portal (the Metadata bank), which aims to gather all relevant metadata in one application. The Metadata bank is based on GSIM and contains the information objects that we have found are necessary to give a full metadata account. This include concepts, variables, quality assessments, classifications and correspondence tables. The Metadata bank will make it much easier for users to find comprehensive and consistent metadata. Equally important it will significantly improve the use of metadata in the production of statistics, thereby enhancing the quality of official statistics.

In this paper, we will present and discuss our efforts to integrate the Metadata bank into the Danish National Accounts' new source data system. By integrating the two, we have managed to build a metadata-driven source data system, that can transform the heterogeneous data sources used to compile the Danish National Accounts into the standardized classifications used in the compilation of the National Accounts. Using specific cases, we will focus on the challenges we met when fitting GSIM to data and describe how the system uses metadata from the Metadata bank to transform the source data.



Integrated metadata for the harmonization of a National Data Infrastructure

Ms Susana Portillo Cruz[1]

[1]Central Statistics Office - Ireland, Cork, Ireland

National Statistical Institutes (NSIs) are facing increased pressures to disseminate timely, accurate statistics with users now demanding data and information faster, more frequently and at a more granular level. This requires National Statistical Institutes to invest in the establishment of a data ecosystem at national level to make full use of secondary and other new data sources which aren't always structured in a consistent, standarised way.

This paper discusses the efforts of the Central Statistics Office in Ireland to provide a technical solution for the use of harmonised concepts, questions and responses using international standards. This solution can be used by the NSI and government agencies collecting administrative data to ensure the consistent representation of variable values, with the aim of using administrative data faster and more efficiently to achieve more timely and accurate dissemination of statistics.

Optimization of Accessibility and Quality of Metadata for Researchers: an Example of Collaboration Between INSEE and CASD

Ms Halima BAKIA[1], Thomas DUBOIS[2], Ifaliana RAKOTOARISOA[1]

[1]CASD, MALAKOFF, FRANCE, [2]INSEE, MONTROUGE, FRANCE

Accessibility and quality of metadata for researchers is fundamental. Metadata is anything that can give some useful information about data sets and help the users to better understand the data produced. It enables the researchers to identify relevant data for their project, before starting the procedures for accessing data, which can be quite long. The detailed documentation outlining the content of data serves as an initial response, acting as the first filter for source selection.

The Secure Data Access Center (CASD) is an organization allowing researchers and datascientists to work remotely and securely with confidential highly detailed microdata. It develops and establishes a secure interface between the research community and the NSS. 507 data sources are made available in a secure way, mainly produced by the French NSI (INSEE) and NSOs, but also from different public institutions. In this context, the exchange of high-quality metadata is crucial.

From 2018, documentation on data content has gradually been made available online by the CASD. Researchers, through satisfaction surveys, applaud this advancement. They are still requesting for more documentation to be made available prior to request of data access.

The process for exchanging metadata between INSEE and CASD is under review with a particular emphasis on the use of a metadata standard for structural metadata documentation (list of variables, codes and their meanings). INSEE and CASD recently carried out an experimentation to exchange files in the DDI standard. The experimentation results shows the expected gains:

- on workload, due to lower metadata entry burden
- on timeliness, the time taken to put documentation on line will be reduced
- on reliability, as online documentation will correspond to the documentation delivered by the producer, thus avoiding re-entry errors.

The use of standard such as DDI demonstrates how it promotes metadata interoperability between the two organizations. To achieve the complete goal of interoperability, the experimentation also shows that the use of standard must be accompanied by best practices definition.

The proper use of data by researchers plays a crucial role in enhancing quality. When researchers have access to standardized and high-quality metadata, it not only facilitates their own analytical work but also enables them to provide constructive feedback to data producers such as INSEE. This virtuous feedback loop creates an environment where data producers can adjust and enhance their processes in response to the specific needs of researchers.

Integrating international standards in the design of reference metadata component for the new Istat system METAstat

Ms Francesca Budano, Giorgia Simeoni[1]

[1]Istat - Italian National Statistical Institute, Rome, Italy

Since the beginning of 2000s, Istat has been equipped with a rich and a well-structured quality and reference metadata system, SIDI-SIQual. It is aimed at documenting and supporting quality monitoring and assessment of statistical production processes. SIDI-SIQual describes the production process and its features: information content; survey phases and operations; activities to prevent, monitor and evaluate sampling and non-sampling errors, standard quality indicators that are both process and product-oriented, etc. The system is well tailored for traditional surveys but less effective for innovative, e.g. multisource or experimental, statistical processes, and it is also technically obsolete, Istat is designing a new metadata system, called METAstat. The new system is going to manage both structural and reference metadata in an integrated way.

With regard to the process documentation, METAstat model relies on well-known international standards like the UNECE Generic Statistical Business Process Model – GSBPM, and the Generic Statistical Information Model – GSIM. The two models are used complementarily to describe the phases and sub-processes of each type of statistical process in terms of inputs, outputs and methods used. Thanks to the flexibility of these models METAstat overcomes the limitation of current system SIDI-SIQual, allowing to describe innovative and complex processes, e.g. detailing different sources and data processing steps to produce different variables or modules in the same process or specifying innovative data and methods used for producing experimental statistics.

In addition, the quality layer of statistical processes, that in GSBPM is described as an overarching process, in METAstat is detailed and integrated in each sub-process, including standard quality indicators, similarly to the current SIDI-SIQual approach. This allows to derive from the statistical process documentation almost all the information needed to produce quality reports according to the European Statistical System Standard SIMS (Single Integrated Metadata Structure).

Finally, METAstat is going to be integrated with many other Istat information systems, from the one managing the National Statistical Programme to the repository of validated microdata, in order to collect metadata and quality indicators as much as possible in an automatic way.

In this way the implementation of METAstat will reduce the documentation burden on production units and at the same time improve coherence and comparability of metadata and quality indicators across different statistical processes.

The paper will describe the model developed for reference metadata and quality documentation in METAstat with a particular focus on how to integrate GSBPM, GSIM and SIMS.

Session 11 - Quality assessment and review I, June 5, 2024, 16:30-18:00

GDP revisions, unemployment and factory gate prices: Regulating the Quality of UK Economic Statistics

Emily Carless[1], Mrs Marianthi Dunn[1]

[1]Office For Statistics Regulation, Newport, United Kingdom

The landscape of economic statistics in the UK is changing, with more new and innovative data available than ever before. There have been multiple economic shocks in the last few years including the Covid-19 pandemic, the war in Ukraine and increases in the cost of living. These have brought increased interest in economic statistics. The regulatory landscape has also changed: the UK's departure from the EU means the Eurostat will no longer verify the quality of UK statistics. As the UK's independent regulator of statistics, we are building on our years of experience of assessing statistics against our Code of Practice for Statistics, to ensure that users can have confidence in the quality of economic statistics in the UK.

Key to this work is the development of a programme of 'Spotlight on Quality' assessments of UK economic statistics. We developed a framework and tools for carrying out these assessments based on our Code of Practice and drawing on international frameworks such as the ESS Quality Assurance framework and the IMF Data Quality Assessment Framework. These assessments consider how the quality of statistics is ensured and communicated through suitable data sources, sound methods, adequate quality assurance and appropriate resources and prioritisation. We will talk through our framework and the key findings from our first Spotlight on Quality assessments.

During Autumn 2023 we carried out rapid reviews of the quality of revisions to the UK's Gross Domestic Product and experimental Labour Market statistics estimates in response to stakeholder concerns. These reviews examined the processes and quality assurance used, the potential improvements that can enhance quality, through use of alternative data collection techniques or additional datasets, and the clarity of communicating uncertainty of estimates to users, particularly in the current economic climate where the landscape is continuously changing. We will explore our findings from these reviews and their wider applicability.

Alongside evaluating the quality of individual sets of statistics we are reviewing the systemic approach to quality of economic statistics in the UK. This review considers both recommendations from previous independent reviews, and our own regulatory work, of economic statistics in the UK, particularly around capability and effectiveness, and the evolving requirements of estimates of economic statistics, such as the implementation of the System of National Accounts 2025, to ensure that the quality of economic statistics in the UK is fit for the future. We will discuss our emerging findings from this work.

Money Makes Statistics

Ms Lidija Brkovic[1]

[1]Croatian Bureau of Statistics, Zagreb, Croatia

The growing need for statistical data requires the Croatian Bureau of Statistics (CBS) to rationally and optimally use the available human resources, as well as the financial resources defined by the state budget.

In order to achieve this, the CBS introduced a system of procedures and tools that enable the calculation of the costs of statistical processes and products, in order to improve its operations, but at the same time contribute to the realization of the principle of economy, from the Code of Practice of the European Statistics. By introducing a system, the CBS gained information on costs of statistical products and processes, comparable to the statistics of other countries; an indication of the need for process changes or improvements and an excellent basis for negotiations on the financing (budget). As part of the a system, two software solutions are developed: Electronic record of working hours and Software solution for cost management (CostMgt).

The first software enables the employees of the CBS to record working hours by products and processes and is the main source of data for the cost management solution.

The CostMgt downloads data from the electronic record of working hours, but also from other systems of the CBS (Central payroll, bookkeeping system, system for monitoring physical presence at work) and calculates costs.

By introducing a system for estimating the costs of statistical products and processes, the CBS will obtain quality data on the basis of which it will be able to allocate its resources, look for savings in operations and determine priorities for improvement. The CBS will receive indicators for which statistical products/processes must undergo significant methodological and technological changes in order to ensure their economic efficiency. Thus, the CBS will be able to establish a high-quality basis for analysis and decision-making on business, statistical products and on the improvement of individual statistical products and processes.

Enjoying cooperation and improvement: Cooperation on quality between Statistics Norway and other producers of official statistics

Senior adviser Grete Olsen[1], Mrs Kari Benterud[1]

[1]Statistics Norway, Oslo, Norway

Statistics Norway (SSB) has collaborated with other producers of official statistics for many years. With the introduction of the new Statistics Act in 2019, cooperation became more concrete.

Collaboration takes place through the Committee for Official Statistics, the method network, the dissemination network and through the system for quality of official statistics. Recommendations are given on quality-enhancing measures, but SSB has no possibility of sanctioning other manufacturers. Our experience is that all producers are eager to improve the production in all aspects, they are very positive to the coordinating role of SSB and find the cooperation useful. Quality evaluations and reviews are parts in the annual report on the quality of official statistics.

The quality evaluation and reviews are based on the quality requirements in the Statistics Act and the quality principles in the European Statistics Code of Practice (CoP).

The first overall quality evaluation of all official statistics was carried out in 2021 and was followed up with a new corresponding survey and data collection in 2022. In this quality evaluation, a questionnaire-based survey was combined with interviews. Based on the answers on the questionnaire and the interviews, SSB recommended 42 improvement areas, asked the producers for improvement actions, and followed up on status for the actions, much like the ESS Peer reviews.

In 2023 SSB conducted one in-depth quality review on one statistics at the Directorate of Fisheries. Our ambition is to carry out 4 quality reviews on single statistics by other authorities in 2024.

The main impression is that the commitment to quality has increased among the producers, and they have effectively implemented some of the actions based on recommendations in the quality reports.

One area where the evaluations have been effective and led to improvements is improved and more accessible documentation. Some producers have established clear responsibility for official statistics in their organisation and improved the information on their role as producers of official statistics both within their organisation and on the website.

In 2023 the evaluation was changed to examine each statistic in the National programme for official statistics, in total 344 statistics (52 from authorities outside SSB). Chosen results from this evaluation will be presented at the quality conference in June.

To improve official statistics according to user needs and to the CoP, a step-by-step approach is often appropriate. Communication, evaluation and follow up on improvement actions contribute and inspire to improvement work

Peer-review during the decentralization of official statistics: experience, lessons and best practices in Abu Dhabi

Dr. Yu Sapphire Yu Han[1], Qais Al Junaibi, Nasser Dayan

[1]Statistics Centre Abu Dhabi, , United Arab Emirates

Statistics Centre Abu Dhabi (SCAD) is responsible for official statistics in the Emirate of Abu Dhabi. In recent years, one of the major trends that reshapes the traditional official statistical system in Abu Dhabi is the decentralization of official statistics. The decentralization in Abu Dhabi is sectorial. This means that certain statistics are mainly compiled by specialized Abu Dhabi Government Entities and SCAD plays an important role. The key question for SCAD is the quality assurance of decentralized official statistics.

To answer this question, SCAD has launched a peer-review program to evaluate and enhance the quality of decentralized statistics since 2023. The peer-review procedure is a formal assessment of the compliance against the Code of Practice by Abu Dhabi Government Entities who produce decentralized official statistics.

An evaluation study was conducted on three entities (selected based on the strategic priorities) that produced decentralized statistical outputs. During the evaluation period, a self-assessment survey has been developed by expanding the principles of the Code of Practice to a set of 55 questions. The results of this survey served as the basis for official meetings and dialogues with the production teams of decentralized entities. After analyzing survey and meetings, assessment reports concluded the current level of compliance, and recommended actions for the relevant entities to enhance the quality of decentralized statistics. For instance, one finding in the evaluation process is the limitation of the production team's understanding of users' need and the lack of sufficient tabulation, which reduce the relevance of statistical output quality.

In result, a peer-review protocol has been developed including three components: a simplified version of self-assessment survey, face-to-face interviews with the production team of decentralized government entities, and SCAD's guideline of quality reporting (an adaptation of international recommendation to the Abu Dhabi Context). Choosing the peer-review protocol as its main quality assurance approach, SCAD is committed to conduct the peer-review to all official statistical products in Abu Dhabi in the coming period. In addition to the peer-review program, SCAD organizes statistical and quality training sessions for the production team in decentralized Abu Dhabi Government Entities to further enhance statistical output quality. SCAD seeks to develop human capital in the Abu Dhabi Statistics Center and the statistical entities by preparing a program for statistical assessors, and this is the first program that is unique locally and regionally in the scope of peer-review.

Bringing users into focus. How focus groups with users in quality reviews contributes to improved statistics

Mr Per Ola Haugen[1], Dr. Frode Berglund

[1]Statistics Norway, Kongsvinger, Norway

Statistics Norway has carried out in-depth quality reviews of individual statistics for many years. The quality reviews are systematic reviews of a statistical production process to identify strengths and weaknesses with the statistics. This results in a report with recommendations and a plan for improvement actions. Focus groups with users are a key element as part of the quality reviews and enable direct feedback from the users on the statistics.

To recruit participants for focus groups, a user and stakeholder analysis is the basis. The statistics officers score the users according to how much influence they have on user needs and how much they are impacted by the statistics. Users with the highest scores are recruited, but it is important that several types of users are represented, for example journalists as representatives of the public.

1-2 focus groups are conducted, each with up to 6 participants. The aim is to create a dialogue between users about user needs. An independent moderator leads the conversation. The moderator has prepared an interview guide that is closely linked to the five quality principles in the CoP.

The focus groups are conducted as on-line meetings, and a video recording of the session is taken to document the views of the participants as input for assessing strengths and weaknesses. The quality team and the statistics officers observe the meeting and are not allowed to interfere in the discussion. Any feedback is discussed with the participants afterwards the focus group.

In the focus groups there are often surprises with views on statistical needs and dissemination of the statistics that were not previously known, and which are not captured by the other elements in quality reviews. An example of such a surprise from one of the reviews is that users do not use the statistics home/webpage to find relevant numbers. Rather, they call the people responsible for statistics and get information directly from them. It may be a sign that the statistics page is not known or not user-friendly. But it can also be a sign of a high-level of service from those responsible for the statistics.

Views from the participants in the focus group always result in recommendations related to user needs and improved dissemination. Focus groups with users thus provide a basis for more relevant statistics and improved dissemination of the statistics!

Quality Reviews in Eurostat

Ilcho Bechev[1]

[1]Eurostat, Luxembourg, Luxembourg

The self-regulatory common quality framework of the European Statistical System (ESS) is built around the principles and the indicators of the European Statistics Code of Practice (ES CoP) and the methods and tools of the Quality Assurance Framework of the European Statistical System (ESS QAF). It is, however, through quality reviews that Eurostat can monitor internally the ES CoP implementation at the level of specific statistical processes and outputs. This assurance mechanism complements the ESS peer reviews, which scrutinise the system mainly at institutional level. The aim of this paper is thus to describe and analyse this operational quality assurance layer that bridges the 'theoretical' quality framework and the 'practical', every-day aspect of developing, producing, and disseminating European statistics. By doing so, this paper shares Eurostat's methods and good practices for further use within the ESS, particularly within organisations that are yet to develop or still in the process of developing and implementing their own quality assessment methodologies. As the current cycle of quality reviews is coming to an end, also some results of the already implemented reviews are discussed and conclusions are drawn.

Session 12 - Online job advertisement, June 5, 2024, 16:30-18:00

Combining Online Job Advertisements with Probability Sample Data for Enhanced Small Area Estimation of Job Vacancies

Dr Andrius Čiginas[1,2], Donatas Šlevinskas[1,2]

[1]State Data Agency (Statistics Lithuania) , Vilnius, Lithuania, [2]Vilnius University, Vilnius, Lithuania

We combine the probability sample data on job vacancies with online job advertisements (OJA) information to improve the estimates of job vacancy totals in small domains like municipalities. We apply domain-level small area estimation models, where aggregated OJA data act as auxiliary variables. Since OJA data is a non-probability sample covering only a limited part of the survey population and its selection mechanism is unknown, we apply non-probability sample integration techniques to use the aggregates properly in the models. We show that OJA is an efficient alternative data source to improve the estimates in small domains.

Innovative Approaches to Enhance Data Quality in Official Statistics: A Case Study on Online Job Advertisement Data

Ms Anca Maria Nagy[1], Eliane Gotuzzo[3], Mr Fernando Reis[2]

[1]Sogeti Luxembourg, Luxembourg, Luxembourg, [2]Eurostat, Luxembourg, Luxembourg, [3]Sogeti, , Luxembourg

The rapid proliferation of online job advertisements (OJA) has shown in a new era of non-traditional data sources, offering the potential to augment official statistics with real-time labor market insights. However, unlocking the potential of OJA data demands careful attention and emphasis on data quality. This paper delves into a pioneering methodology designed to assess and improve the quality of OJA data, with a particular focus on the occupation variable, within the context of official statistics.

The aim of this work is two-fold: to establish a robust quality monitoring procedure and to create a gold standard for OJA data. We leverage natural language processing (NLP) to scrutinize job descriptions and structured fields, utilizing both machine intelligence and human annotation.

Importantly, our methodology introduces the incorporation of large language models, for quality assessment of labelled data and of occupation classification. This addition expands our ability to evaluate and classify occupations for specific samples that may pose challenges for classifiers or human annotators.

Notably, this study unfolds the interplay between human and machine intelligence, highlighting the potential of a combined approach in enhancing data quality. Human annotators contribute their domain expertise to the refinement of classifiers, resulting in a gold standard for OJA data. Evaluation metrics, such as accuracy and consistency, are deployed to assess the quality of the labelled data, while machine learning models are trained on this human-labelled data for automated classification.

The results from this study, focusing on a selection of European countries, point to the need for improvements in occupation classification within OJA data. By providing an analysis of labelled OJA datasets, we offer insights into the accuracy and reliability of classifiers. We envision future iterations of this methodology to extend to more countries, encompass additional variables, and explore the development of refined ontologies.

This paper exemplifies the innovative approaches that official statistics agencies can employ to foster data quality in an era of evolving data sources. It underscores the value of collaborative human-machine efforts and the pivotal role of advanced language models in enhancing data quality. By presenting this work, we hope to contribute to the ongoing discourse on innovation and research in official statistics, shedding light on novel strategies to ensure the highest quality data for informed decision-making and policy formulation.

Experimental OJA based indicators on labour demand changes: opportunities and challenges

Ms Annalisa Lucarelli[1], Kostadin Georgiev[2], Francesca Inglese[1], Renato Magistro[1], Giulio Massacci[1], Giuseppina Ruocco[1], Galya Stateva[2]

[1]ISTAT, , , [2]NSI, ,

Online Job Advertisements (OJAs) offer great opportunities for analysing labour market trends due to the high level of detail of the information they contain and the high frequency with which they are made available.

This work focuses on an attempt to produce new experimental indicators of changes in unmet labour demand based on OJAs, broken down by strategic information not currently available in official statistics. In order to fully exploit the information provided by the OJA data, the main quality aspects of the sources – websites/job portals from which OJA data are scraped – have been taken into account, in particular their relevance and stability over time.

The attempt to construct experimental indicators on the basis of OJAs has been mainly shared between two countries – Italy and Bulgaria – within the activities of the ESSnet web intelligence network project and on the basis of CEDEFOP data.

The main effort concerned the definition of an harmonised methodological framework covering several aspects such as: the choice of the indicators to be produced (levels or changes); the distinction between the concepts of stock and/or flow variables and their definition for the purposes of implementation on the basis of OJA data; the choice of the type of breakdown to be considered (economic activity, occupation, skill, territorial area); the level of detail (to which digit of the International Standard Classifications); and the choice of the reference period.

The first provisional indicators based on OJA are the year-on-year changes in the monthly stock and flow of OJA – i.e. the stock and flow in each month compared with the same month of the previous year – broken down by section of activity (NACE Rev. 2 classification), by major and minor groups of occupations (ISCO-08), by skill (ESCO Skills and Competence Pillar), and by region (NUTS 2 level) over the years 2021-2023.

This first attempt was useful in that the results obtained were to some extent consistent with the economic framework, particularly during the period of the health emergency. It also allowed methodological issues to be addressed and shared between the three countries involved in order to move towards an harmonised framework for the production of these indicators.

The production of a new experimental OJA indicator - based on a harmonised European production framework - is particularly relevant as it could enable new requirements of EU Regulation to be met in the future, addressing unsatisfied/emerging information needs.

Session 13 - Innovative methods & machine learning, June 5, 2024, 16:30-18:00

Updating of Statistical Register by Web Scrapping

Mr Jaroslav Sixta[1], Michal Čigáš[1], Jan Fojtík[2]

[1]Czech Statistical Office, Prague, Czech Republic, [2] Prague University of Economics and Business, Prague, Czech Republic

Statistical offices are responsible for maintaining and updating of statistical registers that form the baseline for official statistics. In the Czech Republic, the Statistical Business Register counts about 2.7 mil units with different statistical attributes. However, the group of users contain not only statisticians and government authorities but also everyday users who are interested in the statistical attributes of companies, bodies etc. The motives comprise different reasons but one of the most important attribute is the classification of economic activity coded by NACE classification. Long-run emphasis of modern government bodies consist in the decrease of administrative burden and therefore it is still more and more difficult to conduct regular survey about actual economic activities of all units in the economy when available administrative sources offer limited options that have been already used up. In this context, the Czech Statistical Office started the cooperation with the Prague University of Economics and Business to develop automatized procedure based on web scrapped data and machine-learning classification of extracted information into NACE codes. The paper provides information about the project and successful classification of information for the update of the Statistical Business Register.

Estimating non-sampling error due to miscoding of groups, and implications for automating coding at NSI's

Boriska Toth[1]

[1]Statistics Norway, Oslo, Norway

Classifying items into categories (also called "codes") is necessary for the production of various statistics (for instance, a labor force survey that publishes unemployment statistics in various occupation categories). There is a major push at statistics bureaus to replacing human coding with machine learning for this process. Despite evidence that misclassifications in these codes can be a significant source of non-sampling error, published statistics do not typically give estimates for a resulting error term. This paper studies how errors in coding affect the quality of published statistics. We give a framework to estimate errors due to misclassification in various groups, to inform decisions on whether and how automated coding can replace human coding, and to indicate for what types of items improvements in coding can translate into more precise statistics.

Our method gives an empirical estimate of the magnitude of error in various groups for a descriptive statistic (such as a total or mean) that can be attributed to miscoding of the groups. We compare the statistics observed when a "gold standard" is used to define the groups vs when an error-prone alternative is used (typically comparing manual vs machine learning-based coding). Bootstrapping estimates can be made of the variance of the errors. A confusion matrix is estimated that shows for each true group how the groups predicted by the alternative classifier are distributed. The confusion matrix allows one to easily see which misclassification errors adversely impact a group-level statistic. In addition, one can perform simulations to vary the location and frequency of misclassification errors in the confusion matrix for the purpose of estimating the impact of an improved classification method on reducing error in published statistics. This method can be used with the presence of, and adjustment for, other types of non-sampling error, such as non-response weightings.

We present a case study of applying these methods in which we estimate non-sampling error due to the misclassification of COICOP codes in Norway's 2022 household budget survey. We reflect on the potential of various machine learning methods to improve the accuracy of HBS statistics published at the 5-digit COICOP category level. Such an analysis can help inform decisions about whether investing resources in developing more modern machine learning workflows that use technologies such as large language models is likely to improve statistics quality.

Retraining strategies for an economic activity codification model

Mr Tom Seimandi[1], Thomas Faria[1], Nathan Randriamanana[1]

[1]Insee, Montrouge, France

The French company registry, SIRENE, lists all companies in France and assigns them a unique identifier, the Siren number, for use by public institutions. As part of the registration process, companies must provide a description of their economic activity. Since the end of 2022, SIRENE leverages a simple text classification model to code each description into an industry from the French classification of activities (NAF).

Using a machine learning model in a production environment comes with challenges, in particular regarding model monitoring and maintenance. In this talk, we will first present the monitoring system developed to track model behavior and detect potential drifts for SIRENE. Then we will address the question of model retraining, including the following considerations :

- Evaluation data : how should evaluation data be collected (data quantity, frequency, sampling strategy) ?
- Training data : what training data should be used for retraining ? For example, should historical data be used systematically or does the model perform better when only trained on recent data ? To what extent should data classified automatically by the model in production be part of the training set to keep it balanced ?
- Retraining strategy : at what frequency ? What are the differences between fine-tuning and retraining from scratch ?
- Algorithmic adaptations : does a more complex text classification model allow performance gains on new real-world data ?

This talk aims to equip practitioners with an improved understanding of the technical and practical considerations involved in retraining text classification models. As such models are becoming an essential component of official statistics, it is crucial to ensure the quality of their outputs in production environments.

A Potential Quality Assurance of the Re-coding to NACE Rev. 2.1, Combining LLMs and Manual Coding

Jacob Kasche[1], Wictoria Widén[1], Gustaf Strandell[1], Kira Gylling[1]

[1]Statistics Sweden, , Sweden

Implementing NACE Revision 2.1 is demanding for many European countries. A major part of the transition is the re-coding of units in the Business registers. Previously, the re-coding process has mostly been done using surveys and manual coding, which often result in large costs. Quality demands on NACE are high; hence quality needs to be high in the re-coding process. In previous quality assurance processes, several coders repeated the re-coding i.e., reconciliation. Because of budget restrictions, it may not be feasible to perform this process on the entire nomenclature.

Because of the increased performance of large language models (LLMs), several countries investigate the possibilities of using LLMs to decrease manual coding. However, the model approach does not only increase the possibilities to lower the use of manual resources. It also facilitates the development of effective quality assurance.

In this paper, a potential quality assurance process, which focuses on combining manual labour with models e.g., LLMs, is presented. The quality assurance process includes: 1. Model inference; 2. Design inference with auxiliary information; 3. Manual coding supported by models; 4. Re-use of manually coded data. Methodologies necessary for each step are presented and the workflow is illustrated with examples from Statistics Sweden. Lastly, the paper discusses the quality assurance process and how it may facilitate an effective transition in the current and upcoming revisions of NACE for an NSI.

A topic modelling approach to estimate relevance of Twitter data to monitor the debate about immigration

Dr Elena Catanese[1], Clelia Romano, Gerarda Grippo, Mauro Bruno, Francesco Ortame

[1]Istat, , Italy

Relevance is the degree to which statistical outputs meet current and potential user needs. In the context of big data, especially on Social Media, it is sometimes not so clear whether the data contain information related to the intended statistics.

A typical social media pipeline consists in choosing a set of key-words and then filtering all the texts that contain at least one of these . Ideally, filters should be able to - eliminate off topic messages since the beginning. For this purpose, the choice of the filter, should be split into a top-down and bottom up approach. First a list of words is chosen by experts according to the intended statistics, then the list should be validated by some data-driven analysis that ensure the relevance of the sampled texts. These data-driven analysis can be performed through a variety of either machine learnings methods for semantic analysis, or Topic modelling, a frequently used text-mining tool for discovering hidden semantic structures in a text body. Also, there are several reasons why in practice a two-step filter must be used. Aim of the works to study public opinion toward immigrants in Italy by exploiting Twitter.In practice, the introduction of a new filter for new studies does not allow to have such a long time-series. For this reason, in practice a sub-sample from the “generic”set must be drawn, i.e. a two-step filter.For this scope an initial set of key-words related to “immigration” was applied to a larger sample of tweets (280,000 daily tweets on average) collected by Istat since 2016, and whose messages sampled through a wide filter (278 key-words) are meant to represent a small scale model of the overall population of messages which are potentially relevant for Official Statistics purposes. The so obtained sample, 24 millions of tweets for the period 2018-2022, has been analysed by means of Latent Dirichlet Allocation, which is a Bayesian Topic Model, that usually aligns much better with human interpretation and is relatively fast and feasible even in presence of a large set of data compared to other Topic Modelling techniques. This analysis enabled to evaluate clusters out of scope and therefore redefine the filter. In addition, we try to identify if present a selection bias induced by this second level filter with respect to a direct filter by evaluating coherence in terms of the discussion about the topics for a period of few months.

Session 14 - Communication and Statistical literacy, June 5, 2024, 16:30-18:00

Supporting the 4th Pillar of Democracy: Dissemination of Official Statistics through the media

Emma Castillo Schiro[1]

[1]Statistics Norway, , Norway

"The state authorities shall facilitate an open and enlightened public discourse" says paragraph 100 in the Norwegian Constitution.

Today's media landscape is increasingly fast-paced, requiring short-format content in an ever-growing ocean of disinformation and biased representations of reality. Journalists are under pressure to deliver quality content that captures the public's attention within short timeframes. How do we support and utilize the media in disseminating statistics and contribute to an enlightened public discourse?

This paper aims to discuss following questions:

- Which role should NSI's take in outreach and dissemination of statistics through the media?
- How do we translate statistics into stories that can be told?
- How do we train and support subject matter spokespersons for NSI's and NSO's?
- Is the European Statistics Code of Practice sufficiently covering the quality of dissemination of statistics?

The discussion draws on the authors professional experience from being a spokesperson for Statistics Norway's Media Barometer by performing live TV- radio- and podcast interviews, holding press conferences and being in frequent contact with journalists and other users, along with training new spokespersons of official statistics.

Quality in dissemination and our contribution to democracy requires a discussion on what role should producers of official statistics take in contributing to an enlightened public discourse. The paper suggests several approaches and presents practical examples.

At last, the paper discusses if the principles in the European Code of Practice sufficiently provides guidelines for the communication and dissemination of official statistics into the future. Perhaps is there need for an additional principle?

Data Democratization and Official Statistics: The Greek Paradigm

Mr Apostolos Kasapis[1], Violetta Ntounia, Christina Pierrakou

[1]Director, President's Office of The Hellenic Statistical Authority (Elstat), Piraeus, Greece

This paper explores the strategic initiatives undertaken by the Hellenic Statistical Authority (ELSTAT) in the pursuit of data democratization, with a focus on the dissemination of official statistics to a wider audience through the utilization of social media platforms, infographics, and infovideos. As the demand for accessible and comprehensible statistical information continues to grow, ELSTAT has recognized the importance of adapting to evolving communication trends to make statistical data more approachable and relevant to the general public.

Drawing upon recent experiences from the 2021 Population-Housing Census and various other endeavors, this paper showcases successful examples of how ELSTAT leverages modern communication channels to bridge the gap between complex statistical data and the broader population. By embracing social networks and visual storytelling tools, ELSTAT has not only enhanced data accessibility but has also facilitated a deeper understanding of the significance and implications of official statistics.

This research contributes to the broader discourse on data democratization by highlighting the challenges, strategies, and outcomes of ELSTAT's efforts in Greece. Furthermore, it underscores the potential for similar initiatives in other regions, emphasizing the pivotal role of data democratization and statistical literacy in fostering informed decision-making and civic engagement in the 21st century.

Istat experiences and new challenges with statistical literacy at central and national level

Ms Patrizia Patrizia[1]

[1]Istat, Rome, Italy

Taking the lead from the title of the session “Statistical literacy: challenges and success stories” the paper will present Istat recent experiences with statistical literacy trying to report with data both aspects of successes and new challenges.

In the Success stories section reports and data on ongoing statistical projects for students and schools will be presented, with a temptative approach, data based, on what counts more: promotion of activities or the topic of activities itself. To better describe it: is it statistics a self-standing and relevant topic which attracts indipendently from dissemination and promotion, or is it largely widespread since promotion and dissemination activity in itself is massive and benefits from the agreement with the Ministry of Education both at national and territorial levels? A hint seems to point to the fact that it is a mix of both but we will see the results from participation data with some comparisons over time.

In the challenges part new projects on the way for 2024 are presented. After a comprehensive and methodological book explaining and illustrating methods for carrying out statistical literacy and with chapters presenting in-depth analysis of single projects in 2023, in 2024 the efforts will be concentrating on statistical literacy for women and adults. Among the adult target a privileged position will be devoted to teachers, since they are for us the trainers to be trained to spread the value of data. First of all they need to be aware of the wealth of data that could enrich their work and then they can in turn disseminate this same data.

Space will also be given to partnerships with non profiworking in the scientific and educational sectors, and illustrating the work being done with them.

The first case is a project with the Bank of Italy for which data on prices and money will be presented to the general public data within the Exhibition “The adventure of money” in the form of conversations with thematic statisticians.

A second case is the cooperation with universities for project of secundary school education and a temptative federation fo organization to foster youth enterpreneurship, for which official statistical data represent a value added.

A third case is the revamping of stduy visita t the central office of Istat, testifying a renewed interest in statistics and thematic issues of statistics.

Statistical literacy in an age of misinformation

Mr Jeremy Heng[1]

[1]Ministry Of Manpower, , Singapore

The new digital age has brought about a wide range of information that is readily available to the masses. Providing the public with reliable statistical information can empower them to make informed decisions in their everyday lives. On the other hand, misunderstanding or misinterpreting the statistics provided can also bring about poor decision making and lead to the proliferation of fake news.

Statistical literacy, known as the ability to understand and interpret statistics, is increasingly important in this new digital age. Statistics should also be supplemented with sufficient contextual information and be made simple to understand, otherwise there is a risk of misinterpretation regardless of the level of statistical literacy.

As statistical information becomes more widespread, a lack of statistical literacy among different users and stakeholders can be problematic on various fronts. Firstly, policymakers are relying more on data for evidence-based policy development in recent years. A lack of statistical literacy among policymakers could be detrimental to the quality and efficacy of policies implemented in a country.

Secondly, institutions are responsible for the dissemination of information to the public, and when unsubstantiated and misleading statistics is circulated in the public domain, it undermines trust and confidence within the larger community.

Thirdly, individuals misinterpreting statistical information can have a viral effect, where the wrong information gets spread through different channels and eventually reaches the masses.

To improve statistical literacy within society and to prevent the spread of misinformation, the Singapore Ministry of Manpower has put in place a quality framework, consisting of (i) verification, (ii) regulation, and (iii) education. For the framework to be effective, a comprehensive series of checks and procedures has to be implemented by the various stakeholders. Individuals, institutions and governments all have a role to play in being vigilant when consuming and disseminating information. Only when there is trust in official statistics, can there be a conducive exchange of ideas, and a platform for effective policies to be developed.

The paper discusses the challenges of a lack of statistical literacy within society, how it can be overcome, and the initiatives that Singapore has implemented to improve statistical literacy and the way statistics is disseminated.

Session 15 - Smart Surveys Implementation, June 5, 2024, 16:30-18:00

Data Quality using Smart Survey in Statistics Norway's Household Budget Survey 2022

Ms Nina Nina[1], Aina Holmøy, Gezim Seferi

[1]SSB, Oslo, Norge

In the process of modernizing data collection for the Household Budget Survey (HBS) 2022 Statistics Norway (SSB) went from traditional data collection using a combination of visits and telephone interviews to self-completion in a web app. This new web app seamlessly integrates the

diary with a receipt scanning function and the survey questions in one solution. Sensor technology in digital or smart devices enables the receipt scanning functionality, hence we call this data collection method smart survey. In this survey method interviewer assistance is to a large degree replaced with interface guidance of respondents. We have gone from interviewer administration to self-administration, but still use interviewers for recruitment and support.

This paper presents our first data on use of smart survey in data collection for official statistics. We will describe para data and survey data and analyse response rates and nonresponse bias; device preferences; share of scanning versus manual receipts registration (which is optional); quality of scanning and the granularity of product item descriptions obtained through manual registration and the need for editing. Our focal point will be what method; automated scanning or manual registrations, the respondents choose in general and for type of receipts and product categories. This will allow us to discuss the impact of app based smart surveys on data quality.

By sharing our first data from HBS 2022 we contribute to the collective understanding of the implications and effectiveness of use of smart surveys in official statistics.

Household Budget Survey within a new EU legal framework – towards higher quality and more harmonization, promoting innovative approaches

Eniel Ninka[1], Jakub Hrkal[1], Erika Taidre[1], Teodora Tchipeva[1]

[1]Eurostat, , Luxembourg

Since 1988, Eurostat publishes statistics in the consumption domain every five years using data from the national household budget surveys (HBS). Up to the 2020 wave data were obtained on a voluntary basis. Starting with the 2026 wave, HBS will be implemented under the EU Regulation 2019/1700 and the related legislative acts. The new legal basis is expected to enhance harmonization of HBS among European countries and with other EU social surveys. Moreover, it would bring improvements in various aspects of quality such as relevance, accuracy, timeliness, accessibility, and comparability.

Since 2017, Eurostat has supported European countries, via various projects and exchanges of practices, in improving the methodology and developing innovative and shareable solutions aiming at modernisation of HBS data collections. To this aim, Eurostat, with the support of a task force on innovative tools and sources for the HBS, has supported via specific grants the development, maintenance and testing of several innovative solutions to collect household budget data including scanning and processing of shop receipts via optical character recognition and machine learning algorithms. As a result, fully fledged web and/or smartphone applications are available, and many countries plan to use them in the 2026 data collection wave either as the only mode or in a multimode setting. More recently, the focus of projects and grants has shifted towards the conceptualisation, proof of concept, and implementation of (trusted) smart surveys. The expected impact on data quality would be a reduction of recall and underreporting errors, real-time data checks and pre-validation performed in-app or in the back-office, and reduced post-fieldwork data processing time that would result in improved overall data quality and in terms of timeliness and accuracy, while maintaining data relevance and preserving privacy and confidentiality.

This paper first describes the potential for improvement of the quality of future HBS by comparing the current state of play and anticipated changes in view of implementing HBS under the new legislative framework. It also informs about the development of methodological and technical documentation for implementation of HBS and about a Eurostat project on the development of a new data production system, activities which will impact the quality of implementation, processing, and dissemination of HBS. The paper then provides a complete overview on Eurostat activities aiming at supporting countries to improve the HBS methodology and further develop innovative and shareable solutions to modernise the HBS data collection.

Enhancing the quality of the prediction of activities in Time Use Smart Survey using a microservice exploiting GPS data

Ms Claudia De Vitiis[1], Fabrizio De Fausti[1], Francesca Inglese[1], Marco Dionisio Terribili[1], Barry Schouten[2], Jonas Klingwort[2], Joeri Minnen[3], Manuela Michelini[1], Tania Cappadozzi[1], Pieter Beyens[3]

[1]ISTAT, Roma, Italia, [2]CBS, Den Haag, Netherlands, [3]hbits CV, Bruxellex, Belgium

Smart surveys offer new opportunities for developing social surveys, especially those based on burdensome compilation of diaries (Household Budget Survey, HBS and Time Use Survey, TUS), as they aim to exploit new data sources through personal devices (smartphones, tablets, wearables) that use sensors and provide information about themselves or their surroundings.

On this topic, the European Statistical System (ESS) have financed two projects: presently the second project is ongoing, ESSNet Smart Surveys Implementation, starting in 2023, that aims to involve and engage households and citizens, to define and operationalize a new/modified end-to-end data collection process, and to test microservices, solution/component for Time Use and Household Budget surveys.

This paper has a focus on Time Use Survey (TUS) carried out making use of web and mobile applications and the inclusion of a microservice for geolocation. The microservice is seen here as middle part software that is supportive to the household in reducing their burden to complete a time diary. The geolocation microservice is developed as an independent service to platforms.

The collection of geolocation points (based on GPS sensor data), allows obtaining relevant information for TUS, predicting the HETUS activities as tentative data for the respondent during the filling of the diary. In this context, in fact, the role of respondents and the interaction with them during data collection is a crucial issue. Starting from the GPS coordinates and some additional information the geolocalized points are segmented into stop and motion; then, by adding geographical-spatial information, the places (Points of interest, POI) that can provide information on the activity are identified. One or more activities (with assigned probabilities) are associated to the POIs through an algorithm that exploits diverse information, such as place categories/activities taxonomy, timing of the stop, country-specific indicators and user characteristics.

A central concern in the development of this microservice is the assessment and the enhancement of data quality, both raw data and processed data. The quality of sensor measurement can be affected by sensor inaccuracy due to imprecision, time inequivalence, device inequivalence. Depending on sensor quality, sensors may produce systematic and random measurement errors. The quality/ accuracy of the predictions depends not only by the quality of raw data bat also by the choice of the location/map features and the use of paradata and contextual data for the classification algorithms.

The integration of traditional surveys with new data produced from smart surveys

Mr Peter Lugtig[1], Danielle McCool[1], Bella Struminskaya[1], Barry Schouten[1]

[1]Utrecht University, , Netherlands

Sensors and device intelligence functions available to a typical smartphone, such as cameras, accelerometers, GPS receivers, ambient light sensors, and predictive algorithms, have become embedded in users' everyday life tasks with the goal of making things easier, faster, and more accurate. Users have become accustomed to the ways in which these devices can improve their lives in very practical ways. The last decade has been marked by an increase in "smart surveys" seeking to augment existing survey methodology by using the tools readily at hand within smartphones (Link et al., 2014; Couper et al., 2018; Struminskaya, Lugtig, et al., 2020).

Although smart surveys can be deployed in isolation, researchers are often interested in integrating results from smart surveys with historical data sources and ongoing, established surveys. In addition, there may be a need to continue traditional surveys for specific sub-groups in the population, meaning that smart surveys and traditional surveys will be used alongside each other, much like in mixed-mode survey design. This presents an unfortunate conundrum, as data gathered from sensors and data acquired via survey questions are often very different with respect to their format, and sensor data can require considerable cleaning and processing before the two can even be directly compared.

In this presentation we will outline different methods that can be used to integrate data from traditional surveys with new smart surveys. We will use examples from several projects at Statistics Netherlands focusing on Travel, Household Budget and Time Useconducted in 2018, 2020 and 2022 to document areas of concern in the integration of smart and traditional surveys, and subsequently develop methodologies to account for these concerns as part of the integration process. We will further test these methods on data collected within the ongoing Smart Survey Implementation project (2023-2025), demonstrating their practical implementation using examples from different European countries.

What methodology should be used in practice depends mostly on the extent to which smart surveys and traditional surveys have different selection effects (different people in different methods) and measurement effects (differences in how the concept of interest is measured).

Decision Criteria to participate in Smart Surveys

Mr Johannes Volk[1]

[1]Federal Statistical Office Germany (Destatis), Wiesbaden, Germany

In official statistics, the development of new survey methods is seen as an important step towards modern data collection in order to reduce response burden, assure high response rates and guarantee high data quality. The German Federal Statistical Office (Destatis) is continuing to develop its data collection instruments and is working on smart surveys in this context. By smart surveys we mean the combination of traditional question-based survey data collection and digital trace data collection by accessing device sensor data via an application (GPS, camera, microphone, accelerometer, ...)

Unlike traditional surveys, smart surveys not only ask respondents for information but also require them to download an app to their personal smartphone and allow the app to access sensor data. These additional hurdles can have a negative impact on participation rates.

Destatis conducted three focus groups with a total of 16 participants to learn more about the attitudes, motives and obstacles of respondents regarding their willingness to participate in smart surveys. This was done as part of the European Union's Smart Survey Implementation (SSI) project, in which Destatis is participating alongside other project partners.

Overall, it became clear that participants are more willing to take part in a survey, to download an app and to grant access to sensor data if they see a purpose in doing so on the one hand and have trust on the other. The purpose can be given if the respondents understand why it is important to conduct the survey, why they should participate, why access to the sensor data is desired, what happens to the data and what it is used for by whom. However, the purpose can also be the respondents' own interest in the topic.

In order to motivate people to participate, it seems particularly important against this background to provide transparent information explaining what it is all about, as well as why and how the survey and sensor access are needed and what is being done to ensure a high level of data protection and data security. The discussions showed that a short invitation letter with the most important information was rated as good and sufficient, but that further information should also be provided, e.g. on an additional information sheet or flyer.

In the further course of the SSI project, a quantitative field test for recruitment is planned. The results of the focus groups will also be used to prepare this test.

From a smart travel-survey proof of concept towards an official statistic

Barry Schouten[1], Jonas Klingwort[1], Yvonne Gootzen[1], Mike Vollebregt[1], Daniëlle Remmerswaal[1]

[1]Statistics Netherlands/Utrecht University, Den Haag, Nederland

The need for smart surveys in official statistics is increasing. This is caused by the fact that traditional diary surveys, such as time use, travel, or household budget surveys, are burdensome for respondents. Statistics Netherlands has taken the first step towards implementing smart surveys and developed a state-of-the-art smartphone app to measure mobility behavior in the Netherlands. In order to move to official statistics, the design of the smart survey must be mature on four levels: methodology, IT, logistics, and legal. The design dimensions methodology and IT are especially new and challenging in a mobility survey; the general population must be willing to participate in location tracking, and IT must guarantee that such tracking runs without interruptions.

The app was tested recently (2022-2023) in a large-scale field test (initial sample n=3,200). This test aimed to analyze and evaluate aspects such as data collection strategy (response rate and participation duration), app technicalities (location tracking and battery management), respondent interaction (active-passive trade-off), data quality and validation (quality requirements), and AI-ML predictions (travel mode/motive prediction). The test's data were analyzed concerning these aspects, which we will report on in this paper.

Firstly, we report on the effect of the length of the reporting period (1 or 7 days) on the response rates and participation duration. Secondly, we will provide detailed information about the technical dimensions of the app and which points are central to its development. Here, we discuss aspects of different development kits and the differences between smartphones. Thirdly, we will provide information about the respondent's interaction with the app and how this influenced the data quality. Different app versions allow us to state whether a lot or little interaction should be preferred. Fourthly, a record-linkage study was conducted to validate the app measurements with external survey data. Various app algorithms that differ in complexity are evaluated in this context. Finally, we report preliminary results on travel mode and travel motive prediction. Especially the last two points can be considered cutting-edge research in Europe. These five analyses together give a picture of the maturity of the methodology and technology, which are important prerequisites for official statistics.

This paper comprehensively presents aspects of smart surveys that play a central role in their development and logistics. Consequently, this work contributes substantially to the currently active field of smart survey implementation at European NSIs.

Session 16 - Census & Multisource registers, June 6, 2024, 09:00-10:30

Behind the Scenes: Crafting Hungary's New Census Database

Dr. Melinda Oparin-Salamon[1]

[1]Hungarian Central Statistical Office, Budapest, Hungary

The Hungarian Central Statistical Office launched a new census database in 2023. The paper delves into the technical aspects of this in-house development, accentuating the use of open-source technologies and the benefits of adopting SDMX 3.0 for data description.

A departure from the old methods of disseminating Hungarian census data through Excel files and a constrained array of multidimensional datasets within an outdated database sets the stage for the shift. The inherent limitations of past practices hindered the ability to provide comprehensive insights to users. In pilot phase previous census data were seamlessly integrated into the new system, undergoing user testing. In response to user feedback and the identification of other issues, the system underwent refinement. As a result, the census database now including the latest census data provides a user-friendly interface that significantly elevates overall usability.

Noteworthy among its features is the absence of limitations on dimensions in multidimensional datasets (althought there is a non-technical limit of dimensions which can be comprehensible for users). The pre-execution of all the calculations (including statistical disclosure control) ensures rapid query execution. Users benefit from instantaneous presentation of chosen dataset data, allowing subsequent application of filters—a functionality that further enhance the system's advantages.

The usability improvements and a notable increase in user downloads underscore the success of the new census database. It presents now an inspiring model for renewing the Hungarian dissemination database as well. This paper narrates the story of a transformative endeavor, highlighting the practical implementation of a user-focused and data-driven approach to statistical dissemination.

Main innovations of the 2021 Spanish Census

Mr Jorge Luis Vega Valle[1], Cristina Casaseca Polo

[1]Demographic Unit, Ine, Madrid, Spain

Following the steps of the most developed countries, Spain completed in 2021 a long journey that began several decades ago and has conducted the first register based Census in its history.

This methodological change entails multiple innovations in different features and represents an increase in the quality of the information produced. Using information from administrative registers has not only allowed Spain to conduct a Census despite the challenges posed by COVID but has also significantly increased the amount of information collected on various topics.

Throughout this document, we will analyze in detail some of the most significant advantages that this methodological change has brought to Spain.

Construction of synthetic data from Population and housing census 2021

Boris Frankovič[1], Lukáš Lednický[1]

[1]Statistical Office Of The Slovak Republic, Bratislava, Slovakia

Population and housing censuses are getting attention from among students and the whole research community. They often apply for access to confidential census data to meet the goals of various scientific projects or master's thesis. Whereas in most of the cases provision of highly detailed confidential data is essential, there are some scenarios where also less precise data could be beneficial. To that end, statistical offices are increasing their efforts to produce publicly available synthetic data from various statistical domains. Although the limitation of disclosure risks reduce the data usability, they can still serve as an instrumental realistic data source for educational and learning purposes as well as for researchers preparing for their work in the Research data centre. In this paper we present a method of construction of synthetic data from Population and housing census 2021 at the Statistical Office of the Slovak Republic with the use of R software, along with information loss and data utility measures.

Census, Revision, ProjectionsExperiences from the Republic of North Macedonia (2018 -2023)

Mr Jane Krsteski[1]

[1]SSO Republic Of North Macedonia, , Macedonia

Census, Revision, Projections

Experiences from the North Republic of Macedonia (2018-2023)

The purpose of this paper is to show the path taken by the SSO in the period between the preparation of the 2021 Census and the preparation of Population Projections until the year 2070. The focus will be on highlighting the specifics and challenges that SSO faced in each of the stages and acquired knowledge in this process.

2021 Census was first census conducted using the combined method. It was the first time a PIN was used to identify persons throughout the registers.

Based on the possibilities offered by the use of PIN, the SSO developed a Pre-Census Database that used data from six administrative sources.

With the Census 2021, 2,192,778 persons were enumerated, of which 1,855,249 were enumerated in the field and 204,804 was self-enumerated by a special web application.

During the editing process, Pre-Census Database were used for quality control, this time as a source for evaluating the coverage and identifying the persons who are part of the total population.

As a result, total number of population in the Republic of North Macedonia in the 2021 Census was 1,836,713 people. In this process 258,932 enumerated persons was identified as living abroad for more than 12 months and were not included in the total population. At the same time, 132,725 people were identified as not enumerated during the field enumeration and were included in the total population.

Next step undertaken by the SSO was a revision of the series of population estimates between Census 2002 and 2021. This activity used data from Censuses and the birth, death and migration surveys owned by the SSO.

As a result of this activity, about 230,000 people were identified by PIN according to official address who were enumerated with the 2002 Census or were born in the period between the censuses and who are not enumerated by 2021 Census. Therefore, SSO assumed that these persons are out-migrants.

In the absence of relevant data on the time of migration for out-migrants a customized method of Linear interpolation based on Mirror statistics was used for revision of estimated population .

Based on the previous data, SSO prepared population projections until 2070. Seven variants where included in the projections based on four hypotheses about fertility, two hypotheses about mortality and three hypotheses about migrations.

Mixed-mode Census 2021 survey with voluntary part in Estonia to improve quality of the estimates

Ms Kristi Lehto[1]

[1]Statistics Estonia, , Estonia

The 2021 census in Estonia was mostly based on administrative data. All EU-mandatory characteristics were collected from administrative data. However, the purpose of the sample survey was to collect information on persons living in Estonia that is not available in the registers (religious affiliation, knowledge of languages and dialects, existence of a long-term illness or health problem and health-related limitations on daily activities).

In 2011 census, when Estonia used the first time of self-enumeration (CAWI - Computer Assisted Web Interview) the response rate of CAWI was 67%. This gave an idea to use the voluntary CAWI part in 2021 census survey.

The probability sample included approximately 40,000 dwellings (around 30,000 of these inhabited),

i.e. around 60,000 persons for whom participation in the population and housing census was mandatory according to the law. In CAWI mode all those who wished could respond voluntarily even outside the sample. CAWI response rate was 43,1% which is very high concerning that answering was voluntary.

CAWI respondents are different from CATI/CAPI (Computer Assisted Telephone Interview/ Computer-Assisted Personal Interview) respondents. They are younger, healthier, less religious and know more foreign languages based on Census 2011. In order to obtain the unbiased estimates, it is necessary to skillfully combine the data of different modes. The estimate extends the proportion of the surveyed characteristic found in CAWI respondents to CAWI population, and the proportion found in CATI/CAPI respondents to the rest of population.

To mix probability and non-probability (voluntary) sample helped to improve quality of the census estimates and publish more detailed breakdowns.

Quality assessment in the Istat Integrated System of Registers: An application to the estimation of the Attained Level of Education

Romina Filippini[1], Sara Giavante[1], Gaia Rocchetti[1]

[1]Istat, Roma, Italy

The potentiality of using new data sources has led National Statistical Institutes (NSIs) to reorganize their production system towards a more structured use of administrative sources, able to provide detailed information while reducing costs and response burden. Exploiting administrative and survey data, the Italian production system of statistics has moved towards a register-based statistics production system, built upon an 'Integrated System of Statistical Registers' (ISSR) composed of base registers and satellite registers.

In this context of modernization, to improve process efficiency and monitor the accuracy of results, the Italian National Institute of Statistics (Istat) introduced a shared framework for the quality assessment and documentation of the production process, able to capture the characteristics and specificities of the new paradigm of statistical production.

One of the main Italian base registers is the Base Register of Individuals (BRI), a comprehensive statistical register storing data gathered from different administrative sources. Core variables like place and date of birth, gender, citizenship are associated to each unit. Moreover, a variable denoting people usually resident in Italy is attached. This subset of data is the basis of the new Italian census that is as much as possible register based. According to this idea, the Attained Level of Education (ALE) is a variable for which a prediction in the register for resident population is obtained through a model based on the integrated use of administrative and survey information.

This document describes the application of the quality framework to the estimation process of ALE in BRI 2022. Specifically, the metadata model is applied for a structured description of the entire production process and a system of quality indicators is computed to monitor each process step during its implementation. The application highlights the importance of the framework from various perspectives .In particular, besides providing crucial information about the quality of the process and the resulting output, the computation of relevant indicators allowed us to monitor quality aspects during the production phase and consequently to immediately recognize suspicious data, facilitating the timely implementation of appropriate corrections. Additionally, the initial implementation of the framework's metadata model resulted in a revision of the sequence of some process steps, leading to an enhancement in overall efficiency. Finally, the structured description of the process through the metadata sheets included in the framework ensures an understanding of the process even by non- experts and guarantees the reproducibility of the process in subsequent years.

Session 17 - Assuring quality in statistics: academia & staff development, June 6, 2024, 09:00-10:30

Using cloud based flexible environment for training on open-source tools and development of new statistics using data from the web

Mátyás Mészáros[1]

[1]Eurostat, , Luxembourg

Open-source tools play a pivotal role in the realm of official statistics, particularly when it comes to harnessing new data sources. They provide a foundation of transparency, collaboration, and accessibility that is essential in a rapidly evolving data landscape. These tools can ensure that statistical agencies can adapt to emerging data sources swiftly, enhancing the accuracy and timeliness of their insights. Moreover, open-source tools encourage peer review and innovation, fostering a community of experts who collectively improve the quality and reliability of official statistics in the ESS.

To use such open-source tools in an enterprise IT environment has several challenges like to install, configure, maintain, and troubleshoot these tools and train users for them. To overcome these challenges, it requires a strategic approach and investment in a flexible environment for training.

Eurostat together with the central IT services of the European Commission, started a project to develop a “cloud agnostic datalab” that can be easily adjusted to the changing needs and environment. The datalab is based on open-source technologies so that it can be replicated in other NSIs and organisations and it is easy to launch containerised open-source tools and combine them as needed.

The presentation aims to provide an overview of this flexible cloud environment, about the available tools and how it is used for the training of the next generation of ESS statisticians in a standard virtual class environment and how it was used for upskilling in the ESS during the 2023 EU Big Data Hackathon. In addition, other use cases will be described to show how these open-source tools contributes to the quality improvement and development of new innovative statistics based on data sources from the web in the Web Intelligence Hub.

The Istat strategy to support quality by strengthening collaboration with academia

Dr Orietta Luzi, Dr Stefano Falorsi, Dr Marco Di Zio, Dr Mauro Scanu

[1]Department of Methodology, Italian National Statistical Institute (Istat), Rome, Italy

The need for developing methodological solutions for more efficient and high-quality official statistics in the modern information context, characterized by new (non-probabilistic) data sources and new technological assets, requires National Statistical Institutes to strengthen research collaborations and knowledge exchange not only among them, but also with the academia.

Over the last few years, Istat has developed a strategy for fostering effective collaboration with the academia, considering several forms of partnerships and research collaborations, and the participation in education and training of students and researchers. This strategy also exploits some of the existing Istat "infrastructures for research", established since 2017 in order to ensure that the methodological developments are in line with the current national and international academic research, so that robust and efficient methodological innovations actually improve the quality of statistical processes and products.

In the paper, we discuss the role of these infrastructures in facilitating collaboration with academia, such as the Advisory Committee for Statistical Methodologies, which involves experts from national and international Universities who are in charge of revising and discussing methodological research projects, teaching master classes and, since 2022, collaborating in the scientific organization of the annual Istat Workshop on Methodologies for Official Statistics. Also the Istat Innovation Laboratory is a virtual and physical place where Istat researchers can collaborate with academia on advanced methodological research projects, mainly in the area of big data usage and Machine Learning methods. Besides the traditional participation in education and training of students and researchers, we also illustrate the role of the Istat Research Committee to support joint research projects with the academia, e.g. by supporting the revision of the Istat administrative rules to allow for funding PhDs. In the paper we will present meaningful results achieved thanks to joint research activities, main problems encountered and open issues.

Establishing Skills and Competences for the Modern Statistical Office in a Learning Organization

Dr Mark van cer Loo[2]

[1]Statistics Netherlands, Den Haag, Netherlands, [2]University of Leiden, Leiden, Netherlands

Like many other Statistical Offices, Statistics Netherlands (CBS) is facing external challenges, including an increasing number of new and volatile data sources; the increasing demand for more current, detailed, and faster output; and the expanding tasks of NSIs towards becoming governmental 'data hubs'. To face those challenges, CBS is on one hand modernizing its data management, production, and dissemination systems according to centralized architectural principles that aim for a standardized, modular and metada-driven way of working. These changes have a profound impact on statisticians' day-to-day work, and therefore raises the question on what competences are necessary in the Statistical Office of the future.

At the same time CBS is facing an outflow of retirees and a labor market that is increasingly competitive. For these reasons CBS has adopted an HR strategy that expresses its desire to be progressive in its recruitment strategies; that adopts continuous learning as a standard; that supports management development in a CBS-wide context; that stimulates a vital work force, and that interprets outflow of staff as an opportunity, not a threat.

CBS' internal educational institute, the CBS Academy (CBSA) plays a central role in fulfilling the HR strategy. Established in 2019 as an internal training institute, it initially focused on setting up a curriculum supporting methodology and data science. Currently, the CBS Academy has four main curricula focused on Personal Development, Management Development, Organization Development and Substantive Skills. Each curriculum has a dedicated project manager who is responsible for the development and planning of trainings in their respective areas. The Academy has developed a vision on skills and competences that support the modernization of statistical production and is implementing this vision into its curriculum. Moreover, the Academy is the driving force behind CBS becoming a 'learning organization'. That is, an organization where its vision is shared by all staff, where there is room for (informal) learning and learning from mistakes (safe learning environment), where a growth mindset is the standard and where we turn our gaze outward to learn from new developments.

In this presentation we focus on the vision and mission of CBS Academy and its role at CBS. We detail its main achievements, the connection between the curricula and modernization, and lessons learned so far. We also look forward to recent developments where we broaden the scope to collaboration between (inter)national statistical institutes in the area of learning and teaching in official statistics.

Competing with the Private Sector: Intangible Employee Motivation Tools at the Czech Statistical Office

Ms Petra Kuncová[1], Marek Rojíček[1], Egor Sidorov[1]

[1]Czech, Prague, Czech Republic

Competent and highly motivated people is the basis for high quality statistics production. Shift to data science paradigm in the production routine increases requirements on human resources. NSIs face ever growing challenges of competing for employees with very dynamic and progressive financial and IT sectors, whose resources are far beyond budgetary possibilities of state institutions. In this respect, the topicality of intangible employee motivation tools to attract and sustain statisticians in the NSIs grows very rapidly. The presentation will focus at the approach to intangible motivation system and its recent developments within the Czech Statistical Office.

Session 18 - Special Session: Implementation of Quality Management Frameworks/ Systems in the Enlargement and European Neighbourhood Policy (ENP)-East Countries, June 6, 2024, 09:00-10:30

Special Session: Implementation of Quality management frameworks/systems in the enlargement countries

Claudia Junker[1]

[1]Eurostat, , Luxembourg

Based on improvement recommendations from different assessments, the selected enlargement countries (Turkey, Georgia, Moldova, Ukraine, Albania) are working on quality matters. They concern the setting up of systematic quality management and measurement, creating an appropriate institutional framework for it, reflecting on existing models of quality management and planning their implementation. The countries have chosen different ways of setting up and developing quality management systems. The objectives of the session will be to present these different ways to implement and improve quality management systems as well as to present elements of the quality management system extended to other national authorities producing official statistics. An exchange of existing experience, innovative and interesting practices will also be fostered.

The outcome of the session will be a better spread knowledge about the way these countries have embarked on setting up, implementing and improving quality management systems and experiences and practice that are particularly relevant for these countries but could also be applied in other countries.

This session will be unique in a way that for the first time three countries, which have recently become candidate countries, will join this session and therefore, it will be an excellent occasion to implement political decisions taken by the Council, also in a session in the Q2024 conference.

Due to a very high interest to actively participate in the conference and from their own sessions as expressed at several occasions by the colleagues from the NSIs of these countries and as witnessed in the 2016, 2018 and 2022 quality conferences and due to a considerable level of attendance from these countries in previous conferences, it can be expected that the session will be attended by around 40-50 participants. . A specific session at the Q2024 will provide an extra impetus to the countries to further improvements in the adopted quality management systems.

This session will be unique in a way that for the first time two countries from the ENP East region, which have become/will become candidate countries, will join this session and therefore, it will be an excellent occasion to implement political decisions taken by the Council, also in a session in the Q2024 conference.

Due to a high interest to actively participate in the conference and form their own sessions as witnessed in the 2016, 2018 and 2022 quality conferences, it can be expected that the session will be attended by around 40-50 participants.

Quality framework and implementation aspects in TurkStat

Dr Serdar Cihat Goren[1], Asila Koçak[1]

[1]Turkish Statistical Institute (TurkStat), ,

Quality frameworks and quality management have become important topics for the National Statistical Institutes (NSIs) of the enlargement countries. The main objective is to provide high-quality products and services to meet user needs. Quality is, however, not only needed and valued for products and services; it also relates to the institutions as a whole as well as to the institutional environment. Hence, it requires actions for the overall management, organization, and governance.

Turkish Statistical Institute (TurkStat) has declared to apply the European Statistics Code of Practice (ES-CoP), the European Statistical System Quality Assurance Framework (ESS-QAF), and the ISO 9001:2015 Quality Management System in the implementation of quality frameworks. TurkStat is aware of the necessity of continuous improvement within the national statistical system to elevate the quality of statistics. The PDCA cycle (Plan-Do-Check-Act) is taken as the basic philosophy for ensuring continuous improvement. In line with those aspects, it is aimed at producing high-quality statistics and ensuring user needs.

TurkStat carries out quality assessments under the name of Quality Logo for the official statistics produced by the institutions and organizations within the scope of Official Statistics Program (OSP) and under the name of Quality Monitoring and Assessment Tool (QMAT) for the official statistics produced by TurkStat. In addition, TurkStat has “ISO 9001:2015 Quality Management System” certification. To implement systematic quality management, quality-related tasks are consolidated under the ISO 9001:2015 Quality Management System.

In this study, as TurkStat, which has been serving as the producer and coordinator of the Turkish Statistical System since its foundation, it is discussed in detail how the quality framework and quality management are implemented and what the good practices are.

Enhancing Quality and Performance in Statistical Production

Dr Elsa Dhuli[1], Dr Liljana Liljana Boçi[1]

[1]Institute of Statistics Albania, ,

Quality is the foundation for building and maintaining the integrity of official statistics. It ensures that the statistics produced are accurate, reliable, and fit for their intended purposes, supporting the effective functioning of the society and economy. Total Quality Management (TQM) strategy is being employed and several initiatives have been undertaken by the Albanian Institute of Statistics in order to enhance the efficiency and quality of statistical processes. An institutional Performance Assessment Framework (PAF) has been developed and improved over the years to support organizational effectiveness, and continuous improvement, while working on standardisation of statistical processes and production.

This paper will provide information on the development and challenges to further improve quality management. The use of Generic Statistical Business Process Model (GSBPM) as well as the steps taken towards integrating the Performance Assessment Framework (PAF) with GSBPM and Generic Activity Model for Statistical Organizations (GAMSO) will be described.

This integration will ensure the quality and efficiency of the statistical processes but as well as help to systematically evaluate and enhance the overall performance in delivering reliable statistics. This integration contributes to a holistic approach to statistical management, encompassing both process optimization and performance excellence, guaranteeing alignment with the established principles outlined in the European Code of Practice for official statistics.

Quality initiatives in challenging times

Nataliia Pavlenko[1]

[1]State Statistics Service of Ukraine, Kyiv , Ukraine

The State Statistics Service of Ukraine (SSSU) is constantly focused on the development and making improvements to quality management system of the state statistics authorities. In order to comply with the principles of the European Statistics Code of Practice, the SSSU in 2010 has approved the Principles of activity of the state statistics authorities.

Ukraine's national statistical system includes three government authorities: the SSSU, the Ministry of Finance and the National Bank. Article 4 of Ukraine's Law on Official Statistics contains the all principles of the European Statistics Code of Practice, this determines their implementation at the legislation level by the members of the national statistical system.

In 2021 the Commission on coordination of the SSSU activity's quality management system has been set up. Its composition incorporates the representatives from 10 functional and sectoral units.

The documentation dealing with the SSSU activity's quality management system comprises:

- Policy of quality;
- Risks management policy;
- National model for activity of the state statistics authorities and documented activity processes;
- Technological maps for statistical production;
- Questionnaire to monitor the activity quality indicators.

In 2022, Ukraine has been granted the candidate-country status to join the EU, that gave impetus to activate this activity. Ukraine's government makes efforts to bring Ukraine's legislation closer to the EU legislation. In 2023, the Cabinet of Ministers of Ukraine has given the state authorities a list of the EU acquis to analyze and develop measures which are to be implemented in Ukraine of which nearly 700 were received by the SSSU.

Since the beginning of 2023, the preparation of quality reports basing on the European standard SIMS and reports on administrative data quality has been launched.

Every year, the SSSU makes the Inventory (consideration and analysis) of statistical observations. Currently, the issues that deal with an increase in the number of statistical data submitted to the Eurostat are rather relevant. The 2022 Eurostat monitoring report has brought to light the spheres of activity that require significant efforts to be made. At the same time, there is a problem with obtaining data from respondents and administrative data suppliers under the full-scale war.

This problem is planned to be resolved by developing and providing the Cabinet of Ministers of Ukraine with the draft Ukraine's law on making changes to Ukraine's law.

Quality management system at Geostat

Mr. Gogita Todradze

[1]National Statistics Office of Georgia (Geostat), Tbilisi, Georgia

The Adapted Global Assessment Mission, conducted jointly by Eurostat and the UN ECE Statistical Division in 2012, recommended the introduction of a quality management system at Geostat and the use of the ESS quality parameters to monitor the quality of statistical outputs.

In May 2014, a Methodology and Quality Management Subdivision (at present Division) was created at Geostat. The main area of this unit is to handle quality issues within all areas of Geostat activities. Creating a full-time unit dedicated to quality issues was a key institutional element for the systematic monitoring of quality.

Based on desk research on international standards, the EU model as a priority for the Quality Assurance Framework based on the ESS Code of Practice (CoP) has been selected.

Division has taken some active steps towards monitoring and improving quality. Quality mapping was conducted for all processes; standard routine descriptions based on GSBPM have been prepared; and the following policy documents and guidelines have been prepared: Quality Policy document, Response Burden Reduction Policy, Revision Policy and Error Correction, Questionnaire Design, Guidelines for Assuring impartiality and objectivity in the Production and Dissemination of Official Statistics. Geostat produces and disseminates metadata based on EURO-SDMX Metadata Structure (ESMS) 2.0. The guidelines to manage the ESMS-based metadata have also been developed.

A high-level working group on quality management issues, consisting of Geostat's top management, heads of departments, and Methodology and Quality Management Division representatives, has been created. The working group is eligible to cooperate with experts, representatives of administrative bodies, international organisations, and partner countries if needed.

The quality audit and self-assessment have been conducted on a regular basis since 2023. The current self-assessment questionnaire has been developed based on ESCoP. Under the Twinning project, Geostat is aiming to adopt a self-assessment survey based on DESAP.

In February 2023, an international team of experts convened by the UNECE conducted a Sector Review of the implementation of the Generic Statistical Business Process Model (GSBPM) in Georgia with the aim of producing a roadmap for organisational change. A road map for the introduction of a more process-based organisational structure in line with the GSBPM was developed. The following steps are preparatory work to align Geostat's statistical processes with the GSBPM as part of the Twinning project.

Session 19 - Statistics and decision making I, June 6, 2024, 09:00-10:30

The use of data in education policies in Portugal: teacher grades in the presence of external assessment

Pedro Luis Silva[1], Patricia Pereira[1]

[1]Directorate-General of Statistics on Education and Science (DGEEC), Lisbon, Portugal

Education policy in Portugal has seen a transformative shift propelled by a collaborative effort between academia and the Directorate-General of Statistics on Education and Science (DGEEC). Our study evaluates grading standards in Portuguese high schools, specifically focusing on the disparity in teacher-assigned grades within the context of external assessment presence.

Our investigation revealed a striking discrepancy in grading practices between subjects with national examinations and those without, particularly prevalent in annual 12th-grade courses lacking a national exam requirement. Teachers in these non-examined courses tend to assign markedly higher grades compared to subjects with standardized national assessments. Concentration of grades on the top of the distribution grades is potentially problematic for students, institutions and society and also impacts assess to higher education.

This work was crucial to inform Portuguese policymakers. Recognizing the implications of this disparity on educational equity and accuracy, the government was prompted to undertake a policy intervention. Namely, the formula of high school GPA calculation was changed the weights given to the annual 12th-grade courses in high school was reduced (those where concentration of high grades is prevalent).

This collaborative endeavour between academia and statistical office underscored the indispensable role of robust data analytics in shaping evidence-based policy decisions. For this work we used different administrative data sets for the population of students in Portugal from secondary education to higher education for the period 2016 to 2022.

Our presentation at the European Conference on Quality in Official Statistics seeks to delineate the methodology employed, the findings elucidated, and the consequential policy adjustments initiated in response to our collaborative research endeavour. It offers a compelling case study showcasing how data-driven insights can engender substantive policy reforms within the realm of education, thereby ensuring a more equitable and accurate evaluation of student performance.

How to support political decision making in times of crisis

Ms Ilka Willand[1]

[1]Head of communication, Destatis, Wiesbaden, Germany

At the beginning of the Covid pandemic in Germany, high-level political decision makers realized: Statistical data, which they urgently needed, was scattered across various websites and databases, in different formats and without consistent quality standards. There was no single point of access to the most relevant indicators from different data sources to enable decision makers to improve the management of crises. As a result the Federal Statistical Office of Germany (Destatis) was commissioned to develop a tailor made digital product based on the requirements of political decision makers.

Destatis launched the “Dashboard Deutschland” after a few months of development in winter 2021. The product is closely built around the needs of the target group and is therefore unique on the German data market. It is strictly focused on crisis-relevant and frequently updated key-figures (mostly economic indicators) whether official data or data from other sources, compiled by the Federal Statistical Office. As a dashboard it is focussed on interactive visualisations and short texts – without tables. Users can also create customized personal dashboards. It contains a feature called “PulseCheck” to compare time series of economic indicators from different sources to identify connections in the development.

The presentation shows how an NSO can be responsive to the needs of a strategic relevant target group during times of crisis and how to increase relevance and trust. But the new product required a fundamental mind-set-change within the NSO. The Federal Office took the role of a curator for data from different sources for the first time. The presentation provides an insight regarding the challenges that the NSO faced during the conception and development. How does the NSO ensure high data quality? How is the product embedded in established publishing processes? How does the NSO measure the success? How is the product developed further? Do political decision makers have influence on the development of the product?

THE CIRCLE OF VIRTUE. From strengthening the institutional framework to improving statistical quality and from statistics to policy implementation

Mrs Claudia Villante[1], Mrs Maria Giuseppina Muratore[1], Mrs Lucilla Scarnicchia[1]

[1]ISTAT - Italian National Institute of Statistics, Rome, Italy

ISTAT (National Statistical Office) started to study violence against women launching a first national survey on 2006, after an accurate planning phase. The survey has enabled the developing of indicators to analyse the phenomenon in Italy through the production of evidence-based data. In addition, thanks to the replication of the same survey in 2014, it was also possible to observe the evolution of the phenomenon over time (in terms of, for example, incidence, frequency, victims' propensity to report, etc.).

On 2017 the Institutional Agreement between ISTAT and the Italian Presidency of the Council of Ministers – Department of Equal Opportunity has been settled aimed at building the integrated data collection and processing system envisaged by the Action Plan against Sexual and Gender-based Violence, according to the specifications and modalities provided therein, defined as "integrated system of violence against women" .

The Agreement has been not only an institutional effort to implement the provisions of national Law (n.119 of 2013), but also the way to boost the production of information from different sources of data, in line with the 3Ps strategy of the Istanbul Convention (Prevention, Protection and Prosecution). Moreover, the agreement strengthened the cooperation among data producers both public (such as Ministry of Health, Ministry of Interior and Ministry of Justice at national level and the Regions, at the local ones) and private entities (as NGOs), deeply engaging in combating gender- based violence.

From May 2022 the statistical efforts to provide quality data to observe the phenomenon of gender-based violence has been boosted from a new legal framework (Law n.53/2022) which gives even more relevance to the systematic production of statistical data on violence against women. The new legal framework commits ISTAT, Ministry of Interior, Justice, Health to measure violence against women, focusing on its causes, dynamics, consequences, looking at regularly monitoring the phenomenon and the victims' protection. This involves continuous work on improving the quality of both existing administrative and survey data, as well as innovating methodological tools and exploring new fields of research: experimental statistics (big data) and machine learning process has been set up to analyse data base provided by the Department of Equal Opportunity's help-line, or studying the protocols of territorial governance networks to combat violence against women.

The paper describes the impacts of institutional mechanisms on improving data quality and innovation processes and describes some of the best practises adopted and their results on data production.

Communicating Ethnicity data quality

Mr Darren Stillwell[1]

[1]Cabinet Office, LONDON, United Kingdom

The UK government is committed to levelling up opportunity and ensuring fairness for all. The Cabinet Office's Equality Hub is central to this commitment.

One of our priorities is improving the quality of ethnicity data held by UK government departments. High quality data about outcomes for different ethnic groups is vital. It can inform effective government interventions to reduce unjustified disparities.

The Equality Hub effects improvements through developing and overseeing programmes of data quality work. We provide guidance to analysts in other departments collecting and analysing ethnicity data.

This session describes the different ways we communicate this guidance. It focuses on three outputs.

First, our innovative Standards for Ethnicity Data. These standards can help improve the collection and reporting of ethnicity data, to improve understanding of ethnic disparities and increase the value of ethnicity statistics. They can lead to better decisions and outcomes for the public.

Second, our 'Methods and Quality Report' (MQR) series. These look at different aspects of the quality of ethnicity data. They provide guidance based on real world ethnicity data.

MQRs are an important way that we alert users to different data quality issues. One MQR described disparities in a particular approach to policing - "stop and search" - between black and white people in England and Wales. If geographical differences are not taken into account, these disparities can be misleading. Another MQR compared ethnicity data provided by a third party with data provided by individuals. There is also an MQR on why detailed, granular ethnicity data is important.

Third, blog posts are another crucial way we communicate important data quality issues. We have discussed the importance of harmonised standards for ethnicity. One blog post outlined the difficulties of comparing international ethnicity data. There has been much internal and external user interest in our posts.

All these outputs package our own thinking on data quality issues for a wider audience. One of our guiding principles is transparency. It is the root of all data quality improvements. By being transparent and communicating our views about different aspects of data quality, we hope to encourage other departments to be clear on the strengths and limitations of their data as well.

Feedback we have received from internal and external users on our outputs has been positive. They have led to internal and external stakeholders holding us in high regard for our technical expertise.

Statistical Training of Policymakers: A cross-country study

Ms Ghislaine Tegha Ghislaine[1]

[1]Development Economics Group, ETH Zurich, Zurich, Switzerland

Policymakers are increasingly required to integrate data and evidence into their decision-making processes. In this study, I investigate the extent to which policymaker training programs adequately prepare them for an evidence-informed approach. Acknowledging the various approaches countries have towards policymaker training, this study analyses qualitative data from curricula and semi-structured interviews with key informants from policymaker training institutions across several countries with compulsory pre-service training programs, as well as those without such requirements in Africa and Europe. I adopt a twofold approach: first, I conduct a comprehensive content analysis of training curricula; second, I conduct a thematic analysis of interviews from policymaker training institutions to gauge the emphasis placed on statistical and data literacy skills. Preliminary findings from the Democratic Republic of Congo suggest that despite a clear interest among policymakers in adopting evidence-based approaches, there is a noticeable absence of statistical training within the compulsory training cycle. Through these analyses, the study aims to shed light on the level of preparedness of policymakers in effectively using statistical tools and methodologies within their roles.

To mislead or not to mislead – why preventing misuse of statistics is more effective than combatting it

Mrs Elise Rohan[1]

[1]Office For Statistics Regulation, , United Kingdom

As our world becomes more abundant with data, statistics are increasingly used to persuade and provoke discussion. The UK Office for Statistics Regulation's (OSR) vision is that statistics should serve the public good. Statistics play an important role in supporting democracy and a big part of that is encouraging their use in wider debate, but it also involves combating and safeguarding against misleading statistics.

This paper explores the definition of misleadingness in the context of our work as a statistics regulator and how preventative measures can uphold public confidence in statistics, enabling misuse to be more easily called out where it occurs. The paper draws from OSR's experience of investigating the potential misuse of statistics and sets out the different ways misleadingness can present in the production and use of statistics. For example, the challenge with complex statistics being distilled into headlines for press releases and social media posts. The paper shares our experience of investigating concerns in the lead up to the UK government 2019 General Election and how this will underpin our regulatory response to the 2024 General Election.

The repetition of incorrect or unsupported statistics creates a validity through reuse, known as the 'illusory truth effect' or repetition bias. Research on this phenomenon has found that we have a cognitive bias to perceive confidence and fluency as characteristics of truthfulness. Our paper discusses why focusing on reprimanding those that misuse statistics does not drive the system wide changes needed to prevent it happening in the first place.

We share our findings from our 'intelligent transparency' campaign which is formed around three principles: equality of access, enhancing understanding, and independent decision making and leadership. These principles are designed to support statistics producers to consider the potential for misuse throughout the life cycle of statistics. We also highlight the power of collaboration and partnerships to call out repeated misuse by individuals.

Finally, the paper takes stock of our learning to date on championing the effective communication of statistics to support society's key information needs. We share our findings on the common pitfalls of communicating statistics that open statistics up to misuse. This includes the presentation of uncertainty and the use of infographics.

Session 20 - Big Data, June 6, 2024, 09:00-10:30

Improving efficiency in assignment and quality control of NACE codes combining innovative methodologies with human expertise

Ms Athanassia Chalimourda[1], Mr Lorenz Helbling[1], Mr Mathias Constantin[1]

[1]Swiss Federal Statistical Office, Neuchâtel, Switzerland

NOGAuto is an assistance system designed to support coding experts in assigning NACE codes to establishments, based on descriptions of their economic activity. This innovative system, developed by the Business Registers Data Section of the Swiss Federal Statistical Office (SFSO) uses Natural Language Processing and Supervised Machine Learning and exploits the hierarchic structure of the nomenclature générale des activités économiques (NOGA), the Swiss version of NACE.

Bridging domain and methodological expertise from the Business Registers Data and the Statistical Methods Sections, we employ a series of quality measures to assess the overall and by-class performance of NOGAuto by comparing the agreement of its predictions to existing NACE codes. Classes of economic activities are inherently imbalanced; we therefore include measures which account for class imbalance like the balanced accuracy. For the 21 NOGA-Sections in our current test set, NOGAuto achieves overall accuracy, balanced accuracy and Cohen´s Kappa of 90%, 87% and 89% respectively. These values are slightly lower at the next lower NACE level consisting of 88 classes. By-class performance is assessed by measures like precision and recall. While the corresponding work is still in progress, we show in three examples of NOGA-Sections how by-class measures combined with the prediction probability can be used to distinguish areas where automatic classification works well from areas where the expert should rather complete the coding task.

Although NOGAuto was originally developed to assist experts in their coding work, a further application is planned in the context of quality control of existing codes. In a first use case, NOGAuto helps to limit the effort of detecting misclassifications in a series of about 50’000 codes assigned to activity descriptions in French and German, two of the four official languages in Switzerland. Codes which deviate from NOGAuto predictions are prioritized for a review of their coding, thus streamlining the quality assurance process.

Integrating an innovative system into statistical production is a challenging task. High standards on quality must be met while allowing for progress in innovative methodologies. Continuous monitoring of adequate global and by-class performance measures helps to combine these seemingly contradictory aspects. We show how connecting NOGAuto with other expert-systems, like the automatic translation service DeepL and an SFSO-internal rule-based classification tool fosters efficiency and user-friendliness. The coding experts with feedback on the final code and recorded comments support the development of NOGAuto, thus continuously improving efficiency and quality throughout the coding process.

Early estimates of maritime traffic using innovative data sources

Mr Nikolaos Roubanis, Ms Boryana Milusheva

[1]European Commission - DG Eurostat, , Luxembourg

In 2023, Eurostat and the European Maritime Safety Agency established a cooperation agreement to develop methods and produce early estimates of European ports vessel traffic, exploring the use of Automatic Identification System (AIS) and other administrative, and commercial data available in EMSA. As a first step, EMSA data on vessel traffic were aggregated according to the Eurostat vessel type classification and compared to data for selected high-traffic ports submitted to Eurostat under the Directive 2009/42/EC on statistical returns in respect of carriage of goods and passengers by sea. In a second step, quarterly trends in vessel traffic at port and country level for the years 2015 to 2019 were calculated for each dataset. The comparative data analysis indicated a good match, particularly of the trends at EU level. A method was then developed to estimate vessel traffic (port calls) for the most recent quarter with a model integrating Eurostat data of the previous years and the most recent EMSA data. The method was tested on quarterly data of the past 5 years, showing results with a deviation at EU level for all reporting ports of 2-4% from the statistical data Eurostat received a year later. Further work to reduce identified differences includes better understanding of how countries classify certain vessel types reported to Eurostat and improving the aggregation of EMSA data to better match Eurostat vessel classification. It will allow for more accurate estimates and more granular short-term traffic statistics at port and vessel type level across the EU. Furthermore, the project will improve the quality of maritime statistics by improving timeliness and accuracy in the classification of data by vessel type.

AIS-Driven Maritime Insights: Improving Italian Port Traffic Analysis

Dr Angela Pappagallo[1], Luca Valentino[1], Francesco Sisti[1], Norina Salamone[1], Mauro Bruno[1]

[1]Istat, Rome, Italy

The Maritime Transport survey (TRAMAR) provides statistics on the transport of goods and passengers carried out for commercial purposes. The survey has a census character, referring to all arrivals and departures made in Italian ports by ships with a gross tonnage of at least 100 tons for commercial reasons. Fishing boats, military vessels and sailing vessels are excluded from the observation domain. The estimates provided are necessary for the fulfillment of EU obligations, such as the Eurostat-requested F2 table, reporting the number of arrivals in Italian ports. In this regard, Eurostat proposed transmitting the F2 table quarterly instead of annually. Italy, while not responding to this request yet, started a research study to combine Automatic Identification System (AIS) data with traditional sources, to enhance data dissemination timeliness.

AIS is an automatic tracking system extensively used in the maritime world, for safety and management purposes. AIS produces a huge amount of real time data, containing information regarding the navigation status of vessels. The analysis of AIS data can improve the overall quality of maritime traffic statistics. In some research studies, AIS have been used to calculate covariates that allow arrivals to be estimated using a statistical model. Our goal, while ambitious, is to reconstruct the complete routes of all vessels visiting Italian ports.

A route is defined by vessel's identifier and two consecutive port visits, the first at the departure port, the second at the arrival port. Finding a port visit, given by a ship being in a port area at speed zero, is the first step to obtain a route from AIS data. Unfortunately, this simple idea faces several obstacles in practice, mostly due to errors in the data values. Indeed, AIS data, coming from both terrestrial and satellite sources, are merged into a single database through a complex integration and reconciliation process, resulting in some inconsistencies and shortcomings.

Thus, we propose a methodology to overcome these issues to avoid losing vessel's data, which could imply losing port visits, while reconstructing the complete route of the vessel. To this end, we developed an algorithm to detect missing AIS data and attempt deterministic or probabilistic imputation. We assessed the quality of our methodology by comparing the results with statistics produced by traditional sources. We achieved good results for most ports, but we still observe a partial coverage of the phenomenon for smaller ports, characterized by frequent and short vessel travels.

Qualitative Assessment of Wikipedia-Sourced Big Data on Enterprises

Alexandros Bitoulas[1], Mr. Fernando Reis[2]

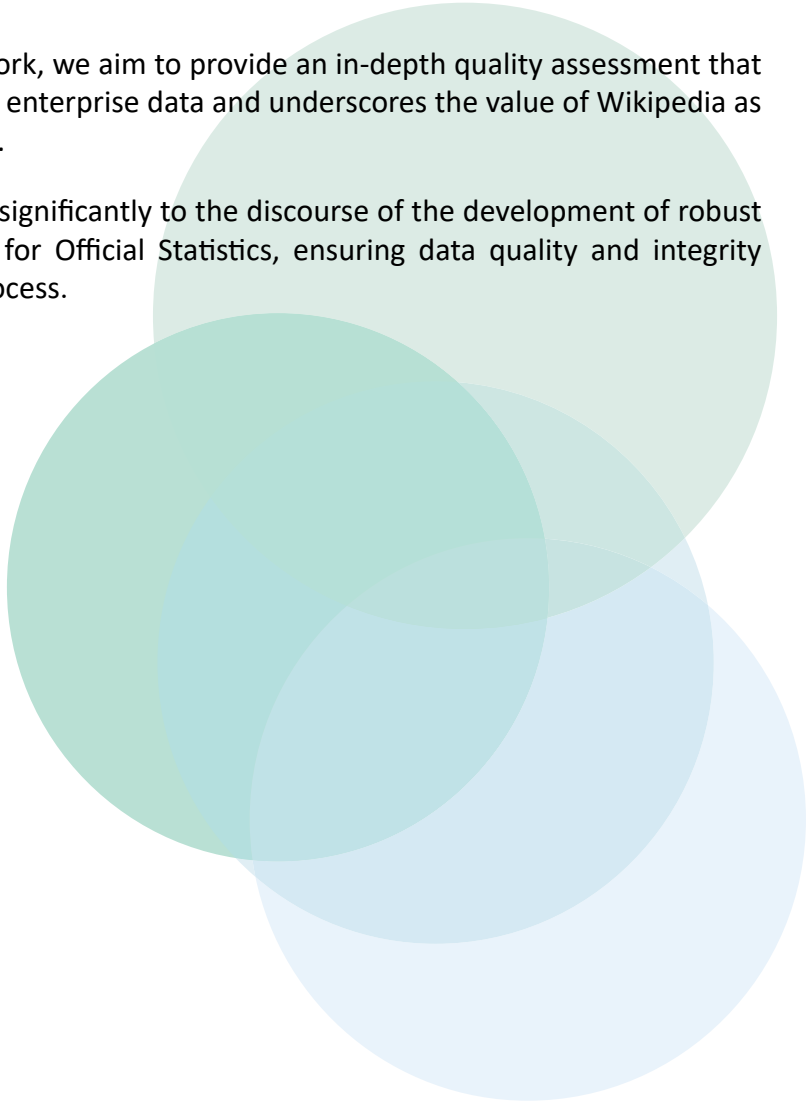
[1]Sogeti Luxembourg, , , [2]Eurostat, ,

In the evolving landscape of Big Data analytics, the integrity and quality of data are pivotal, especially in complex environments like enterprise data. This study embarks on a qualitative assessment of Big Data on enterprise data, with a unique focus on data sourced from Wikipedia, adhering to the suggested Big Data Quality Framework by UNECE (2014).

Our analysis spans the initial stages of the data lifecycle - the Input and Throughput phases - and extends, to a lesser extent, to the Output phase. The Input phase examines the acquisition and pre- acquisition analysis of Wikipedia-sourced data, emphasizing aspects like the institutional and business environment, the complexity of the data, the completeness of metadata, the linkability and the selectivity of the data, all key components of the statistical quality of a data source. In the Throughput phase, we delve into the transformation, manipulation, and analysis of this data, underlining the principles of system independence, steady states, quality gates and discussing how the presence of unstructured information and noise can significantly influence the quality of the data. Additionally, an assessment of the Output phase is conducted, evaluating the reporting and dissemination qualities of the derived Big Data product, including its conformity to standards, coverage and overall relevance.

By applying this comprehensive framework, we aim to provide an in-depth quality assessment that aligns with the intricate requirements of enterprise data and underscores the value of Wikipedia as a source of Big Data for Official Statistics.

Our findings are expected to contribute significantly to the discourse of the development of robust Big Data strategies on enterprise data for Official Statistics, ensuring data quality and integrity throughout the statistical production process.



Impact of the partial time coverage of retail chain data on the accuracy of the price index calculation

MS Petra Mazurekova[1], Helena Glaser-Opitzova[1]

[1]Statistical Office Of The Slovak Republik, , Slovakia

In the framework of modernization and improvement of the quality of price statistics by using new data sources, the Statistical Office of the Slovak Republic (SOSR) has commenced using transaction data from retail chains, also called scanner data. SOSR collects data for food and non-alcoholic beverages sector from the five largest retail chains that contain sales and quantities sold aggregated on weekly basis for each individual items. The implementation of the new data source in the production environment required a significant methodological change. The calculation of price indices no longer involves the price determined at a specific time (as in traditional collection), but the average price per unit of goods for the observed period. This type of price more accurately reflects the prices that consumers pay throughout the entire observed period, taking into account discounts and the impact of these discounts on the quantity of goods sold. Consequently, weekly data sets are aggregate on monthly level due to the frequency of the compilation of the price indices. These monthly files contain aggregated values of sales and quantities sold for individual product items for the selected weeks of the month and subsequently the average price of individual products is calculated. The accuracy of the average price is influenced by the length of the time period considered within the month. Theoretically, the best practice would be to use a complete month for the compilation of Harmonized Index of Consumer Prices (HICP). However, in practice, due to the HICP publication, the time span of price data only covers two weeks for the reference month. The aim of this paper is to assess the impact of insufficient time coverage, i.e. to examine the impact of using price data with different lengths of time span on the values of average prices and price indices. We use data from Slovak market to compare different indices Jevons, Törnqvist and GEKS-Törnqvist.

Session 21 - Special Session: The ESS Innovation Agenda, June 6, 2024, 11:00-12:30

Innovation in AI and quality: the one-stop-on AI/ML for statistics

Ms Francesca Kay[1]

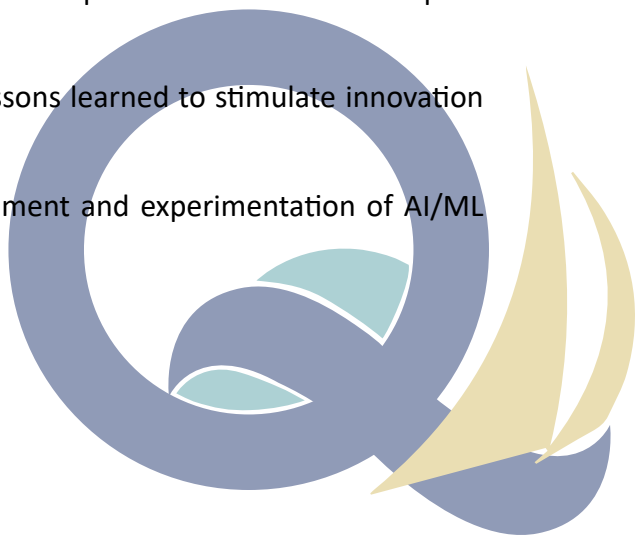
[1]Central Statistics Office, Ireland, , Ireland

The use of Artificial Intelligence/Machine Learning (AI/ML) for the production of official statistics is one of the strategic domains that need to be developed further and where coordinated action is beneficial.

The potential for AI/ML is still being developed and it is timely that a coordinated action to enable systematic learning, sharing of experiences, identification of good practices and reuse of solutions should be undertaken now, with the resultant economies of scale, reduction of costs and improvement of quality at individual NSI level. Taking a cross-European approach will facilitate the scaling up or reuse of existing solutions, the standardising of methodology, improvement of quality and will allow for the coordinated consideration of legal and ethical issues which could slow progress. Eurostat issued a grant call for the creation of a one-stop-shop on AI/ML for official statistics which has led to the creation of a consortium of 14 countries to deliver the project. The One-Stop-Shop for Artificial Intelligence/Machine Learning for Official Statistics (AIML4OS) will play an important role in developing innovative solutions with respect to statistical products and processes, allowing for more timely production of official statistics and the delivery of better responses to user needs. The project will develop knowledge and use cases supporting the use of AI/ML-based solutions for the production of official statistics.

Key to the success of the project will be addressing quality challenges as they arise. Quality will be at the heart of the delivery and will be looked at in a number of different ways by:

- developing, maintaining and evolving a coherent set of relevant capabilities – including methodologies, guidelines, sandboxes, labelled data, processes, methodological, implementation and quality frameworks for implementing AI/ML based solutions in official statistics across the ESS,
- providing support and guidance for the integration and maintenance of relevant AI/ML based solutions in ESS organisations through training and active and efficient support,
- building communities around open-source solutions developed and maintained by ESS members,
- delivering use cases to produce new, innovative statistical products which can be adopted by ESS members,
- sharing ideas, experiences, success stories and lessons learned to stimulate innovation based on the use of AI/ML and
- embedding quality in the transition from development and experimentation of AI/ML based solutions to actual production.



The ESS innovation agenda implementation

Albrecht Wirthmann[1]

[1]Eurostat, Luxembourg, Luxembourg

In February 2023, the European System Statistical Committee adopted the ESS innovation agenda. It aims at organising, coordinating, supporting, and contributing to the development of innovation activities in the ESS. The scope of the ESS innovation agenda is broad on purpose covering new and improved statistical products and processes but also pays the necessary attention to data sources and their methodological, quality, and legal dimensions.

In order to achieve the goals outlined above, the ESS innovation agenda is deployed through activities of two types: 'innovation execution' and 'innovation mechanisms'.

Projects, which can lead to innovative statistical products and services or improved statistical processes are referred to as 'lighthouse projects'. Technical and methodological developments strengthen the generic ESS capabilities required to realise innovation are called 'cross-cutting projects'. The ESS innovation agenda mentions in particular use of new technologies such as Artificial Intelligence, privacy enhancing techniques in the exchange and use of sensitive data across organisations, use of smart devices, advanced methods for data integration and producing multi-source statistics, technological frameworks for integrating geospatial data, capabilities to use cloud native environments for developing and deploying large scale novel technologies, and methods to assess and improve quality in statistical processes and products.

Improving quality requires the extension of methodological and quality frameworks to non-traditional approaches of producing official statistics, for instance, non-probability sampling methods to correct for selection bias, transparency of ML algorithms and other methods, adaptive/rolling survey design, multisource and mixed-mode statistics. It can cover analysis and communication of sources of uncertainty. These developments can contribute to produce a coherent framework for quality assurance and transparency.

The innovation agenda is deployed in a series of activities included in the Annual Works Programme of Eurostat and the ESS. An initial list of 21 projects were included into the innovation action plan resulting in a balanced portfolio of ESS innovation activities with a high potential to produce synergies across the various activities. This is specifically important for improving quality, a cross- cutting capability of relevance for a significant number of innovation projects. A specific project will aim at enriching existing quality frameworks with elements of a quality framework for non-traditional data sources.

The execution of the innovation agenda is supported by the ESS Innovation Network, a new expert group, which will facilitate and oversee the innovation processes across the ESS.

Official statistics of a lifetime - and beyond?

Mr Hans Viggo Sæbø[1]

[1]Statistics Norway, Oslo, Norway

Official statistics are part of the infrastructure of democratic societies. As societies have changed statistics have changed accordingly, not so much in content but in ways of production and dissemination. This paper describes important trends in official statistics over approximately the last 50 years, a period roughly corresponding to the author's professional experience in this area.

The trends considered comprise global trends affecting official statistics and the corresponding changes in these statistics and the way they are produced. The general quality revolution in the last century has affected official statistics. This includes the recognition of such statistics as a public good available for free for everyone, and the development of institutional quality frameworks for official statistics. The Internet caused a shift in the way statistics were spread, while more use of secondary (non-statistical) input data has changed the way of producing statistics.

Based on the history, some thoughts on the future development of official statistics and how to meet new challenges are addressed. Now the data era threatens to replace statistics. Major shifts are difficult to foresee, but there is no doubt that the development of artificial intelligence will affect both the way statistics are produced, and not least how statistics are understood and used in society. Quality frameworks for official statistics may be changed, but the main principles or core values of such statistics should remain. Cooperation and statistical literacy are keys. Official statistics must be associated with trusted institutions.

Aspects linked to the development of European and Norwegian statistics are briefly described. Examples referring to Statistics Norway are believed to be valid beyond the national level.

Special Session Proposal on the ESS Innovation agenda

Mr Albrecht Wirthmann

[1]Eurostat, Luxembourg, Luxembourg

Topic: In a world of constant change, new and improved statistical products have become a necessity for official statistics. On the one side, demand from users is evolving, they ask for more granular, more timely and better integrated data. From a quality perspective that means Official Statisticians have to reconsider what users need and how we can provide “fit for purpose” information. On the other hand, technical opportunities appear at such a pace that official statistics have difficulties to follow, assess and integrate them in the production of official statistics while they entail the potential to keep up the high quality standards that have contributed to make official statistics a trusted source of information.

In February 2023, the ESSC adopted the ESS innovation agenda. It aims to organise, coordinate, support and contribute to sustainable innovation activities of the ESS. These activities seek to create both new statistical products and new capabilities necessary to develop and implement innovative products. In addition, the ESS innovation agenda aims at putting in place mechanisms to enable innovation, by creating conditions for integrating innovation into production and for sharing innovative ideas.

The ESS innovation agenda is recognising quality as an overarching and cross cutting dimensions of all ESS innovation and methods to assess and improve quality will receive a constant attention.

The session will be the opportunity to highlight some of the key features in implementing the ESS innovation agenda, to exchange on the quality challenges at stake when innovating and discuss the way forward.

Session outline:

Paper 1: Albrecht Wirthmann, Eurostat – Highlight of ESS innovation agenda implementation It will present the concept of the ESS innovation agenda highlighting its quality dimensions and challenges

Paper 2 : Hans Viggo Sæbø (Stat Norway) - Innovation and modernisation in Official Statistics.

It will describe the general evolution of official statistics and the effort to continuously maintain the quality of official statistics. It will also elaborate on its future developments.

Paper 3: Francesca Key (CSO, tbc) – Innovation in AI and quality: the one-stop-on AI/ML for statistics. It will outline the new project for accelerating AI/ML adoption in the ESS and show how the quality challenges it raises will be addressed.

Paper 4: Barteld Braaksma (CBS, tbc) – Scaling out innovation.

It will describe CBS experience on the path for scaling innovation in the production of Official Statistics while addressing the need to maintain high quality standards

Scaling out innovation

Barteld Braaksma[1], Dr. Olav ten Bosch[1]

[1]Statistics Netherlands (CBS), ,

Bringing innovative ideas to production, from the very first rough sketch up to a stable implementation in robust production systems, is a big challenge. It requires serious efforts and perseverance, and not all ideas will make it up to implementation. Successful innovation projects may lead to efficiency gains, burden reduction or new products and services. But it is not easy, quality may be affected, both positively and negatively, and has to be managed; and unexpected side effects may occur. In this presentation this will be illustrated by some concrete case studies.

The first use case is the adoption of scanner data for the CPI. The main goals were to increase efficiency and to reduce respondent burden. Starting with a few data providers that had to be convinced one by one of the added value, early 2020 the full transition was complete and CBS abandoned all manual recording of prices in stores. This had the unexpected side effect that, in COVID times, CPI production could continue without hiccups.

A second use case relates to the modernisation of the communication process. A few years ago, CBS reconsidered its whole communication process. The communication department was reorganised, the scheduling of press releases was restructured, the website was redesigned, the use of social media was expanded, among other things. This led to a much higher visibility of CBS and more media coverage, and trust for official statistics in the Netherlands.

Detecting innovative companies with a combination of webscraping, natural language processing and machine learning methods is another use case. Smaller companies (including start-ups) are not observed in the EU Community Innovation Survey (CIS) and there is not enough data for e.g. geographical breakdowns; which limits the use of innovation statistics for policy goals. The new approach is based on a comprehensive data set of 500 thousand companies' websites which is scraped for keywords which are analysed using a machine learning method trained on CIS data. This allows for much more granular data. The experimental results attracted the attention of the Ministry of Economic Affairs, interested in the feasibility of creating new regular statistics based on this approach. A key issue to be solved is comparability of figures over time, since the key words used to characterize innovation change (concept drift).

In this presentation we will briefly dive into such use cases, make general observations and see lessons learnt for future innovation projects.

Session 22 - User needs & expectations, June 6, 2024, 11:00-12:30

Empowering society: How statistics serve the public good

Ms Sofi Nickson[1]

[1]Office For Statistics Regulation, , United Kingdom

Join the UK Office for Statistics Regulation (OSR) as we showcase findings from our research programme, exploring how official statistics serve the public good. In this presentation we delve into how official statistics serve society, and invite you to join the conversation. We share findings from our research, situating this within a wider discussion on what it means for statistics to serve the public good and why this is important to understand.

Our evidence covers a range of perspectives, including views from members of the UK public who shared their perspectives on the public good with us through a series of public dialogues. We also set out evidence on how official statistics serve society through government and policy making, providing new insights and a fresh perspective on a topic that is often taken for granted. Additionally, we speak about ongoing research exploring the role of official statistics in decision making by individual members of the public, shining a light on relatively hidden uses of official statistics that are crucial to explore if we wish to fully understand their role in society.

We use our wide-ranging research findings to present a compelling illustration of how official statistics are an invaluable tool for society. In doing so, we share not only theoretical insights but also practical advice. We invite you to listen to our evidence and share your perspective as well; together we can help official statistics reach their full potential in serving the public good.

Cascais Data: Pioneering Innovation and smart governance

Marta Cotrim[1]

[1]Cascais City Council, Cascais, Portugal

Cascais is positioning itself as a living laboratory for experimentation and knowledge through data, with a mission to be the best place to live for a day or a lifetime. The citizen is at the heart of this innovative approach, which is supported by four key commitments: governance, citizens, talent, and the future.

The use of data for decision-making, both for immediate action and for defining predictive and strategic models, is crucial. In accordance with a business intelligence approach, the municipality has invested in gathering data from multiple sources, implementing modern monitoring systems, and integrating its technological infrastructure. These measures, along with the establishment of information quality standards, enable the efficient implementation of initiatives that enhance the welfare of citizens.

Monitoring and evaluating the municipality's performance in various aspects translates data into relevant information, increasing the speed and effectiveness of decision-making and strategic renewal. Additionally, the dissemination of information, both processed and raw (open data), ensures greater transparency for the community and empowers them to make inquiries and demand answers from the municipality.

The local innovation strategy is disseminated through data on the dedicated communication channel, the Data Cascais portal (www.data.cascais.pt). The portal presents the strategy in a clear, interactive, and intentional manner. The portal allows citizens to track the municipality's progress in implementing projects that aim to improve their quality of life. It also provides information on the municipality's position regarding local targets and strategies in line with the SDGs and other national, European, or international benchmarks.

The portal serves a utilitarian purpose by providing georeferenced information on public facilities, leisure spaces, sports, and other aspects of the municipality. This format differs from typical information websites and aims to simplify complex data while showcasing its potential and interactivity.

Cascais Data presents itself as an interactive platform, reserving space for customized data requests, as well as the dissemination of content produced by users, based on data available on the portal.

The new Istat open source and standards based architecture for high quality web dissemination of official statistical data

Mr Carlo[1], Mr Alessio Cardacino[1]

[1]ISTAT, Rome, Italy

In the context of recent modernization processes, ISTAT has turned on the innovation of its web architectures for data dissemination, by designing and implementing its new corporate reference web architecture for the dissemination of official statistical data.

This architecture, based on international statistical standards for the availability of statistical data and metadata (in particular the SDMX standard - Statistical Data and Metadata Exchange) as well as highly interoperable and compliant to the open data paradigms at the highest levels, was implemented using the most recent and advanced web technologies: by this way it is also able to increase the quality statistical aggregated data dissemination and to foster the standardization and industrialization of statistical data dissemination processes.

Based on this new architecture, in the last two years ISTAT has published various corporate dissemination systems: the new Corporate DataWarehouse (IstatData), the data dissemination portal of the Permanent Population Census and the Public Statistics Hub system for the centralized dissemination of data coming from the institutions that are members of the National Statistical System, the new web system for the dissemination of Foreign Trade data, Have also been developed by ISTAT some systems, such as the new Territorial Statistical Atlas (AST), which have specialized some functionalities of the new dissemination architecture to highlight certain thematic aspects (in the case of the Atlas, the statistics about the territory.

This work aims to highlight the following specific features of the new platform, related to:

- Data modelling according to the SDMX information model
- Data transformation and processing of no compliant SDMX data source
- Industrialised solutions for the periodical data upgrade
- Footnotes and flags (SDMX attribute) management
- SDMX annotations as practical means to drive the user data visualisation experience
- Dashboards for quick synthetic data analysis
- Performance issues related to big size datasets and high volume of concurrent accesses
- Functionalities facilitating organizational aspects related to data migration and corporate data warehouse management

Bringing the data providers closer to official statistics

Isabel Almeida[1], Glória Carrilho

[1]Statistics Portugal, Lisbon, Portugal

Statistics Portugal is being focused on simplifying business data collection procedures, as well as communication and feedback with its respondents to motivate data collection cooperation for statistical purposes and reduce the statistical burden.

The aim is presenting the evolution of the set of initiatives that Statistics Portugal has currently implemented in this area:

WebInq: online service for electronic response. At this moment, all the business surveys are available on WebInq (online response rate is 98,9%).

Voluntary survey about statistical burden, placed at the end of each survey and collects data on the number of people involved in that answer, the time spent, and the respondent perception on the degree of difficulty and usefulness of the information collected; and about WebInq features.

Feedback to Data Providers, carried out through reports, of a macroeconomic or personalized information, including indicators on the relative position of the business in the sector of activity in which they operate, as well as the results of the surveys in which they participate.

In the last Peer Review report, published in 2023, engagement with data providers was recognized as a strength of Statistics Portugal, and the customized feedback reports were considered an innovative practice.

How to serve society through official statistics portals: the Spanish SDG indicators experience

Dr Pedro Revilla[1], Antonio Salcedo[2], Ana Carmen Saura[3]

[1]National Statistical Institute of Spain(INE), Madrid, Spain, [2]National Statistical Institute of Spain (INE), Madrid, Spain, [3]National Statistical Institute of Spain (INE), Madrid, Spain

Official statistical portals can serve society in several ways by providing access to high-quality data, supporting evidence-based decision making and promoting transparency and accountability. The implementation of a Sustainable Development Goal (SDG) indicators portal presents special challenges, given their complexity and difficulty. In the case of Spain, there is an added difficulty since it has a decentralized statistical system, both departmentally and territorially. The National Statistical Institute of Spain (INE) is committed to the production and dissemination of high quality indicators that ensure an appropriate monitoring of the 2030 Agenda.

This paper shows the INE experience in building and managing the National Reporting Platform (NRP) on SDG indicators, and the way in which it tries to follow the principles of the EU Code of Practice and the methods and tools of the EU Quality Assurance Framework. In particular, the paper addresses Coordination and cooperation (principle 1bis), Relevance (11), Timeliness and Punctuality (13), Coherence and Comparability (14), and Accessibility and Clarity (15). It also discusses Accuracy and Reliability (12), a somewhat controversial issue in SDG indicator platforms, because part of the indicators come from outside the official statistical systems.

INE launched the NRP on SDG Indicators in December 2018, in order to disseminate the data corresponding to Spain and thus to facilitate the monitoring of progress towards the goals and targets of the 2030 Agenda. The INE's NRP is wider than just a dissemination database. It is a mean of collecting SDGs data and metadata from data providers. It is used as a tool for the coordination of the national statistical system, by ensuring compliance with methodological standards and the quality of data and metadata. It can also help to improve access to national and subnational data and statistics, identify data gaps, and encourage collaboration. Finally, an additional objective of the Platform is the transmission of data and metadata to international bodies (e.g. custodian agencies) through SDMX files.

The paper also shows how INE NRP contributes to the communication of quality issues. For each indicator, there is a link to its metadata, which includes metadata for the current indicator available from Spanish statistics closest to the corresponding global SDG indicator. Similarly, a link to the UN metadata is provided. In addition, for all statistics produced by the INE and for several of the ministries, a standardized methodological report is provided, according to the ESS Handbook for Quality and Metadata Reports standards.

Recent challenges in Labour Market Statistics and the role of the EU Labour Force Survey in user needs satisfaction

Ms Hanna Strzelecka[1]

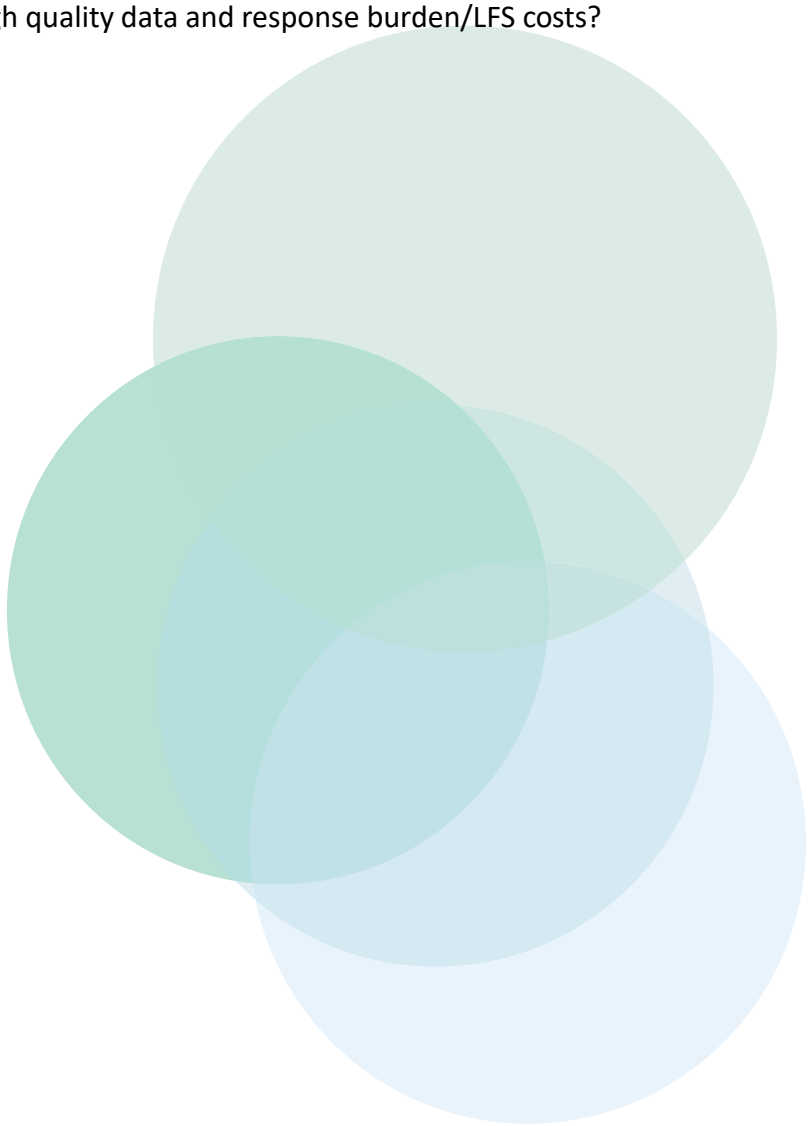
[1]Statistics Poland, Warsaw, Poland

The past years have been challenging for producers of Labour Market Statistics. New phenomena have appeared and gained in importance which should be taken into account in data production by official statistics. There are users who look for data such as digital platform employment, that is difficult to be captured by using traditional survey methods. COVID-19 measures have impacted statistical data collection, while at the same time the interest in high quality and timely labour market statistics has risen, and new indicators appropriate for the situation at hand were requested. The EU Labour Force Survey has been the most important source of data on the labour market, but the users still would like to add more topics to the EU LFS or receive more granular data. Moreover there are new ILO resolutions which should be implemented within the EU LFS. From the other hand response rate in the LFS in many countries is decreasing which only confirms that it is more difficult to conduct the survey.

The presentation will discuss above mentioned issues by trying to answer two below questions:

1. Is the EU Labour Force Survey able to deliver data on changing user needs?

2. How to find a balance between high quality data and response burden/LFS costs?



Session 23 - Machine learning I, June 6, 2024, 11:00-12:30

Quality Dimensions of Machine Learning in Official Statistics

Mr Younes Saidani[1], Dr Florian Dumpert[1]

[1]Destatis, Wiesbaden, Germany

Official statistics distinguishes itself through the legally stipulated requirement to ensure the quality of its publications. To this end, it adheres to European quality frameworks, which are operationalised at the national level in the form of quality manuals. Hitherto, these have been designed and interpreted with the requirements of “classical” statistical production processes in mind. Thus, in order to ensure continued adherence to quality standards, a tailored quality framework must be developed to accompany the increasing use of machine learning (ML) methods in official statistics.

This presentation (1) identifies relevant quality dimensions for ML by analysing the quality principles contained in the European Statistics Code of Practice and (2) fleshes them out in light of the methodological peculiarities of ML. Unlike previous works, (2a) robustness is proposed as a stand-alone quality dimension, (2b) machine learning operations (MLOps) and fairness are discussed as two cross-cutting issues with relevance to most quality dimensions, and (2c) suggestions are made how quality assurance can be conducted in practice for each quality dimension. This work provides the conceptual groundwork for embedding ML quality indicators in the quality management systems used by official statistics for assessment and reporting, thus ensuring that the quality standard of official statistics continues to be met when new statistical procedures are used.

Using Artificial intelligence on the development of official statistics'

Dr Haidy Mahmoud[1]

[1]National Statistics Office (CAPMAS), Cairo, Egypt

Societies are going through many changes. In a rapidly changing world, the knowledge revolution is the ideal tool for predicting future needs and moving societies forward. Artificial intelligence has become a strategic force for many governments around the world, as artificial intelligence methods have witnessed a boom from 2020 until now. Artificial intelligence techniques have increasingly expanded in recent years, especially in migration and mobility surveys, due to the difficulty of collecting their data using traditional methods.

This paper aims to improve the quality of statistical products in Egypt by studying the impact of using artificial intelligence in collecting data for the Migration and Mobility Survey, which is scheduled to be conducted in Egypt in 2024, as well as identifying the extent of the impact of artificial intelligence on the quality of field surveys. Hence the importance of the study in that it addresses a very important topic, which is identifying the requirements of the new stage in the development of the quality of official statistics through the application of modern technologies, the most important of which is artificial intelligence.

The study seeks to test a main hypothesis, which is that there is a positive relationship between artificial intelligence and the quality of field surveys in Egypt. The study used the descriptive analytical approach, in addition to SWOT analysis in analyzing the correlational relationships between the variables in assessing the statistical situation of Egypt, based on the reports of the Statistical Committee. For the United Nations and the World Bank website, the study showed the importance of statistical agencies' interest in using modern technologies in developing the quality of statistical products.

Quality improvements: bringing users along for the ride

Mr Jarle Kvile[1]

[1]Statistics Norway, Oslo, Norway, [2]University of Oslo, Oslo, Norway

Statisticians at NSIs face a growing call for making quality improvements, including using big data and trying more advanced methods. But statisticians also face capacity constraints and have an imperative to work in a way that preserves, or even builds, the public's trust. (See Pullinger 2020)

Working at the intersection of these three imperatives—quality improvement, capacity constraints, and trust—poses serious challenges for statisticians. I argue that faced with this challenge, statisticians often take a 'closed door approach' to making quality improvements. The 'closed door approach' describes a way of implementing quality improvements that keeps the end users out until the data is delivered. A statistician taking the 'closed door approach' might, for example, make improvements without disclosing the changes to end users, or else inform the end users with a many-page long PDF with a titled something like, 'Production changes and quality improvements to tourism statistics.' At Statistics Norway, I usually took the 'closed door approach': explaining my quality improvements in a lengthy PDF, and then being pleasantly surprised when end users had no comments to those changes. In time, I realized that the lack of response was not because users had no issues with the improvements, but rather because they could not understand what they were. In sum, my 'closed door approach' might have improved the quality of official statistics and acknowledged my capacity constraints, but it was not building trust with my users.

I advocate for statisticians to take an 'open door approach,' in which statisticians bring users along for the ride. This paper builds on existing frameworks to describe a framework for an 'open door approach' to quality improvement. Under that framework, statisticians: identify opportunities for quality improvements through a rigorous and independent thought process; collaborate with users to take use of their domain knowledge at the earliest opportunity, e.g., as participants in focus groups; improve the quality and share preliminary results with end users in brief meetings to ensure that most needs are met and understood; and deliver the final results in a transparent format that is digestible to end users. The example used in the paper is self-reported sick days and is written in real time. This paper lays out the framework for this approach and explains how that framework serves the goals of quality improvement and trust-building, while being mindful of capacity constraints.

Citation: Pullinger, John. 'Trust in Official Statistics and Why It Matters'.

Combining deep neural networks, rule-based system and targeted manual coding for ICD-10 cause of death coding of French death certificates

Ms Diane MARTIN[1], Elisa Zambetta[1], Nirintsoa Razakamanana[1], Aude Robert[1], François Clanché[2], Cecilia Rivera[1], Zina Hebbache[1], Diane Martin[1], Rémi Flicoteaux[3]

[1]INSERM CEPIDC, , France, [2]DREES, , , [3]APHP, ,

Cause-of-death (CoD) statistics are key indicators in epidemiology and public health. These statistics come from death certificates (DC) completed by physicians and coded usually by official statistics authorities according to the standards of the WHO International Statistical Classification of Diseases and Related Health Problems (ICD-10) to construct time and cross-country comparable statistics.

Causes of death in DC are usually coded, either by automated rule-based expert systems, or by coding experts. Based on dictionaries of medical expressions, text standardization steps, and on thousands of decision rules in decision tables maintained internationally according to WHO official updates and recommendations, rule-based expert systems ensure homogeneity of ICD coding.

However, the entire process requires significant human resources if expert systems are unable to fully automatically code a sufficient number of certificates.

In France, 38% of DCs in 2018 and 2019 could not be fully automatically coded, and a complementary traditional manual coding campaign could not be carried out due to a lack of human resources.

State-of-the-art deep neural network (DNN) algorithms are expected to perform well for this type of classification task and can be trained on previously labeled data. Several research works showed that if trained on sufficiently large sets, they can achieve very high coding accuracy on most of certificates. Despite these encouraging results, few countries have gone as far as a full production rollout for official CoD statistics. Indeed, having several coding modes cohabitating in production requires developing a strategy to articulate them with specific constraints such as the human resource available.

In this article, we present the new approach developed and implemented for producing CoD statistics of 2018 and 2019 in France in the context of catch-up mentioned above. To code the DCs for these two years, we use the predictions of seq-to-seq DNNs trained (i.e. estimated) on past data (AI, 34%) and manual coding (3%), the latter targeting DCs of particular public health interest and those for which the AI predictions have a low confidence index. A loop of interaction between the three coding modes is introduced. This is the first time that France has used deep learning to produce [part of] official CoD data. We evaluate the performance of the retained approach and its consistency with a traditional coding campaign on a test sample that is representative of the entire population of deaths and is not used in the training of the algorithms.

Artificial intelligence as a support for survey respondents: defining the process of Istat's new AI service

Ms Paola Bosso[1], Gabriella Fazzi[1], Paolo Francescangeli[1]

[1]Istat, Rome, Italy

Currently, responding units involved in the surveys of the Italian National Institute of Statistics (Istat) can request assistance and support in accessing and navigating the data acquisition systems, as well as with legal obligations or doubts about the survey's content. Assistance service is provided through synchronous (toll-free number) and asynchronous channels (dedicated email). The service is managed by a specialized Contact Center acquired from an external supplier.

The current management of service requests is exclusively interactive and is based on two levels: first-level assistance is provided by operators from the external company (Contact Center assistance) who solve the most common problems using FAQs. The second level of assistance focuses on cases with a higher degree of complexity and is provided by Istat experts (Istat assistance). In 2023,

260.000 tickets were managed by phone and email to assist households and enterprises, and were resolved at either the first or second level of assistance.

In 2024, Istat plans to introduce a new integrated assistance service, with the objectives of optimizing resources and simplifying communication with Istat. The new service provides automatic interaction processes with respondents, also through the use of Artificial Intelligence solutions, in a multi-channel perspective (telephone, email, pec, web, webchat, social media). Automation is achieved not only through integrated management of different systems and data flows, but also through the use of machine learning algorithms to process and respond to respondent requests.

Experimental analyzes on tickets acquired over time show that a significant percentage of them concern issues that can be managed using Artificial Intelligence (AI). The use of AI can help reduce waiting times for respondents and costs for Istat.

This paper describes an automated procedure for classifying assistance requests and responding to the most frequent ones by using natural language processing techniques and supervised classification algorithms. Therefore, an initial level of automated assistance is introduced, and a new process is designed to direct requests to the appropriate level of assistance: AI assistance, Contact Center assistance, or Istat expert assistance. Additionally, we address privacy concerns that may limit certain results and applications of AI.

Session 24 - Quality of registers, June 6, 2024, 11:00-12:30

Implementing the quality framework for the Istat Integrated System of Statistical Registers: challenges and solutions

Mrs Cecilia Casagrande[1], Mrs Sara Giavante[1], Mrs Fabiana Rocci[1], Mrs Giorgia Simeoni[1]

[1]Istat - Italian National Statistical Institute, Rome, Italy

One of the pillars of Istat modernisation programme, started in 2016, is the creation of the Integrated System of Statistical Registers (ISSR). Each register of the ISSR is the result of the integration of several administrative data sources and possibly survey results. Thus, the processes underlying the statistical registers are very complex due to their multisource nature and the need for coherence, within and across the registers, of the produced results.

To monitor such a complex system, Istat developed a new quality framework, based on a metadata model that refers to the UNECE standard GSBPM (Generic Statistical Business Process Model) and including several quality measures. The framework assures a structured and detailed documentation for transparency and traceability reasons and allows assessing processes and outputs quality, both while they are in progress and ex-post, in a systematic and standardised way. For each GSBPM sub- process considered relevant for the statistical registers' processes, the set of the possible input, statistical methods and outputs is specified, as well as a set of standards quality indicators for monitoring and evaluating purposes. The metadata are needed not only for documentation purposes but also to provide the information useful to calculate and properly interpret the quality indicators. The framework was tested on two statistical registers of the ISSR, confirming its validity and usefulness but also highlighting the need for a certain degree of customisation when applied in each register. The implementation started in four statistical registers through different working groups in a parallel way: for each of them the processes should be first mapped with the GSBPM, then metadata should be compiled and the applicability of quality indicators should be evaluated. Sometimes quality indicators have to be tailored to the register to make them meaningful and useful.

An informal restricted group of expert of the framework, involved in the different applications, is sharing those experiences, in order to deal with issues or doubts that may arise, as well as to assess ideas for possible improvements of the framework itself. In this way, the coordination and coherence between the implementations is guaranteed through discussion and problem-solving analysis.

The paper will describe briefly the framework, the further challenges that the coordination group is encountering, the solutions identified and how the achievement of the final fine-tune of the quality framework is planned.

Polish experiences from the 2021 census

Dominik Rozkrut[1], Janusz Dygaszewicz[1]

[1]Statistics Poland, , Poland

The population and housing census in the 2020/21 round in Poland took place under the sign of the COVID-19 pandemic, which has spread across the world. Some countries abandoned carrying out the census on time and postponed the survey to the next years. The COVID-19 pandemic affected a wide variety of social and economic phenomena, but Statistics Poland faced this “challenge” and prepared solutions that made it possible to conduct the census while maintaining the highest quality and necessary precautions.

This paper presents Polish experiences from the 2021 census, in particular how logistics and organizational measures undertaken in the difficult pandemic conditions contributed to further work on the development of census solutions in the field of digitalization and improvement of data quality in statistical surveys.

The population and housing census 2021 was conducted using a mixed method, i.e. combination of data from administrative sources and data collected from respondents. As a result of the census carried out during the pandemic, the list of information system operators providing data in the census was expanded (to include providers of publicly available telecommunications services) in order to collect current telephone numbers of respondents. This change reduced the number of face- to-face interviews in favour of telephone interviews.

The course of the census was also influenced by behavioural and psychological behaviour of the population, resulting from health risks and law restrictions. Hence, it was necessary to monitor data collection phase using dedicated IT solutions (i.e. management application, dashboards) even more rigorously than under ‘normal’ conditions, and to take additional measures to ensure the highest data quality.

The electronization of the census in Poland, started in 2011, continued and developed during the 2021 census, as well as the flexible use of CAxI data collection methods, turned out to be an excellent “cure” for random events.

The 2021 census was effectively conducted despite the difficult pandemic conditions. Every possible effort has been made to organize the census in such a way as to collect the highest data quality. The consequence of these experiences and Statistics Poland’s approach to widely use official registers and information systems, is the intensification of work on expanding the use of administrative sources - not only for censuses, but also to substitute and complement statistical surveys. This also demonstrated the great need of agile management in the event of crisis.

Identification of Unregistered Emigration in the Norwegian Population Register

Linn Krokedal[1], Stian Nergård[1], Erling Johan Haakerud Kvalø[1]

[1]Statistics Norway, Kongsvinger, Norway

A precise estimate of the target population is inherently important in population statistics. However, factors such as increased immigration, and few incentives for deregistration after emigration mean that population registers may not always accurately reflect the target population. This study aims to identify unregistered emigration using “signs of life”. That is, detecting historical inactivity of individuals who have emigrated, but are still listed as residents in the population register.

Unregistered emigration contributes to over-coverage, as the number of actual emigrations exceeds the number of registered emigrants. This estimation error affects not only size and composition of the population, but also impacts demographic indicators, such as death and fertility rates. Statistics on households and families may also become skewed due to these discrepancies. There is still no consensus on how to identify or deal with unregistered emigration. Addressing this, we first provide a comparison of methods adapted from the literature for estimating the number of unregistered emigrations. The Zero-Income Approach provides a method with minimal computational and data quality requirements, which serves as a foundation for the estimation. The Household-Income Approach builds upon this by correcting for household income factors. Finally, the Register Trace Approach provides the most comprehensive and detailed picture of unregistered emigration. Our estimates suggest that unregistered emigrants account for approximately 0.44 percent of the adult population in Norway. Second, we analyse the demographic characteristics of the non-deregistration group. We find that the problem of unregistered emigration is not equally distributed across the population, indicating that some subgroups are more prone to discrepancies than the rest of the population. Among immigrants, the over-coverage due to unregistered emigration is substantially higher, accounting for 2.29 percent of the population.

Coherence of integrated data from the Italian Education and Training Register in the framework of the Statistical System of Registers

Ms Giovanna Brancato[1], Mrs Claudia Busetti[1], Mrs Lucia Coppola[1]

[1]Istat, Italian National Statistical Institute, Rome, Italy

Istat is developing a new statistical Thematic Register on Education and Training, namely TRET, based on administrative data from different sources, properly arranged for statistical purposes. The register will provide official yearly statistics on education and training, as well as cross-sectional and longitudinal information on individual educational patterns. Its first release is planned in 2026.

TRET is part of the Istat Integrated System of Statistical Registers. Consequently, it is linked with the registers on the so-called base populations: the Base Statistical Register of Individuals (RBI) and the Base Statistical Register of Economic Units (RBUE). It will also be linkable with other thematic registers, e.g. the Thematic Labor Register, for education-to-work transition analyses.

The integration with the base registers implies constraints on the underlying base populations and their characteristics. Indeed, TRET statistical units (students, graduates …) are a subset of RBI units, which provides core time invariant (gender at birth, birth date and place) and variant (citizenship, residency) variables. Similarly, RBUE provides the most relevant variables on education and training institutions.

The process of building up the annual version of the TRET involves a set of operations that can generate non-sampling errors. Firstly, education and training variables from input data are validated. A student can be present in different sources or show more records in a single source (e.g. school changes). Therefore, inconsistency of individual characteristics have to be checked and edited.

Secondly, the integration with the base registers, performed by deterministic linkage on a pseudonymous code, is subject to linkage errors leading to coverage and measurement problems. Furthermore, inconsistency between base register variables and the original input administrative data have to be taken into account. Finally, the core unit of the TRET is the “education position”, identified by a triplet of units, i.e. the individual, the institution and the program in which the student is enrolled. Further editing and imputation is adopted to ensure cross-sectional and longitudinal consistency of these units.

In this paper, the process of implementation of the TRET is described with reference to the grades up to lower secondary education. All the typologies of errors arising during the process will be explored and, when possible, examples of estimates of the errors will be provided through quality indicators. Finally, the impact of the errors on education and training statistics will be assessed by means of indicators on coherence with other sources of benchmark.

Quality Improvements in the EuroGroups Register Process and Products

Pau Gayà Riera[1], Ioannis Sopranidis[1], Alexandre Depire[1], Dimitrios Chionas[1]

[1]European Commission (ESTAT), , Luxembourg

The EuroGroups Register (EGR) is the European statistical register on Multinational Enterprise (MNE) groups created by the European Statistical System and managed by Eurostat. It receives input data from the National Statistical Institutes of the EU and EFTA countries and a commercial data provider, consolidates them and makes it available for statistical purposes. The EGR is updated annually, and Eurostat disseminates a set of tables and articles as Experimental Statistics.

Profiling is a method to analyse and maintain the legal, operational and accounting structure of an enterprise group at national and world level, in order to establish the statistical units within that group, their links, and the most efficient structures for the collection of statistical data.

To ensure better timeliness and higher accuracy for a limited set of large MNE groups that significantly impact the European statistics, a new strategy for the data quality management is in place. The split of the EGR population in two tiers, a “top-tier” (the largest, most important, and complex MNE groups) and a “bulk-tier” (all the rest). This allows for two different data quality management processes for the update of the two sub-populations, to meet the increasing demand in timeliness and accuracy from users.

Eurostat has set up the Future EGR project to meet the management strategy together with the expressed user needs and the hard work is paying off. The paper will highlight the quality improvements that have been implemented and will provide an overall view on the different possibilities to better use the EGR richness of data, including for dissemination of further Experimental Statistics.

The major quality improvements concern the higher accuracy, frequency, and timeliness of the EGR frame. At T + 11 months for the reference year 2022, using the automatic exchange of the data of the profiled groups, the EGR will already contain high quality information. For the reference year 2023, an additional frame at T + 4 months, to include the starting picture, will be released.

Further automations for the “bulk tier” with standard data checks that are continuously conducted to ensure the highest quality of the data and the correct implementation of the European business statistics regulation.

Finally, relevant efforts are put to enhance the communication with users and the collaboration with other Directorates at the European Commission. Additional data sources to increment the informative value of the EGR especially for the economic variables are also under analysis.

Quality monitoring system for the European statistical business registers

Ms. Isabelle Collet, Ms. Iliyana Iskrenova

[1]Eurostat, Luxmbourg, Luxembourg

The paper will present the quality monitoring system for the European statistical business registers implemented by Eurostat according to the European business statistics quality framework.

“As the statistical office of the European Union, Eurostat’s mission is to provide high-quality statistics for Europe. In fact, quality is the trademark of European statistics which are unique in providing reliable, comparable statistics at EU level.”

In line with this mission, the Regulation (EU) 2019/2152 on European business statistics, applicable from 1st January 2021, strengthens the European framework for the statistical business register, to be used as an authoritative source for deriving high quality and harmonised statistical business register populations to produce business statistics. The European framework for the statistical business register includes a couple of areas the National Statistical Business Registers and the EuroGroups Register.

In keeping with both the regulation spirit and its mission, Eurostat set up a harmonised monitoring system to assess the quality and the compliance of all the domains under the European business statistics regulation, including the European framework for the statistical business register domain.

The paper will present the different components of the quality monitoring system for the European framework for the statistical business register. It will also provide and discuss details on quality indicator selected for this monitoring system and the automation of the assessment process.

A strengthening of the European framework for the statistical business register quality is crucial for the European Statistical System to provide business statistics that can be trusted.

Session 25 - Metadata Quality II, June 6, 2024, 11:00-12:30

Assessing fitness for integration – a meta-data driven approach

Dr. Thomas Gottron[1], Andrea Novello[1], Ilias Aarab[1], Bernadette Lauro[1]

[1]European Central Bank, Frankfurt, Germany

Modern data landscapes are composed of a large number of diverse but complementing datasets. For insightful analytics these complementing datasets need to be combined, i.e., semantically and technically integrated. Integration, however, poses several challenges, including determining which datasets are compatible for integration, understanding the technical methods for achieving integration, and assessing the extent of linkability and coverage among various datasets.

To support users in using and combing a large variety of datasets, we maintain comprehensive metadata repositories at the European Central Bank (ECB). Metadata describing datasets, storage and access roles are used to support data discoverability and accessibility. Metadata describing concepts, data models, transformation rules and mappings are used to support users in dataset integration and analysis.

By leveraging extensively such metadata, we designed a fitness for integration dashboard. This dashboard aims to inform users about available data, illustrating how it can be integrated and the degree to which data aligns across common dimensions. However, the dashboard represents just the visual component of a broader solution. The centrepiece of the solution is the metadata driven and fully automated four-step process for populating the dashboard with relevant information:

1. We utilize metadata describing data models to identify semantic dimensions and suitable identifiers for data integration and aggregation.

2. By leveraging logical inferencing, we ascertain which datasets can be integrated, determine their technical storage locations, and identify attributes for slicing the data into semantically valuable aggregates.

3. The underlying actual data is queried and employed to compile various Key Performance Indicators (KPIs) to assess linkability, coverage, and integrability. These KPIs are calculated at different levels of detail and aggregation, such as per country, over time, or based on other relevant breakdowns sourced from taxonomy-driven reference information.

4. The interactive dashboard retrieves and visualizes these pre-computed KPIs, along with additional business metadata related to the datasets. This dashboard efficiently serves ECB analysts and researchers, enabling them to make more informed decisions on navigating and utilizing the ECB's data for their analytical and research purposes.

To showcase the feasibility of our approach, we implemented a prototype leveraging the ECB's data dictionary, the in-house Hadoop based data and analytics platform, and RShiny to construct the dashboard. Furthermore, the prototype demonstrates the benefit of high-quality and semantically well modelled metadata for supporting users in exploring and understanding the data landscape.

Semantic and ontologies of data sets along a data production process

Dr Michele Riccio[1], Mauro Scanu[1]

[1]ISTAT, Rome, Italy

National Statistical Institutes (NSI) organize data production processes according to models, as GSBPM, that represent all the relevant steps. Some of these steps are characterized by the presence of data and their transformation.

To achieve the purposes of each step NSI need different concepts and semantic data structures. NSI need to describe data transformations for each step, to achieve Lineage.

Our goal is to represent these concepts and data structures by a meta ontology for each step. To describe data transformation we combine meta ontologies, SparQL queries and data access based on ontologies. By this way we get a formal represention - machine readable - of semantic structures and data trasformations.

The main “milestone” data structures on which we have worked are the following.

1.- The starting data set for a statistician usually takes the name of “design matrix”. A design matrix appears as a unit data set, consisting of a rectangle of microdata with units as rows, variables as columns, and the set of rows that is either a population or representative of a population.

2.- A design matrix is slightly modified up to a data set which is still a unit data set: the validated data, on which statistical analyses can be performed. A milestone where a metadata structure substantially changes happens when the first statistical products (in terms of aggregates) are computed directly from the validated data set. These products are usually parameters or characteristics of the distribution of a univariate or multivariate distribution of variables computable from the validated data set: totals, percentages, conditional percentages, means, medians and quartiles, interquartile ranges, Gini indices, are typical examples of this activity. These elements characterize mainly the “measure component” of a dimensional data structure.

3.- Statisticians in NSIs usually do not only produce aggregates that refer to a distribution of one or more variables. Other products take information from different already available aggregates in order to produce indicators that compare statistics along time or space, between populations or between variables. This is the case of specific measures, that will be generally called indicators in the rest of this paper (as ratios, percentage changes, index numbers) and that need specific semantic structures that, again, affect the “measure component” of a dimensional data structure.

The purpose of this paper is to investigate the corresponding semantic structures that characterize each milestone data set and to represent it by means of ontologies.

CSO's new metadata portal - Promoting standards, promoting consistency

Mr Ken Moore[1]

[1]Central Statistics Office, Cork, Ireland

In order to promote the use of standardised metadata across the Irish Statistical system, CSO Ireland is developing a new, public facing metadata portal. This portal will contain resources to support users, researchers and the general public to make our standardised metadata more visible and accessible to allow all users to better understand the statistics we produce.

The portal is currently under development with a planned release date of Q2 2024. It will contain 2 key elements. Firstly a data standards resource hub where a number of data standards for key concepts will provide information on standard questions and response options, related classifications and code lists, general descriptions of the concepts being documented and API's to facilitate usage across the Irish Statistical system. Secondly it will also contain an external view of our internal metadata portal (Colectica) which allows users to see the key metadata elements for all of our Business and Social statistics inputs in a searchable format.

This paper will describe the journey to date in the development of the portal, highlight the content, outline the rationale for developing the portal and discuss potential next steps and future plans.

Visualizing Survey Flow and Improving Data Collection Through Paradata

Mr Gezim Seferi[1], Bengt Oscar Lagerstrøm

[1]Statistics Norway, Oslo, Norway

It is generally known that Survey design elements such as questionnaire length, topic, and wording can adversely affect response rates, quality, and overall lead to measurement errors. The Total Survey Error (TSE) paradigm, outlined by Biemer & Lyberg and Groves (see Biemer & Lyberg, 2003; Groves, 1989; Groves & Lyberg, 2010), encompasses various error sources and offers a comprehensive framework to analyze and address systemic deficiencies throughout the survey process. In practical terms, TSE serves as a tool to assess and mitigate errors from the conceptual stage to data collection and processing, aiming to minimize overall measurement errors. Considering the Total Survey Error (TSE), the use of self-administered web surveys has placed increased emphasis on safeguarding against measurement errors. On the other hand, web surveys have provided new data often referred to as Paradata, which pertains to data related to the data collection process during the field period.

This paper presents how we use paradata at Department for methodology and data collection, Statistics Norway, to develop tools that help us identify challenging questions, bottlenecks, underreporting, and the overall flow in the questionnaire. The main focus will be on showcasing developed tools that visualize the flow in the questionnaire using Sankey diagrams visualizing the flow between multiple questions or sections (Schmidt, 2008), uncovering straight-lining (A. Regula & Jerald G., 1981) and underreporting. The case in the paper will be Statistics Norway's cross-border trading survey, illustrating how the use of paradata has contributed to changes in the questionnaire to avoid measurement errors. Additionally, it will demonstrate how visualization aids in understanding respondent behaviour during the questionnaire completion.

By sharing our python code and experience with paradata associated with the cross-border trading survey, we contribute to the effort of understanding respondent behaviour through new tools and minimizing pitfalls in light of total survey error.

Introducing a NACE classification index to enhance transparency, user- friendliness, and foster uniformity in its application across the European Statistical System

Matthias Flohr[1], Denis Dechandon[2]

[1]Eurostat; unit B1, Luxembourg, Luxembourg, [2]Publications Office of the European Union; unit A1, Luxembourg, Luxembourg

Eurostat and the Publications Office of the European Union (OP) are collaborating on a ground-breaking initiative to streamline user interaction with statistical classifications within the European Statistical System (ESS). This project aims to significantly enhance transparency, user understanding and the harmonised application of the Statistical Classification of Economic Activities (NACE), which is one of the most widely used classifications within the ESS , serving as the foundation for over 40 European statistics products. The key information asset rendering this enhanced harmonisation possible is the NACE index, which breaks down the content of each NACE position into granular entries, drawing upon the NACE explanatory notes (and agreements on their interpretation). Each index entry corresponds to a specific economic activity. This simplifies future revisions of the NACE, ensuring a streamlined and adaptable system.

Further harmonising the interpretation of NACE across the ESS will increase the quality of those products, in particular:

- accuracy, thanks to the more accessible detailed descriptions of each NACE position, reducing the risk for erroneous encoding of an economic activity
- comparability across countries (as these descriptions further operationalise the common standards set out in NACE) and over time (since the granular index entries facilitates the task of constructing accurate correspondence tables between successive NACE versions)
- accessibility and clarity, thanks to the user-friendly presentation of what a specific activities NACE position comprises, and the possibility to look up under which position a specific activity of interest to the user is classified.

Technically, this project revolves around the development of a dynamic online application that empowers users to efficiently navigate, search and export statistical classification data published by the OP. This user-friendly interface will enable users to input specific search terms or combinations of terms, making it easier than ever to find the correct categories in the relevant classifications.

Additionally, the application will offer the ability to explore various economic statistical classifications within the ESS, facilitating the export of classification data. Furthermore, the application will elucidate the structural relationships and correspondences among different statistical classifications, fostering a comprehensive understanding of the international statistical classification system. By providing granular details, the NACE index elevates the performance of the search function within the online application.

The introduction of the NACE index, supported by this innovative online application, promises to revolutionise the accessibility and utility of statistical classification data showcasing the ESS commitment to facilitating data-driven decision-making in the European Union.

Session 26 - Coordination, June 6, 2024, 14:00-15:30

Coordination within the national statistical system – experiences from Denmark

Ms Karin Blix[1], Janne Solgaard Kruse

[1]Statistics Denmark, Copenhagen, Denmark

Statistics Denmark started working on developing national guidelines for official statistics for other national authorities (ONAs) producing official statistics in 2017. This paper will describe the process of developing and finalising the national guidelines for official statistics. The national guidelines for official statistics are prepared as a shortened version of the ESS CoP in order to give the ONAs guidelines that are more accessible to someone who does not have statistics production as their core task. The development of the guidelines has been in progress over a number of years, as they have been continuously revised in relation to the experience the ONAs have had in dealing with them. This paper will also describe the process of monitoring compliance with the national guidelines for official statistics and the follow-up processes. We will conclude with a discussion of the challenges the ONAs have expressed in dealing with the guidelines and complying with them and the way forward.

Good practices based on the OECD Recommendation to ensure an efficient coordination of the national statistical system

Mr Julien Dupont[1], Nora Bohossian, Adrian Zerbe

[1]Oecd, , France

Coordination is a key element of an efficient national statistical system. While international statistics guidelines emphasize the coordinating role of national statistical institutes, there is little guidance to set up an efficient coordination and to monitor this implementation. The statistical programmes are the main instruments in place to ensure this coordination. Their development and implementation rely on specific mechanisms and a strong cooperation between the producers of official statistics, while users should also be involved in their elaboration. The statistical programmes also require guidance developed by the coordinators. However, statistical programmes are increasingly insufficient to ensure an efficient coordination of the national statistical system. A strong legal basis is required, as well as appropriate institutional settings and additional tools. In this area, the move to more granular and more frequent statistics requires stronger mechanisms to ensure the production of high-quality statistics. Introducing new dimensions in the quality management framework to cover administrative sources and privately held data and extending the quality management to the whole national statistical system are important to overcome new challenges and ensure trust in official statistics. This paper presents the main trends in the area of coordination and highlights selected examples of good practices across countries adhering to the OECD Recommendation on Good Statistical Practice (hereinafter the Recommendation). On this basis, the paper aims to derive a set of good practices to advance the state of coordination in existing international guidelines such as the OECD Recommendation.

Develop and maintain the quality competences: the role of networks of quality experts in the French Official Statistical Service

Ms Typhaine Aunay, Nathalie Garrigues

[1]Insee, , France

The Quality policy of the French official statistical service (SSP) aims to “integrating quality into processes, with the aim of securing and improving efficiency”. This cannot be achieved without disseminating a Quality culture among all staff, which is an important issue given the territorial and functional organisation of the SSP. In addition, quality is a cross-functional area for which maintaining skills is specific.

To face these challenges, the Quality unit has set up two networks of Quality experts (one in the regions and the other in the Ministerial Statistical Offices), who have two major missions.

On the one hand, they are responsible for promoting the Quality culture within their entity, through training or communication actions. To do this, they benefit from the training system designed and regularly updated by the Quality unit, and are provided with ad hoc support and materials that can be adapted to local situations and the target audience.

On the other hand, they lead the development of their entity’s Quality strategy and support teams in implementing the resulting actions. The active participation of agents in concrete Quality initiatives helps to maintain their skills in this area. To this end, the Quality unit provides ongoing support for the networks, based on the sharing of experience.

The effectiveness of the Quality networks is enhanced by the strong coordination of the SSP by INSEE and by the policy of staff mobility, which encourages the exchange of practices and the dissemination of knowledge.

After outlining the challenges and specific features of disseminating a Quality culture, this article will present the organisation of Quality experts into networks, the cornerstone of developing and maintaining the quality skills of SSP staff.

Slovak National Statistical System - assessment of selected coordination tools

Albert Ivancik[1]

[1]Statistical Office Of The Slovak Republic, Bratislava, Slovakia

An effectively functioning and integrated national statistical system is a prerequisite for decision-making, policy-making and sustainable development of any country.

The Slovak National Statistical System has undergone significant changes in terms of coordination over the last two years. The latest amendment to the Act on State Statistics has strengthened the position of the Statistical Office of the Slovak Republic as the coordinator of the NSS. With regard to the main objective of coordination - to build a metadata-driven, quality-oriented and product-driven NSS – such conditions and tools were created to facilitate the standardisation of selected activities of the statistical production process, as well as the management of metadata and statistical outputs.

Our paper seeks to further specify and assess selected coordination tools, namely the Coordination Council for State Statistics, which acts as an expert coordinating body for issues related to the performance of the tasks of state statistics by other national authorities performing state statistics; then the Uniform Statistical Information System, a new information tool designed to simplify and, more importantly, unify specific statistical activities across the NSS; and finally, methodological audits, the implementation of which serves to assess the quality and performance of both individual members of the NSS and the NSS as a whole.

From the point of view of the Statistical Office of the Slovak Republic, the role of the coordinator and the performance of coordination brings a number of challenges, but also interesting opportunities. At a time when quality and accurate official statistics are competing with fast data, it is crucial that NSIs are able, through promoting reliability, impartiality and objectivity, professional independence, and quality, to act as trustworthy institutions serving society, while at the same time, through innovation, continuously moving forward and seeking new approaches and new users of statistics.

Irish Statistical Code of Practice (ISSCoP) Coordination and Certification of Other National Authorities (ONA'S) across the Irish Statistical System

Mrs Caroline Barrett[1], Mrs Maria Looney[1]

[1]Quality Team, Central Statistics Office, Ireland, , Ireland

The Central Statistics Office (CSO) is mandated under EU Regulation 223/09 with the coordination of Official Statistics compiled by other National Authorities (ONA's) who operate within the Irish Statistical system. These statistics play an essential part in creating an informed society and are indispensable for making evidence-based policy decisions.

In line with the framework, the CSO developed a set of quality guidelines under ISSCOP (Irish Statistical System Code of Practice) in 2013 which are a subset of the ESCoP (European Statistics Code of Practice) principles.

While some progress has been made to date, the recent 2022 Peer review process recommended that the ONA's certification process should move to a unified benchmark by aligning the ISSCOP process more closely to ESCoP. Following these recommendations, the CSO has recently provided dedicated staff resources to review and strengthen the ISSCOP certification and coordination roles by working with other producers across the Irish statistical system (ISS) to support quality improvement across the national statistical system.

This paper will describe the steps taken by the CSO in order to successfully develop and implement a quality support service across the Irish Statistical System. It will outline the current state of play by discussing the existing framework in place for the ISSCOP process, the level of engagement by the ONA's, the challenges experienced in moving to the ESCoP process and the progress made to date. It will also set out how the certification and quality coordinator team are promoting and providing guidance to the ONA's on advancing the ISSCOP certification on the national code of practice (Irish Statistical System Code of Practice (ISSCoP)) while also setting out how quality is assured and monitored by the coordinator. Finally, the paper will also detail the progress made to date in building effective relationships with the ONA's and the challenges and opportunities that have been encountered to date.

Dialogue with users in defining the official statistics activities programme: experience of the French National Council for Statistical Information (CNIS)

Ms Dominique Francoz[1]

[1]Insee and CNIS, Montrouge, France

The CNIS is one of the bodies of the French Official Statistical System. It ensures dialogue between producers and users of official statistics. It highlights new needs, using a forward-looking approach. In this way, it helps to ensure that each year a statistical work and surveys programme is drawn up in line with the needs of those involved in understanding society in the social and economic spheres.

Every 5 years, the CNIS carries out a prospective analysis of medium-term needs and the changes to be made to the information system in the medium term. A year-long process of gathering requirements from different users and producers' projects on the 2024-2028 medium term programme has led to the desired guidelines for the five-year programming of official statistics work and surveys.

This paper aims to present the process for designing the 2024-2028 medium-term programme that will both guide the future statistical work and frame the forthcoming consultations in the CNIS committees over the next five years. It will focus on how user needs have been gathered, how they have steered the recommendations in the various committees and how they will be implemented in the future CNIS activities in various ways (committees' agendas, setting up of working groups, organizing seminars or conferences).

In a second part the paper will show how this programme fits in with the CNIS's annual activities and how users also interact between two medium-term programmes.

Finally, based on the 2018-2023 period experience, the paper will assess the extent to which the user needs expressed both annually or during medium term programmes' preparation influence the official statistics works.

Session 27 - Geostatistics II, June 6, 2024, 14:00-15:30

New data sources in spatial surveys

Dominika Rogalińska, Ms Magdalena Skalik

[1]Statistics Poland, Warsaw, Poland

This paper discusses the direction of spatial surveys in the Statistics Poland, in regards new data sources and new processing methods. The social and economic processes taking place in the modern world determine the need for new information, which is provided in a short time. New sources and methods are an opportunity to provide the needed data while reducing survey costs, reducing respondent burden and improving data quality.

The article presents the results of work that the Statistics Poland doing in the context of Big Data (i.e. Earth Observation methods, internet data, mobile network operators' data) and obtaining data from administrative sources (systems and registers), while indicating the possibility of their application in statistical production. These methods and sources are increasingly used but there is still area for improvement.

The use of new data sources and methods in spatial surveys brings both opportunities and challenges. These challenges are multidimensional and solving them requires multiple actions at different levels. In the paper, these challenges were identified along with directions for solution.

Capturing user needs at local levels: a pillar for a listening architecture

Arnaud Degorre[1], Magali De Raphelis Soissan[2], Sonia Besnard[1], Mathilde Gerardin[1], Fabrice Hillaireau[1], Loup Wolff[1], Marie Sala[1]

[1]Insee, Paris, France, [2]Conseil national de l'informatique statistique, Paris, France

Driving future statistical production according to user expectations is a key concern for official statistical institutes. As stated in principle 11 of the European Statistics Code of practice, relevance is based upon NSIs capacity to meet the needs of users, hence the necessity of "procedures [...] to consult users, verify the relevance and usefulness of existing statistics with regard to their current needs and to examine and anticipate their new needs and priorities." (indicator 11.1).

In this perspective, the French NSI Insee has enlarged the range of its listening channels, relying on both quantitative and qualitative measures, and addressing a large range of users, like public administrations, researchers, journalists, private companies... till the civil society as a whole. Internet surveys, focus groups, satisfaction surveys, prospective interviews with public stakeholders, monitoring of social networks are all facets of this system. Proper consideration of geographic scales and regional interests in these listening devices is also a priority identified by INSEE, in order to best guide the production of local statistics, when the demand for geospatial information is rapidly growing.

To critically examine the listening channels set up for regional users, a review was carried out for the year 2023. It was based on the works of Unece (report " measuring the value of official statistics", 2022) and the Committee on Statistics and Statistical Policy (20th meeting, 2023, conference on "how could statistical organizations become better listeners"), and more generally on academic works studying listening architectures in organizations ("Creating an 'architecture of listening' in organizations", Macnamara, 2015). The review had various deliverables, first of them being a global view of regional users diversity, described through a dozen of "personas". The latter were used to design user stories and better contextualize use cases of regional statistics. The review then identified the channels for capturing local needs, whether through direct exchanges (interviews, focus groups, surveys) or indirect (monitoring user expression channels such as social networks). A typology of capture channels has been established, so as to identify contributions and limits for each of them. The review focused especially on the necessary distinction between promoting channel, usually "product-driven" way of interacting with users, and listening channels, which are "need-driven" way to achieve engagement with users.

This regional review is part of new national strategic plan to strengthen quality at INSEE and ONAs, consistent with the recommendations of the last peer review carried out in France.

Earth Observation Data and Artificial Intelligence to Quality Assure Construction Statistics

Ms Maren Köhlmann[1]

[1]Federal Statistical Office Of Germany, , Germany

To support the construction statistics by quality assurance measures, the project 'Earth Observation and AI for Construction Statistics' (EO4ConStat) is set up of a consortium of the German National Statistical Institute (Destatis), the Federal Agency for Cartography and Geodesy (BKG) and the German Aerospace Centre (DLR). The objective of developing a method to quality assure to construction activity statistics using remote sensing data and artificial intelligence is relevant for high quality construction statistics.

Due to the German government's ambition to build new, affordable and climate appropriate housing, construction statistics are currently in the political focus. The aim of the project – which is financed by the European Commission – includes developing algorithms to detect buildings and building construction sites and possibly define new building starts as well as construction dismounts from earth observation data. This is to be achieved by using and adapting an open source segmentation algorithm. Furthermore, a methodology to compare these results of the change detection analysis with the data collected through the traditional statistical channels shall be contrived.

A pathway towards better integration of statistics and geospatial information with the power of standards

MS Rina Tammisto[1], Maurice Brandt[2], Panu Muhli[3], Rico Santiago[4], Sara Stewart[5]

[1]Statistics Finland, , Finland, [2]Destatis, , , [3]National Land Survey of Finland, , , [4]Ordnance Survey of Northern Ireland (OSNI), Land & Property Services, , , [5]UNECE Consultant, ,

The European Commission funded an action led by the United Nations Economic Commission for Europe (UNECE) to develop capacity in the integration of geospatial and statistical information across the UNECE region. The aim of the action is to foster stronger links between the statistical and geospatial communities across the UNECE region, encouraging greater integration of geospatial and statistical information by promoting stronger institutional partnerships and the use of common standards. As part of this action, UNECE established a task force on standards issues relating to the integration of geospatial and statistical data (INGEST), bringing together representatives from national statistical and geospatial organisations across the UNECE region to discuss the current use of standards, to explore any present issues and constraints, and to identify priorities and future actions to be undertaken regarding the use of standards to improve the harmonisation and interoperability of statistical and geospatial information. The INGEST Task Force has undertaken activities to understand the current use of standards across member organisations and share use cases and best practice, determine the suitability of operating environments to support the use of standards relating to the integration of statistical and geospatial information across the data lifecycle, and identify domains where data integration is hampered by the lack of common standards. Within the context of wider policy/reporting requirements, the Task Force INGEST identified priorities for standards harmonisation work and recommended related actions that would improve the harmonisation and interoperability of statistical and geospatial information, and thus support the greater integration of such data.

The Task Force aims to (during spring 2024) produce a report that contains recommended actions and methodological guidelines to support the implementation of these actions at the country level. In our presentation, we will showcase the main results of the Task Force's work and experiences in carrying out the work.

The evolution of the spatial data production model in Istat. New perspectives for the analysis of population socio-economic phenomena

Dr Giancarlo Carbonetti[1], Raffaele Ferrara[1], Gerardo Gallo[1], Mariangela Verrascina[1]

[1]Italian National Statistical Institute (Istat), Rome, Italy

Over the years, official statistics have shown increasing attention to the strong need for statistical information referring to sub-municipal levels. These data are useful for spatial analyses, business objectives and social, economic and environmental decision-making processes. In particular, users demand data with greater frequency, consistency and richness of information detail.

In this context, the Italian National Statistical Institute (Istat) has been engaged in a modernisation process that includes a significant revision of the production processes for official statistics. In particular, two important pillars of Istat's new organisational and production structure should be mentioned: the construction of statistical registers and the Integrated Register System (IRS); the design and implementation of the Permanent Population and Housing Census (PPHC).

The development of statistical methodologies and the adoption of advanced IT solutions represented the main topic to integrate administrative data with the results of sample surveys and to geocode information at a very fine level over the territory (enumeration areas or addresses). This has made it possible to define an integrated and georeferenced database for the territory, rich in annually renewable information on the main demographic, social and economic variables of individuals, household types and some characteristics of dwellings.

The new process meets high quality standards, allows for a high timeliness in the release of outputs and guarantees maximum data consistency. The results offer new perspectives for analysing the main socio-economic phenomena of the population at sub-municipal scale, both on spatial and temporal dimensions. It is possible to focus attention on specific issues (e.g. "household deprivation") through the definition and calculation of appropriate indicators.

Istat is also planning the release of sub-municipal data and the design of a user-friendly platform for the dissemination and representation of results using GIS tools.

Geographical data quality for spatial analysis and geospatial statistics

Dr Julien Gaffuri[1]

[1]Eurostat, Luxembourg, Luxembourg

Together with official statistics, geographical information science provides crucial inputs to inform citizens and policy makers on various thematic domains. The specificity of GIS analyses is the spatial dimension, and provide advanced quantitative information on the spatial distribution of real world phenomena and their potential correlations – the most significant progress for pan-European spatial analyses is the recent publication of the Eurostat Census 2021 population grid. Like all computer-based analyses, the reliability of GIS analyses depends on various factors, such as the methodology used and also the input data and their quality.

Geographical information used in GIS software are stored vector GIS databases represent and classify individual real world entities as features with some properties and a geometrical representation.

Vector GIS databases content is diverse: There is no unique way to represent the world, and each GIS dataset represents only some aspects, with some generalisation, depending on its utilisation. Quality, as the capacity to address user requirements, is of course a crucial notion to consider for GIS data.

The way to describe the quality of GIS vector data is specific to the GIS domain, and not always duly addressed by the statistical community.

Common GIS quality components are formalised in ISO 19157:2013 standard: Positional accuracy, completeness, thematic accuracy, temporal accuracy, selection criteria, level of detail, minimum mapping unit (MMU), etc. The importance of controlling GIS data quality for the production of spatial analyses is illustrated with two Eurostat geospatial statistics products :

A first example on residential occupancy indicator. The input GIS data represent the location and size of residential buildings and allows assessing the available residential space at 1km² grid level. This information is then combined with population statistics, also at 1km² grid level, from the Eurostat Census 2021. This indicator informs on the region with a high population pressure, and also touristic regions with a predominance of secondary residences.

A second example on accessibility indicator to healthcare services. The input GIS data represent the road transport network and the location of main healthcare services. GIS routing algorithms based on graph theory allow computing travel time from populated cells to healthcare services and assessing their accessibility at 1km² grid level.

For both examples, different input GIS datasets of various quality are used (National topographic data, TomTom-MultiNet(C) dataset and openstreetmap). The comparison of the outputs illustrates the importance of using reliable input GIS data, with properly controlled and documented quality.

Session 28 - Cooperation with academia, June 6, 2024, 14:00-15:30

III “How to serve society” – Engaging with academia and the Scientific Community

Tobias Thomas[1], Thomas Burg[1]

[1]Statistics Austria, ,

Capacity Building and Cooperation of NSOs with the Scientific Community

The scientific community can play a key role as a partner for National Statistical Offices (NSOs), as the exchange on methodological issues as well as the cooperation on projects and beyond can create added value for both official statistics and academia. Among others, the Austrian case can provide interesting insights.

In line with its Strategy 2025, Statistics Austria has in recent years increased efforts to build capacity and strengthen cooperation with the scientific community. This entails the collaboration with – among others – the Austrian Academy of Sciences (ÖAW), including a dedicated ÖAW-Statistics Austria lecture and research series. In addition, Statistics Austria turned out to be the driving force in establishing a new research data infrastructure in Austria:

First, the Austrian Micro Data Center (AMDC), created in 2022, enables researchers of accredited research institutions to work with pseudonymised and linkable microdata on the level of individuals and firms. This is crucial, as access to high quality microdata is a precondition for the empirical investigation of various interdependent developments in economic and social sciences and to identify causal effects. The AMDC now offers an international state of the art solution, linking microdata deterministically via unique pseudonymised identifiers across various data sets of Statistics Austria, as well as administrative registers, and to microdata provided by researchers. With the AMDC, Statistics Austria caught-up with European forerunner NOSs like CBS from the Netherlands or Statistics Denmark. Meanwhile, Statistics Austria granted accreditation to more than 50 national and international research entities.

Second, the Austrian Socio-Economic Panel (ASEP), launched in 2023, establishes a longitudinal household panel. While such household panels are well known internationally, , in Austria such kind of panels were not existing so far. In contrast, the ASEP is based on survey and register data, demonstrating how a new and innovative data infrastructure enables basic research as well as policy evaluation from an interdisciplinary perspective.

Based on the AMDC and ASEP, in 2023 the Center for Science was established within the organisational structures of Statistics Austria – representing a Single Point of Contact for Science. Statistics Austria considers enhanced cooperation with the scientific community as key to our work and provision of services. The proposed panel seeks to present the Austrian example (and others) and to extract lessons learned from cooperation with academia and research institutions at the benefit of the wider statistical community.

Nearly 10 years of EMOS from an NSI point of view

Mr Markus Zwick[1]

[1]Federal Statistical Office, Wiesbaden, Germany

The European Master in Official Statistics (EMOS) is a story of success. 10 years after the feasibility study “Towards a European Master in Official Statistics” offers 33 universities in 17 countries of the European Statistical System (ESS) an EMOS programme. A major contribution to the success of EMOS lies in the cooperation between the universities and the official data producers. The National Statistical Institutes (NSIs) offer lectures on data production, offer internships and topics for Master’s theses and make their data available for this purpose. Of course, this is not without self-interest of the NSIs. On the one hand, it is one of the tasks of the statistical offices to present their data production transparently and openly to the users, and the scientific community is a major user.

EMOS is meanwhile for official data producer also important in the context of data literacy. On the other hand, EMOS offers official data producers’ access to the next generation of scientists and can train them in their subjects at universities.

Nevertheless, EMOS also poses a number of challenges for official data producers. In Germany, five universities offer an EMOS programme. This means that there is a high demand for lecturers in the field of official statistics. Linked to this, the demand for internships and topics and supervision for final theses is also high. NSIs with a statutory research mandate have a slightly easier time meeting these needs. Germany is not one of them, but due to its federal structure, it has 14 state statistical offices that also offer EMOS services.

The presentation will go into more detail on the points mentioned. This with reference to the current ‘Study on the future of EMOS’, which will have undergone an initial evaluation by the NSIs in the ESS by the time of the conference.

The ethics self-assessment; a tool for the ethics of data use and re-use

Dr. Mayra Pamela Ambrossen[1]

[1]UK Statistics Authority, , United Kingdom

There are many elements playing a role in the aim of being able to make the most of the power of data to serve society. In this presentation we will focus on ethics for the reuse of data.

Before academics, governments, research students and charities reuse data to conduct research, a data access journey takes place. Part of it is making sure the projects in question will benefit the public, are lawful and ethical. The Centre for Applied Data Ethics (CADE) at the UK Statistics Authority (UKSA) plays an important role in advising and helping this to flow. We facilitate this by giving researchers confidence in the ethics of their research and encouraging mitigations to be put in place to ethical risks that may arise in their research journey (from accessing data to publication). We do this by making available a self-assessment tool, producing guidance and data ethics training.

Because researchers were not always trained to consider the ethics of their research and it is a relatively new topic; The CADE makes the most of its resource to make the self-assessment tool efficient and helpful rather than being a barrier or another step to access data for reuse. We have received over 1,000 self-assessments for review since 2021, when the CADE was established, and we come back to researchers shortly with feedback, inviting them to consider mitigations to ethical risks if needed.

This tool will be the focus of the presentation. It was developed in accordance with the UKSA's ethical principles and the UK legislation in mind (Digital Economy Act, General Data Protection Regulation and Statistics and Registration Service Act). In the presentation we will introduce the tool, talk about its users and its importance in the data access journey for researchers and for the UKSA. We would like to open conversation about it and hear and learn from the audience. We are interested in hearing about the data access journey for others and the role ethics plays in it.

Advancing statistical higher education – case of the European Master in Official Statistics

Maja Islam[1], Dario Buono[1]

[1]Eurostat, Luxembourg, ,

Official statistics are undergoing a period of profound transformation. Fueled by the urgent need to monitor and evaluate emerging global trends, manage crises and measure impact and societal progress, the demand for high-quality official statistics has intensified.

The emergence of new data sources, including privately held data, big data and the Internet of Things (IoT), has fundamentally altered the landscape of official statistics. These vast and heterogeneous data sources present both challenges and opportunities for the production of reliable and timely official statistics. Traditional data collection and analysis methods, often ill-equipped to handle the complexities of these new data sources, are being challenged, requiring the development of innovative approaches that guarantee the trustworthiness, accessibility and interpretability of official statistics in the data-driven world.

In this context, the European Master in Official Statistics (EMOS) programme plays a pivotal role in preparing future statisticians for the complexities of the modern data landscape. EMOS is a unique path in graduate-level education in official statistics building on study programmes that exist at European universities. The EMOS label is awarded by the European Statistical System Committee and is at present carried by 33 master's programmes in 17 countries and their partners in statistical offices.

In response to the above-mentioned transformative trends in official statistics, Eurostat in collaboration with all EMOS partners carried out a comprehensive study to critically evaluate the programme's strengths and weaknesses. The study draws insights from in-depth interviews with students, university representatives, potential employers and other stakeholders involved in the programme, as well as the latest developments in higher education policy and official statistics. These insights will be used to identify opportunities for improvement and ensure that the EMOS programme remains at the forefront of training skilled statisticians.

Initial feedback from students underlines that access to practitioners enhances the learning experience and is a key motivator in choosing the EMOS specialisation track at graduate level. While generally positive, feedback from EMOS-labelled master's programmes and their partners in organisations producing official statistics brings certain challenges to the fore that can be grouped as challenges linked to the utilisation and engagement with EMOS and those linked to brand visibility and awareness.

This presentation will take stock of the results of the study on EMOS and invite for a reflection on the challenges and future development opportunities for the programme.

Collaboration with Academia in Statistics Portugal

Mrs Sofia Rodrigues[1], Pedro Campos[1], Mr José Pinto Martins[1], Mrs Maria Joao Zilhao[1]

[1]Statistics Portugal, Lisbon, Portugal

In response to the evolving challenges within the contemporary ecosystem, national statistical offices, exemplified by Statistics Portugal, have undertaken multifaceted initiatives aimed at fostering innovation and efficiency. This paper delves into the integral aspect of collaboration with academia as a pivotal component of these initiatives. The landscape is marked by a surge in available data, coupled with the imperative to reduce costs and streamline statistical processes. Statistics Portugal has strategically engaged with academia to navigate these challenges.

The paper explores the paradigm shift in Statistics Portugal's approach to handling the escalating volume of data from administrative sources. The recruitment of specialized researchers, equipped with diverse skills, contribute significantly to the processing and analysis of intricate databases in collaboration with Statistics Portugal staff. Specific examples highlight collaborations in diverse projects, such as the Survey Income and Living Conditions and Mortality Table.

Mentorship emerges as a critical aspect, facilitating the development of the National Data Infrastructure. This mentorship role extends beyond individuals to project-centric contributions, exemplified by a transformative case study involving the birth of a child. Training initiatives play a crucial role, and Statistics Portugal has invested significantly in specialized training courses offered in collaboration with prestigious universities. Examples include programs in Artificial Intelligence for Business and Macroeconomics, ensuring the staff stays abreast of the latest developments.

The paper also sheds light on the pivotal role of research access to official microdata in promoting a comprehensive and interconnected data ecosystem. Statistics Portugal, recognizing the unique needs of the academic community, has established procedures to facilitate researcher access to statistical information for research projects and theses preparation. Accreditation processes, dataset availability, and safe centers for accessing microdata are meticulously detailed.

Additionally, Statistics Portugal's involvement in the European Master in Official Statistics (EMOS) network is highlighted. Through collaboration with the Faculty of Economics of the University of Porto, the institution actively engages in post-graduate education in official statistics at the European level.

In summary, this paper provides a comprehensive overview of Statistics Portugal's holistic approach to modernization and innovation, emphasizing collaboration with academia, recruitment strategies, mentorship programs, training initiatives, and facilitating research access to official microdata. These initiatives collectively contribute to navigating the intricate landscape of contemporary statistical challenges.

Session 29 - Data integration, June 6, 2024, 14:00-15:30

Traffic and Mobility Indicators

Ms Miriam Blumers[1], Ms Evangelia Ford-Alexandraki[1], Mr Matyas Meszaros, Mr Nikolaos Roubanis[1]

[1]European Commission - DG Eurostat, , Luxembourg

The availability of new data sources and privately held data has increased exponentially in recent years and accelerated the transition towards using innovative data sources in the ambit of transport statistics. Correspondingly, Eurostat is currently exploring their use and developing methods to produce statistical indicators based on these; as well as implementing automated data collection methods. The aim is to enhance the overall quality of mobility statistics by improving their completeness and relevance, i.e. meeting emerging data needs. Concretely, three use cases are being investigated: the (1) density and distribution of recharging stations, the (2) availability and efficiency of public transport and (3) traffic and air quality. The data sources considered are to a large extend publicly available or crowd sourced. In a next step, Eurostat will develop the methodologies for producing indicators for each use case. All the technical implementation is made in the cloud based datalab which provides a flexible environment for prototyping new indicators. These indicators will for example be the maximum and average distance between recharging stations within a defined region; the percentage of a region's area and percentage of population that can be reached within a certain amount of time by public transport; and the average concentration of air pollutants during rush hour.

Technology, innovation and economic performances: a microdata integration strategy proposal

DR Stefano Stefano[1], Dr Davide Di Cecco[2], Dr Roberta Di Stefano[2]

[1]Istat, , Italy, [2]Sapienza University, Roma, Italy

The relationship between technology use, innovation, and economic performance is a crucial topic in the empirical economic literature on growth. Our work proposes an integration strategy of surveys and administrative data in order to obtain a micro-level dataset that incorporates all available information on these subjects. The sources at our disposal are two business surveys (Information and Communication Technologies - ICT, and Community Innovation Survey - CIS), available annually and

3-yearly respectively, and the extended business register derived from the integration of several administrative sources. Our proposal allows for the construction of a comprehensive firm-level dataset, which provides the basis for analysing the relationship between investment in technology and innovation, and economic performance in a continuous time framework, without additional respondent burden.

The integration utilises deterministic and statistical linkages.

Firstly, a simple record linkage is performed which use CIS as pivot survey, retrieving the respondents present in the same three-yearly period in the corresponding ICT surveys (harmonisation of reference period).

Secondly, a Statistical Matching (SM) approach is used, to provide joint statistical information for missing units in ICTs:

a) an exploratory analysis is conducted to eliminate redundant predictors and focus on the most influential variables (matching variables). b) naive micro approaches (hot deck imputation techniques: nearest neighbour distance, random or rank) are employed to obtain the synthetic dataset, taking into account the sampling design underlying the recipient dataset and the corresponding survey weights; c) adjustment for missing data; d) the results of the matching are evaluated in terms of the preserved marginal distribution of the variables imputed in the synthetic dataset.

Finally a complete micro-data was obtained, where data on all the variables is available for every unit.

This integration proposal represent a zero-burden solution to provide an important source of information for research/policy purposes and to support official statistics in exploring complex causal effects and gain new insights into these fundamental economic phenomena.

Results of a pilot study towards an early labour cost index: successes and failures in the multi-source Italian production system

MS Francesca Romana Pogelli[1], Eleonora Cimino[1], Maria Cirelli[1], Donatella Tuzi[1]

[1]Istat, , Italy

In 2019, in the context of the Labour Market Indicators (LMI) Review task force, Eurostat launched a pilot exercise aimed at studying the feasibility of a flash estimate (t+45, i.e. 45 days from the end of quarter t) of the hourly labour cost index (LCI), anticipating the current official timeliness (t+70). Italy was included in the subset of testing countries.

The LCI is one of the Principal European Economic Indicators and measures the cost pressure arising from the production factor “labour”. In Italy it is compiled into an integrated multi-source statistical system that, exploiting Admin, Survey and National Accounts (NA) data, produces a number of timely and broad coverage labour market indicators. Nevertheless, the sources contributing to this system are only partial or not available at all for this new early deadline. In this context, the extension of the current methodology for an anticipated estimate, as suggested by Eurostat, is not achievable.

This paper describes strengths and weaknesses of a number of experiments aimed at an early estimate of the LCI exploiting the available information at the new time constraint. At a first stage, a time series approach was pursued, based on the full information available until t-1, supported by auxiliary early signals on t. A macro deterministic extrapolation was initially tested, followed by forecasting solutions, improved progressively introducing exogenous predictors on t. As an experimental evolution, a more elaborated micro approach was subsequently followed, based on the exploitation of the partial Admin, Survey and NA sources available at the early timeliness. Tests were carried out on a number of quarters, along a period characterized by a frequently chancing economic situation and a very significant outbreak (the recent pandemic emergency) and results were evaluated looking at revision errors. The micro approach turned out to be the most promising in terms of quality, but much more dispendious as far as time/resources involved, burdens that strongly limit its implementation. In both cases, the partial availability of data implied appreciable results in condition of steady state. Weak and difficulty interpretable outcomes emerged in situations of breaks, changing legislation etc., not predictable events that affect differently the variables involved in the LCI compilation.

The main conclusion of the tests performed is the evidence that the quality of flash estimates cannot be assured in a context of partial information and changing economic/legislation situation, with the consequent risk of providing misleading early signals.

An innovative approach to using a variety of data sources and statistical methods for micro-level statistical collection

Mr Anders Grönvall[1], Ylva Olsson[1], Elin Lundh[1], Jesper Fransson[1]

[1]Swedish Board Of Agriculture, Jönköping, Sweden

In the framework of agricultural statistics, a most important statistical survey is the Farm Structural Survey that has been conducted in EU since 1966 and in Sweden since 1927.

The surveys of the structure of agriculture in the EU are regulated in (EU) 2018/1091 on integrated farm statistics, which, among other things, describes what data the member states should collect as well as how and when to collect it. In the regulation it is stated that the member states shall use one or more of the following sources or methods for the purpose of obtaining data on microlevel:

- statistical surveys;
- administrative sources (specified in regulation);
- other sources, methods or innovative approaches.

Traditionally, a statistical survey has been conducted and since 2000 Sweden has developed methods where some of the data collection is done by using administrative registers. The third option was new for this regulation and the guidelines did not give any specific help on how this could be done.

In the preparatory work for the 2020 census we evaluated the different possibilities to collect data for each variable. One part of the survey covered the field of animal stables and manure management. In total it included xx variables. From previous experiences we knew that these types of variables were difficult to collect with good quality through questionnaires in a statistical survey It is difficult to make questions regarding these issues understandable in a questionnaire, which results in a high partial non-response and adversely affects other parts of the survey and the total willingness to participate in it. For national purposes, Sweden also conducts a fertilizer and manure survey every third year which covers the national need for statistics in the field.

From the preparatory work we knew that no single administrative register would cover the entire need for information regarding these aspects. However, there is a lot of data and information available in different places, like administrative sources, information at advisory organisations, other statistical sample surveys, legislation etc. Our solution was to combine all these data sources, link them to our frame of holdings and use statistical methods to create data on micro level for each holding. The large number of variables made it a big challenge where a lot of different methods were used. The experience from this project will have a significant/great impact on the work with production of statistics in the future.

The Future of Sampling Frames in Official Statistics

Pedro Campos[1], Mr Carlos Marcelo[1], Mrs Sílvia Mina[1], Mrs Ana Santos[1], Mr António Portugal[1], Mr Jorge Magalhaes[1]

[1]Statistics Portugal, Lisbon, Portugal

One of the most crucial tasks in today's Official Statistics is related to the need for timely, rapid, and high-quality collection of information from individuals, families, and households. The high degree of updating in the universes and sampling frames is instrumental in achieving this objective, aiming to prevent low error rates. Errors may arise, for example, due to vacant housing or outdated personal contacts. It is essential to contemplate, on one hand, the integration of administrative information in updating sampling frames and, on the other hand, the frequency of such updates.

This work reflects on how administrative sources can continually enhance information about dwellings and individuals. Specifically, the focus is on updating contact information (phone, email, address), as well as sociodemographic information that aids in the sample selection for various statistical surveys. The study also delves into the opportunities provided by geographic information systems and their ability to define territorial patterns, facilitating the visualization of dwelling distribution and assisting in sample selection. For example, in the Household Financial Situation Survey, the sampling frame benefited from the fact that it was visually possible to represent the information for more specific strata.

We aim to address the following questions:

1. What information should be used to update sampling frames in family surveys?
2. What is the impact of sampling frame updates, especially on the calculation of selection probabilities?
3. How to monitor territorial dynamics to determine the optimal timing and frequency of sampling frame updates? In other words, what metrics should be employed to measure the inflow and outflow of dwellings/individuals and their respective sociodemographic changes?"

Session 30 - Emerging innovations, June 6, 2024, 14:00-15:30

An Innovative Framework for Analyzing Official Statistics: Symbolic Data Analysis

Prof. Paula Brito[1], Prof. Pedro Duarte Silva

[1]Faculdade de Economia, Universidade do Porto & LIAAD - INESC TEC, , Portugal

In classical Statistics and Multivariate Data Analysis data is typically represented in a data array where each row represents a statistical unit, for which one single value is recorded for each variable. This representation model is, however, too restricted when the data to be analyzed comprises variability. That is the case when the entities under analysis are not single elements, but groups formed from the aggregation of the original statistical units. Then the observed variability within each group should be explicitly considered. To this aim, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or distributions over a given domain. Symbolic Data Analysis provides a framework for the representation and analysis of such complex data, taking into account their inherent variability. This framework is of particular relevance in the analysis of official statistics, where the interest often lies in statistical units at a higher aggregated level, and where confidentiality issues prevent the dissemination and analysis of the microdata.

In this work we consider data where individual units, resulting from the aggregation of large amounts of microdata, are described by distributions of numerical attributes. We assume parametric models for numerical distributional variables based on the representation of each distribution by a central statistic, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles.

Note that this comprises interval-valued data as a special case. Multivariate Normal distributions are assumed for the whole set of indicators, considering alternative structures of the variance-covariance matrix. This model then allows for multivariate parametric analysis of distributional data.

Applications to official data put in evidence the added value of the proposed approach in this context.

Industry 5.0 and official statistics: opportunities for quality and process efficiency

Pasquale Papa[1], Paola Bosso, Giovanni Gualberto Di Paolo, Diego Distefano

[1]Istat - Italian National Statistical Institute, Rome, Italy

About a decade ago, a new era began in management of industrial production processes, named Industry 4.0. A further step followed, introducing Industry 5.0. While industry 4.0 includes principles connected to digital transformation of production processes through use of most advanced technologies such as AI (Artificial Intelligence), IoT (Internet of Things), cloud computing, blockchain technology, cybersecurity, industry 5.0 adds other aspects, among which the most relevant concern human factors and sustainability.

At the same time, official statistical systems are progressively moving towards exploitation of alternative sources, multisource data collection techniques, and new technologies to support survey processes, such as AI and M2M.

Objective of this work is to delve into opportunities offered by Industry 5.0 to official statistical systems, notably relating to quality of outputs produced, burden on respondents and efficiency of survey processes. They do not concern exclusively large companies, which traditionally represent the most suitable environment for the application of new technologies, but also small and medium-sized industrial entities.

The proposed analysis consists of two macro-phases: the first one concerns the review of characteristics and main functions of the most widespread software tools designed for management and support of industrial processes 5.0. In general terms, these software tools are second generation ERP (Enterprise Resource Planning) which, through native cloud, allow 360° company management (active cycle, passive cycle, warehouse management, accounting, production management, inventory management, reporting tools, sales and CRM tools, documents). A main advantage of a native cloud ERP management software is integration with third-party systems such as industrial machinery for Industry 5.0. In fact, an ERP system compatible with Industry 5.0 has the possibility of integration with production execution systems (MES-Manufacturing Execution System). As a result, it will be possible to trace all steps in transformation of raw materials into finished products. MES software allows managing the activities of the production departments, acquiring data on quantities, times and production costs in real time.

The second macro-phase consists in a specific analysis of needs of official statistics that identified software tools are able to satisfy, in terms of structural and short-term variables, focusing quality and process efficiency issues. It involves also the design of a short structured interview to collect data on attitudes and opportunities by a limited number of small or medium-sized enterprises adopting 4.0 or 5.0 technologies. The macro-phase includes, finally, a cost-benefit analysis related to possible implementation of the new approach in real survey processes.

Enhancing Quality in Official Statistics: A European Perspective of Open Source Technologies

Mr Antoniade-Ciprian Alexandru-Caragea[1], Mrs. Ana-Maria Ciuhu[2,3]

[1]Ecological University of Bucharest, Bucharest, Romania, [2]Statistics Romania, Bucharest, Romania,

[3]Institute of National Economy, Romanian Academy, Bucharest, Romania

This presentation explores the pivotal role of open source technologies, notably R and Python, in enhancing the quality of official statistics in Europe. We emphasize how these technologies are revolutionizing data collection, analysis, and dissemination, crucial for quality improvement in statistical processes.

Central to our discussion is the integration of open source tools within the European statistical system. We illustrate how leveraging these technologies can lead to more efficient, transparent, and robust methodologies in statistical data handling. The adaptability and versatility of open source tools are key in meeting the complex requirements of contemporary data science in official statistics, thereby directly contributing to the improvement of statistical quality.

Additionally, we delve into the broader implications of open source technology in the European context. This includes its role in fostering innovation, educational development, and building a skilled workforce. We argue that open source tools are instrumental in enhancing the capabilities of statisticians and data scientists, a vital aspect for maintaining and improving the quality of statistics in Europe.

Furthermore, we examine the impact of open source technologies on the operational strategies of statistical agencies. This encompasses a detailed look at how these technologies streamline operations, enhance statistical methodologies, and lead to more effective practices.

In conclusion, our submission will not only highlight the current applications of open source in European official statistics but also propose strategies for its broader integration in the future. We aim to provide insights into the effective utilization of open source technologies for advancing the quality, reliability, and accessibility of statistical data in Europe, thereby contributing to the ongoing discourse on the strategic importance of innovative methodologies in the field of official statistics.

Multi-Party Secure Private Computing and Quality: a prospective alliance

Fabio Ricciato[1]

[1]Eurostat, , Luxembourg

Privacy-Enhancing Technologies (PETs) is a general umbrella term referring to a diverse range of emerging technological approaches aiming to reconcile data confidentiality and data utility. Multi-Party Secure Private Computing (MP-SPC) methods represent a sub-group of PETs that enable two or more organisations holding confidential data sets to cooperate and compute statistical results based on the integration/combination of their respective datasets without revealing their input data to each other or to an external third party. MP-SPC systems may facilitate cooperation among statistical organisations and with external data providers. MP-SPC systems, may contribute to overcoming the barriers – particularly concerning legal compliance and public acceptance – to the (secondary re) use for statistical purposes of data collected primarily for non-statistical purposes by other entities.

In this contribution, starting from the definition of statistical “quality” as defined in the European Statistics Code of Practice (CoP) and ESS Quality Assurance Framework (QAF), we elaborate on the relationship between MP-SPC and statistical quality. We show that such relationship is one of mutual reinforcement and alliance: MP-SPC effectively contributes to strengthen statistical quality along several dimensions, while a proper consideration of the quality dimensions spelled out in CoP and QAF may guide the design and implementation of effective MP-SPC systems, both at the technical and organisational level.

Elaborating on the mutual relationship between MP-SPC and quality aspect is an important exercise that comes very timely as Eurostat is preparing to launch a new project [1] on the specification, feasibility analysis and prototype demonstration of a MP-SPC system in support of statistical innovation. If successful, this project will advance towards the deployment of a shared MPSPC-as-a-service infrastructure for the whole ESS, in line with the provisions of the recently adopted legislative proposals by the European Commission [2,3] .

Footnotes

[1] Call for tenders ESTAT/2023/OP/0004.
[2] See the EC proposal for amending Regulation (EC) No 223/2009 on European statistics (link), and specifically Art. 17f and Recital 15 therein.
[3] See the EC proposal for a new Regulation on European statistics on population and housing (link) and specifically Art. 13, Art. 14 and Recital 30 therein.

Transitioning to the New Editing Process at Statistics Sweden

MR Fredric Nyström[1], Karl Sahlberg[1], Pia Hartwig[1]

[1]Statistics Sweden, , Sweden

In 2021, the management of the Statistics Sweden (SCB) made a central decision to restructure the manual editing process with the aim of shifting focus from editing data at the micro level to the macro level. The new editing process is essentially divided into two parts; respondent editing (editing that occurs before data submission) and macro editing (editing that occurs after data submission).

The decision was made due to the need to streamline resource allocation and prioritize development projects in other areas. The main goal was to ensure cost-effective work with data collection and editing. The new approach centered around respondent editing, reducing the need for manual editing after data submission. Instead, respondent editing in online surveys became the primary means to correcting data, using prompts to guide respondents to submit correct data.

To implement this new editing process, SCB provided guidelines and checklists for designing and evaluating prompts. The focus was on identifying errors that could have a significant impact on the quality of the statistics. The target group for the guidelines is primarily the competencies within SCB that collaborate in the survey.

Having a macro perspective with selective elements is already a consideration in the selection and design of prompts. This is to facilitate work in the subsequent steps of macro editing. The subject group working with the Production Value Index is an example of a product that has successfully managed these challenges and maintained quality while transitioning to the new editing process. To facilitate this transition, the group implemented measures such as prompts and an automated correction system. In addition, macro levels selectors were introduced to identify specific industries requiring manual editing, thereby reducing the risk of errors at the industry level.

Finally, the team assessed unedited and edited data, comparing microdata that had not been edited or processed at all with microdata that had undergone assessment by both the collection and subject teams. Drawing from this data, the project team could conclude that nearly all incorrect values impacting their respective industries would be identified during the subject team's macro level editing. With this conclusion, the project team felt confident in moving towards a more macro-based editing approach.

Our presentation will describe how the implementation of the new editing process has worked in practice. We will provide examples from surveys where the new editing process has had a significant impact.

Enhancing Data Quality: A DataOps approach with R and GitLab

Mr. Alexandre Cunha, Mr. Bruno Lima, Mr. João Poças

[1]Statistics Portugal , , Portugal

DataOps has emerged as a fundamental practice, essential for optimizing management and delivery of data-products. The DataOps framework is designed to enhance collaboration between those involved in the data lifecycle. Using R for data analysis facilitates the implementation of streamlined and reproducible data workflows. Its integration with GitLab, a version control and collaboration platform, empowers teams to efficiently manage and version data projects, fostering reproducibility and ensuring that analyses can be recreated and validated by other team members. The R package

{targets} is a key tool for the definition of Data pipelines that help to streamline the movement of data from source to destination. The integration of data pipelines with R and GitLab facilitates the automation of data extraction, transformation, and loading (ETL) processes, reducing manual effort and enhancing data quality.

Using GitLab facilities like “Issues” and “Issue Boards” in a DataOps context can significantly boost project management, collaboration, and tracking within data science workflows. GitLab Issues provide a structured way to define, assign, and track tasks or objectives within data projects, where each issue can represent a specific piece of work, such as data cleaning, model development, or report generation. Meanwhile Issue Boards provide an at-a-glance view of project progress. Likewise, team members, stakeholders, and managers can quickly see which tasks are pending, in progress, or completed, promoting transparency.

Data on Portuguese exports and imports are collected monthly, on each trade exchange, through mandatory online business surveys implemented by Statistics Portugal. This international trade data is not error-free and as such the reported values need to be checked in an automated way to detect potential errors. Using a DataOps approach, we implemented an ETL process to produce a score model that returns the likelihood of error for each reported value. To achieve this, we defined issues that could be translated into functions to be implemented through an R package developed specifically for this project. In this way the project team was able to monitor and discuss the tasks using a kanban board to manage these issues.

In conclusion, the convergence of DataOps, R, GitLab, reproducibility, and data pipelines represents a powerful paradigm shift in data science and in statistical processes. Here, we show some insights into how these elements can be effectively combined to foster collaboration, increase the efficiency of data workflows, and ensure the reproducibility of results in the dynamic and fast-paced world of data-driven decision-making.

Session 31 - Special Session: Communicating Quality, June 7, 2024, 09:00-10:30

Communicating quality to different types of users

Ms Wendy Schelfaut[1]

[1]Statistics Belgium, , Belgium

National Statistical Institutes have different types of users, to which different types of communication apply. At Statistics Belgium, the overall communication strategy focuses on giving a 'Quality Label' to all of our statistical publications. Three main objectives are put forward in this strategy:

1. Assume the role as a reliable hub and increase trust through objective and pertinent communication.

2. To further modernize the statistical process and be a leading statistical agency in communicating it, nationally and internationally.

3. Ensure a customer-friendly and efficient government organization through accessible and transparent communication.

Within these objectives, the focus will be on broadening the communication, connecting with the users and renewing the communication approach.

To make sure that these goals are met, Statistics Belgium has developed a set of user personas. Eight types of users, assembled through in-depth user research in the past years, make sure that we keep all types of users on the top of our mind, at all times.

In the three objectives as well as in the three focus points, showing and communicating quality is the overall goal, to be acknowledged as a qualitative and trustworthy institution by all of our users, no matter what type. In the presentation, different examples of concrete quality communication actions will be given, such as workshops for journalists, a conference on 20 years of EU-SILC in Belgium, methodological analyses for advanced users, and the conference on statistical literacy.

Quality assurance and user centred design in dissemination

Susanne Taillemite[1], Ms Julia Urhausen

[1]Eurostat, , Luxembourg

Today more than ever, statistical organisations compete for attention in an environment saturated with information. They can only accomplish their mission if statistics are effectively disseminated and reach their target audiences. In this context, ensuring the attractiveness and quality of our dissemination products and services becomes crucial.

The paper will present two essential perspectives: the imperative of ensuring quality in the dissemination process, and the significance of user experience (UX) through a user-centered design approach.

The first section looks into mechanisms for ensuring quality dissemination products, which should cover all stages of the dissemination process, from conception and design to implementation and quality review.

The evolution from traditional one-way communication to a dynamic two-way process is highlighted, emphasizing the need to consider diverse target audiences and channels for effective communication. The application of the 5 Cs of communication—clear, correct, complete, concise, and compassionate—is discussed, with an emphasis on collaboration between statisticians and dissemination experts.

The second section will highlight how Eurostat is constantly working on further refining its user-centred design approach, showcasing how the two pillars - user research and usability testing - enhance the quality of our dissemination products.

For Eurostat, UX is a key priority when creating our dissemination products. The paper will elaborate how our qualitative and quantitative user research aims to identify key user segments of users of European statistics, their needs, paint points, and expectations. The results of this user profiling study will also be presented.

This work is complemented by usability testing of our products and tools. Direct interaction with our users enables us to gain in-depth insights into their practical use of our statistics and dissemination products. These insights are then used to improve the quality and relevance of existing dissemination products and for the design of new products targeted at users with different profiles and needs.

By gaining an in-depth knowledge about our users and their evolving habits for accessing and using data we can design statistical products that respond even better to specific user segments or audiences, thereby creating a positive user experience and maintaining Eurostat’s position as the trusted provider of high-quality statistics and data on Europe.

Communicating the third round of ESS peer reviews – a Member State experience

Ms Petra Kuncová[1]

[1]Czech Statistical Office, Prague, Czech Republic

Third round of the ESS peer reviews, aimed at reviewing compliance and alignment with the European Statistics Code of Practice by the national statistical systems, was a key activity in the field of quality management in the Czech Statistical Office in 2022 and 2023. The actual visit of the peer reviewers took place in March 2023 and was preceded by intensive preparatory activities such as filling in the self-evaluation questionnaire, organizing the visit, securing participants, etc. Likewise, after the visit, the work continued with cooperation to create a final report and to propose improvement actions. All these phases were in the Czech Statistical Office accompanied by intensive communication, both external - intended for users, partners and respondents of the CZSO, and internal - employees of the office. In the period before the visit, the CZSO explained to the public the purpose of the Peer review, the meaning of the European Statistics Code of Practice, benefits for the public, etc., and the intensity of communication increased as the date of the visit approached. After the end of the visit, when we already had a draft recommendation, communication took place mainly with the Czech Statistical Council, and of course also with the office's employees.

For communication, all communication channels that the CZSO normally uses were used – CZSO websites, intranet, the magazine Statistika&My, twitter, presentations, exhibitions etc. During the campaign, all documents prepared by Eurostat were used to a great extent. Practical examples will be shown during the presentation.

Quality as a Part of the Brand of a Statistical Organization

Ms Hanna Ikäheimo[1]

[1]Statistics Finland, ,

Organizations often summarize their core tasks in a mission that describes the purpose of the organization's existence. The missions of statistical institutes vary from country to country, but often the mission involves producing high-quality reliable data and statistics.

A mission creates one perspective of an organization's brand. Other elements of the brand are, for example, the organization's core competencies, its relationship with stakeholders, and its style of expressing itself. By defining the elements of the brand, we crystallize an organization's identity: what we want to be as an actor and what we want to be known for. By acting and communicating in line with the brand, we aim to lead the organization's reputation in the desired direction.

Official statistics are based on common manuals, methods, and the Code of Practice. In the CoP the quality of statistics is described by means of relevance, reliability, and comparability of the data. The wishes of the users are very much in line with these values.

In society, there is a demand for high-quality information. How can we use this demand in building the brand? The importance of high-quality data production can be included in the core of the brand and utilized consistently in communication. But the brand comes to life also in the organization's actions and expressions.

The CoP creates the basis for statistical production. Therefore, the NSIs have a strong foundation for building a comprehensive and quality-based brand. When communicating about statistics, we can systematically highlight the sources and methods used in statistics. This helps to highlight the uniqueness and quality of the official statistics.

The statistical offices have in-depth knowledge of society's data reserves. It is possible to use this expertise more widely in society. For example, Statistics Finland has proposed improving the quality of Finnish registers and has coordinated data quality work aimed at society's register keepers. The purpose of the work has been to increase the understanding of the quality of information and to develop the quality of registers, so that their use can be more effective.

Official statistics are produced with open procedures that have the quality of products in mind. With a solid foundation, it is good to build a unique brand for a statistical agency, a brand that sets the agency apart from other actors.

‘Communicating quality’

Lukasz Augustynia[1]

[1]Eurostat, European Commission, Luxembourg, Luxembourg

Effective communication and promotion of National Statistical Institutes’ and Eurostat’s products and services is indispensable for making their high-quality statistical work known to EU citizens and policymakers. In fact, it is often the quality of this communication that determines whether a user will click on and work with a product developed by a European Statistical System (ESS) partner or on one of many often less reliable products prepared by their competitors from the private sector.

Promoting our values and the principles governing our work further strengthens the image and reputation of ESS members as producers of reliable figures on Europe, helping to underpin our democracy. Within this context, communication on how we ensure quality in our work, including the use of peer reviews as well as our increasingly innovative practices, is central if we are to continue to earn the trust of current and future users.

The National Statistical Institutes and Eurostat have been investing in the communication of their products and services, also strengthening their cooperation in the area of communication. This session will provide concrete examples of this work. The session will feature five presentations dealing with different aspects of communication of statistical quality.

Between 2021-2023, Eurostat has been coordinating the third round of the ESS peer reviews. Two presentations in the session will be devoted to this topic, with Eurostat sharing its experience in communicating this complex exercise from the European perspective and Czechia presenting an ESS country perspective.

Next, Statistics Finland will present its communication work centred on quality as part of their brand. The presentation will consist of a number of elements, including the relationship between quality, communication and statistical work (f.ex. the improvements in the data quality of the country’s registers and Statistics Finland’s strategy).

Then the Belgian Statistical Office’s presentation will focus on communicating quality to different types of users. This will include working with ‘user personas’ and presenting examples of quality communication actions aimed at them, such as workshops for journalists, methodological analyses for advanced users, and the conference on statistical literacy.

Finally, in their presentation, Eurostat will talk about how dissemination tools and products need to remain attractive, interesting, and relevant for users. With this in mind, the presentation will outline the two pillars of Eurostat’s user-centered design approach in creating its dissemination products – user research and usability testing – and how they enhance the quality of its dissemination products.

“Communicating the third round of ESS peer reviews – the Eurostat experience”

Ms Claudia Junker[1]

[1]Eurostat, Luxembourg, Luxembourg

The third round of ESS peer reviews took place between 2021 and 2023, coordinated by Eurostat. The peer reviews aimed at reviewing compliance and alignment with the European Statistics Code of Practice by the national statistical systems and at providing forward-looking recommendations for a further development of the national statistical system.

A novelty in this round of peer reviews was to accompany it by a communication campaign, based on generic communication material developed by Eurostat. The idea behind the communication campaign was to use the peer reviews as an additional opportunity for communicating the value of official statistics to various categories of users but also to data providers/respondents. For this purpose, several materials were developed like key visuals, videos and facts sheets. This kind of communication was supposed to target a more general target group of users.

The other idea behind the communication campaign around the peer reviews was to enhance communication on the peer review process in a given country to some more specialised users such as the government and policy-makers. It had the aim to make government authorities responsible for the National Statistical Institute to be aware of such a unique process as peer reviews and the recommendations resulting from it as a number of recommendations would be addressed to the relevant government authorities, e.g. recommendations to change the statistical law, to ensure proper resourcing of the statistical office, enhance the coordination role of the National Statistical Institute in the national statistical system and others. For such recommendations the support and willingness to act from those government authorities would be needed for their implementation.

Therefore, specific communication material was prepared to address those authorities, to explain the process of the peer review, the content of the European statistics Code of Practice and the continuous drive for improvement and better services to users. The paper will describe in more detail what kind of communication material was available and how it was used in different countries to indeed accompany the peer review process with a targeted communication campaign. Good practices and examples will be presented – as inspiration for other countries and to learn from each other.

Session 32 - Statistics and decision making II, June 7, 2024, 09:00-10:30

Climate change: a statistical short-term answer

Isabelle Remond-tiedrez[1]

[1]Eurostat, Luxembourg, Luxembourg

Climate change is a high priority on the European Commission political agenda and call for relevant economic indicators. On our way to an EU climate-friendly economy, the European Statistical System needs to closely evaluate and monitor the impact of climate change on EU economic growth, on jobs and on the investment needed to achieve targets.

There is not yet standardized framework to follow up climate change statistics and in particular investments. However, Eurostat, with the help of the Member States, has set up a classification of environmental purposes that will allow the environmental accounts framework to provide statistics on climate change in a medium term. This approach is part of a methodologically defined and internationally recognised framework. In the short-term, there are already some relevant statistics such as Structural Business Surveys, PRODCOM data, Labour Force Surveys that form the basis for setting up macro-economic measures on climate change. Eurostat has investigated this path to provide some key indicators' estimates on climate change. The approach starts with defining the climate change policy in question and then delineates economic activities related to the climate change within the classification systems of NACE and Prodcom. The methodological approach uses top-down modelling techniques, building on a cascade of information already existing. Results of the estimation procedures may vary due to the definition itself, whether downstream and upstream activities are included in the policy of climate change but, as well, due to the evaluation of how much an economic activity is related to the policy area.

The approach has applied quality principles such as a sound methodology for the statistical process where standard concepts and classifications are applied. It does not involve any burden on respondents, neither national statistical institute as the approach uses established and available Eurostat data. The approach will allow high quality standards, especially comparability and timeless and more importantly relevance of the estimation provided.

Engaging with External Partners; Experiences and Perspectives of the Turkish Statistical Institute (TurkStat)

Ms Fazilet Devecioglu[1]

[1]The Turkish Statistical Institute (TurkStat), Ankara, Türkiye

As indicated in the European Statistics Code of Practice (ESCoP), maintaining cooperation with academic institutions and international bodies has a significant influence on the effectiveness and credibility of a statistical authority. Furthermore, academia and international bodies are pivotal stakeholders in addition to government authorities, with academia generating academic products based on statistical outputs and international bodies making vital decisions and implementing projects for the betterment of the global community. Consequently, fostering collaboration, considering their specific needs, and engaging in joint projects contribute significantly to elevating the quality of statistics and bolstering the credibility of statistical offices.

The Turkish Statistical Institute (TurkStat) has placed considerable emphasis on fostering interaction with the leading international, regional organisations such as UNECE, OECD, the World Bank, ILO, UNICEF, UNFPA, SESRIC, and more. Recognising the reciprocal nature of these relationships, TurkStat has actively pursued a two-way dynamic interaction. TurkStat, with the aim of strengthening relations based on partnership and mutual trust, has become increasingly receptive to collaboration activities in many statistical areas. To formalize and structure these cooperative efforts, TurkStat has signed memoranda of understanding and working plans with numerous international and regional organisations, providing a robust framework for cooperation and streamlining collaboration processes.

In addition to the collaboration activities with international organisations, TurkStat has strengthened its ties with universities. Within this framework, TurkStat sponsored various congresses, and its staff actively participated by delivering presentations on diverse subjects, fostering partnerships with the scientific community. Moreover, TurkStat extended its hospitality to university communities, hosting them at its premises to introduce new statistical trends, introduce the organisation, and attract interest. Additionally, TurkStat played a role in promoting statistical literacy by disseminating the ''Indicators of 100 Years'' and "Statistical Literacy" products developed by TurkStat, to further support statistical literacy.

This paper aims to present TurkStat's experiences in collaborating with academia and international and regional organisations, share the perspectives related to quality, its achievements, and coordination efforts. The focus will encompass TurkStat's initiatives aimed at bolstering the credibility of statistical information, providing more comprehensive and qualified statistical data, enhancing statistical literacy, developing professional competence of TurkStat personnel, and strengthening relations with external partners. The first section of the paper will provide overview of TurkStat's collaborations with international and regional organisations. Secondly engagements with academia and tools will be elaborated. The concluding section will outline major achievements.

The contribution of Citizen Generated Data (CGD) for measuring gender- based violence (GBV) in Italy. A quality issue for official statistic

MR Francesco Gosetti[1], Ms Alessandra Battisti[1], Ms Alessandra Capobianchi[1], Ms Federica Pellizzaro[1]

[1]Istat - Italian National Institute of Statistics, Rome, Italy

The measurement of GVB, even if much progress has been made alongside, is still hampered by difficulties and limitations, concerning content, methodology, regularity and timeliness of existing data. At the same time, the data revolution, with a greater volume of data coming from different sources and a changing relationship of individuals and communities with data, represents an exceptional challenge for the official statistics and to fill gender-related data gaps. This claims also for greater citizen and stakeholder's participation in planning and data-related activities and National Statistical Systems (NSS) have to broaden their scope to a "data ecosystem" that include, adopting a participatory approach giving due consideration to the voice of citizens, all actors that generate data. In this regard new and alternative data sources may complement existing gaps providing exceptional insights on underreported phenomena, enabling additional analysis and reducing the statistical burden. Nevertheless, most of the time these data are not produced for statistical purposes and, given their operational nature, NSOs face challenges to leverage such an opportunity.

Within this framework, Istat, in collaboration with the National Department for Equal Opportunities, developed a multi-source approach to collect, analyze and publish data from different sources on GBV, including administrative and alternative ones, devoting particular effort to enhance value also of non-traditional data sources such as citizen generated data (CGD) on the topic. To this aim, data underlining services offered by specialized agencies for women victims of violence, such as antiviolence centers and shelters and calls directed to the national helpline 1522, have been integrated in the NSS enabling a mutual win-win outcome for both data producers and users as well as for the statistical ecosystem as a whole. Such a process required a great variety of competences and activities and has to undergo a constant monitoring in order to ensure data quality, comparability and usability.

The paper aims at discussing 1) the main steps that led to the integration of the above-mentioned data into the integrated informative system on GBV in Italy, with particular emphasis on those activities that most affect the quality along the whole statistical value chain (needs assessments, collection, processing, analysis and interpretation) 2) the role of the continuous involvement of respondents as data providers and users in order to ensure data collected comply with the principles governing official statistics 3) the role of Istat in improving data collection capacity of data producers by coordinating the whole process.

Location of physical assets – addressing one of the main data gaps in assessment of climate-related risks

Malgorzata Osiewicz[1], Leslie Yvonne Pio[1], Giuseppina Borea[1], Carlos Mateo Caicedo Graciano[2], Augustin Lion-Atlan[2], Loriane Py[2], Leïla El Kaissoumi[3], Léopold Gosset[3]

[1]ECB, Frankfurt, Germany, [2]Banque de France, Paris, France, [3]ACPR, Paris, France

The identification of the exact location of non-financial companies as well as their physical assets is one of the main challenges for the accurate assessment of the exposure to physical risk such as floods, wildfires, or draughts.

National data sources on business statistics might contribute to addressing this data gap, in particular data collected on enterprise and their local units under the European business statistics (EBS) framework looks very promising. Apart from the address, it also offers information on the economic activity and number of employees of each local unit, and in selected countries additional variables such as revenues and total assets. Those attributes can be used to estimate the value of physical assets at different locations – a key element for the assessment of potential losses caused by natural catastrophe.

To assess the data availability the Statistics Committee Expert Group on Climate and Statistics (STC EG CCS) launched fact-finding among national statistical institutes (NSIs) and national central banks (NCBs).

First, the paper summarises the findings from the exercise elaborating on the key elements: available information, collection modes, potential data quality issues and access modalities across European countries. The initial analysis is re-assuring: overall, the business population seems well covered in most countries. Importantly, enterprises and local units can be identified by standard company identifiers for most of the euro area countries which enables linking information with the ESCB's datasets on financial exposures such as loans, debt securities and equity.

Second, we illustrate a usability of the dataset on the example of French datasets. We quantify the potential mismeasurement of physical risk based solely on the registered address of a company versus assessment based on multiple locations of a company.

Finally, the paper proposes an avenue for expanding the analysis to other countries to better capture the climate-related physical risk in financial and economic studies.

Session 33 - Data Visualisation, June 7, 2024, 09:00-10:30

How to Communicate and Visualise the Quality of Short-term Business Statistics Indicators to Users?

Hanna Dieckmann[1]

[1]Eurostat, , Luxembourg

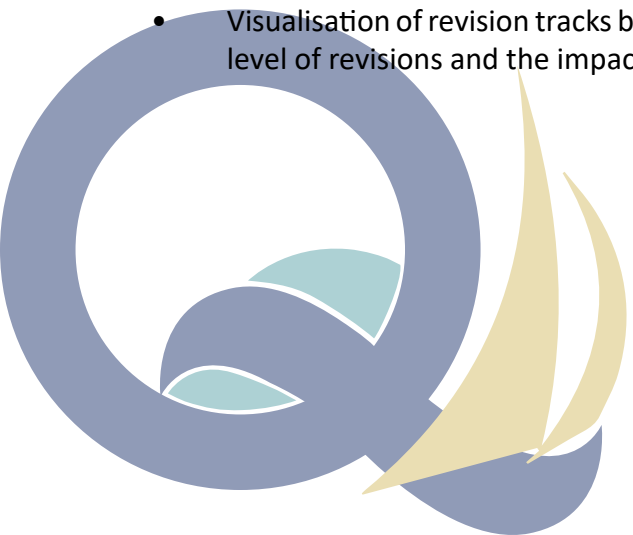
Economic policy implementation relies heavily on the timely availability of economic indicators. Short-term business statistics (STS) indicators are published as monthly, quarterly, and annual indices. The deadline to publish the data is short, therefore, revisions constitute an integral part of production and publication process of STS indicators. Different types of revisions can be distinguished. Apart from routine revisions due to late incoming data and regular benchmarking, revisions can be caused by methodological changes and the correction of errors.

In view of STS, revisions are necessary to provide good quality data. Revisions are not always a sign of insufficient quality unless they are systematically very large. At the same time, the revision indicators give guidance to users what is the expected level of revisions for the data and allow for an assessment of the reliability of the STS indicators.

The aim of the paper is to promote the existing ESS standards for quality reporting and present different ways of communicating the results of the revision analysis to the users. In doing so, it presents new ways to visualise revision tracks to show possible future developments in the field of quality reporting of STS indicators.

The paper is structured as follows:

- Introduction to aspects of quality of STS indicators.
- Overview of the release and revision policy of STS.
- Explanation of different reasons for revisions.
- Illustration of different reasons for revisions using examples of revisions tracks from STS European aggregates.
- User-oriented revision analysis of STS Principal European economic indicators using different quality indicators like the mean absolute revision (MAR), the mean revision (MR) and the relative mean absolute revision (RMAR). The focus is on differences between the first and second publication as well as the first and last publication of the industrial production index, the construction production index and the retail trade volume index.
- Visualisation of revision tracks by different means to give guidance to the users on the expected level of revisions and the impact of revisions on their estimates and models.



Social Media: Bring statistics to life

Ms Ilka Willand[1], Kerstin Hänsel[1]

[1]Head of Section, Federal Statistical Office Of Germany, Wiesbaden, Germany

As a result of digitalization society has changed very dynamically in recent years. The field of communication plays a central role in this development: It is as well changing dramatically by the increasing importance and worldwide use of social media, which enables also statistics institutions for the first time to get in direct contact with a high number of their users. Social media offer the opportunity to establish official statistics in an unprecedented dimension as a trustworthy data source and to fulfill NSIs legal mandate which is to enable citizens in forming their political opinion. In particular, the biggest opportunity lies in direct communication with younger age groups, who had not been reached in a direct way ever before. On channels like Instagram young adults or even kids can be reached to deliver them with reliable data about society and to strengthen their data literacy. The other side of the medal is that fast and direct contact between everyone (non-filtered and lead by personal opinions) opens also opportunities for antidemocratic and manipulative parties to influence and split the society to an enormous extent – first of all younger generations are in danger to become victims of fake news, framing and disinformation.

Due to the social changes, the German statistics office Destatis has fully digitalized its communication in 2020 and operates since 2021 an Instagram channel with the strategic goal of increasing young people's data literacy. Simplification and storytelling are playing a central role to run this video-based channel successfully.

The presentation provides approaches on how to spread information and increase data literacy among people 14 years plus. How does the platform influence the way of presenting statistical results? Which formats achieve the greatest reach? And what limits and challenges does the shift towards social media bring?

Statistical Quality in Data Visualization

Mr Steven Klement[1], Rosemary Byrne

[1]U.S. Census Bureau, Washington, United States

Today's powerful graphics software packages can create intricate static or interactive visualizations of data that were unheard of a short while ago. Many statistical agencies are embracing the newest and most powerful techniques to make intriguing visuals hoping to entice people to use their data products. Meanwhile the literature is trying to catch up to determine which types of visualizations get the true story of the data across to the user and which are confusing or worse, misleading. The quality of the visualization can affect the reputation of the statistical agency and the data it produces. This means the agency should produce visualizations that are eye-catching to attract people to the data product but also ensure the data are accurately represented.

The U.S. Census Bureau has been developing statistical standards for data visualizations for a decade. The current version has been in beta mode for about three years as development was slowed during the COVID shutdowns. The key issue we are dealing with is that a standard is normally something that you either shall do, or something you shall not do. But we have found that creating a good data visualization is more an art form that does not easily lend itself to strict rules. Therefore, good practice in data visualization is fostered by a combination of formal standards and best practices.

Combining both in one place allows for single-source guidance that ensures the visualization creator is well-informed to produce the best possible products.

This paper will lay out the issues we dealt with in getting our standards to their current state, how we resolved them, and what may lay ahead in the future. We hope to promote discussion in the international statistical community as to how others are dealing with statistical quality in data visualization.

Interactive Web Visualisation of Eurostatistics via R: Enhancing the Quality of Data Presentation through Storyboarding

Piotr Ronkowski[1], Ms Rosa Ruggeri Cannata, Ms Anette Sundstroem

[1]European Commission / Eurostat, , Luxembourg

In today's digital landscape, there is an increasing demand to make statistics both more accessible and more meaningful to the general public. The transformation of Eurostat's "Eurostatistics" from a static PDF document to an interactive web visualisation based on R represents a significant advance in the quality of data presentations which have to be produced in a short time or be updated frequently. This initiative, led by Eurostat Unit C1, is not merely a format change but a strategic enhancement of how statistical data are communicated. Central to this transformation is the innovative use of the Flexdashboard package in R, particularly its storyboard feature.

Storyboarding organises data into an engaging narrative, effectively drawing out the main trends in the various indicators. The integration of the Plotly package further enhances this storytelling characteristic by incorporating captivating interactive charts. These charts not only highlight the most relevant macroeconomic indicators but also add an interactive dimension to the narrative. This synthesis of narrative clarity and interactive features plays a pivotal role in making statistical data more accessible and comprehensible to a broader audience, regardless of their level of expertise.

The strengths of the interactive web visualisation tool reside in its open-source foundation in R, built upon web standards such as CSS, HTML and JavaScript, ensuring universal operability across devices and leading browsers. It produces self-contained output without dependence on additional servers. Moreover, the tool seamlessly integrates data from varied providers and in different formats, offering a streamlined, narrative-style presentation that is well-suited for publications involving high data volume and frequency. With low security risks, an intuitive learning curve, and a swift production process, this tool offers a fresh perspective on traditional publishing techniques and is particularly fitting for the publications requiring regular updates. Eurostat is investigating the possibility to generalise this tool, with the view of sharing it with interested national statistical authorities. This would permit a step further towards innovation with a low investment cost.

At the conference, Eurostat will present how this approach to data visualisation, especially through the use of storyboarding, can improve the quality and efficiency of data presentation.

Session 34 - Privately-held data, June 7, 2024, 09:00-10:30

Working with a mobile network operator (MNO) to create a privacy- conform method for a better access to MNO-Data

Lorenz Ade[1], Maurice Brandt[1]

[1]Federal Statistical Office Of Germany, Wiesbaden, Germany

Mobile network operator (MNO) data has huge potential for official statistics. Commercially available data is currently created and processed in a kind of black box, as the aggregation and extrapolation of the data is a business secret. MNOs currently can't share their processing steps, as they employ confidential information like cell tower positions and local market shares for their algorithms. For the usage of MNO data in official statistics, this black box has to be opened if the quality criteria of transparency and comparability are to be achieved.

To solve this problem, DESTATIS has partnered with T-Systems (a subsidiary of the MNO "Deutsche Telekom"). In this collaboration, T-Systems provides DESTATIS, for the first time at all, with access to the secure environments of an MNO to work with anonymized raw signal data. The goal of the collaboration is to create a standardized, transparent and privacy-conform method for an access to MNO data. This work is a part of a large German research cluster on the anonymization of georeferenced data (AnigeD).

The collaboration between DESTATIS as national statistical institute and T-Systems as MNO guarantees that the necessary expertise in relation to statistics, data protection and mobile network is present in the project. Additionally, the collaboration ensures that the business interests of the MNO are maintained, as they are free to offer their data commercially based on their own algorithms.

Key focus besides achieving the necessary quality criteria for official statistics is the compliance with privacy standards. This is not alone a legal necessity. Ensuring data privacy is fundamental to maintain the trust of the public in official statistics. To achieve this goal when handling the very sensitive MNO data, privacy by design has to be included in the processing of the data from the very beginning. Additionally, data protection officers and institutes are consulted regularly.

Towards the usage of Mobile Network Operators' data for European official statistics production: improving quality through standardisation

Tiziana Tuoto[1], Matthias Offermans[2], Ricardo Herranz[3], Margus Tiru[4], Florabela Carausu[5], Dr Erika Cerasti[1], Roberta Radini[1], Cristina Faricelli[1], Paolo Mattera[1], Gabriele Ascari[1], Giorgia Simeoni[1], Luca Valentino[1], Edwin de Jonge[2], Miguel Picornell[3], Villem Tonnison[4], Cristina Escribano[3]

[1]Istat, Rome, Italy, [2]Statistics Netherlands CBS, , Netherlands, [3]Nommon Solutions and Technologies, Madrid, Spain, [4]Positium, , Estonia, [5]GOPA Worldwide Consultants, , Germany

In recent years, the widespread adoption of personal mobile devices has opened new opportunities to collect rich spatio-temporal data about human activity. In this context, governments are embracing data-driven approaches as a powerful tool to inform policy-making. As an example, during the COVID-19 pandemic different European governments used data obtained from Mobile Network Operators (MNOs) to monitor the mobility and presence of people across the territory. These data proved to be useful for a variety of purposes, such as epidemic modelling and mobility planning, which fostered the interest of National Statistics Institutes (NSIs) in using MNO data for the production of official statistics. To leverage the proved richness of MNO data while acknowledging that the processing of these by private companies, through an undisclosed methodology, may not reach the quality and reliability standards requested by official statistics, the European Statistical System (ESS) is investing in creating the conditions to integrate MNO data in the production of statistics featuring increased timeliness and information content. In this paper, we present the work done in the context of the Multi-MNO project funded by Eurostat with the aim of developing a standardised pipeline for MNO data processing. We define the high-level principles followed by the proposed pipeline — which has been designed to be modular and general enough to fit the needs of different countries, multiple MNOs, and several statistical purposes — and describe the methodological steps of the proposed pipeline, with particular focus on possible quality assessment measures to be implemented throughout the different processing phases.

Applying the extended Total Survey Error approach to statistics based on new data sources: the case of MNO data

Mr Gabriele Ascari[1], Giorgia Simeoni[1]

[1]Istat, Rome, Italy

Among the new data sources that National Statistical Institutes (NSIs) have begun to explore, data from Mobile Network Operators (MNOs) occupy a relevant position. The informative potential of these data covers many domains, from population statistics to mobility and tourism. However, as with other data sources that originate from non-statistical purposes, MNO data should be thoroughly manipulated before statistical offices can extract meaningful information from them. Similarly to administrative data, MNO data are generated out of the control of NSIs. Furthermore, in order to fully exploit the potential of these data, integration with other sources – even traditional ones – becomes a crucial passage. Additional difficulties derive from the fact that the initial preprocessing operations have to be carried out by the MNOs themselves to preserve confidentiality and also because a relevant infrastructure is needed to elaborate such high volume “big” data. Therefore, the issues that may affect the overall process, along with the ones affecting the source itself, can be hard to identify and assess. In this work we propose a structured approach to explore the quality of statistics based on MNO data, focusing in particular on the quality issues arising during data processing. The errors arising during the process are studied under the lens of the well-known Total Survey Error approach: although developed in the context of sample surveys, this approach has already been extended for processes based on administrative data and combined data and, in our opinion, can offer valuable insights for processes based on non-traditional data.

Quality assurance of official statistics based on privately held data: the use of reference methodological pipelines

Peter Struijs[1], Fabio Ricciato[1]

[1]Eurostat, Luxembourg, Luxembourg

In anticipation of the revision of the framework regulation of European Statistics, the European Statistical System (ESS) is preparing for the future (re)use for statistical purposes of new data sources, including data produced and held by the private sector. In this way statistical authorities intend to produce better, richer, and timelier statistics to respond to the growing information needs of their statistical users, from public policy organisations and decision makers to citizen and business actors.

Basing statistics, partly or completely, on privately held data requires a novel approach to quality assurance. New data sources have characteristics that differ from more traditional data sources. Importantly, whereas NSIs are used to receive the data from surveys and administrative sources and then do the processing internally, this practice cannot be assumed to be feasible or anyway convenient when dealing with new data sources. On the contrary, under certain circumstances processing the data by the data holders at their premises may be preferable to transferring the raw data to the NSIs. In fact, the data holder and the NSI may share the production process of the statistics in several ways. Nevertheless, the NSI remains fully responsible for the statistics produced and their quality. This may lead to challenges in quality assurance.

The ESS is currently exploring a novel approach to quality assurance for the case of using Mobile Network Operator (MNO) data for official statistics. The core of the approach consists of data holders and NSIs agreeing on the use of a so-called reference methodological pipeline, that is, a detailed and modular description of the data processing flow, with clear indications of what each module does and how it operates, expressed in a formal, non-ambiguous language. This pipeline is a prerequisite to keep NSIs in effective control of the quality of the statistics produced.

The paper looks at the potential of the approach to the MNO case for application to other privately held data sources. It will be argued that even if other data sources have distinctive characteristics and the data processing pipelines will differ from source to source, the adoption of reference methodological pipelines will be necessary to enable the ESS to exercise quality assurance when dealing with new data sources.

Session 35 - Quality assessment and review II, June 7, 2024, 09:00-10:30

A novel Asymmetry Resolution Mechanism for solving asymmetries in International Trade in Services: methods and practices.

Marios Papaspyrou[1], Georgios Papadopoulos[1]

[1]European Commission (Eurostat), , Luxembourg

Asymmetries in trade statistics and in particular International Trade in Services statistics (ITSS) data is a long-standing phenomenon, reducing the credibility of official statistics and thus hindering the user from making good use of them. Asymmetries occur when there are differences in the mirror flows between countries. The DGINS 2019 Bratislava conclusions underlined the importance of work to reduce asymmetries: "…implemented via more coordination and cooperation between domains, between countries, and between the ESS and ESCB". Under this mandate, Eurostat introduced the Asymmetries Resolution Mechanism for ITSS data (ITSS-ARM) in April 2022.

The core aim of the ITSS-ARM is to resolve major cases of intra-EU bilateral asymmetries, thereby improving the ITSS data quality. A balanced approach was adopted, aiming at participation of all EU Member States. Eurostat developed a data-driven scoreboard methodology to prioritize the most prominent asymmetry cases, by building a score of "importance". The score takes into account both the magnitude of the asymmetry as well as the trade volume involved. Therefore, the selected cases are the most significant from the point of view of the member state, in the sense that a country will have a greater incentive to dedicate resources in resolving the largest asymmetries with their larger trading partners. An IT tool was implemented to automate the process. The selected priority cases are followed up in trilateral meetings (Eurostat and the two member states involved) and the progress is regularly reported to relevant expert groups.

During this process, several findings of a broader interest (recommendations, best practices, practical guidelines) came to light. For the benefit of all ITSS data compilers, this know-how is collected in a technical document, while selected best practices are presented in dedicated workshops. Thus, the mechanism should help reinforce the "cooperative" culture among data compilers in the member states, as well as to resolve (or prevent) bilateral asymmetries. The improvements achieved on the published data from the first year of operation of the ARM show a successful mechanism with positive impact on reducing observed asymmetries.

Unemployment – Monthly data – Sources and Methods

Boyan Genev[1], Nevena Cholakova[1]

[1]Estat, , Luxembourg

The Monthly Unemployment Rate (MUR) has proven to be both relevant and reliable. It is frequently commented on by the wide public, the media as well as by analytical experts. From its onset, the unemployment data presented by Eurostat has followed the definitions of the International Labour Organisation (ILO), applied in a time consistent and harmonized way by the EU Labour Force Survey (EU-LFS). The higher frequency of MUR data (compared with annual and quarterly data) introduces additional techniques, when combined with additional sources that deserve special attention.

Different strategies are in place at the level of EU Member States to deal with these challenges, linked to the need of a systematic overview of the quality of the statistics. In some cases, only national definitions are followed, leaving some room for misinterpretation or doubt in relation to quality. Nine EU Member States do not compile MUR statistics according to the ILO standards, while an additional five countries do so only after using auxiliary information from the national employment authorities. While Eurostat takes full responsibility for the compilation and communication of ILO unemployment for the former group of countries, for the latter group not all methodological elements are transparent. Combining data sources requires indeed addressing the quality impact on the compiled statistics. For the rest of the countries, additional methods of smoothening are often needed. Finally, seasonal adjustment is applied at the end of the time series compilation. Here again, qualitative descriptions of the statistical operations are to a big extent missing at country level.

This presentation explains the introduction of the legal requirement for the provision of information on sources and methods for MUR data and describes the process of fulfilling the requirement. After specifying the qualitative framework in the Commission Implementing Regulation (EU) 2019/2241, Eurostat combined all the relevant elements in a structured and understandable way.

The work done in this area supports the hypothesis of a high comparability among countries and underlines the statistical relevance at EU level. Many statistical elements are now clearer to the wide public and to the advanced statisticians. The final document enables analysis of time series methods and enhances the qualitative comparisons among them. The paper also outlines the quality challenges ahead and sketches out the possible interlinks between the qualitative descriptions already published and the description of time series features subject to future investigations.

Enhancing Quality Management in Mortality Surveillance: A Comprehensive Audit and Evaluation

Dr Júlia Martinho[1], Afonso Moreira[1], Ana Paula Soares[1], Daniela Freitas[1], Liliana Bernardo[1], Sofia Pimenta[1], Pedro Pinto Leite[1]

[1]Direção de Serviços de Informação e Análise da Direção-Geral da Saúde (Directorate of Information and Analysis of the Directorate-General of Health), Lisboa, Portugal

Mortality surveillance plays a pivotal role in identifying disease trends and evaluating risk factors, enabling early detection and targeted preventive interventions. The data derived from observed causes of death, collected through death certificates, empower policymakers to effectively allocate resources, prioritise health programmes and assess the impact of interventions. Furthermore, this mortality information contributes to improving official vital statistics. In Portugal, the Death Certificates Information System (SICO), managed by the Directorate-General of Health (DGS), fosters coordination among entities involved in death certification and reporting. SICO promotes data accuracy, accessibility, and citizen privacy, aligning with the World Health Organization's (WHO) strategies for the enhancement of vital statistics and health policy implementation.

An in-depth internal audit of 408 death certificates from SICO was conducted, focusing on compliance with essential criteria crucial for accurate documentation. Adopting a quality management systems approach, the audit aimed to assess the quality of the death certification process and integrated risk management practices to identify potential errors or inconsistencies. Criteria included adherence to established WHO and United Nation standards, encompassing the completeness of the health service user number, Part I and Part II sections, accurate recording of the basic cause of death, logical sequencing of the causes of death, singular cause per line, specification of cancer type and location, detailing relevant microorganisms in infectious causes, specifying heart failure etiology, recording the duration between disease onset and demise, maintaining consistent time intervals between different sections, and avoidance of abbreviations.

Conformity proportions were computed to gauge adherence to the aforementioned criteria. The audit results revealed both exemplary aspects and areas requiring improvement within the death certificate completion process. Identified challenges included inadequate completion of Part II, disorganised causes of death, unspecified cancer types and microorganisms, lack of information regarding heart failure etiology, and incomplete disease onset information. Additionally, when autopsies were waived by the public prosecutor's office, issues arose concerning unknown causes of death without access to supplementary information.

This comprehensive audit highlights the importance of stringent quality management systems for death certificates, crucial for accurate cause-of-death statistics. These findings underscore the need for refining existing protocols, prioritising continuous quality improvement initiatives, and implementing robust risk management strategies to enhance accuracy, reliability, and completeness in death certificate documentation. Such an approach ensures better data integrity, facilitating more informed decision-making processes.

Improving the quality of balance of payments statistics via granular bilateral analysis - Trilateral comparison: Austria, Italy and Spain

Jorge Diz Dias[1], Nadia Accoto, Erza Aruqaj, Jorge Diz Dias, María García del Riego, Milena Matteo

[1]European Central Bank, , Italy

The statistical recording of cross border transactions and positions between two countries should match like a mirror image. However, the presence of sizeable differences between bilateral external statistics has been a growing concern for compilers and users. The bilateral differences adversely affect the quality of official external statistics and their usability as a basis for sound policy analysis and decision making.

Our paper contributes to the existing literature and institutional initiatives on improving the quality of external statistics by presenting a novel framework for improving the quality of external statistics via an in-depth comparison and reconciliation of external statistics. First, while existing studies and initiatives mostly focus on bilateral data comparisons for a specific subset of the external statistics (e.g., trade in goods; trade in services; foreign direct investment), our exercise covers all balance of payments statistical categories. This provides a comprehensive approach and overview of the bilateral data quality in external statistics. To do so, we present a way to prioritise the in-depth analysis and reconciliation for the categories that contribute most to the overall bilateral differences. Second, our framework proposes an institutional framework involving a trilateral exercise, where three countries pairs actively engage and compare their data. This helps detecting systematic patterns on the bilateral data, as one country's bilateral data is analysed in detail against the information provided by two other countries. The trilateral setting provides also valuable learning experiences for the participants thanks to information sharing, reconciliation discussions and network building.

Finally, our exercise suggests a practical way on how countries can organise such data comparison exercises, emphasising the importance of inter-institutional cooperation and microdata sharing.

Sharing microdata in a secure fashion is key to solving bilateral differences due to the possibility for precise follow-up investigations.

In the paper we show an application of this framework to the country trio Austria, Italy and Spain. We summarise the main features of their external statistics sources and methods, investigate the main reasons for the largest bilateral differences, and present the reconciliation results obtained from the investigations.

Regulative, structural, and detailed improvements of national accounts data quality

Maria Dimitriadou[1], Ms Christine GERSTBERGER[1], Ms Martina PATONE[1], Mr Enrico INFANTE[1], Ms Malgorzata SZCZESNA[1]

[1]Eurostat, Luxembourg, Luxembourg

This paper presents three initiatives on how data quality for national and regional accounts is being improved in recent years. The paper firstly shows how the quality reporting exercise brings benefit to the effectiveness of the national accounts' regulation; secondly, discusses how the coordination at European level for structural benchmark revision improves the quality of data; and thirdly, presents a method to define thresholds for reporting metadata on revisions in regional accounts as an example of quality improvement at the detailed level.

Regulation (EU) No 549/2013 establishes the European System of Accounts 2010 (ESA 2010) which is the internationally compatible EU accounting framework for a systematic and detailed description of an economy. It is the source for a multitude of key economic indicators, including gross domestic product (GDP), that are used for policy making and economic analysis. The quality of ESA 2010 data transmitted by the Member States is systematically monitored during the annual quality reporting exercise set out in Regulation 2016/2304. The paper outlines the role of the quality reporting exercise in enhancing the quality and reliability of national accounts data, together with the various projects initiated during the last five years.

Benchmark revisions, carried out at least once every five years, are vital for incorporating new data sources and aligning with the latest international statistical methodologies. They improve the quality of data and ensure consistency within national accounts and across Member States. The paper outlines the coordinated efforts for the 2024 coordinated benchmark revision, the coordinated communication by the European Statistical System (ESS), the benchmark revision requirements and main expected changes.

A robust methodology is provided to effectively identify revisions that significantly deviate from typical values in regional accounts. A dynamic, percentile-based thresholding method is used, adjusting thresholds according to the standard error (SE) of historical revisions for three-unit measures (national currency, persons, hours worked) across all NUTS 2 or 3 regions within each country. This approach mitigates disparities emerging from the different region sizes and measurement scales. The thresholds are further tuned so that if the SE is high, indicating less reliability, the threshold lowers to flag more revisions; conversely a low SE raises the threshold, targeting most extreme revisions. This ensures that the thresholds are pertinent and applicable across diverse contexts.

Session 36 - Confidentiality and data protection II, June 7, 2024, 11:00-12:30

Implementing a Data Inventory Catalogue to Enhance Data Governance and GDPR Compliance in the Central Statistics Office (CSO), Ireland

Mrs Miriam O Reilly[1]

[1]Central Statistics Office Ireland, Cork, Ireland

The Central Statistics Office (CSO), Ireland, is implementing a data inventory catalogue to effectively manage its data holdings, enhance data governance, and streamline business processes. This centralised metadata repository provides a comprehensive overview of data assets, including data inputs, outputs, programmes used, and storage locations. The inventory also plays a crucial role in complying with the General Data Protection Regulation (GDPR) by facilitating data retention and ensuring data privacy adherence. Each piece of data in the inventory is classified according to a standardised scheme, aligning data storage and processing practices with GDPR requirements and other legal obligations.

The initial implementation of the data inventory has already yielded significant benefits, including increased transparency into data flows, enhanced data accessibility, and improved overall data quality.

This paper will document the progress made to date, the benefits achieved, and the challenges encountered. It will also highlight our future plans for the application including our aim of further integrating it with the CSO's data retention policy to assist in improved compliance with our GDPR requirements.

In conclusion, the data inventory catalogue serves as an indispensable tool for cross-functional collaboration, data governance, data discoverability, and GDPR compliance. It streamlines data usage and accessibility, ultimately contributing to enhanced data quality and overall organisational efficiency.

Asking about private and sensitive attributes using item count techniques - methodological and theoretical challenges

Dr. Barbara Kowalczyk[1], Dr. Robert Wieczorkowski[2]

[1]SGH Warsaw School of Economics, Warsaw, Poland, [2]Statistics Poland, Warsaw, Poland

Quality data about private, stigmatizing, socially unaccepted or illegal features and attributes is extremely difficult to obtain via traditional questionnaires and surveys. Some special statistical techniques have been developed to meet these challenges, including randomized response techniques, non-randomized response techniques and item count techniques. Among applied researchers item count techniques seem to gain the greatest interest and have been applied quite successfully in individual surveys. But still methodology and theory of item count methods is not fully developed. In particular, many questions arise as far as the type of the control variable is concerned, degree of privacy protection and the method of estimation. In the article we address all these issues. We also put a special emphasis on the method of estimation. For this purpose we conduct a comprehensive Monte Carlo simulation study in which we distinguish between theoretical item count models and their real life counterparts.

What Do We Mean by “For Statistical Purposes Only”?

Ms Danielle Neiman[1], Phd John Eltinge[1], Phd Michael Hawes[1], Steven Klement[1], Paul Marck[1], Dr Sallie Keller[1]

[1]US Census Bureau, Washington, United States

Data subjects are often told their information will be used for “statistical purposes,” and statistical agencies are legally required to use these data “for statistical purposes only,” but what does this actually mean? The information collected should be used and accessed in a way which maximizes the usefulness of the information but also maintains privacy, confidentiality and use restrictions. In today’s data-driven world, statistics is a far reaching and expansive discipline, actively used across virtually all scientific fields, and extensively leveraged by financial firms, social media platforms, and public agencies, with myriad daily implications, both large and small, for the average person. With statistics, as a concept, being so broad a discipline, one might expect that the term “statistical purposes” (intuitively, those actions taken in pursuit of the generation, use, or interpretation of statistics), would be similarly expansive. To the contrary, “statistical purposes” has become a term of art in law, ethics, and practice that defines and constrains the legal, ethical, and appropriate uses of certain data. This paper will explore how the term “statistical purposes” has been variously defined in United States Federal laws and regulations, and how it has been interpreted in practice by the U.S. federal statistical system. We will discuss how the legal and ethical guardrails of “statistical purposes” align with the core objectives and mission of a statistical agency, and how some of the inherent ambiguities of what may constitute a statistical or non-statistical purpose get at the heart of some of the most vexing challenges facing statistical agencies today. Finally, we will touch on research into how other countries’ statistical agencies define the term and compare to what U.S. agencies are currently using.

The new wave of privacy concerns and its impact on official statistics

Ms Yolanda Gómez Menchón[1], Ana Canovas

[1]INE, , Spain

One of the basic principles of all statistical offices is to guarantee data confidentiality. The first horizontal statistical law in the EU from 1990 was related with the statistical confidentiality, recognised as the main statistical principle in United Nations in 1994 and in the EU in 1997, then, before the existence of the first Data Protection Directive in 1995. The Treaty of Amsterdam “constitutionalize” statistical confidentiality principle in 1999 and it was implemented in Regulation 223/2009 on European Statistics and further elaborated in the European Statistics Code of Practice. In addition, the principle of statistical confidentiality works together with the other statistical principles of impartiality, reliability, objectivity, scientific independence, cost-effectiveness and non- imposition of excessive burdens on economic operators. Therefore, since the very beginning, statisticians have been treating individual data from natural and legal persons with the highest degree of protection in the EU. The question now is, what has changed in the last eight years?

A new wave of privacy concerns arrived as a consequence of the challenges coming from the rapid technological developments and globalisation. This end in a new regulation in the European Union aimed at harmonising personal data protection rules in the Member States (GDPR). Citizens and governments are nowadays more aware and sensitive on this issue. The European Data Protection Supervisor and national and European lawyers interpret the law and activities in a restrictive way and, at the end, official statistics are made pay for the sins of others.

In this paper we address this issue with the aim to make evident how all GDPR data protection principles are already covered by the EU statistical principles and what could we do to show this reality increasing trust in official statistics. We will also analyse our principles in the light of the use of innovative methods for statistical production, is it our code enough?

The Principle of Minimization of Personal Data under the GDPR in Official Statistics in the European Union

Mr Apostolos Kasapis[1], Artemis - Dimitra Kritikou[1]

[1]Hellenic Statistical Authority (ELSTAT), Piraeus - Athens, Greece

This paper examines the interplay between official statistical production in the EU and the General Data Protection Regulation (GDPR), aiming to pinpoint particular effects of the GDPR on official statistics. It discusses the legal premise of conformity with GDPR provisions and identifies the legal basis for processing personal data for official statistical purposes, emphasizing the particular conditions set forth by the data minimization principle and its implementation in statistical production.

The paper delves into the concept of data minimization, and its specific requirements, setting as a basis the exploration of the principles of necessity and proportionality as integral components of the GDPR framework. It analyzes how these principles guide the collection and processing of personal data, ensuring that only data essential for the specific statistical purposes are processed and that such processing does not excessively infringe upon the data subject's rights and freedoms. By scrutinizing these principles, the paper aims to identify their role in achieving a balance between the need for comprehensive data and the necessity of upholding privacy and data protection standards. The paper concludes that adherence to data minimization does not imply the use of non-personal or anonymized data but rather the use of personal data to the extent necessary for the data controller to fulfill their legal purposes in a lawful manner. It asserts the need to articulate the relevance and necessity in official statistics of data emerging from the particular methodological descriptions of each statistical product, ensuring compliance with the GDPR's data minimization principle.

Session 37 - Machine learning II, June 7, 2024, 11:00-12:30

Applying Machine Learning to Longitudinal Administrative Data: A Case Study in Education

Dr Fabrizio De Fausti[1], PhD Romina Filippini[1], PhD Simona Toti[1], PhD Marco Di Zio[1]

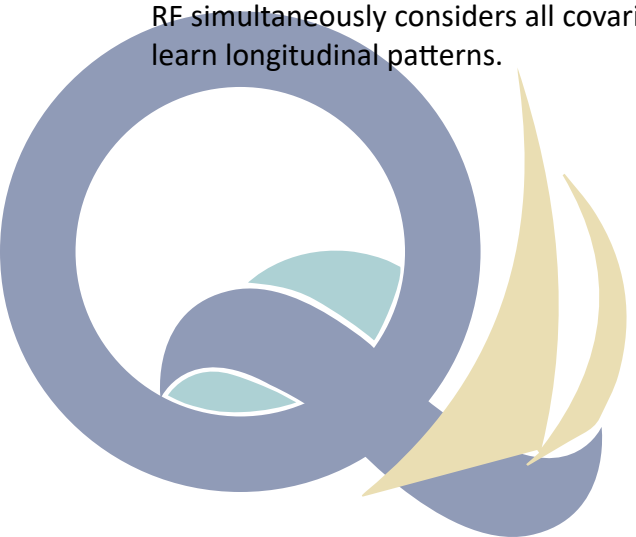
[1]Istat, , Italy

The increasing availability of administrative sources has significantly changed the official statistical production system, moving towards a register based approach. This involves the annual construction of updated registers, achieved through the integration of various types of data sources, with a particular emphasis on administrative sources. Advantages are expected in terms of reduction of costs and of response burden, and the possibility of having micro data at disaggregated levels enhancing the production of detailed statistics. On the other hand, typical issues with administrative sources are emerging, such as delays in data availability and coverage problems. Although administrative data generally completely enumerate units of the population they refer to, such a population can be a subset of the statistical target population.

In this context, the production of a complete and coherent dataset becomes a crucial activity, making necessary the application of various procedures for predicting delayed data and estimating missing data. A significant advantage lies in the fact that, once the data time lag is overcome, updated administrative information becomes available, providing the opportunity for an evaluation and refinement of the procedures. Among other, an important topic for which many information are available from administrative sources is Education: information on students, school attendance and educational level are available from the Ministry of Education since 2011. The official ALE estimates for all the Italian resident population are produced by the Italian National Institute of Statistics (Istat) adopting a mass imputation approach that integrates administrative, survey, and 2011 census data (DiZio et al.,2019). Variable ALE is produced at a micro level, using log-linear models, with the goal of reproducing the ALE distribution obtained from the Permanent Census Survey, properly weighted.

Estimates result to be accurate, however, they lack longitudinal coherence. To fully leverage the opportunities presented by longitudinal administrative sources, ML technique are applied for the ALE prediction task with the goal to assess the potential of ML technique in making accurate predictions and improving micro-level estimation accuracy while addressing longitudinal incoherences.

In this paper, we experiment a prediction of the Attained Level of Education (ALE) using longitudinal administrative information with the use of ML techniques. A comparison is made between Random-Forest (RF) and Long-Short-Term Memory (LSTM) technique, a type of recurrent neural network (RNN) architecture designed to capture and learn long-term dependencies in sequential data. While RF simultaneously considers all covariates, LSTM uses a neural network architecture to analyze and learn longitudinal patterns.



Data mining techniques on the administrative data system to enhance the accuracy of the population census counts

Antonella Bernardini[1], Nicoletta Cibella[1], Giampaolo De Matteis[1], Gerardo Gallo[1], Mr Antonio Laureti Palma[1], Fabrizio Solari[1]

[1]ISTAT, ,

Since 2020, the National Institute of Statistics (ISTAT) has been producing fully register-based population size estimates integrating the population register with 'Signs of Life' (SoL) derived from administrative data. The recent development of data mining applications, as well as the increasing availability of large amounts of data, suggests new methodologies for estimating and assessing population counts [1]. In this study, SoL are used to implement first a supervised classification strategy to distinguish between usual resident and not usual resident population in Italy and after an unsupervised model to assess the real usual place of residence.

In the first step, no specific location was assumed, although location variables, as place of work or tax domicile, were used as SoL identifiers. Supervised training and testing data sets were built using the 2021 census sample survey data. A machine learning model is used for classification based on the Support Vector Machine (SVM). Estimator overfitting and underfitting were checked using the cross- validation and the validation curve, while the size of the training set was controlled by the learning curve. Quality assessment of ML results was performed, as well as an evaluation of the importance of the 2021 census sample survey data as the training set. The assessment shows the discriminant role played by a population register in a register-based population count estimation, distinguishing the Italian situation from countries where population registers are not available.

In the second step, in order to assess the real usual place of residence, we use utility consumption, electricity and gas, as data sources to identify the monthly consumption patterns associated with each point of delivery. Through a cluster analysis of the consumption patterns in association with the information on households included in statistical registers it is possible to assess the usual place of residence, i.e. where a household, or part of it, actually lives, reducing the possible misplacement errors. Furthermore, under certain conditions of unicity between services provided and households served, through each home's energy consumption model, it is possible to estimate the number of people who live there. This estimate contributes to improve the quality of census statistics, especially on Italian household characteristics, and to improve the overall evaluation of populations affected by misplacement errors.

Enhancing data quality controls on money market transactional data: a comparative study of anomaly detection techniques

Gianluca Boscariol[1], João Oliveira Ferreira[1]

[1]European Central Bank, Frankfurt Am Main, Germany

This paper presents four machine learning anomaly detection techniques, regularly applied at the ECB to a transaction-level dataset, to improve data quality, providing results on the practical effectiveness of these techniques to the Money Market Statistical Reporting (MMSR) dataset. The anomaly detection techniques are one component of the thorough daily data quality management (DQM) applied to the MMSR dataset, among others actively searching for anomalies on the basis of both traditional quality checks and innovative techniques. The DQM involves verifying selected transactions with reporting entities, keeping track of all feedback in a structured form, and requiring resubmission of corrected or missing data as needed. The structure of the DQM process is essential to analyse the effectiveness of the anomaly detection techniques used, contrasting advanced machine learning tools such as isolation forests, HDBSCAN, and XGBoost against a more traditional two-step general least squares regression. The paper concludes that the latter two techniques offer significantly higher effectiveness than the first two when applied to the MMSR dataset for the latest 4 years, offering guidance to discriminate among innovative techniques applied to official statistics data quality.

Population Trajectories Survey Methodology: Improving Coverage Error by Clustering of Freguesias and Dwelling Segmentation using Census Data in Portugal

Dr. Afshin Ashofteh[1], Dr. João Lopes[2], Dr. Pedro Campos[3]

[1]Nova Information Management School, Nova University Lisbon and Statistics Portugal (INE), Lisbon, Portugal, [2]Statistics Portugal (INE), Lisbon, Portugal, [3]University of Porto and Statistics Portugal (INE), Lisbon, Portugal

Official statistics on living conditions and access to goods are crucial in monitoring inequality based on racial and ethnic origin and population decline as a critical contemporary demography challenge. However, data collection for minorities and sensitive information can be restricted. In this study, an independent sampling survey was designed, and machine learning algorithms were used to overcome these issues. We used clustering methods and Census 2021 data to identify essential variables and homogenous freguesias to distribute the sample size. In addition, dwellings were segmented, and the clusters were analyzed and discussed to minimize the non-response and coverage errors. The proposed methodology provides a comprehensive final survey with proper target population coverage.

Session 38 - Microdata level perspective, June 7, 2024, 11:00-12:30

Organic farming in Italy: comparison and integration among sources for improving data consistency

Dr Roberto Gismondi[1], Chiara Gnesi[1], Pietro Nurzia[1]

[1]Italian National Statistical Institute, , Italy

Organic farming is an important feature for agricultural holdings in the European Union. It indicates the propensity of farmers to apply production techniques that guarantee sustainable agriculture and preserve food safety for consumers. A sustainable food system is at the heart of the European Green Deal. The European Commission has set a target of at least 25% of the EU's agricultural land under organic farming by 2030. The area used for organic production in the EU continues to increase. It expanded from 14.7 million hectares in 2020 to 15.9 million in 2021, the 9.9% of the utilized agricultural area.

In Italy there are three main data sources regarding organic farming. The first one derives from administrative data managed by the Ministry of Agriculture. Control offices check whether the agricultural lands are managed using organic practices or not. Through these data, the Ministry is compliant with the ESS Agreement on organic production statistics ESSC 2020/42/6/EN. The second source derives from the Integrated Farm Statistics Regulation IFS (EU) 2018/1091, which ruled the last agriculture census. The census 2020 picked up data on organic farming (lands and animals) directly from farms. The third source is the Farm Accountancy Data Network (FADN), established by the Regulation (EC) 1217/2009. The FADN collects data on organic production: yield per hectare and quantity of milk per cow. EUROSTAT and DG AGRI request consistency among administrative data and those ruled by the IFS and the FADN Regulations.

As regards 2020, all three data sources were available. Comparisons were carried out both at the macro and micro level, in order to assess to which extent discrepancies may be due to different microdata. The census estimated that the relative weight of organic surfaces in 2020 was equal to 6.7%, while the ministry estimated 5.6%. Though record linkage, the administrative microdata were compared with the correspondent census microdata, as regards the feature "the farm is organic or not". The 94.5% of linked farms had the same feature (concordant farms). On average, discordant farms are larger (as regards surfaces and labor force) and they are mainly located in the South.

Furthermore, organic surfaces and organic cows were multiplied by the correspondent yield coefficients supplied by FADN, with the purpose of estimating organic production (crops and milk). This last attempt is very useful in view of the entry into force of the SAIO Regulation (EU) 2022/2379, which requests for organic production as well.

Estimating Non-Regular Earnings for Small Firms: A Micro-Data Based Approach

Mr Gergely Attila Kiss[1], Beáta Horváth[1], István Balázs[1]

[1]Hungarian Central Statistical Office, , Hungary

This paper is about a new method for estimating non-regular wage mass for organizations with less than 5 employees. Our method is based on an outlier filtering procedure on a job level panel data. The original data is administrative and comes from the National Tax Authority. The panel data is created by connecting monthly observations from cross sectional samples by job id. This job id based connection results in a dataset where we only observe the set of occupations where there were no change in either personnel or key factors regarding the job, such as type of relationship, working hours and FEOR (Hungarian version of ISCO). The current method is about projecting the behaviour of larger organizations to the small ones on aggregated level. Thus our method provides an opportunity for the published statistic to attain more details.

The new method will provide aggregated numbers based on micro data. As such, it makes all kind of disaggregation attainable that were not possible before. Furthermore, the promising results give an opportunity to generalize the method to the population of all organizations. This generalization would result in a consistent approach of estimating non-regular earnings on national level.

The outlier filtering is carried out in two stages. In the first stage, a mechanical approach flags all candidates of outliers in each time series from the panel sample. This is facilitated by several machine learning techniques combined with some rules of thumb approaches on defining outliers. In the second stage additional rules are introduced to filter out candidates based on domain expert knowledge that can grab the economic insight behind events where massive pay outs happened. A good example is March of 2021, during the COVID pandemic those who worked in healthcare received a retroactive raise, that would seem as a massive non-regular earning pay out in the aggregated level but should not be accounted as such.

Our results are promising as the conducted heterogeneity test and time series comparisons show little to no difference in several sample characteristics. While comparing time series we set the requirements to be matching dynamics in total and by the previously published strata. The minimal requirements for heterogeneity test were to reproduce the same ratios to total in the cross-sectional and our panel sample. In case of differences we decided to do sample correcting weighting to make the panel aggregation and the previous statistics match in macro level.

Design and evaluation of a micro-integration strategy for pay inequalities

DR Stefano Stefano[1], Dr Alessandro Martini[1], Dr Davide Di Cecco[2]

[1]Istat, , Italy, [2]Sapienza University, Roma, Italy

Addressing potential discrimination in earnings is a priority for policies at both EU and national levels. The Gender Pay Gap (GPG) is the most significant wage differential, and several statistics, including the unadjusted GPG, are used to monitor imbalances in earnings between men and women.

To analyse the GPG and wage differentials, in accordance with Regulation (EC) No 530/1999, Eurostat uses microdata from the Structure of Earnings Survey (SES), which is a large enterprises survey aiming to provide accurate and comparable data on earnings across countries and over time. The survey examines the relationship between employee characteristics (such as sex, age, occupation, length of service, and highest educational level attained) and employer characteristics (such as economic activity, size, and location) with the level of pay. SES microdata are available every four years for reference years 2002, 2006, 2010, 2014, and 2018.

This paper proposes a strategy for integrating several administrative sources with Labour Force Survey microdata, taking into account the sample design. The aim is to provide annual datasets of microdata, instead of every 4 years, without additional statistical burden, while also including information on individuals not covered by the SES survey (e.g., people employed in enterprises with fewer than 10 employees).

The process consists of two phases, each with several steps:

1. the creation of a representative sample, defined as: a) union of the different FOL waves according to a criterion of proximity to each October (survey reference month), retaining information on in- scope individuals (employees only); b) intersection of the union sample with the reference universe; c) definition of a stratified random sample of employees with allocation guided by the target variable (monthly salary); harmonization of the probabilities of inclusion of the joint sample

2. record linkage with administrative data (social security data and Comunicazioni Obbligatorie from Italian Ministry of Labour and Social Policies) in order to verify and validate FOL information and to integrate the missing variables.

Finally, sampling and non-sampling errors are evaluated to ensuring data quality.

Our proposal provides a zero-burden solution to offer valuable information for research and policy purposes. It also supports official statistics in exploring pay gaps, particularly in longitudinal series, and covers the micro firms domain. Additionally, this methodological proposal has the potential to be extended to self-employed domains.

Using microsimulation to improve the quality of the official Austrian population projection

Hannah Diethard[1], Philip Slepecki[1], Dr Martin Spielauer[2], Hannah Diethard[1]

[1]Statistics Austria, Vienna, Austria, [2]Austrian Institute of Economic Research, Vienna, Austria

Population projections in official statistics are generally produced using the cohort component method. It is computationally simple, does not require a broad range of input data, and is well-established among demographers. However, it cannot account for complex and dynamic demographic processes, model interactions, or produce detailed results for individual-level characteristics. To overcome these limitations, Statistics Austria has implemented a microsimulation model for its official population projection that builds on the characteristics of individuals instead of entire cohorts and allows for the simulation of realistic life-courses. While most dynamic microsimulations in the social sciences include some modelling of demographic processes and several national statistical offices have microsimulation models in their “toolbox”, Statistics Austria is the first European statistical office using microsimulation to produce its official population projection. To mitigate the effect of this methodological break on the comparability of projection results over time, we start by replicating the results of the cohort component method in a microsimulation and gradually introduce new model features. As a first step, we incorporate a model of international migration which explicitly accounts for the relationship between the emigration risk and the duration of residence. In addition, the place of birth is included as an individual-level attribute in the form of detailed country clusters. As an ex-post validation, we compare the results of the microsimulation with the cohort component method and the observed data for Austria for the years 2012 to 2021.

We show that the microsimulation projection matches observed emigration patterns more closely, especially following the increase in immigration to Austria in 2015 and 2016. The model will be gradually refined and it can be extended with additional modules for education, employment, health and other socioeconomic characteristics.

Session 39 - MNO data, June 7, 2024, 11:00-12:30

Enhancing Official Statistics with New Data Sources – Methodological Developments for Integrating Mobile Network Operator (MNO) Data with non-MNO Data

Ms Gloria Deetjen[1], Maurice Brandt[1]

[1]Destatis, , Germany

The use of new data sources in official statistics offers not only the opportunity for new analysis but also to improve and to complement regular statistics. Mobile Network Operator (MNO) data are amongst the most favourable types of new data sources for this purpose because of their high relevance in terms of time and location. The combination of the so-called MNO data with other traditional or new types of data are a great way to enrich existing statistics as various studies have shown. However, the integration of MNO data into regular official statistical production requires profound solutions and innovative methodologies because of potential errors and quality issues which can mostly arise from either the data itself (e.g. data configuration) or from the way the data is used (e.g. direct or auxiliary). Moreover, it is the role and responsibility of statistical offices to maintain high quality outputs in a world of new digital data.

The ESSnet Trusted Smart Statistics project on “methodological developments based on new data sources” aims to address these needs by proposing a reference frame for methods regarding the combination of MNO and non-MNO data for official statistical production. These new methodologies alongside with guidelines are crucial for a successful linkage of MNO data with non-MNO data in the European Statistical System. Therefore, intermediate results on a methodological proposal with respect to total error frameworks, data configuration, and the role of MNO data are presented. In addition, first results on a landscaping analysis of the most promising non-MNO data sources to be integrated with MNO data show the great application potential which new digital data contain.

Further, intermediate results on a proof-of-concept of an ad-hoc survey to improve MNO data are introduced. The intermediate results of all three aspects provide valuable insights regarding quality and the use of new data sources.

Enhancing the Quality of Mobile Network Operator Data with a traditional survey with the right questions

Dr Alexander Kowarik[1]

[1]Statistics Austria, Vienna, Austria

In the evolving landscape of new data sources, the use of Mobile Network Operator (MNO) data has become increasingly significant for official statistics. However, the heterogeneity in mobile phone usage throughout different population groups causes biases in MNO data and therefore creates substantial challenges. This work, conducted by Statistics Austria in collaboration with Destatis, ISTAT and Statistics Sweden aims to address these challenges by developing a comprehensive questionnaire designed to measure and correct biases in MNO data.

Our survey aims to cover areas such as service provider details, number and types of devices, foreign SIM cards, usage patterns (stationary, on-the-go, travel behavior, nighttime usage), and contract types. This targeted survey should enable a deeper understanding of the coverage of specific mobile phone network operators and the behavior of their users. The developed model questionnaire can either be used in a stand-alone survey or be integrated into existing surveys. The aim is to provide a robust method for improving the quality of MNO data in official statistics.

The paper presents the questionnaire and the motivations behind the different topics covered by it. Subsequently an outlook on how the results of such a survey could be used for estimation is provided.

This paper is based on work of WP4 of the ESSNet on “Trusted Smart Statistics – Methodological Development Based On New Data Sources”. It is a research project to develop methods for the integration of MNO and non-MNO data (e.g., survey data, census data, other big data sources etc.). Work package (WP) 4 task is to develop a Proof of concept of an ad-hoc survey to improve MNO data.

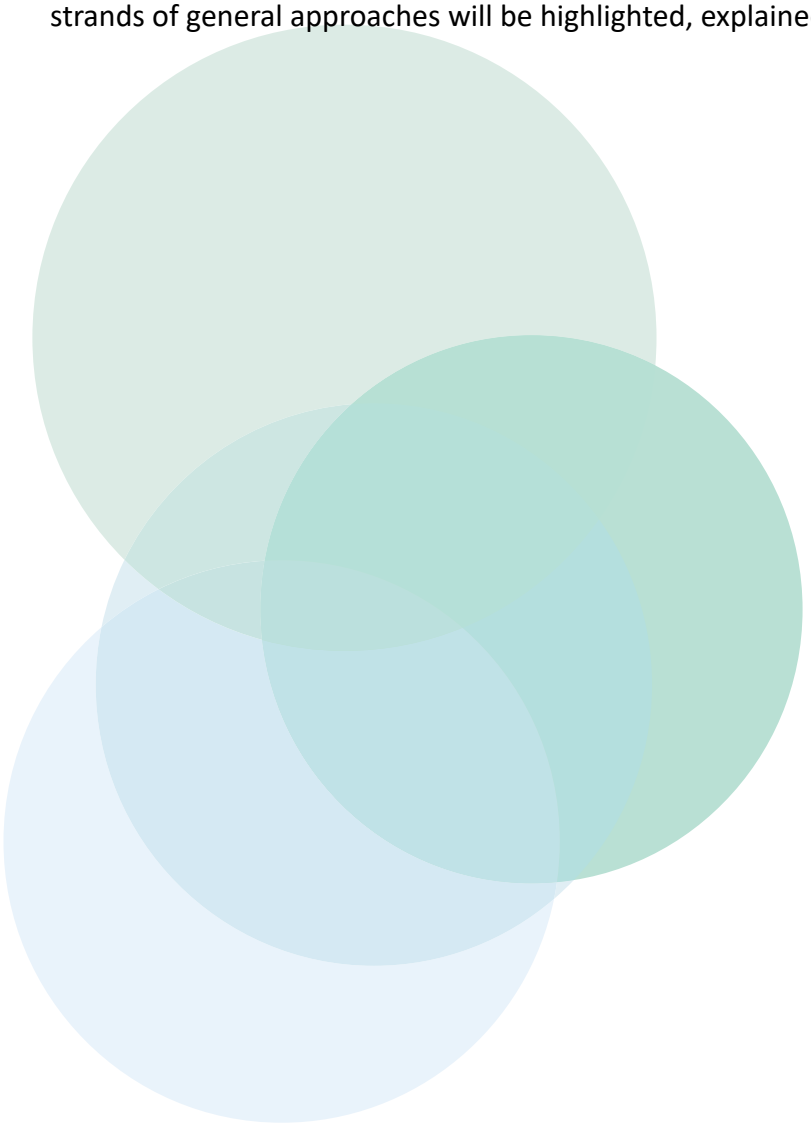
Methodologies for integrating MNO and non-MNO data

Dr Li-Chun Zhang[1]

[1]Statistics Norway & Univ. Southampton, Oslo, Norway

Mobile network operator (MNO) data comprise of call detail records as well as high-frequency passively-collected signalling position data. Combining MNO and non-MNO data holds obvious promises for many statistics of interest, possibly with a much more detailed temporal-spatial resolution than what is otherwise possible. In reality, however, there exist many obstacles to achieving the quality that is necessary of official statistics. In particular, the National Statistical Office (NSO) typically does not have access to the micro-level MNO data that can be linked to population registers or other relevant sources, whereas the MNO aggregates may be subject to measurement errors, population domain misclassifications, device duplication noises, user ambiguity and target population coverage errors. Despite the great attention it has received in the recent years, few official statistics based on MNO data exist today.

This ongoing ESSnet project aims to guide and facilitate the integration of MNO data with other non-MNO data to produce regular official statistics. Considerable methodological development is needed in this respect. In this talk, we will present an overview of the relevant progress, ranging across super-population modelling, randomisation and quasi-randomisation perspectives. Several selected strands of general approaches will be highlighted, explained and illustrated.



Session 40 - Special Session: Best Practices in Quality Management in the Southern European Neighborhood Policy countries, June 7, 2024, 11:00-12:30

Challenges in assessing and assuring the quality of new data sources for population and housing census 2025

Dhafer Al-shawawreh[1]

[1]Department Of Statistics, amman/aljbyha, Jordan

The General Population and Housing Census 2025 is the most important statistical product that serves the public and private sectors in precisely knowing the current situation and planning for the future. The General Population and Housing Census 2025 is designed to be a hybrid census which inputs consist of administrative records and data collected from households and establishments.

Using administrative records as input will provide many benefits, such as reducing time, effort,and cost, more accuracy, and reliability,but at the same time it presents many challenges for the quality check of the census, which makes it necessary to ensure the quality of the administrative record before using it as input in the population census.

The stages of the quality check of the administrative record are divided into four stages: The source stage (understanding and evaluating the source), The data stage (receiving the actual data and evaluating it's quality), The processing stage (processing administrative data and using it in the census), and The output stage (assessing the quality of the census that uses administrative data).

Both of the source stage and the data stage represent the quality of administrative record inputs Before accepting them as input to the census.

We applied the Statistical Code of Practice for the European (ENP-south) to evaluate the quality of each stage separately for the new administrative record. Many challenges appeared in evaluating the source stage, the data stage, and the processing stage. However.

In this paper, we will discuss the general plan for examining the quality of the record as a good input for the census then revealing the problems and challenges that faced the quality team, measuring the problem, and trying to solve it in the processing stage to accept it or merge it with a new record from another source so that it has a value commensurate with the conditions of the quality of the census inputs, or to accept a part of it As inputs for the census or for the purposes of confirming the quality of the census or rejecting the record

Quality in Official Statistics (Best practices in Quality Management in the Southern European Neiborhood policy countries (ENP-South)) Experience of Lebanon

Maria Nalbandian[1]

[1]1, , Lebanon

1. CAS mission according to the law
   - Collect, process, produce and disseminate social and economic statistics at the national level.
   - Technical supervision of statistics produced by ministries and other public administrations.
   - Improving methods and harmonizing statistics among all statistical producers.
   - Respecting the confidentiality of information provided by individuals, households and institutions.
   - Equal access to all user's statistical releases at the same time.
2. CAS challenges
   - Lack of National statistical system.
   - Lack of Legislative decrees for the Law.
   - CAS by law is not the responsible of conducting census of population.
   - Within the Structure of CAS, there are no Quality, communication and Human resources Units.
3. Current situation of the Quality of Statistics produced by CAS and Ministries

3.1. Social statistics: GSBPM

Overarching processes

- Specify needs.
- Design.
- Build.
- Collect.
- Process.
- Analyze.
- Disseminate.
- Evaluate

Challenges

- Surveys are not conducted on a regular basis.
- Old sampling frame.
- Lack of human resources.

3.2. Economic Statistics

a) Consumer Price Index
b) National Accounts

4. Principles of communication and dissemination
   - Transparency: Clearly marked corrections and changes.
   - Accessibility: Statistics and analyses are public, free and accessible to everyone (Anonymous raw data LFHLCS 2018/2019).
   - Comprehensibility: Published definitions and methodology
   - Independence: All users have access to the statistics and analyses at the same time.
   - Confidentiality: Protected data Privacy on individuals and establishments.
5. Expectations from the sector
   - Implementation of quality of statistics tools.
   - Reporting on quality measurement.
   - Benefit from the experiences of participating countries.

Best Practices in Quality Management in the Southern European Neighborhood Policy countries (ENP-South) - Experience of Algeria

Soraya Khamer[1]

[1]Office National Des Statistiques, Alger, Algeria

Given the importance of quality management as a guarantor of the reliability of official statistics, it was decided to establish a Quality approach within Office National des Statistiques (ONS), referring to the Euro-Mediterranean manuals. The stages of the gradual implementation of our quality approach were outlined through a five-year roadmap (2015-2020). At the end of 2020 and as agreed, each technical department had completed its mapping of the GSBPM model on at least one application per structure. During 2021, in order to ensure continuous improvement in the quality of work, after carrying out the updates, it was decided to extend the mapping of the GSBPM model to other applications. In 2022, ONS trained 12 focal points in various sectors (key user sectors: Bank of Algeria, commerce, interior, industry, energy, housing, finance, transport, agriculture, etc.).

Furthermore, the ONS will continue to participate in discussions on the challenges and opportunities of the quality approach for the ENP South region through the work of the group, currently managed by MEDSTAT.

Impact of the Application of Total Quality Standards on the development of official statistics'

Dr Haidy Mahmoud[1]

[1]National Statistics Office (CAPMAS), Cairo, Egypt

This paper aims to improving the quality of statistical products in Egypt through examining the impact of applying the total quality standards on migration and mobility surveys, as one of the most important surveys carried out by statistical agencies worldwide.

The study uses descriptive analytical method, and Statistical methods to evaluation of the statistical status of Egypt, based on the European code of best practices, and Generic Statistical Business Process Model, in addition to the SWOT analysis, Statistical methods were used to identify the correlation between the various variables.

The study showed that the standard (accuracy) is the most affect by 38% correlated with the other quality standards; then the (availability) is 21%; that both the standard (accuracy) and (availability) affect 59% on the quality of the output of the statistical survey, It was also found that 99% of the sample frame design affects the quality of the statistical product, 61% of the response rates are due to the accuracy and clarity of the statistical form used in the data collection form, only 7% of the researcher's training on the data collection form affects the fieldwork method, it was found that 55% of the analysis of the survey results is due to the accuracy of published data.

Finally the study recommends the need to improve quality reporting through the use of measurable quality dimensions based on the GSBPM.

Assessing GSBPM Implementation at the High Commission for Planning, Morocco : A context of modernization and digital transition

Adil Ez Zetouni[1]

[1]HCP - High Commission for Planning of Morocco, , Morocco

This abstract belongs to the session on (Best Practices in Quality Management in the Southern European Neighborhood Policy countries (ENP-South):

Our presentation delves into the implementation and ongoing application of the Generic Statistical Business Process Model (GSBPM) at Morocco's High Commission for Planning (HCP), with a specific focus on integrating innovation into the statistical system. This initiative forms a part of a larger modernization drive, which includes enhancing IT infrastructure, data management strategies, collaborative tools, innovative data collection methodologies, and the use of alternative data sources.

The presentation assesses how different divisions within the HCP have adopted and adapted the GSBPM, involving key stakeholders in the process and improving process and sub-process descriptions. This has been instrumental in enhancing documentation quality and overall process effectiveness. We also introduce the impact of GSBPM on process standardization and efficiency, comparing time management and quality outcomes before and after its implementation.

Speed Talk Session 1 - Quality frameworks, June 5, 2024, 12:45-13:30

Consistent Quality Reporting while reducing reporting burden: a case study of SIMS implementation

Ms Susana Portillo Cruz[1]

[1]Central Statistics Office - Ireland, Cork, Ireland

Transparency in statistical production is a fundamental principle to foment trust in official statistics. As such the use of Quality Reporting is a core requirement for all producers of official statistics to inform users on the methodologies, processes and key quality indicators associated with the disseminated outputs. Through these reports users can evaluate and make informed decisions on the correct use and assess the quality of the disseminated statistics. With the advent of Eurostat's Single Integrated Metadata Structure (SIMS) in 2015 this obligation was standardized across Europe.

This paper discusses the efforts of the Central Statistics Office (CSO) in Ireland to harmonise different types of quality reports and ensure a unified message across various requirements. In addition it highlights the technical aspects of the migration to SIMS using Colectica as a central repository of reports in order to achieve a more streamlined quality reporting workflow for all our disseminated statistics, while simultaneously decreasing the reporting burden on our Statisticians.

Implementing process mapping to support the quality of official statistics - experience from Ireland

Michael Quinlan[1]

[1]Central Statistics Office, Cork, Ireland

At the Central Statistics Office, (Ireland), CSO, we implement process mapping as a powerful tool to significantly ensure and improve the quality of Irish official statistics. Our process maps provide a clear description of statistical production processes.

A concise CSO Process Documentation policy is in place that communicates corporate goals and individual roles and responsibilities. This policy ensures good governance of process maps, supports maps to be maintained as current stores of corporate knowledge and promotes quality benefits of process mapping. All our process maps are standardised, linked to the Generic Statistical Business Process Model (GSBPM), embedding the GSBPM as the office process model.

This paper will outline the many benefits we gain from process mapping, to support quality. These include optimising the potential for continuous process improvement, supporting training and induction for new staff, promoting statistical literacy, embedding key knowledge of statistical processes in the organisation, linking process maps to additional survey documentation, and supporting resilience and risk management. The paper will also highlight how we share the quality benefits of process mapping across the wider Irish Statistical System (ISS) - supporting engagement with ISS organisations, supporting improvement of standards, and supporting ISS organisations to gain Irish Statistical System Code of Practice (ISSCOP) certification, The paper will also include any lessons we have learnt, and plans for gaining additional benefits to further support the quality of Irish official statistics.

Statistical Quality and New Data Sources – Towards a New Quality Framework for OECD Statistical Activities

Mr Julien Dupont[1], Nora Bohossia, Adrian Zerbe

[1]Oecd, , France

The quality of the OECD evidence-based analytical works depends on the quality of data and statistics collected and compiled by the Organisation. Over the past few years, the digitalisation of economies and societies and the COVID pandemic has generated an explosion in the need for new data, coupled with increasing computing capacity to exploit them. Many of these data have the potential to generate new, timelier, more granular evidence and more trusted information for a wide range of users including citizens, analysts, and policy makers. With the emergence of new and varied data sources that are increasingly used by OECD statisticians and analysts in their data, analytical, and quantitative work, there is a need to revisit the Quality Framework of OECD Statistical Activities established in 2011. The new data landscape poses complex new questions about access to, and use of high-quality statistics, at national and international level. Along with many opportunities, evidence based on new data sources also entails challenges by way of adherence to existing and new quality dimensions, interoperability, linkability, reproducibility, and security, but also co-ordination of projects, data and knowledge sharing, and development of new skills. This paper draws some important lessons about access to new data sources, how they blend in with more traditional activities, and how the scope of quality assurance can be extended to newly derived statistics and indicators over the data life cycle in order to continue providing data with trusted quality to support sound, evidence-based analysis and policy advice along with a broader information function for society. This paper describes the development of a revisited quality framework for OECD statistical activities that will make explicit and detailed reference to new data sources, based on the main lessons the Organisation has learned so far from its work to develop evidence from new data sources while respecting the need to ensure that its data and statistics remain of high quality and continue to command public trust, as well as the new tools developed to monitor its implementation.

Applications of Statistical Methods and (Inter)national Standardization

Ms Maria Maria[1], Mafalda Costa[2]

[1]ESTG|IPP, Felgueiras, Portugal, [2]DRAP Norte, Amarante, Portugal

Standardization is an activity that, in an organized and recognized way, makes it possible to draw up reference documents (standards), intending to increase the levels of quality, safety, efficiency, and interoperability of processes.

In this context, the Portuguese Quality Institute (IPQ), as the national standardization organization, welcomed the creation of the Portuguese Technical Commission 225 (TC 225) for standardization on the Applications of Statistical Methods to follow the developments of the International Organisation for Standardisation - Technical Commission 69 (ISO/TC 69 Applications of Statistical Methods) and to disseminate this knowledge at national level.

One of the main objectives of CT 225 is to support the drafting of normative documents on various technical-scientific topics that make it possible to establish clear rules on the use of statistical methods and consequently become benchmarks that can help to strengthen and extend the metadata system, improve the quality of the products and services disseminated and improve the statistical production process (official or not).

CT 225 works on topics such as Terminology and Symbols; Application of Statistical Methods in Process and Product Management; Acceptance Sampling; Measurement Methods and Results, Statistical Measurement Methods and Results, Statistical Tools for Implementing the Six Sigma Methodology, and BIG DATA.

The National Statistics Institute (INE), the Bank of Portugal (BP), the Academy, and many other official institutions actively participate in the standardization activities of TC 225 and recognize the use of ISO standards as having several advantages, namely: they are international standards that are accepted globally among peers, facilitating interoperability and compatibility between products and services from different countries and organizations; they establish guidelines to guarantee the quality, safety, and efficiency of processes; they promote the improvement of systems, reduction of errors and optimization of resources. They promote credibility and trust. Complying with ISO standards demonstrates a commitment to quality, safety, and sustainability. This positively influences customers, suppliers, and partners. ISO standards are reviewed periodically to keep up to date with technological, regulatory, and market changes. This encourages innovation and continuous improvement within organizations.

In this work, TC 225 aims to make known the technical-scientific standardization projects underway, the opportunities opened up in this field, and their influence on the quality of these projects.

Statistician as a Pastry Chef making statistical cakes using GSBPM

Ms Žaklina Čizmović[1]

[1]Croatian Bureau Of Statistics, Zagreb, Croatia

Everyone is aware that for making tasteful cakes the grandmother's cooking book, cookware, utensils, tools and the Pastry Chef are needed. On the first, it sounds very simple and realistic, but how this scenario can be implemented in the statistical world is an on-going issue.

Obviously, the main tool in achieving sweet statistical cooking result is planetary known GSBPM solution which offers every statistician to be a Pastry Chef in doing statistical desserts as much as possible tasteful and attractive for the final user on a long-term.

Getting acquainted with all GSBPM main processes and its subprocesses requires time, knowledge, organisation, coordination, monitoring, patience and objective assessment not only within the kitchen but also in the pastry shop. Dealing with the nitty-gritty GSBPM facts is never ending story which encourages the statistician to work on own development continuously.

In order to achieve that, many years ago the Croatian Bureau of Statistics created the national version of GSBPM 5.1 with necessary explanatory notes for each process and its subprocesses in form of a paper document and in 2022 had created the GSBPM module as a part of the POMI quality application and database which beside the quality reporting offers monitoring and development of statistical products.

Talking about the statistical products and its improvement, we have noticed that for users the statistical way of working is not easy to understand, therefore we would like through this paper demonstrate the implementation on a well-known Portuguese product called Pastéis de Belém; simulate the registration of the sweet in the Croatian Annual Implementation Plan and use the GSBPM philosophy by describing main processes in POMI. Identify critical points in the product development and describe how they were improved.

The GSBPM tool offers statistician never to be angry at any point in the statistical subprocesses, but to be happy and do what is at certain point of time possible.

How this journey started and how it was developed during last two decades is presented in this paper. Bon appetite!

Implementing GSBPM – experiences from Denmark

Ms Karin Blix[1]

[1]Statistics Denmark, , Denmark

In 2009, Statistics Denmark decided that GSBPM should be the process model for the institution. GSBPM was translated to Danish and activities were started to implement GSBPM in the organisation. The implementation had a rough birth. Only parts of the organisation got off to a good start. Many measures have been initiated over the years, but it has been difficult to get a proper hold on the organisation. The latest measures taken in the organisation has partly been the result of a large staff turnover and partly of pressure from outside the organisation.

This paper will describe some of the measures taken to implement and make GSBPM our process model and discuss how an organisation can benefit from adapting GSBPM as its own process model – without modifying the model itself.

Speed Talk Session 2 - Smart Survey Implementation, June 5, 2024, 12:45-13:30

How does the general population think about surveys with smart features?

MS Monica Perez[1], Janelle van den Heuvel[1]

[1]Statistics Netherlands/Utrecht University, , Netherlands

NB: THIS PAPER IS PART OF AN INVITED PAPER SESSION SMART SURVEY IMPLEMENTATION SUBMITTED EARLIER

Smart surveys potentially offer new features to make the utility of surveys more salient and leverage any objections against surveys. The Smart Survey Implementation (SSI) project aims to involve and engage citizens in the policy design and evaluation which potentially contribute to the gain of trust and participation of the citizen. This can only be achieved by diving into the respondent perceptions within realistic and legitimate smart survey settings. Therefore, as part of the SSI project, the cross-national survey on smart survey perceptions was introduced which aims to provide empirically supported understanding of how citizens feel about surveys with smart features, including how well they understand what is being measured and what they consent to. This survey was conducted in three countries (Italy, Slovenia and the Netherlands. The perception survey consists of a paper questionnaire and in addition respondents are asked to conduct an online questionnaire sequentially. The latter includes four smart feature tasks. During the presentation, results of the survey on smart survey perceptions will be presented, in particular the cross-country differences.

When are smart surveys mature?

Mr Remco Paulussen[1], Marc Houben[1], Barry Schouten[1]

[1]Statistics Netherlands, , Netherlands

The new types of data and the new types of methodology in smart surveys imply investments in almost all stages of the statistical process; from data collection case management and back office systems to real-time processing procedures to post-survey editing and adjustment methods. Given the rationale of reducing measurement error, time series shifts are anticipated, implying that a change to smart data collection needs to be well prepared and introduced.

For official statistical institutes, surveys are often repeated and large-scale, so that solid and robust logistics and architecture are imperative. The business case for 'going smart' must be extra strong. In Eurostat-funded project Smart Survey Implementation, several case studies are investigated, tested and evaluated on their maturity to go smart. In this presentation, we present so-called maturity criteria that have been put forward in all design levels: methodology, IT architecture, logistics and legal. The criteria are illustrated with examples and open for discussion with the conference attendees.

This subsession is part of the larger session on Smart Survey Implementation.

Smart surveys - what are they?

Mr Peter Lugtig[1], Barry Schouten, Remco Paulussen, Joeri Minnen, Florian Keusch, Hannah Bucher, Bella Struminskaya, Danielle McCool, Theun Pieter van Tienoven, Claudia de Vitiis

[1]Utrecht University, , Netherlands

Paper 1 in invited session: smart survey implementation.

Smart surveys employ features of smart devices. Keeping respondents at the heart of data collection, they form a bridge between survey and big data. The incentive to go smart is especially strong for surveys that are considered burdensome non-central to respondents and/or certain topics for which questions form weak proxies for the concepts of interest In smart survey data collection AI and machine learning methodology play a prominent rle in te transformation of ‘smart’ data to statistics. But what are smart surveys exactly? In this paper a taxonomy is presented of smart feastures and illustrated by a demo of two services included in the Eurostat funded project ‘Smart Survey Implementation’.

Business process in the context of smart surveys

Marc Houben[1], Mr Remco Paulussen[1]

[1]Statistics Netherlands, , Netherlands

Applying smart solutions in surveys has an impact on the logistical process. Additional attention is needed in several process activities of the statistical business process. In this paper we will look into this. Which additional capabilities are needed? Which process activities need adaptation? What actors are affected? Which activities can be supported by e.g. micro services and machine learning modules? We will use GSBPM as a framework and give examples from e.g. HBS and/or TUS. The goal of this paper is to give you – as a statistical office – some guidelines on what is needed, in the business process, to start using smart solutions in production.

This paper is part of the 'Smart Survey Implementation' topic, for which different papers have been submitted.

Speed Talk Session 3 - Improving household surveys, June 5, 2024, 12:45-13:30

Understanding the biases of wealth surveys: evidence from housing wealth of French households

Mr Olivier Meslin[1]

[1]Insee, PARIS, France

Wealth surveys were initially designed to provide high quality data on households' wealth by measuring a broad set of assets for a representative sample of the population. It has however repeatedly been shown that the top tail of the distribution is partly missing in survey data because of four types of bias: miscoverage of the population of interest and non-random unit nonresponse (inducing underrepresentation of wealthy households), and item nonresponse and reporting error (resulting in underreporting of assets). However, few papers were able to measure precisely the relative importance of the biases, mostly because of "the lack of a benchmark measure of the true outcome" (Meyer et al 2015).

In this paper, I contribute to a better understanding of the biases of wealth surveys in a case where a benchmark measure is actually available. I link the 2017 French wealth survey to a new statistical database on housing wealth developed by the French statistical institute (Insee). This new source is based on cadastral, fiscal and commercial register data and covers all resident households and all properties located in France. Moreover, it provides an accurate estimate of the market value of all privately-held housing units based on machine learning algorithms. Finally, sampled households can be linked to this database using fiscal identification numbers and personal information (names, place and date of birth). The linked data contains information on assets and wealth reported by respondents, along with information on assets and wealth recorded in the administrative database for all sampled households (including non-respondents).

I first decompose the discrepancy between survey-based estimates and benchmark estimates into five sources of bias: population miscoverage, sampling error, unit non-response, weight adjustment and respondents' reporting behavior. This approach makes it possible to measure the effect of each source of bias on the distribution and concentration of housing wealth. I then document the specific non-response behavior of wealthy households, distinguishing the patterns in the contact rate and in the cooperation rate. Finally, I analyze the reporting behavior of respondents, identifying the type of housing assets that went massively unreported. I show that respondents use a conceptual framework that differs in important ways from that intended in the survey design. Finally, I estimate the relative importance of intentional under-reporting, unintentional under-reporting and unintentional over-reporting and show that intentional under-reporting plays a major role in the underestimation of the upper tail of the housing wealth distribution.

Innovative methodologies to improve quality of official statistics – the Nigeria Labour Force Survey Methodology Revision as a Case Study

Mr Adeyemi Adeniran[1,2], Mr Kumafan Dzaan[1,2]

[1]ISI-IAOS, Abuja, Nigeria, [2]Chartered Institute of Statisticians of Nigeria (CISION), Abuja, Nigeria

Labour market statistics play a crucial role in shaping decisions on job policy, guiding employment generation interventions, and providing a comprehensive understanding of a nation's labour market dynamics. This paper delves into the dynamic realm of statistical methodology, focusing on the significant strides made by Nigeria's National Bureau of Statistics (NBS) in enhancing the quality of labour market data through the recent revision of the Nigeria Labour Force Survey (NLFS). According to the UN Handbook of Statistical Organization, National Statistical Offices (NSOs) exist to provide information essential for development and for mutual knowledge and trade among the States and peoples of the world. The NLFS serves as an interesting case study, illustrating how innovative methodologies can address existing challenges and significantly elevate the quality of statistics in Nigeria, thereby ensuring that the NBS to attain its mandate.

In revising the methodology of the NLFS, the NBS adopted a multifaceted approach, encompassing advancements in sampling techniques, survey design, definition, and concepts, as well as data processing. Leveraging cutting-edge technologies and statistical tools, this initiative was aimed to not only capture a more accurate representation of the labour force but also to ensure timeliness and relevance in a rapidly evolving socio-economic landscape. Key components of the methodology revision included the integration of advanced sampling methodologies, such as stratification and cluster sampling, to improve the survey's representativeness across diverse demographic and geographic dimensions. Furthermore, the adoption of electronic data collection methods and real-time monitoring mechanisms serves to enhance data accuracy, reduce response bias, and expedite the overall survey process.

This paper also explores the experience of undertaking this innovative and critically important exercise. Additionally, efforts to engage stakeholders through increased collaboration and data transparency are highlighted in the paper, as they are crucial elements in ensuring the reliability and credibility of the revised NLFS.

The outcomes of this innovative methodology revision are anticipated to yield a more robust and nuanced understanding of the Nigerian labour market. By presenting a detailed examination of the challenges faced, the methodologies employed, and the impact on data quality, this paper contributes to the broader discourse on advancing official statistics methodologies. The lessons drawn from the Nigeria Labour Force Survey can serve as a valuable blueprint for other nations seeking to enhance the accuracy and relevance of their own statistical frameworks in an ever- changing global landscape.

Inferences to a sample of volunteers in a household survey

Ms Soonpil Kwon[1,2], Heeyoung Chung[1], Youngmi Kwon[1]

[1]Statistics Korea, seo-gu, Republic of Korea, [2]University of Seoul , Dongdaemun-gu, Republic of Korea

An increase in non-probability samples from different sources and the development of IT for data processing require a change in the paradigm of statistical production based on probability samples. Because it's difficult to select and maintain a probability sample due to a decrease in the coverage of the sampling frame, an increase in non-responses and survey costs, and a worsening survey environment such as COVID-19. However, for the inference of finite populations, the problems of selection bias, under-coverage and unknown sampling probability of non-probability samples must be addressed. For this purpose, data integration of non-probability samples and high-quality reference probability samples, and the specification of a model connecting the two datasets are essential. It's possible to borrow the unbiasedness of probability samples.

This study examines four estimators of the propensity score weighting method(ipw), the calibration weighting method(cal), the mass imputation method(reg), the doubly robust method(dr). The variance of each estimator is estimated using bootstrapping.

For the simulation study, public master datasets of the 2021 Survey of Household Finances and Living Conditions are used as the population. The value of interest is the average annual current income of households, and the auxiliary variables are the demographic characteristics of the household head and the number of household members. We assume various scenarios and estimate the average of non-probability samples and the confidence interval for each scenario.

All four estimators show a drop in relative bias, mean square error. The probability of including the population mean in the 95% confidence interval is approximately 80% on average. Among the four estimators, the doubly robust estimator and the calibration estimator show the most stable estimation results. When a non-probability sample is treated as if it were a SRS, serious selection bias problems arise.

This simulation study shows that stable selection bias correction is possible if appropriate auxiliary variables are used even for variables of interest that fluctuate greatly, such as average annual current income of households. Therefore, it is expected that it will be possible to expand application to a variety of surveys and variables of interest.

Correcting for time series breaks in the Swedish Labour Force Survey at the micro level - combining models and calibration

Frida Videll[1], Thomas Önskog[1]

[1]Statistics Sweden, Solna, Sweden

The Swedish Labour Force Survey (LFS) is the foundation for official statistics on the Swedish labour market. It is the only source of statistics that continuously provides a coherent picture of the labour market in terms of employment, unemployment, hours worked, etc. and the interest from users is therefore considerable. As from January 1, 2021, the LFS must comply with the new EU framework regulation on social statistics. To do so, several changes has been made to the survey, both regarding the population and the questionnaire. In addition to the changes caused by the new framework regulation, a revision of the auxiliary information used in the estimation was implemented at the same time as the new framework regulation. To the serious concern of the users, the changes made to the procedure of the LFS has caused breaks in many of the time series.

To estimate and correct for the breaks in the time series of the LFS, Statistics Sweden performed a parallel run with both the old and the new questionnaires during 2021. The result of the parallel run has been complemented by data from a series of other sources, such as data on respondents that were affected by a change in the definition of employed, estimates based on both the old and the new auxiliary information, respectively, as well as flow data describing how the employment status of respondents vary from one quarter to the next.

In this paper, we describe how the available data has been analysed and combined to correct for the time series breaks and how we have derived new time series for the LFS that are comparable back to 2005. Using a combination of imputation models for the micro data and calibration, we have derived a set of recalibrated weights for all the respondents of the LFS during the time period 2005-2020. The recalibrated weights constitute a link at the micro level between the old and new procedures. Time series calculated using the recalibrated weights and the imputed micro data during the time period 2005-2020 are fully comparable with time series calculated during 2021 and onwards using the new procedure.

Indirect estimation of selected characteristics of the working and unemployed population in the functional areas of provincial capitals

Mr Tomasz Jozefowsk[1]

[1]Statistical Office in Poznan, Poznan, Poland, [2]Poznan University of Economics and Business, Poznan, Poland

Nowadays, regardless of how functional urban areas of provincial capital cities are delineated, there is a lack of detailed data from sample surveys, including information about the labour market. This is due to small sample sizes for such domains. Based on such data, direct estimates typically produced by official statistics, are characterised by poor precision. This problem can be remedied by employing methods of small area estimation (SAE), which can be used to obtain reliable estimates at lower levels of spatial aggregation, i.e. in small spatial domains characterised by small sample sizes. SAE methods can even produce estimates for zero sample sizes.

While small area estimates tend to have better precision, they often do not sum up to direct estimates at higher levels, which are regarded as reliable. One way of solving this inconsistency is the use of Structure Preserving Estimation (SPREE). These estimators exploit the known association structure of auxiliary variables for target domains from administrative registers or censuses, while preserving the consistency with estimates at higher levels of aggregation. The method seems to be particularly attractive for estimating non-standard domains which were not planned at the design stage. This is important because it can help to meet expectations of users of statistical data, who often need consistent information about domains which are of interest at a particular moment.

The main goal of the paper is to present the usability of the SPREE estimator and its generalizations (GLSM, GLSMM, MSPREE, MMSPREE) for estimating the number of people in employment in selected functional areas of provincial capital cities by sex and age. Emphasis will be placed on the assessment of the quality of estimates produced using a number of various sources, such as surveys and administrative registers.

Utilizing Integrated Data Sources for Small Area Estimation of Poverty Indicators

Michele D'Alò[1], Danila Filipponi[1], Francesco Isidori[1]

[1]Istat, Roma, Italia

In recent years, ISTAT has conducted comprehensive studies aimed to apply Small Area Estimation Methods (SAE) for computing Sustainable Development Goals (SDGs) indicators related to health, occupational status, gender equality, and poverty using data collected by means the main social surveys.

The importance of applying SAE methods in the production of official statistics is emphasized by the recent approval of Regulation IESS (Integrated European Social Statistics) by the European Parliament and Council (No. 1700 dated 10/10/19). The regulation establishes a common framework for European socio-economic indicators on individuals and families based on integration of data from multiple sources along with innovative methodologies, including small area estimates, and accuracy levels. Clearly, this effort is driven by the growing demand from decision-makers for granular statistical information, which is essential to ensure that decisions, resource allocation, and policies are grounded in accurate and detailed data.

This study focuses on findings related to poverty indicators, with a particular emphasis on the "At Risk of Poverty or Social Exclusion" (AROPE). The AROPE rate is the proportion of the overall population facing the risk of poverty or social exclusion. The indicator includes individuals classified as At Risk of Poverty (ARP), those experiencing Severe Material and Social Deprivation (SEVDEP), or residing in households characterized by very Low Work Intensity (LWI).

The measurement of this indicator relies on data sourced from the European Survey on Income and Living Conditions (EUSILC) and the estimates are disseminated for five macroregions (NUTS1), twenty regions (NUTS2) and ten age classes. Despite the regions being planned domains, the direct estimates of the indicator do not meet the precision requirements set by EUROSTAT. Consequently, alternative estimation methods are necessary to enhance the overall precision of the estimates.

The study aims to delineate the production process for producing Small Area Estimation (SAE) estimates of the AROPE and its component indicators, even at the level of unplanned domains such as provinces and metropolitan cities (NUTS3).

Several design and model based SAE methods has been applied, exploiting the increasing availability of administrative data in ISTAT organized in a System of integrated statistical register, which can be linked with socio-economic surveys data. The work provides insights into the implementation of established SAE methods that utilize both area-level and unit-level models and issues related to key aspects such the model selection, model diagnostics, and benchmarking.

Using paradata to assess the quality of the questionnaire design in the Adult Education Survey

Ms Sara Grimstad[1], Katharina Rossbach[1], Elise Alstad[1]

[1]Statistics Norway, Oslo, Norway

Statistics Norway (SSB) conducted a Eurostat financed Grants project related to the Adult Education Survey (AES) from 2021 to 2023. One of the project's two objectives was to improve the user experience to increase the data quality. To do this, we used several qualitative methods, such as user testing the survey questionnaire, as well as conducting focus groups and explorative interviews to map the respondent user journeys. These qualitative methods helped us to gain insight into potential challenges with the questionnaire flow and the questions themselves. Specifically, we identified problems with naming non-formal education and training activities that could affect the data quality and risk of break-off for specific questions. Hence, we adjusted parts of the questionnaire flow and questions before the data collection for the Norwegian AES 2022 started. Despite having two rounds of user testing, several focus groups and explorative interviews, challenges in the questionnaire persisted. Specifically, results from the AES 2022 indicated there were still problems with naming activities, and the non-response rate for these questions increased from 2016 to 2022.

As Statistics Norway has recently set up a system to make paradata easily accessible, this article will focus on how these data, in addition to the qualitative methods mentioned, can contribute to further insight into presumed pain points in the questionnaire. Paradata is a helpful quantitative tool for analysing and identifying areas to improve the questionnaire, which in turn can contribute to better data quality. We use paradata from web responses in the Norwegian AES 2022, together with survey data, to analyse how the identified problems from the qualitative analyses manifested. The paradata indicators we use to analyse this are error messages, previous page actions, change of answer, break- off rates and survey data with information about the use of "Don't know" and "Refusal". We then combine our quantitative and qualitative findings to provide a holistic picture. Finally, we suggest better ways to structure the questionnaire to reduce non-response and the respondent burden, to improve the overall data quality in the statistics.

In sum, we want to illustrate how paradata can be used to assess if problems in a survey still persist or not, and how one can improve the survey. Thus, the main objective of the paper is to offer ideas on how paradata can help us to assess and improve the quality of questionnaire designs and its resulting statistics.

Speed Talk Session 4 - Data processing, June 5, 2024, 12:45-13:30

Improving statistical registers' quality through attribute-driven spatial matching

Dott. Damiano Damiano[1], Researcher Daniela Ichim[1], Researcher Luisa Franconi[1]

[1]Istat, Rome, Italy

In response to the increasing demand for information in modern society, official statistics often rely on multi-source production models. Since statistical registers have lower production costs compared to field surveys, these registers are frequently integrated to derive the estimates of interest. This contribution proposes a data reconciliation method through spatial correspondence to integrate registers of residential addresses and buildings, enhancing their quality and increasing the capability of linkage operations. The method leverages housing ownership conditions, owners' residential addresses, and the proximity between buildings and addresses. Implementation and evaluations were conducted using one Region (NUTS 2 level) as the study area within the Integrated System of Statistical Registers of the Italian National Institute of Statistics. A significantly larger share of the addresses has been identified as belonging to both the address and building registers. Moderate precision values and high recall values serve as key performance indicators for the proposed method. Our approach is sufficiently flexible to allow various extensions related to parameterizations (urban/rural), statistical units, or complex proximity distances.

Establishment and comparison of predictive models for oil and petroleum products, electricity and gas: A Cross-Border Analysis

Msc student Aglaia Papakyrillou[1,2], Ms Anna Mpoukouvala[1,2], Professor George Tsaklidis[2]

[1]ELSTAT, Athens, Greece, [2]Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece

This study investigates and compares five different predictive models for three critical parameters: oil and petroleum product deliveries, electricity availability and inland gas consumption. The research spans six consecutive months, from October 2023 to March 2024, providing a comprehensive overview of predictions during periods of both low and high energy consumption for five countries of EU, each differing in various parameters.

Various model types were developed, capturing a broad range of approaches and reflecting methodological diversity. Specifically, AutoRegressive Moving Average models with exogenous variables were applied to stationary time series. Two indicators were utilized for the first and second models, while more than two indicators were employed for the third model. Corresponding indicators for all combinations of countries and parameters were used in the implementation of the first and second models, in contrast to the third model, where indicators were selected considering the specificities of each case. The fourth model was based on a Recurrent Neural Network (RNN), specially designed for analyzing time series data. In the last predictive approach, linear models were employed, individually tailored for each combination, capitalizing on their flexibility in adaptation.

In conclusion, both the performance and advantages of the selected models are presented, highlighting significant differences and similarities deemed crucial for comparing the provided approaches.

Reducing Manual Editing at Statistics Sweden – an Agency Wide Approach

Ms Jenny Hjort[1], Mats Bergdahl-Kercoff, Magnus Sjöström

[1]Statistics Sweden, , Sweden

Like most NSIs, Statistics Sweden allocates a significant proportion of its resources regarding business statistics for the manual editing of data and output. Over the years, Statistics Sweden has made several attempts to make the editing process more efficient, with some success for some specific surveys but with limited total effect.

In late 2021. Statistics Sweden's top-management team decided that another approach was needed to drastically reduce manual editing so that resources could be used for much needed development work in other areas. The approach was labelled, "The New Editing Process", with the main focus on ensuring that the data coming to Statistics Sweden was correct by means of well-developed respondent controls as well as a clear focus on macro level editing to determine questionable values. The importance of editing was also clearly stressed if justified based on the stated purpose of the statistics, although not with the aim to catch every single value that was not fully correct.

The statistics were divided into three categories based on previous knowledge of hours spent on editing. There were around 20 statistical products in category 1 and 2 respectively which were the focus of the central implementation efforts as they together comprised about 80 percent of the data editing. It was a clearly stated goal for these products that manual editing should be reduced as much as possible unless clearly justified. The responsibility for making changes were placed with the line organisation, but a central implementation organisation was also put in place to coordinate, provide support and to make sure that common development needs were taken care of.

Initially, training regarding the quality concept for official statistics was carried out as well as workshops to better understand and define user needs. Resources, including expert support in methodology and measurement issues, were allocated to the statistical products to review the current editing and develop implementation plans. The approach was initially met with much scepticism but was also considered long overdue by many. This meant that progress was quick for some statistics but significantly slower for others. Two years on, Statistics Sweden has reduced resources spent on manual editing by about 50 percent and there is indication that more potential exists for a few statistical products.

Our presentation will describe the overall approach, results, the main features of the new editing process as well as the identified success factors.

Time-Series Exhaustive Automatic Modeling: a new methodology for model identification

Dr Carlos Sáez Calvo[1], Luis Sanguiao Sande[1], Félix Aparicio Pérez[1], María Teresa Vázquez Gutiérrez[1], José Fernando Arranz Arauzo[1]

[1]Ine Spain, , Spain

Seasonal adjustment of time series plays a pivotal role in modern official statistics, ensuring accurate and reliable data analysis. However, due to resource constraints and time limitations, the models identified in an automatic way using the current software may not be optimal. This leads to a worse performance of seasonal adjustment, since these models must be maintained for a year.

We present a new R package, Time-Series Exhaustive Automatic Modeling (TEAM), which aims to automate and enhance the yearly model identification phase. The goal is to provide in an automated way a list of optimal models, where the optimality criteria can be specified by the users to meet their specific needs.

The methodology employed in TEAM is characterized by an exhaustive search and ranking of models. Initially, an exhaustive search of specifications is conducted for each time series, testing all possibilities for parameters such as data transformations (logarithms or levels), the order of the ARIMA model, inclusion of outliers, and calendar regressors. Subsequently, each specification is processed using the JDemetra+ software in a parallelized way, yielding diagnostic information to construct four indicators assessing the model's performance across distinct areas.

The four indicators and their respective areas of evaluation are as follows:

1. Model Diagnostics: Measures the model adequacy by using the statistical tests on the residuals of the RegARIMA model and considering the statistical significance of the model coefficients and their autocorrelations.

2. Signal Extraction: Measures the model's efficacy in signal extraction using SEATS (via canonical decomposition).

3. Revisions: The magnitude of revisions when new data is available is captured by this indicator.

4. Residual Seasonality: This indicator considers statistical tests on residual seasonality after the seasonal adjustment process is performed.

To rank the models effectively, a final score is computed by appropriately combining the four indicators. Importantly, users retain the flexibility to adjust the weights assigned to each area according to their specific requirements. For instance, users could prioritize models with minimal revisions based on their preferences. Moreover, an alternative approach based on the Pareto boundary is also explored. Finally, TEAM presents the user with a selection of the best models based on the final score, enabling them to choose the most suitable model according to their needs.

A small step for EMIR, a giant leap for transparency - strategies for boosting data quality in granular derivatives data

Grzegorz Skrzypczynski[1], Gemma Agostoni[1], NIcola Calabrese[1], Marco D'Errico[2], Lukas Henkel[1]

[1]European Central Bank, Frankfurt, Germany, [2]European Systemic Risk Board Secretariat, Frankfurt, Germany

The European Market Infrastructure Regulation (EMIR), a significant piece of the EU post-crisis legislation, introduced several requirements for EU-based counterparties entering derivative contracts, including the obligation to report granular transaction-level information daily to Trade Repositories (TRs). This information is then shared with over one hundred EU authorities.

The ECB and the ESRB staff have been intensively using this data, leveraging a dedicated, high-performance IT infrastructure. The information has been used to work on a set of diverse tasks, facilitating timely monitoring and analyses of derivative markets, progressively becoming a building block for risk analysis and policy development and assessment.

However, nearly a decade since mandating the reporting requirements, the use of the data remains challenging, due to persistent and significant data quality problems. While there has been some improvement over years, data reported under EMIR continues to fall short of being fully satisfactory for scaling and automation. Despite employing advanced IT tools and having accumulated a vast expertise, it still requires considerable time and effort to clean, aggregate and analyse. Given these persistent problems, conclusions drawn from the data carry considerable uncertainty: is a specific pattern indicative of a relevant financial development or merely a by-product of poor data quality? This paper describes how EMIR data is used at the ECB and ESRB Secretariat, highlighting the primary data quality challenges the experts encounter when working with the data. It delineates the most common and significant data quality issues inherent to EMIR and proposes a comprehensive taxonomy. This classification enables us not only to identify the root causes of these issues, but also explore the reasons for their persistence over time.

The paper then offers recommendation for enhancing data quality. Although it is unrealistic to anticipate the immediate resolution of all issues in a dataset of this scale and complexity, we propose that the implementation of certain systematic and potentially automated solutions could significantly improve the quality of the data. Remarkably, this improvement could be achieved with minimal effort from reporting entities and regulators.

Finally, we argue that enhanced data quality will boost the use of this information by regulators and policy makers. This progression would facilitate market monitoring, leading to further data quality improvement. Even more importantly, this would bolster financial stability of the EU and its member states. We also believe that streamlined procedures and better data quality management could benefit the industry itself.

Speed Talk Session 5 - Statistical leadership/HR management, June 5, 2024, 12:45-13:30

Taking on the challenge of developing statistical leadership to build organisational capability and equip for the future

Ms Heather Bergdahl[1], Ms Anna-Maria Kling[1], Ms Marie Haldorson[1]

[1]Statistics Sweden, Örebro, Sverige

Statistics Sweden has in recent years been challenged by a team of external experts, who annually evaluate the quality of a selection of the agency's important statistical products. The challenge presented to us is to develop our statistical leadership in order to deliver statistics that are fit for purpose. Statistics Sweden has taken on this challenge. We recognize that we are professionals with different areas of expertise such as in data, methods, analysis, as well as understanding the statistical quality aspects and the implications of these for the prioritised uses of the statistics we produce. As professionals we adhere to scientific principles and the quality framework of the European Statistics Code of Practice which strengthens the reliability of and trust in the statistics we produce. We also have a responsibility towards our users to meet their present and future information needs such that the statistics suit their intended uses.

To be effective in this task we need to show statistical leadership in our different areas of expertise. This leadership extends outside the organisation to our users, with whom there must be strong engagement. Our statistical expertise and thinking will bring added value to the dialogue with users which we can utilise in translating their needs into statistical solutions. It also requires of us within our agency to be partners in assuring that standards and best practices are being applied across the organisation.

We believe that the extent to which we can develop our statistical leadership as a comparative advantage will help us to strengthen our position in the new and challenging environment that National Statistical Systems meet in the face of an increasing number of actors on the information market.

In this paper we will develop our thinking on what the difference is between leadership in general and statistical leadership as well as why statistical leaders are needed more than ever. We will also highlight what opportunities we see to demonstrate statistical leadership externally and internally. Finally, we will share challenges we experience in the process of building statistical leadership as an organisational capability.

New statistical resilience or how to survive in the data ecosystem

Ms Ana Carmen Saura[1], Yolanda Gómez[1], Ana Cánovas[1]

[1]National Statistics Institute of Spain, Madrid, Spain

As declared by the Commission “the European data strategy aims to make the EU a leader in a data- driven society. Creating a single market for data will allow it to flow freely within the EU and across sectors for the benefit of businesses, researchers and public administrations”, it means that this Strategy necessarily crosses its path with the statistical functions. This could be a threat or an opportunity for the ESS.

Aware of this situation and taking into account all the previous initiatives already carried out by the ESS, the revision of the Statistical Law has been boosted to tap the potential of the new data sources for official statistics available in the current digitalised society.

This paper scrutinizes the opportunities, challenges and potential roles for the statistical community in the future implementation of the amended Statistical Law within the context of the new digital and legal ecosystem. It also looks for synergies with the European Data Strategy and related legal acts (Data Governance Act, Artificial Intelligence and Interoperable Europe Act).

Human resources management by deliveries: the experience of the Brazilian Institute of Geography and Statistics

Mrs. Ana Cristina Martins Bruno[1], Mrs Paula Leite da Cunha Melo

[1]IBGE - Brazilian Institute of Geography and Statistics, Rio de Janeiro, Brazil

The Management and Performance Program (PGD) was established by the Brazilian Federal Government in 2020, after the experience with emergency remote work during the pandemic. More than a Program that formalizes teleworking in Federal Public Organizations, the PGD is today a driver of performance improvement in the public service, focusing on the articulation between the work performed by the participants, the units' deliveries and organizational strategies. The Program authorizes and institutionalizes teleworking, but it is much more than that, it acts as a people management instrument that changes the logic of work in the public service. Through PGD, the management focus becomes the units' deliveries, compliance with the individual work plan and the articulation between what is produced and organizational strategies. The PGD no longer considers presence and attendance, compliance with the working day at a fixed time, frequency and permanence in the body's physical facilities as control elements. And it starts to focus on managing deliveries and results, based on dialogue and cooperation between leadership and its team. At IBGE, in particular, in the Planning and Management Coordination of the Executive Board, employees joined the Program and today work full teleworking. The Unit Delivery Plan is the central management instrument that records deliveries, goals, deadlines, demands and recipients, being the reference for preparing individual work plans. Deliveries are linked to strategic projects or value chain processes, leading to cohesion between daily work and the strategic vision. Delivery assessment results are analyzed and considered for subsequent planning cycles, in a continuous organizational learning approach. This article aims to present the benefits of PGD (improvement in team management, alignment of results with strategy, transparency in unit deliveries, better use and retention of talent, reduction of expenses, mainly with the maintenance of physical spaces and quality participants' lives) , their main challenges (management support and engagement, internal communication and strengthening organizational culture) and the main lessons already recorded, such as the tools used and the good management and control practices adopted and the construction of an environment of trust between leaders and their teams.

Innovation Management System at the Czech Statistical Office

Ms Petra Kuncová[1], Marek Rojíček[1], Egor Sidorov[1]

[1]Czech Statistical Office, Prague, Czech Republic

Innovation is essential in today's fast-changing world. This of course also applies to national statistical offices. The management of the Czech Statistical Office places great emphasis on innovation. As part of the project on the introduction of quality management into the CZSO, among other things, activities were carried out leading to the creation of basic documents for the innovation management system - Innovation Policy and related internal regulations. The presentation will deal with the innovation management system in the Czech Statistical Office from the point of view of its organization, collecting and identifying change ideas and their subsequent implementation. Special attention in the presentation will be devoted to ensuring support and active approach of employees and their motivation during the initiation and implementation of changes.

Building the future through collaboration in “half-an-houR”

Mr Almiro Moreira[1], Mr Alexandre Cunha[1], Mr Álvaro Combo[1], Mr Bruno Lima[1], Mr João Lopes[1], Mr João Poças[1], Ms Paula Cruz[1], Mr Pedro Campos[1], Mr Pedro Guerreiro[1], Ms Rita Santos[1], Mr Rui Alves[1], Mr Pedro Sousa[1]

[1]Statistics Portugal, , Portugal

The 1/2 houR (read: half-an-hour of R) sessions, initiated in 2023 at INE (statistics Portugal), feature brief presentations followed by a short discussion period. These weekly 30-minute sessions serve as an informal meeting between colleagues, encouraging the exchange of knowledge and work experiences involving developments using the R language. Despite the physical distance between INE’s headquarters in Lisbon and its various remote branches, 1/2 houR facilitates closer collaboration.

1/2 houR targets INE technicians using R in their routine work (useRs). Each month, a designated curator selects a theme for the monthly presentations, recruits the presenters and moderates the sessions for that month. The sessions take place weekly, on Wednesdays at 12h00, broken down into around 15 minutes for the presentation, followed by 15-minute discussion period. The initiative aims to strengthen collaboration between INE technicians who use the R programming language. Through presentations that show practical examples of using R in the workplace, participants share their tips and tricks, promoting valuable discussions for both presenters and participants. These sessions emphasize the willingness of all participants to contribute their time, share experiences, and seek assistance in the context of statistical production tasks at INE.

Although R is a powerful tool for statistical production and data science, the success of this language is also measured by the involvement of its users’ community. 1/2 houR probably plays a significant role in the professional development and learning of its useRs at INE, but its impact becomes even more pronounced in fostering connections among individuals from diverse units and teams. Without it, these people might encounter greater challenges in getting acquainted or discussing the projects they undertake collectively for a common good: the production of official statistics.

Speed Talk Session 6 - Experimental analysis & sources, June 6, 2024, 12:45-13:30

Evaluating the Accuracy of Official Statistics and Survey Data: The Case of Covid-19 Vaccination Rates in Germany

Ms Karolina von Glasenapp[1]

[1]Gesis - Leibniz Institute For The Social Sciences, Mannheim, Germany

The recent global health crisis has highlighted the critical role of statistics for society and policymakers. To guide effective public interventions, it was essential to ensure that the collected data were of high accuracy. At the same time, the unprecedented situation and the need for rapid data delivery posed additional challenges for data providers. This context raises the question of to what extent the requirement for accurate data was met.

In practice, the empirical evaluation of accuracy is often difficult due to the unknown true value and the lack of external benchmarks. This paper utilises the unique availability of different data sources on the same measure, namely the Covid-19 vaccination rates. Using a meta-analytical approach of a random-effects model, the study compares the official health statistics in Germany (considered a benchmark with a known amount of error due to reporting issues) with data from surveys of different designs (N ~ 30). First, the paper assesses the accuracy of survey estimates as the deviation from the benchmark data. Second, the key characteristics of the survey design, sampling and mode, are examined as potential determinants of accuracy. Hereby, the paper focuses on the difference between probability-based and nonprobability-based surveys, as well as between interviewer-administered and self-administered surveys.

Altogether, the paper compares two distinct data types, official statistics and survey data, and highlights the strengths and limitations of each approach. For survey data, it further identifies the survey design leading to the most accurate estimates. The results of the paper can help practitioners ensure efficient and high-quality data collection in the future.

The evolution of immigrant groups in Luxembourg -A Symbolic Data Analysis Approach

Catarina Melo[2], Paula Brito[2], Pedro Campos[1,2]

[1]Statistics Portugal, , Portugal, [2]Faculty of Economics, University of Porto, Porto, Portugal

Luxembourg is distinguished by its demographic dynamism among European countries.

From 2010 until 2020, around 80% of the population growth was due to migration. In general, the main reason why people immigrate to Luxembourg is labour. Considering the immigration phenomenon, this work examines the different immigrant groups in the labour market from 2014 to 2022 by analysing data from the Labour Force Survey (LFS)

of Luxembourg with Symbolic Data Analysis (SDA) [1] techniques.

Microdata was aggregated and 21 symbolic objects were created based on birthplace and length of residence in Luxembourg. They were primarily described by 16 modal variables which are multi-valued variables with a frequency attached to each category. Clustering algorithms were applied and the hierarchical using complete linkage demonstrated the greatest separation between clusters and homogeneity within clusters. The Heuristic Identification of Noisy Variables (HINoV) algorithm [3], suggests that with six variables the objects may be separated into groups with similar labour market profiles. The Monitoring the Evolution of Clusters (MEC) framework [2], monitors cluster transitions over time by identifying temporal relations between these structures. This was used to track the movement of population groups between clusters. The results show that by using just six variables it is possible to break down the 21 objects into groups with similar labour market profiles.

Furthermore, it is also possible to verify that people from the European Union (EU) and Neighbouring countries have similar profiles while the Portuguese have opposite characteristics. The Luxembourgers are in between. Profiling people from non-EU countries is challenging. Lastly, the MEC framework revealed significant object movements.

By combining the LFS and SDA, this work can be easily replicated in nations that use the LFS, enabling comparison of results and monitoring in the future.

A major part of the present work was developed through an internship at the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (Statec), at

the ‘March´e du travail et Education’ unit, under the European Master in Official Statistics (EMOS) programme.

Take the most of vital statistics with value for global health: the case of place of death

Dr Barbara Gomes[1,3], Sílvia Lopes[1,2]

[1]Faculty of Medicine, University of Coimbra, Coimbra, Portugal, , , [2]NOVA National School of Public Health, Public Health Research Centre, Comprehensive Health Research Center, CHRC, NOVA University Lisbon, Lisbon, Portugal, , , [3]. King's College London, Cicely Saunders Institute of Palliative Care, Policy & Rehabilitation, London, United Kingdom, ,

Vital statistics are valuable to improve health for all. One statistic is overlooked: place of death. Information on where people die is recorded in death certificates/registrations in all EU countries and others. It is crucial for health policy and planning, particularly to ensure good end of life care. We lead a European Research Council funded project aimed to put this statistic under spotlight and find ways to improve it.

We first studied place of death data from 32 countries (26 EU), provided by statistics institutions or health agencies. Analysing where 100 million adults died from 2012 and 2021, we discovered that home deaths rose during the pandemic in most countries. The rise was highest in women and cancer in most countries. Overall, 30.8% died at home (age-standardised 31.3%), ranging from 13.1% in Malta to 65.1% in Uganda. The results have important implications for how we organise future care, and drew attention of the media, patient/carer organisations and policymakers.

Home is indeed a central dying place yet not where most people die. This is important because most patients prefer to be cared for and to die at home, as shown in an umbrella review we published this year. Home is also the most consistent place of death category across countries, though there are nuances within (e.g., own home vs. home of a relative/friend) and limitations in some countries, where "home" includes assisted living facilities (e.g., USA) and some care homes (e.g., Portugal), depending on the interpretation of the death certifier. However, the scope of improvement of this statistic lies fundamentally elsewhere. Hospital is often merged with other health institutions, and places within are rarely specified (e.g., it can be very different to die in the Emergency Department, Intensive Care Unit or a Palliative Care Unit). This group accounted for 47.5% of all deaths (24.2% Netherlands, 74.5% Republic of Korea). Nursing homes and similar places are a heterogeneous group yet critical in ageing populations. This group accounted for 17.8% of all deaths in 20 countries (2.7% Slovakia, 35.7% Sweden).

We are now conducting ethnographic fieldwork in four contrasting countries: Portugal, the Netherlands, Uganda and the USA to examine policies, views of relevant stakeholders, and experiences of patients with advanced conditions and their families in the last months of life. This will lead to recommendations on how to improve the quality of place of death statistics towards an international classification of dying places.

Hungarian Truck Toll Mileage Index as a business cycle indicator

Dr Klaudia Máténé Bella[1], Gergely Attila Kiss[2]

[1]Hungarian Central Statistical Office, Budapest, Hungary, [2]Hungarian Central Statistical Office, Budapest, Hungary

The National Toll Payment Services Plc. in Hungary is a state-owned economic organization whose basic task is toll collection (e-vignette, e-toll), provision of related services, and control of road use rights. Heavy goods vehicle with a total weight over 3.5 tons are required to pay an e-toll. The control system of National Toll Payment Services Plc. operates 24 hours a day and 365 days a year. The toll control cameras installed on the toll road network recognize and record the data of vehicles passing through the intersection, which is important from the point of view of toll payment. The Hungarian Central Statistical Office receives the camera data of the National Toll Payment Service Ltd. in an anonymized form on monthly basis from 2019. The dataset contains the camera ID, GPS coordinates of every camera, recording time (year, month, day, minute, second), vehicle type code (e.g. car, truck, bus, etc.), the anonymized code of the license plate, as well as the vehicle's country code.

Based on the database examined between March 2019 and June 2023, the data on trucks were selected for every month. The anonymized codes of license plates were sorted by time, and then the code of the camera making the next recording was assigned to a given record. Thus, at a given record level, the distance covered can already be determined based on the two cameras. Based on the GPS coordinates of the camera data, a distance matrix was created which contains the distance of each camera from the other cameras. This distance matrix can be used to calculate the distance travelled for every anonymized license plate of trucks.

These calculated distances for trucks were aggregated for every month and a monthly time series, namely Truck Toll Mileage Index (TTMI) was created with a basis of monthly average of 2020. Finally, this index was seasonally adjusted using JDemetra+ software. As a result, it can be stated that both the raw and the seasonally adjusted TTMI show a high correlation (0.8) with the corresponding fixed base index of industrial production. This research succeeded in forming a business cycle indicator with high explanatory power.

Twitter as a data source for Official Statistics: analyzing Italian conversations on the Russia-Ukraine war

Maria Clelia Romano[1], Ms Fabrizio De Fausti[1]

[1]Italian National Institute of Statistics (ISTAT), , Italy

The large amount of information contained in conversations on Twitter (now X) makes them a very attractive source of data for studying opinions and attitudes about key topics in the public debate. The Italian National Institute of Statistics (ISTAT) has been considering as strategic the use of Twitter data in order to enrich statistical production and produce more timely statistical products. The project we will present is an initial step just in a plan for using Twitter data sources for Official Statistics. It has been designed since its initial stages with quality aspects in mind, following concretely a quality-by design approach. In the paper, we will present indicators and methods aimed to synthesize the information complexity of data collected from Twitter analyzing the conversations about the Russia-Ukraine war. We focus on changes in the amount of conversations and their contents. Through the application of descriptive analysis, we observed a substantial convergence of results in terms of a declining number of conversations about the war just after the first weeks since the starting of the conflict and a renewed interest seven months later (in September 2022), related to the occurring energy crisis and the consequent increase in prices.

We present also first result of a topics analysis showing main themes emerged in the debate on the war. We used topic-analysis techniques which highlights the impact of the war-related topics, among all the topics discussed by the users of Twitter, as well as the evolution of such an impact in time.

Both the quality controls and the results show that it is important to continue experimenting with various methodologies. Indeed, it has been shown that it is useful to combine different methods (words frequency, semantic filtering, topic analysis, etc.) for the study of the dimensions and trends of the phenomena of interest, as well as for a validation of the results. Studies and experiments like the ones presented here are very promising and poses initial foundations for the definition of a quality framework (e.g. by considering integration with surveys, evaluating the quality of keyword-based filters by computer assisted exploration tools, etc.). Sound and controlled methods as well as a full-fledged quality framework are necessary to properly exploit the results in official statistics production.

Validation of crypto-asset on-chain transactions - relevance, risks, and challenges for official statistics

Eldin Delić [1], Urszula Kochanska [2], Salim Talout Zitan [3], Laura Tresso [4]

[1] Deutsche Bundesbank (Germany), [2]European Central Bank, [3]Banque de France (France), [4]European Central Bank

With the steady evolution of crypto-assets into a more mainstream phenomenon, various crypto-asset services and related business models have emerged in this field. The validation of crypto-asset transactions, especially mining/staking pools and validation-as-a-service, provides an interesting case in this context. These constructs can be seen as entities raising capital from the public to generate a common return for investors. The validation of crypto-asset on-chain transactions requires and generates the equivalents of millions of USD daily – the amounts that so far are not captured in any official statistics. This paper aims to provide insights into the latest developments in this phenomenon and to progress towards closing the statistical gap.

The first part of the paper examines the current trends and specificities of the validation of crypto-asset transactions covering mining and staking on selected blockchains as well as in Decentralised Finance (DeFi).

As the validation of transactions is largely dominated by mining/staking pools, the second part of the paper analyses these based on their geographical incorporation, business model, and governance. The objective of the analysis is to explore the relevance of the mining/staking pools and their inherent risks.

The third part of the paper elaborates on the challenges to incorporate the validation of crypto-assets in macro-economic statistics. In this context, the paper elaborates on diverse approaches to measuring country specific output of validation services and the geographical distribution of the validation fees, also providing the respective first estimates. In the collection of new data required for the estimates, the paper utilises multiple data sources, and provides a detailed analysis of the related challenges.

Keywords: Distributed Ledger Technology (DLT), Blockchain, crypto mining and staking, Decentralised Finance (DeFi), official statistics

Speed Talk Session 7 - Quality & Data management, June 6, 2024, 12:45-13:30

The terminololy module in a centralised metadata system: a crucial contribute to quality

Ms Claudia Brunini[1]

[1]Istat, Roma, Italy

An effective centralized metadata system is based on three key components: the referential metadata module, the structural metadata module and the terminology module. Referential metadata module describes statistical processes and provides all the elements for quality assessment; structural metadata module describes the data involved in the processes assigning them the role and the relationships between them; the terminology module documents the semantic aspects. This work focuses on this last module.

The terminology resource has the role of detailing the semantics of the metadata, allowing in this way its identification. The availability of accurate and strict terminological resources guarantees an improvement in the quality of the statistical information. It also contributes to a clear identification of objects (units, populations, variables, etc.) and allows to overcome the constant problem of semantic overlap.

In particular, the terminological component represents a fundamental asset for preventing and managing the risk of specification error, which is relevant in the initial phase of the statistical production process. In this phase the concepts and dimensions, previously identified, are made operational in terms of detectable characteristics (variable), statistical units, target population and classifications as well as territorial and temporal dimensions. These aspects, if not carried out correctly, can cause severe consequences on some components of quality such as the relevance of the data and accuracy, affecting also the distortion of the estimates produced.

To support the phase of preparing the conceptual framework of the survey, the terminological module allows to search for existing semantic resources, providing documentation of the statistical objects to which they are associated. It also allows to analyze which processes use them and with which specific definition. For a better analysis, additional information is also made available to researchers, such as the process responsible for its maintenance and updating, the dates of validation and any end of validity, the list of user processes, regulatory references and any other changes. In this way, the researcher can implement a conscious selection (or possibly a new formulation) of the semantic resources, proceeding with a correct attribution to its statistical objects. The contribution illustrates in details how the functionality of the terminological component can effectively support the phases of the statistical processes where the specification error is particularly risky. How it can facilitate the researcher's work and promote the construction of metadata internally coherent, harmonized with national and international sources and complete from the point of view of quality indicators.

Data Management Process Roles in The Quality of Official Statistics

Mr Ali Abdelhamid[1], Mrs Rawia Wagih Ragab[1]

[1]CAPMAS, Cairo, Egypt

Official statistics play a pivotal role in informing policy decisions, guiding public discourse, and fostering societal development. The quality of these statistics is inherently linked to the effectiveness of data management processes employed throughout the statistical lifecycle. This research aims to investigate the nuanced roles that various data management processes play in shaping the quality of official statistics. Through a comprehensive review of existing literature, the study will establish a theoretical framework that outlines key data management processes, including data collection, validation, cleaning, storage, and dissemination. Employing a mixed-methods approach, the research will assess the impact of these processes on data quality by developing and applying relevant metrics. Case studies and examples of successful data management strategies will be analyzed to identify best practices and potential areas for improvement. The study also aims to explore challenges associated with implementing robust data management practices and investigate opportunities for innovation, such as the integration of emerging technologies. The findings of this research will not only contribute to the academic understanding of the interplay between data management processes and statistical data quality but will also offer practical insights and recommendations for policymakers and statistical agencies seeking to enhance the reliability and credibility of official statistics.

How Statistics Sweden uses the Department of Data Management to Ensure Access to High-Quality Data

Ulf Durnell[1], Kristina Strandberg[1]

[1]Statistics Sweden, , Sweden

In 2021 Statistics Sweden underwent a complete re-organization. At the wake of a pandemic a change of that magnitude could have been seen as a big risk, but it was the firm belief of management that some changes were crucial for Statistics Sweden to be successful in its mission – to keep supplying Sweden with accurate and reliable statistics. One of the core changes in the new organization was the establishment of a Department for Data Management – a department with the responsibility to secure and streamline the process of data management for the entire organization, and to ensure the continued access of high-quality data.

It might sound like a contradiction, but one of the biggest challenges facing statistical institutes today is the rapid growth of available digital data. From responders, there is pressure to submit data digitally, through machine-to-machine solutions. Swedish authorities also are tasked to make it easier for companies to respond. Digitalization of business systems and registers, constantly adds to the list of potential data sources, from both private actors and other government agencies. Add to that the possibility to use data from other administrative sources, such as mobile network data. Considerable development of data collection methods and data management processes is needed to stay on top of the digital transformation, and ultimately – to stay relevant in an ever-changing world.

During the first two years of operation, the work at the Department of Data Management has been concentrated to two things: developing automated and efficient processes for data collection and data management, and to enabling wider use of existing data. This involves creating technical solutions, as well as straightening out legal and data security related problems. Statistics Sweden now has standardized processes for Monitoring of Potential Data Sources, Approval of new Data Sources and a Process for Register Production. To support data management, a new IT platform with related metadata system has been designed and is taking form.

In conclusion, giving the Department of Data Management the mandate and responsibility to supply the organization with high quality data has helped Statistics Sweden to bring about several necessary changes in our processes for data collection and data management. In this paper, we present the new department, and highlight the important work that has been fundamental to achieve these changes.

Speed Talk Session 8 - Statistics and decision-making / user engagement, June 6, 2024, 12:45-13:30

Assessing the quality statistics on Catalonia by meeting user needs

Dr Enric Ripoll[1], Mr Jordi Galter[1]

[1]Statistical Institute Of Catalonia, Barcelona, Spain

In this presentation we analyse the available empirical evidence on user ratings of official statistics in Catalonia. Assessing the quality of official statistics, a public good, requires measuring its multiple dimensions, therefore it is essential to rely on the assessment of its users. According to the principles of the European Statistics Code of Practice, relevance, accuracy, timeliness, punctuality, comparability, coherence, accessibility and clarity are the dimensions of the quality framework of the European Statistical System. On the other hand, the current concept of quality reinforces the consideration of official statistics as a public good and, as such, it is essential to have the opinion of its users.

To obtain this valuable information, that provided a broad external assessment of the content of Catalonia's Statistical System, a structured consultation was carried out during the spring of 2021 with the participation of more than 130 different types of users who had to evaluate each statistical action and provide advice regarding the identification and prioritization of possible new contents to be included in the next Statistical Plan of Catalonia. We also consider analogies with the consultation carried out by the Spanish High Council on Statistics in 2011 and with the annual user satisfaction surveys conducted by Eurostat.

The consultation collected suggestions for new statistical operations recommended by experts. In this context, the bulk of new economic statistics focus on economic accounts, emphasizing foreign economic relations and new satellite accounts. In turn, measures of mobility, increasing spatial disaggregation and raising the timeliness of data focus the proposals on demographic statistics. On the other hand, specific suggestions in the fields of education, health and social protection, where new statistics would focus on opportunities to enrich current statistical information with a combination of other non-statistical sources, also deserve attention. Finally, the suggestions made in the attention to the research activity of third parties emphasize the convenience of accessing (micro) data of statistical and administrative origin, from speeding up their availability and data resulting from processes of integration of multiple sources.

DEPP@Scribe project: Better structuring and documenting education data & taking action for research and innovation

Axelle Charpentier, Alexis Lermite, Ms Axelle Charpentier[1]

[1]Depp, Paris, France

The DEPP@SCRIBE project was born out of the desire to better structure and document the statistical information systems of the DEPP (which is the statistical departement of the French ministry of education) in the light of their growing importance in recent years, particularly with regard to monitoring the careers of pupils and staff and measuring skills via standardised national assessments. It is also a project that makes it possible to meet the commitments on the quality of administrative data in the official statistical service, the DEPP's missions and the challenges posed by the evaluation of public policies.

DEPP@SCRIBE is part of a European and international drive to promote access to administrative data in order to improve the quality and quantity of studies and research on the education system and evaluative work.

It is temporarily supported by the IDEE programme (Innovations, Données et Expérimentations en Education), supported by the National Agency for Research (ANR) and led by a team from the J-Pal Europe/ENS laboratory.

DEPP@SCRIBE has several objectives:

- Increasing the value of data and make better use of statistical production by facilitating access to the wealth of data produced by the business offices;
- Meeting quality requirements in line with European statistical best practice;
- Encouraging studies on education carried out by the DEPP (and other players such as researchers) thanks to easier access to data;
- Responding more effectively to requests from the scientific community and thus save time for the production of studies;
- Helping to inform public debate on education and political decision-making.

To achieve these objectives, the project plans to develop two new digital tools which will be integrated into a single portal:

1. Depp@logue, the DEPP's catalogue of education data, to showcase the wealth of statistical information systems, provide a better understanding of study and research opportunities, and build innovative and relevant study protocols (the catalogue will not be the point of access to the data);
2. Depp@thèque, the DEPP's education data library, a digital platform for accessing data and carrying out studies based on the connection of administrative sources, or even on the enrichment of survey data collected independently (it will be an information warehouse equipped with calculation servers)

The project should be completed by the end of 2024. Access to the platform will be free of charge and open to any research teams that have signed an agreement with the DEPP.

Inter-agency collaboration - how does it result in an improvement in the quality of child welfare statistics?

Unni Beate Grebstad[1], Svetlana Beyrer

[1]Statistics Norway, Oslo, Norway

Inter-agency collaboration - how does it result in an improvement in the quality of child welfare statistics?

Statistics Norway (SSB) has constructed a novel data design for the compilation and storage of child welfare statistics through a system known as DigiBarnevern (DBV). This innovative system enhances the frequency of data flow pertaining to child welfare services by leveraging cloud computing infrastructure, specifically Google Cloud.

Collaborative efforts between SSB and The Norwegian Directorate for Children, Youth and Family Affairs (Bufdir) have driven the development of the advanced data reporting and storing. This paper aims to present SSB's experiences in inter-agency collaboration particularly regarding the professional input to the acquisition of data for statistical purposes. The most important aspect regarding the quality of data from this new reporting system is the continuous reporting that includes validation response in real time as well as the follow up of mistakes that the municipalities are reporting. Understanding the new reporting system and getting control of their own statistics have been important for the child welfare services.

Furthermore, this study will delve into broader aspects of the DBV project, encompassing both its limitations and accomplishments. The exploration of SSB's role as an information technology (IT) system supplier for an administrative authority (Bufdir) and, indirectly, for municipalities will be a focal point. Additionally, the paper will examine the experiences related to managing security, legal considerations, and confidentiality issues within the DBV framework.

Anticipated for 2024 the production of official statistics will encompass data from both the old and the newly developed reporting system. SSB acquires data from many other municipal services, not only from the child welfare services. The paper will therefore highlight the possible adaptability of the DBV project's products for application in other municipal services.

Keyword: child welfare statistics, cloud computing, collaborative efforts, pilot to production, data quality.

Towards modern user-oriented dissemination – chances and challenges for statistics

Ms Magdalena Ambroch[1]

1Statistics Poland, Warsaw, Poland

Robust information has become the most desired good in the modern world. In the flood of easily available data, official statistics is still distinguished by high quality, transparency and credibility. However, without a new approach to data dissemination and communication, it may be difficult to maintain the leading position in this area. Modern dissemination and communication of statistics should answer the wide scope of the needs of current data users: the demand both for open specialized data, and for straightforward and attractive information addressed to the wider audience. The presentation will show how Statistics Poland tries to modernize its practice to keep the interest of data users and meet their diverse needs. Statistics Poland provides more and more user-oriented products, like profiled databases (including newly launched Knowledge Databases), digital tools and publications. Our experience shows how important it is to determine a strategic plan for modernizing dissemination, including the need to build the understanding and openness to change within the staff.

Urban public policy models as a reaction to the analysis of natural flows.The example of gravity field andUrban circular economy

Michał Kudłacz[1]

[1]Statistical Office In Krakow, Krakow, Poland

This article will concern the analysis of urban public policies. There are many phenomena that currently play a great role in the context of programming urban development policies: globalization, metropolization, digitization of the economy, climate change, economic crises and shortage crises, migration crises and the development of artificial intelligence. These are the main challenges also for urbanized spaces. Urban policies can also be understood as an attempt to complement and strengthen the natural processes occurring in the economy.

This article addresses the issue of how to respond to changes in urban economies, which are, among others, the effect of the previously mentioned megatrends and which are described through quantitative indicators. The aim of this article is to indicate to what extent the authorities of selected cities demonstrate awareness and willingness in urban policies to face contemporary challenges, thanks to information about changes taking place in local economies and challenges, information about which comes from quantitative indicators. The article will present indicators describing the phenomenon of urban circular economy in selected Polish cities, as well as the impact of these cities on the development of the regional environment. The key issue is to compare quantitative indicators describing these phenomena with the development policies of these cities, which will allow the verification of the following research hypothesis: Urban policies take into account key, changing development challenges that result from natural processes described using quantitative data.

Quality of short-term labour market indicators

Ole Villund[1]

[1]Statistics Norway, , Norway

Recently, the wide-ranging economic consequences of major events, such as financial crises, pandemics, and wars, have underlined the need for speedy and robust labour market information. The present study compares different series of short-term statistics in order to discuss their practical quality, i.e. how useful they are for monitoring the Norwegian labour market. Specifically, we compare monthly Labour Force Survey data consisting of trend figures, seasonally adjusted and unadjusted figures, register-based employment statistics consisting of preliminary and final figures, each seasonally adjusted and unadjusted. Additional relevant data are production figures from national accounts as well as vacancy statistics, which are published quarterly.

Arguably, different quality criteria such as frequency, timeliness, and accuracy are all important to the users, but typically there is a trade-off between them. This trade-off is shaped by the interactions between statistical producers and users. Statistics Norway keeps regular contact with key users such as governmental bodies and non-governmental organisations, to discuss results as well as methodological developments. The recent pandemic and its impact on the Norwegian labour market accentuated the need for useful short-term indicators.

This study aims to discuss quality from the perspective of users, notably the additional information gained from monthly figures compared to quarterly (frequency), and from preliminary figures compared to final figures (timeliness). Furthermore, we compare multiple different time series, rather than trying to measure “error” by assuming there is a “correct” value (accuracy). Also, the merits of seasonal adjustments and smoothing methods are discussed to some extent. The goal is to contribute to a more user-friendly quality reporting that also benefits expert users.

Speed Talk Session 9 – Coordination, June 6, 2024, 12:45-13:30

Coordination within the NSS: challenges and opportunities in national statistical system

Gjurgjica Miloshevska[1]

[1]Statistician On Energy Statistics, skopje, Macedonia

Coordination within the NSS: challenges and opportunities in national statistical system

The National Statistical System is organized through written documents that cover procedures, processes that should be conducted while collecting statistical data, analyzing ang publishing. NSS has an obligation to create a system that will provide a continuous influx of statistical data from numerous sources, administrative bodies, private companies, etc. All these should be determined via one written procedure established by methodologies, regulative, programs, strategies, etc. The same system consists of National Statistical Institution, national governmental organizations stakeholders of statistical system, like National Banks, Ministries, other governmental bodies, all responsible for conducting, processing of statistical data.

When we talk in current days about challenges, we need to see it in the frame of current societies equipped with modern technologies that will enable to all of us participant in statistical system more efficient surveillance of conducting of data, noticing on time deficiencies, reacting on time to provide continues flow and providing data to users.

When it comes to challenges, we need to stress that statistical data are of interest to governmental establishments and they have interest in providing data that could be always interpreted in a way suitable for them, i.e., present current economic situation via statistical data adapted to their need. Therefore, is always important as a leading head of statistical institution to be put professionals and not politically adequately chosen persons. Having in mind that each institution is a hierarchical organization with vertical and horizontal systematization, putting professional and skill full person will be always of great importance.

The challenge is also to make quality planning in staff, financial expenses, good long-term planning is of essential importance for providing sustainable data.

To foresee all changes that could occur in societies and will have influence on future production of data also need as a prerequisite to have analytical tools, skillful persons capable of doing analysis, participation in strategies on national and international level related to long-term planning and creating strategies.

When it comes for using informatics technology, we could mention development of artificial intelligence, its fast development, we need always to be cautious about possibilities of losing a significance of official data, considering that various private organizations exist that collects data on various terms.

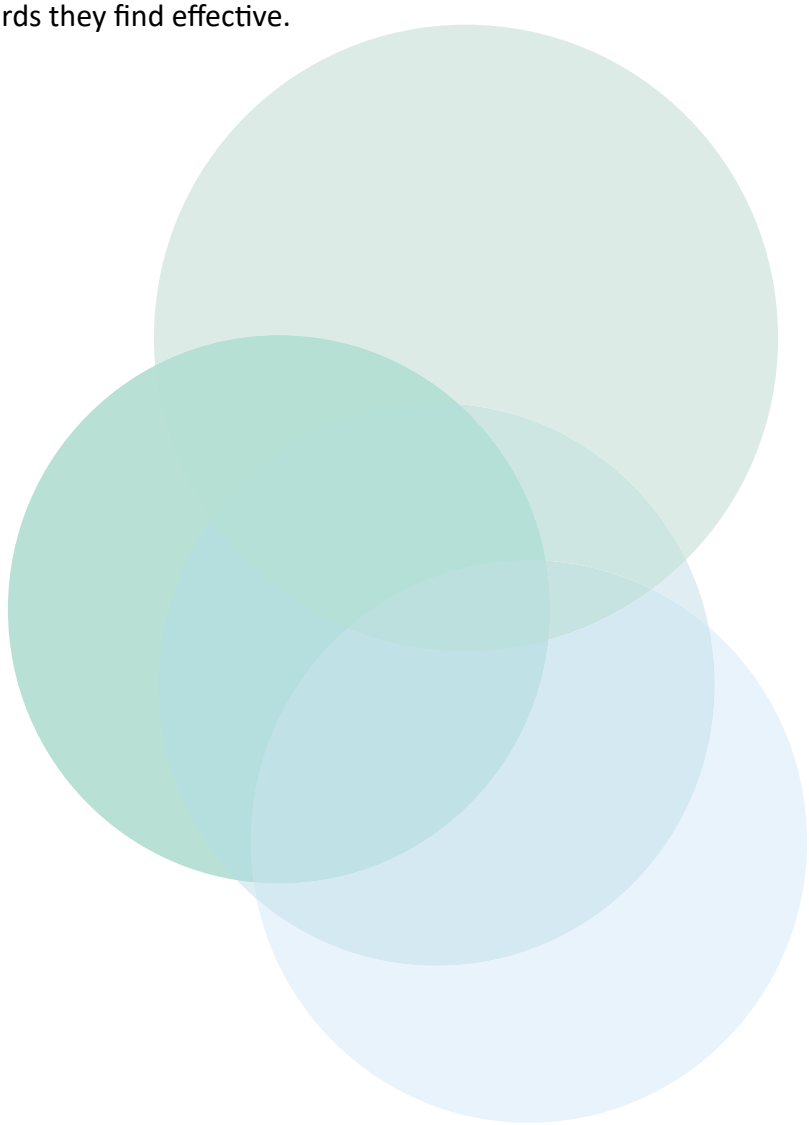
U.S. Census Bureau Quality Standards

Paul Marck[1], Sallie Keller[1], Danielle Neiman[1], Steve Klement[1]

[1]U.S. Census Bureau, Washington, United States

Statistical legislation and quality standards are evolving to be more customer focused. The U.S. government's 2018 Evidence Act brought about significant changes to coordinate the U.S. Federal Statistical System including creating a council of Chief Data Officers, a presumption of accessibility for data, and expanded access to data. As part of the agency response to these changes, the 2022 revision to the U.S. Census Bureau Quality Standards introduced categorizations for Core, Experimental, and Research based statistical products. These changes are shifting a paradigm toward users producing their own products. Statistical agencies may not be in complete control of all the products that are produced.

This paper will explore the evolving landscape and future updates for quality standards in progress with more of a customer focus. Our intent is to refocus standards to put statistical products first and allow for more rapid product development. The goal is to be able to go from idea to published statistics in months instead of years. We will discuss issues we need to navigate include controls for what is considered a statistical purpose and how to streamline our processes while still maintaining product quality. The goal of this paper is to elicit feedback from other statistical agencies in multiple countries on what legislation and standards they find effective.



Quality-important “brick” in the national statistical system in the Republic of North Macedonia

Ms Mira Todorova[1]

[1]State Statistical Office, Skopje, Macedonia

State Statistical Office of the Republic of North Macedonia in the past years worked on establishing framework for quality management and significant progress was done in this field. But to be recognized as producer and coordinator of official statistics it is of crucial importance to ensure the quality to be recognized in other producers of official statistics.

The starting process was checking compatibility of their production with Eurostat Guidance note concerning Other National Authorities (ONAs) in the process of adoption of Five-year Programme for Statistical Surveys.

Within IPA 2019 Project for Quality management, SSO has started the work for communication of quality to ONAs and production of first quality reports in ESMS format.

Quality reports were selected as staring point for quality work with ONAs, because these Reports are milestones for assessing the statistics produced in national statistical system. In addition, the quality is seen as a very important mark for differentiation of reliable statistics from “fast” statistics available now in digital society.

The work done within this Project showed us that we must work more intensively on developing quality framework in ONAs in the preparations for the next round of Peer Review of official statistics in the Republic of North Macedonia.

A shared pathway for quality in the Italian National Statistical Network

Ms Paola Giordano[1], Mr Andrea Bruni[1]

[1]Istat, Rome, Italy

In recent years, Istat, in its institutional role of coordinator of the National Statistical System (NSS), has undertaken several initiatives aimed at fostering the quality of official statistics produced in the italian data-ecosystem-driven environment, such as the release in 2022 of a new edition of the "Italian Code for the Quality of Official Statistics", a document mirroring the European Statistics Code of Practice (ES CoP), whose target recipients are the NSS entities other than the Other National Authorities (NSS-not ONA).

With the ambition of pursuing full compliance with the quality framework of the European Statistical System, Istat has followed the recommendations of the NSS key institutions, namely the Committee for Policy and Coordination of Statistical Information (Comstat) and the Commission on Quality Assurance of Statistical Information (Cogis): a supplementary document, the "Guide for the implementation of the Italian Code", has been defined.

Indeed, the Guide can be considered as the Italian version of the ESS Quality Assurance Framework (ESS QAF), adapted to the national context, being the NSS an highly decentralised composite network with more of 3,000 entities. Therefore, the purpose of the Guide is to support the statistical offices of the NSS-not ONA bodies (mostly Municipalities, but also Ministries, Regions and Provinces) in the interpretation and implementation of the Code, with a set of 301 methods: recommendations, practical suggestions, best practices.

The national Quality Assurance Framework will serve as the operational layer, just as the release of the ESS QAF was intended to assist the implementation of the ES CoP. In fact, for the generic NSS statistical office the method may represent a benchmark to be achieved and also an opportunity to stand its role within the Public Administration it belongs.

Istat researchers working in the Quality team and in the NSS coordination division jointly developed the Guide, and the building up of the document has followed a plan made of several steps:

1. assessment of the internationally existing Quality Assurance Frameworks;
2. analysis of the NSS protocols, Comstat directives, Cogis opinions, and current national legislation;
3. a strategic open approach (circular top-down and bottom-up process), with a massive involvement of the NSS-not ONA relevant stakeholders: draft versions have been shared, devoted workshops have been organized.

The Comstat finally approved the final draft of the Guide in May 2023, then published on both the Istat and the NSS website last October, in italian language only, for the time being.

Engaging external entities towards a better data ecosystem in Abu Dhabi using the statistical maturity index

Dr. Yu Sapphire Yu Han[1], Mahmood Belselah[1], Nasser Dayan[1]

[1]Statistics Centre Abu Dhabi, , United Arab Emirates

Statistics Centre Abu Dhabi (SCAD) is responsible for official statistics in the Emirate of Abu Dhabi. SCAD provides technical supervision of statistics and statistical data ecosystems for government entities. In addition, SCAD is transitioning towards using admin sources to produce official statistics in recent years. A generic challenge in using administrative registers for statistical purposes is that the data in these sources are collected and maintained by other government entities for non-statistical purposes. This not only makes a statistics office highly dependent, it may also affect the quality of the statistical output. The key question for SCAD is the quality assurance of external entities' statistical data system as well as their delivered admin sources.

To answer the question, SCAD has been carrying out the statistical maturity index (SMI) program. The SMI is an annual self-evaluation model that contains 44 items. It aims to measure the compliance of Abu Dhabi Government Entities in terms of applying the laws and legislation regulating statistical work, building statistical capacity and applying statistical quality standards. Within the annual SMI's comprehensive approach on the statistical system, special attention has been paid on the granularity of admin sources provided by entities. Each quarter, SCAD applies quality measures that evaluate the accuracy, timeliness and accessibility dimensions on received admin sources. The SMI is weighted average of quality scores (out of 100%) and assigned to participating entities as a KPI. The SMI program provides assessment reports concluding the findings, improvement points and timeline for entities. Entities' progress towards quality improvement is closely monitored and supported by SCAD. To better educate and engage entities in the field of statistical quality, SCAD also conducts workshops to discuss relevant aspects of the project including the criteria and the timeline.

At this moment, there are 28 Abu Dhabi Government Entities participating in the program. The overall SMI has reached 10% increment in 2023 compared with 2022. Moreover, the overall admin source quality has reached 3% increment in the second quarter of 2023 compared with the last quarter of 2022. SCAD has received positive feedback from external entities about the usefulness of the SMI program for a better Abu Dhabi data ecosystem. In addition, the engagement with external entities also enhanced the utilization of the admin sources among SCAD production teams.

Challenges and considerations in implementing quality management institutional framework

Ms Held Curma[1], Mr. Denis Kristo[2]

[1]Institute of Statistics, Tirana, Albania, [2]Institute of Statistics, Tirana, Albania

Over the recent years, the Albanian Institute of Statistics has undertaken comprehensive initiatives to implement the European Code of Practice across various dimensions. A pronounced emphasis on quality has been instated, transcending the purview of the National Statistical Institute (NSI) to encompass the broader National Statistical System (NSS). Particular attention has been directed towards advocating for and implementing a robust quality management framework. The institute is actively pursuing the strategic imperative of leveraging administrative data for statistical production as a means to enhance cost-effectiveness.

This paper endeavours to expound upon the challenges entailed in advancing quality management within the NSS and extending its reach to Other National Authorities (ONAs) engaged in the production of official statistics. It will systematically explore considerations pertaining to the requisite institutional framework for ensuring quality assurance and cultivating trust among users.

While the initiative may encounter impediments owing to divergent institutional frameworks, the involvement of many stakeholders, heightened workloads, or shifts in organizational culture, its implementation is envisaged to yield enduring benefits for both data producers and the use of data in official statistics.

Italian National Statistical system: Quality Reporting matters at every level

Andrea Bruni[1], Ms. Paola Giordano[1]

[1]Istat, , Italy

In recent years, Istat, in its legislative role of coordinating the National Statistical System (Sistan), carries out initiatives to support the improvement of the quality of statistics produced within the Sistan, integrated into the quality policy adopted by Istat and consistent with the international reference framework on data and metadata quality in official statistics.

Regarding metadata quality, in 2018 the Steering and Coordinating Committee for Statistical Information (Comstat), Sistan's main governance body, released the Guideline Act No. 3 and in particular Article 4, which states that the clear and comprehensible dissemination of data requires metadata describing both the process of production and the characteristics of the statistical results obtained.

Additionally, in 2021 both the Comstat and the Cogis, short for Commission for the Quality Assurance of Statistical Information (Sistan's main vigilance body), suggested that Istat should implement two documents related to quality, the target recipients being the Non-Other National Authorities entities. The Italian version of the European Statistical System (ESS) Quality Assurance Framework has been published in October 2023, and the current work is aimed at defining an Italian "Handbook for Quality and Metadata Reports", version of the "ESS Handbook for Quality and Metadata".

Such Handbook may be considered a strategic operational document to guide Sistan entities toward the production of standardized quality reports on statistical processes and outputs, for facilitating the comparability of the meta-information among statistical processes and among Sistan entities.

Therefore, the Handbook will serve as a reference, within the Sistan framework, for the proper collection of the set of metadata, accompanying the dissemination of aggregated data for their proper use and interpretation.

The drafting of the Handbook is going to be accompanied by the definition of the customized template serving the reporting purpose. Such template, based on the ESS standard SIMS (Single Integrated Metadata Standard), is planned to be used as a Metadata Structure Definition in the Istat SDMX Toolkit, so to publish quality reports as SDMX-compliant messages.

For all these work strands, a Task Force has been set up. It is following several steps: 1) analysis of existing practices at Sistan entities on quality reporting, in particular structural and reference disseminated metadata; 2) mapping of the metainformation available at Sistan Entities against the SIMS concepts; 3) design of the items to be included in the standardized quality report; 4) preparation of the Handbook with the operational guidelines for the Italian quality report to be compiled.

Speed Talk Session 10 - Improving business statistics, June 6, 2024, 12:45-13:30

Implementing Nowcasting Techniques for Timelier Publications

Mr Gergely Attila Kiss[1], Beáta Horváth[1]

[1]Hungarian Central Statistical Office, , Hungary

The goal of this project is to be able to provide timelier information to policymakers and other stakeholders about three key indicators of the economy Producers Prices in Industry index (PPI), Producers Volume in Industry index (PVI) and Services Producers Price index (SPPI). The current Short Term Statistics (STS) publication dates at the Hungarian Central Statistical Office (HCSO) come with high latency (ranging between t+30 and t+90). Whereas, this project is to produce new estimates with using the available and new data sources to produce indices in month (t+0 days) up to t+20 days for the above three. This could make stakeholders increase their agility, as the estimated indicators, would come significantly sooner.

For this we developed a nowcasting tool that we could test on the European statistical Awards' Nowcasting competitions. The tool was used to nowcast monthly PPI index and without any specifications for the needs of the countries. It produced good and consistent results for several of them. Also, this tool was developed to be easily applicable for other time series. It could already produce PVI estimates in the last months of the competition and further modifications to reach SPPI estimations can be easily made. Our idea was to use standard supervised machine learning models (random forest, ridge, lasso, etc.) that use the World Bank Commodity Price Data as explanatory time series to nowcast the EU countries' indices.

The tool's algorithm works as following: enriches the data with standard transformations of lagging and differencing the time series to create a large pool of possible explanatory variables. Then, it uses several different feature selection procedures to create a more concise pool of variables. After the pool is complete, it starts the hyperparameter tuning using different sample sizes, estimation window types and one-step ahead forecasts for cross validation.

We plan to further enhance the tool's efficiency with domain expert knowledge to tailor for the Hungarian needs. Our expectation is that it will dramatically increase the efficiency of the algorithm while decrease it's timeliness. The change in both is expected due to the reason that in the competition we had to provide in month (t+0) estimates and that created a strong constraint on the available data. Analog to the balance-variance trade-off, we plan to find the balance between timeliness and accuracy that provides the optimal solution.

Developing price statistics for internationally traded services – practical experience from Statistics Sweden

Mr Mikael Nordin[1], Björn Forssell, Marcus Fridén, Erik Hillström

[1]Statistics Sweden, , Sweden

Sweden, as a small open economy heavily reliant on international trade, has witnessed a significant surge in the trade of services over recent decades. In 2019, the ratio of trade in services to GDP rose to 29%, up from 10% in 1980, highlighting the growing importance of this sector. Consequently, the interest in expanding the coverage of the economic statistics has increased, specifically considering the area of price statistics. Producer price indices for internationally traded services is a relatively unexplored part of the economic statistics, which only a few national statistics institutes compile and publish on a detailed basis. Statistics Sweden have compiled service producer price indices (SPPI's) for exports and imports since 2020.

This paper aims to share Sweden's experiences and success in the development and compilation of SPPI's, focusing on the challenges encountered in creating a robust data frame for sampling and weights calculations from multisource statistics. A major obstacle faced was the lack of classification concordance between SPPI and the trade value data source, which adhered to the EBOPS classification, while the SPPI followed a national version of CPA. Bridging this classification gap was crucial for establishing a high-quality data frame, aiding in the classification of data at the lowest possible level.

The paper also addresses the fundamental challenges in identifying services and establishing representative price measures for SPPI's. Key emphasis is placed on the intricate process of linking the national versions of CPA and EBOPS classifications, providing a detailed description of this crucial step.

Insights shared in this paper contribute to the broader discourse on improving statistical methodologies, particularly in the context of international trade in services and the quality of official price statistics. Official price statistics are also widely used as input in other areas of the official economic statistics. For instance, SPPI's are used to calculate GDP in constant prices, enabling users to separate between price and volume changes in the nominal GDP figures. Consequently, high-quality price statistics extends beyond the intrinsic value for the individual statistical survey, as it also is an improvement for the overall quality of official economic statistics.

On the use of Value Added Tax data in creating a timely monthly Production Value Index

Ms Annika Lindblom[1]

[1]Statistics Sweden, Örebro, Sweden

The Production Value Index (PVI) is a survey that measures the development of the production of goods and services in the business sector. Turnover is the variable which serves as an indicator of production value (except for two industries where production volumes are used).

Detailed definitive PVI are published quarterly, approximately 45 days after the end of the reference quarter. A preliminary (and more aggregated) monthly PVI are published no later than 40 days (29 in the retail trade) after the end of the reference month.

Statistics Sweden strives to improve quality in the statistics, reduce response burden as well as production costs. Definitive PVI are, since reference month April 2015, based on turnover values for small and medium sized enterprises taken from the Swedish Tax Agency's Value Added Tax (VAT) register. However, when it comes to the preliminary monthly PVI the problem of timeliness arises. Many small and medium sized enterprises declare VAT too late to be part of the production process needed to meet the fixed publication dates. In addition, small and medium sized enterprises often are allowed to declare VAT quarterly which means that distribution by month must be modelled.

During recent years focus on the use of administrative data has increased at Statistics Sweden. Consequently, a project started to explore the possibility of replacing direct collection with VAT-data also in the preliminary monthly indicator. In a first step, the manufacturing sector were studied since the structure of the service sector with many small enterprises makes it more complicated in this context.

Conclusions from this project led to a preliminary PVI where turnover values from small and medium sized enterprises in the manufacturing sector is taken from VAT-data. Preliminary monthly estimates, based on this new method, have been published since reference month April 2023. This paper will focus on experiences and results from the project as well as on an evaluation of the quality of estimates produced by the new method. Furthermore, Statistics Sweden has decided to continue with the service sector and hopefully be able to identify industries where replacing direct collection with VAT-data would be feasible. Experiences from the service sector will be included in this paper.

Improving quality in seasonal adjustment in Short-Term Statistics using JDemetra+ regressors and TEAM R-package

Cruz Gómez[1], Ms Elena Rosa Pérez[1], Dr Carlos Sáez Calvo[2], Luis Sanguiao Sande[2], Félix Aparicio Pérez[2], María Teresa Vázquez Gutiérrez[3], José Fernando Arranz Arauzo[3]

1S.G. Short-Term Statistics, Statistics Spain, , Spain, 2S.G. for Methodology and Sampling Design, Statistics Spain, , Spain

Short-term business statistics (STS) are the earliest statistics released to show emerging trends in the European economy. Monthly and quarterly STS provide data for the main economic sectors: industry, construction, trade and services, excluding financial and public services.

STS Regulation requires data that are calendar adjusted and calendar and seasonally adjusted, in addition to unadjusted data. Seasonal adjustment (SA) procedures eliminate the estimated seasonal and calendar effects from the original time series and obtain SA-estimates that are likely to reveal what is new in a time series.

JDemetra+ is the seasonal adjustment software officially used in the European Statistical System. It uses a model-based TRAMO-SEATS approach for performing seasonal adjustment of time series. In this approach, a RegARIMA model is fitted to the series. It also offers the chance to calculate regression variables to model calendar effects, including trading days regressors that take into account the composition of the days of the month.

ARIMA models used to adjust STS time series play a very important role to obtain accurate adjusted data. But sometimes the work of updating the ARIMA models can become a burdensome task as the manual identification of a suitable model can become complex and time-consuming and automatic procedures can provide a model without taking into account some restrictions considered as essential for the domain expert.

Time-Series Exhaustive Automatic Modelling (TEAM) is an R package developed by Statistics Spain, and based in the JDemetra+ ecosystem, that can help in the process of updating ARIMA models by providing a list of models ordered according to a global score.

Methodologically, we can set a priori specifications (outliers, calendar regressors, maximum/minumum values of the ARIMA parameters), and the local scores by hierarchical levels can help us guarantee the quality. In this sense, one of the main advantages of TEAM is that the ARIMA models provided can be subject to some restrictions specified by the domain expert.

To carry out calendar adjustment at Statistics Spain we have been using customized working days regressors. However, for time series of specific activities the residual effects of trading days were not being completely removed. Several analyses have been undertaken and improvements have been achieved when using JDemetra+ regressors.

Due to the increase in quality, the JDemetra+ regressors and the use of the TEAM package to update the ARIMA models are included in the seasonal adjustment process as from January 2024 in STS at Statistics Spain.

Quality aspects in a re-established system of inter-connected multi-source construction activity statistics

Carsten Schumann[1], Marianne Schepers[1]

[1]NSI Germany (destatis), Wiesbaden, Germany

The European Systemic Risk Board has identified data gaps in construction activity statistics and new European legislation is underway to meet this demand: Quarterly statistics on construction starts and works completions will complement existing statistics on building permits. The coverage will be extended to commercial real estate in addition to residential real estate. The target variable is usable floor area in new buildings. In Germany, this has to be incorporated in the existing system of construction activity statistics, which also needs to provide structural information for National purposes, such as social aspects (share of social housing, number of rooms per dwelling, number of dwellings per building), environmental/urban planning aspects (building material, energy sources for heating, sustainability measures, footprint, type of building, construction in new or existing building) and economic aspects (construction cost, time to start and complete the individual construction, institutional sector of the builder). Furthermore, monthly data-points are requested and temporal precision needs to be improved (better assignment of events to the precise reference periods).

This presentation outlines the necessary combination of different approaches to ensure a comprehensive quality framework for this inter-connected system of business cycle indicators and structural statistics: (1) a new seamlessly digital dataflow with stronger focus on availability of administrative data, which also includes influencing the harmonisation of local administrative data structures to comply with statistical needs and reduce burden on respondents; (2) a better integration of official statistical institutes of City-administrations into the data flow for data validation purposes and to harvest local expertise, which also includes a single-source-of-truth-database for all statistical publications from cities, states and the NSI to publish consistent results; (3) an enriched annual survey of all granted permits to identify undeclared construction starts and works completions; (4) regular matching with a newly designed statistical register of addresses and the statistical business register where new registrations hint to undeclared works completions; (5) artificial intelligence to identify undeclared construction sites on satellite/aerial images; (5) data mining in privately held big data on real estate advertisements to identify undeclared works completions; (7) a smart system of automated regular reminders to respondents that point out the obligation to report their construction start and works completion, including a reference to possible fines; and finally (8) a published revision calendar to include all of the above mentioned data sources in the respective reference periods, which may lead to methodological improvements to reduce revisions after systematic analysis.

Development of an application to assure the quality of the monthly production of the Austrian HICP

Mr Manuel Koller[1], Alexandra Schindlar[1]

[1]Statistics Austria, Vienna, Austria

Consumer price indices (CPIs) are key indicators for the monetary policy and hence reported to the national banks and the central banks. The National Statistical Institutes provide each month the inflation rate, which is the outcome of the consumer price survey and hence the input for the monetary policy.

Starting from the task to develop quantitative indicators for the CPI price survey the authors developed an application for validation of the price survey and the computation of the inflation rate with the aim to minimize errors in the price survey and embedded the work in the framework of quality assurance as theoretical basis. The quantitative indicators were developed within and classified to four different aims: The overview over the price observation per month and over time, the increase of the quality of the price survey, the detection and minimisation of errors and the improvement of the sample itself.

The resulting application allows to monitor the validation process within the production of monthly results on the one hand. Moreover, the defined quantitative indicators are developed on different stages of the production process: The price observation itself, some transformations of the observed prices and on the resulting index numbers, as well as on monthly and annual rates of change and on contributions to the inflation rate. The application itself mimics the validation process that was done with less automatic support prior to its introduction.

In the end, conclusions for the future and open issues are addressed as the application of time series models or the introduction of new variables into the price survey, which are needed to improve the development of quantitative indicators.

Delineation of complex statistical units in Greece through the implementation of manual and automatic profiling

Ms Adamantia Georgostathi[1], Mrs Christina Karamichalakou

[1]Hellenic Statistical Authority (ELSTAT), Piraeus, Greece

Economic globalisation has set new rules on the operation of Multinational Enterprise Groups (MNEs), which tend to organise themselves in a more complex way, engaging more legal units across more countries than before and frequently reorganise their internal structure. Given this development, the approach followed by the National Statistical Institutes (NSIs) for business statistics, related to the legal unit (LeU) cannot reflect properly anymore the economic reality.

Consequently, the introduction and the implementation in business statistics of a more complex and relevant to real economy unit, such as the statistical unit “enterprise” was imperative.

To this end, the Hellenic Statistical Authority (ELSTAT) is running since reference year 2018 a large project with special focus on the creation and updating of complex statistical units in Greece, through automatic and intensive profiling and the parallel implementation of the profiling results in the Statistical Business Register in Structural Business Statistics and in other business statistics, in order to ensure the harmonized and consistent use of the statistical units in the different statistical areas.

In this paper, the different sources and methods used, related both to manual and automatic profiling, are presented, along with a detailed description of the survey for the direct collection of data on the intragroup / intra-enterprise flows for main economic variables. The impact of the implementation of the statistical unit enterprise on the Structural Business Statistics results is presented in a dedicated section of the paper. Emphasis is also given on the synthesis and work of the Large Cases Unit (LCU) of ELSTAT and its interrelation with the profiling process.

Poster Session 1, June 5, 2024, 15:45-16:30

Information system on occupation

Alessandro Capezzuoli[1], Maria Grazia Mereu[2]

[1]Istat, ROMA, Italia, [2]INAPP, ROMA, Italia

The Italian information system on occupations is an inter-institutional partnership between the National Institute for the Analysis of Public Policies (INAPP) and the Italian Institute of Statistics (ISTAT) with the intent of creating a network of people who at various levels produce information regarding professions and the accessibility on the web of information regarding each Professional Unit thanks to the availability of statistical indicators and a variety of administrative data sources generated by the activities of institutional partners.

Data connected by a kind of Ariadne's thread - the occupation code - that guides the user along the intricate path of a large amount of information variously located. The distinctive feature of the system, compared to similar tools, consists in it being a "distributed information system" in which each of the partners' websites is a gateway to the system.

Data from different sources are put in connection each other, but always independently, using effortless software applications that direct the user through the various links.

The heart of the information system is a widget shared by the institutions participating in the system with a simple "copy and paste. A wide range of statistical and administrative data continue to reside in the institution where they are produced, maintaining the visibility and responsibility of the rights holder unaltered.

The structure of this system requires that each partner is able to "answer the call", referring to a page that contains the data of the specific professional unit identified by the code provided in the link.

A necessary condition to participate in the system:

the data can be organized according to the current edition of the "Nomenclature and Classification of Occupations" (http://cp2011.istat.it)
at the same time, each partner must be able to refer to other institutions, providing an updated list of addresses
a complex system that can be defined web linked open data.
similar and non-homogeneous data are correlated with each other through the web exchange of a unique key.

The classification and defining adopted by the occupation information system is becoming the most used tool on the labor market in Italy.

The system promotes the adoption of standards for the production of information, querying databases, organizing and sharing information, avoiding to add new construction processes to the already existing systems, release and maintenance of the information itself.

It is oriented according to open data philosophy.

Reflections on ESG data quality strategy for Sustainable Development -an integrated Value at Risk approach-

Paola Casciotti[1]

[1]Italian National Institute of Statistics, Roma, Italy

The path towards the Sustainable Development Goals set by the 2030 Agenda, given the now short time horizon, highlights challenges that require innovations or better organization in data production and interpretation, also to neutralize misleading greenwashing phenomena.

The most characterizing and challenging aspects of the SDG framework in relation to data quality are the high degrees of intertwined relationships between goals and between intermediate targets relating to the different and complex dimensions of Sustainability. The several dimensions of Sustainability, in turn, involve, with their respective needs of data and data production, numerous categories of stakeholders: supranational, institutional, academic, political, economic, sectoral, individual and so on. Consequently, it is strategic to develop a new integrated and balanced ecosystem of analytical data coming from different sources in compliance with data quality principles.

The reflections contained in this work intend to provide a contribution to the discussion in relating the challenges posed by the dimensions of sustainability on one hand and the dimensions of data quality on the other hand, with particular reference to the aspects of comparability, consistency, completeness and further aspects such as flexibility and modularity.

Based on the results of a previous study, it will be represented the opportunity to adopt the Environmental-Social-Governance (ESG) criteria and their combinations to develop an integrated datawarehouse for the different dimensions of sustainability.

Furthermore, a Value at Risk based approach will be proposed, like advanced frontier, to define coherent metrics for realistic impact estimates, through the construction of matrices according to the double materiality principle.

Uses and quality of the OpenStreetMaps network for the massive calculation of routes and travel costs

Begoña Admirable, Cristina Rovira, Mireia Farré

Statistical data contribute to decision making, research and improvement of public policies; moreover, the use of statistical data allows to respond to urgent demands of society.

The aim of this presentation is to show an example of the reuse of statistical data in relation to issues where mobility and travel time costs are key factors, such as school planning or the location of emergency medical transport bases.

This has been achieved by combining geolocated population data with the use of information derived from the OpenStreetMaps (OSM) application, which is a collaborative project for creating free, editable maps. The main output of the project is data generated in the form of raster images and vector data, creating a database (planet.osm) in the PBF binary file format.

From the OSM data (the streets and roads where it is possible to walk and drive and the transport network data) a graph has been created to model the networks in order to perform the calculations of getting from a point of origin to a point of destination.

The pgRouting extension of postGIS has been used to create the graph and to assess the quality of the OSM project data for each type of network.

School pathways: key indicators for primary and secondary school pupils in Portugal (Poster)

Mrs. Ana Paula Ferreira[1], Patrícia Pereira[1], Joana Duarte[1]

[1]Directorate-General of Education and Science Statistics - Ministry of Education, , Portugal

The poster focuses on the analysis of the situation after 3 years of students who entered the 3rd cycle of basic education or secondary education in mainland Portugal, and aims to present information on how many students completed their courses within the expected time (three school years), how many remained on the courses without completing them, how many moved on to another education and training offer and how many were not found in the education and training system in mainland Portugal. The assessment of the situation after three years of students in a given school year is based on the following categorization: a) students who have completed the 3rd cycle or secondary education b) students who, having started on a particular pathway, completed the cycle/level of education in another education and training offer, c) students who remained enrolled in the cycle/level of education without having completed it, d) students who did not complete any cycle/level of education and were enrolled in other education and training offers and, finally, e) students who did not complete any cycle/level of education and were not found to be enrolled in secondary education in mainland Portugal.

This analysis already has a series of eight periods, in which the cohorts of students who entered at the beginning of the 2012/13, 2013/14, 2014/15, 2015/16, 2016/17, 2017/18, 2018/19 and 2019/20 school years were analyzed, and the situation of these students was determined three years after their entry, i.e. at the end of the 2014/15, 2015/16, 2016/17, 2017/18, 2018/19, 2019/20, 2020/21 and 2021/22 school years, respectively. In addition to the temporal evolution of the indicators, we also looked to see if the completion rates depended on the following variables: a) the training and education offer/course the student entered; b) the geographical location of the educational establishment - Region (NUTS II) and Intermunicipal Community/Metropolitan Area (CIM/AM) (NUTS III); c) the student's gender, d) the student's age in the year of entry; e) the student's School Social Action (ASE) level in the year of entry; g) the nature of the educational establishment, public or private.

These studies, carried out on the basis of information collected by official statistics, allow for the construction of new statistical indicators on the situation after 3 years of primary and secondary school pupils in mainland Portugal.

Metadata integrated system in Eustat

Ms Marta Mas Moreno[1], Marina Ayestaran Arregi[1]

[1]Basque Statistics Office - Eustat, Vitoria-Gasteiz, Spain

For years, in the Basque Statistics Office (Eustat) statistical operations have been documented by means of so-called “technical projects”. These projects serve as comprehensive procedure manuals, and they not only compile tasks and methods used in the statistical processes, but also serve as a transparency tool for users of official statistics.

In recent years, these technical projects have become the basis of an integrated metadata system consisting of several data description products as definitions, classifications, and methodological sheets. This system connect several subsystems and internal applications of Eustat with the dissemination system. In addition, European standards for documentation and metadata exchange such as the Euro SDMX Metadata Structure is used for the methodological reports and classifications.

In this contribution we will explain the phases of this integrated metadata system, we will give details about already developed modules, such as the technical projects and methodological sheets, those that are in the implementation phase, such as the definitions and classifications modules, as well as those to be developed in the coming years.

Improving Labour Market statistical literacy

Boon Heng Ang[1], Ms Mavis Lim[1]

[1]Manpower Research and Statistics Department, , Singapore

As the National Statistical Agency for manpower statistics, the Manpower Research and Statistics Department (MRSD) of Singapore's Ministry of Manpower collects, analyses and disseminates essential statistical information on Singapore's labour market. Helping people understand and utilise manpower statistics is just as important as our primary statistical activities. To improve labour market statistical literacy, we have different approaches to reach out to the relevant stakeholders.

This paper highlights the steps we adopt to improve labour market statistical literacy and reduce statistical misconceptions among relevant stakeholders, whom can be categorised into organisations (i.e. educational institutions, private corporations, the media outlets and the public sector) and individual users.

To assess the level of labour market statistical literacy among the relevant stakeholders, we conduct informal polls to gauge users' awareness of the latest manpower statistical information and the concepts behind them. Our approach to improve manpower statistical literacy and reduce statistical misconceptions is tailored for each stakeholder. Our outreach efforts include liaising with various organisations to bring statistical information to them. An example would be MRSD's collaborations with tertiary institutions. We conduct sessions with students at educational institutions to raise their awareness of the labour market situation and show them how such information could be used for their job search. For individual users, we publish regular papers and infographics that inform users of the latest labour market happenings and explain confusing concepts such as the concept of unemployment.

Istat user survey 2023: new features and main findings

Ms M. Francesca Loporcaro[1], Mrs Giuseppina Pica[1], Mrs Roberta Roncati[1]

[1]Istat, Rome, Italy

Principle 11 of the European Code of practice states that user satisfaction should be monitored on a regular basis and is systematically followed up. Starting from 2013 Istat carries out an annual "User satisfaction survey". The purpose is to measure the satisfaction or dissatisfaction with the products offered and the perceived quality of data and metadata available.
Users of www.istat.it website are asked to indicate the level of satisfaction or dissatisfaction with relevance of content, accuracy and reliability of information, timeliness of updates, accessibility and clarity of presentation and possibility of comparison with other data. To measure satisfaction, is used a scale based on four categories of statements that express a negative to positive attitude: "Very dissatisfied", "Dissatisfied", "Satisfied" and "Very satisfied" (Likert scale).
In the last edition of the survey, held in 2023, a user profile self-assessment was added. It is based on evaluating three principal features: computer skills, statistical literacy, and frequency of data use.
Users could choose between four profiles ranging from light users to advanced user (this kind of self- assessment is quite similar to what is detected in Eurostat user survey). All the information collected on users - trust in statistics, frequency of use, statistical literacy and computer skills, files and software used - are extremely useful for creating homogeneous user groups which Istat can use in designing new products and services.
This work aims firstly at presenting the main findings of the survey: satisfaction with products offered by Istat, perceived quality of official statistics and metadata, trust in statistics and user profiles; secondly the focus will shift to the strengths and weaknesses of the survey and to its use for improving statistical production, highlighting possible future development.

Framing effects in surveys: Experiences from the local election survey in Norway 2023

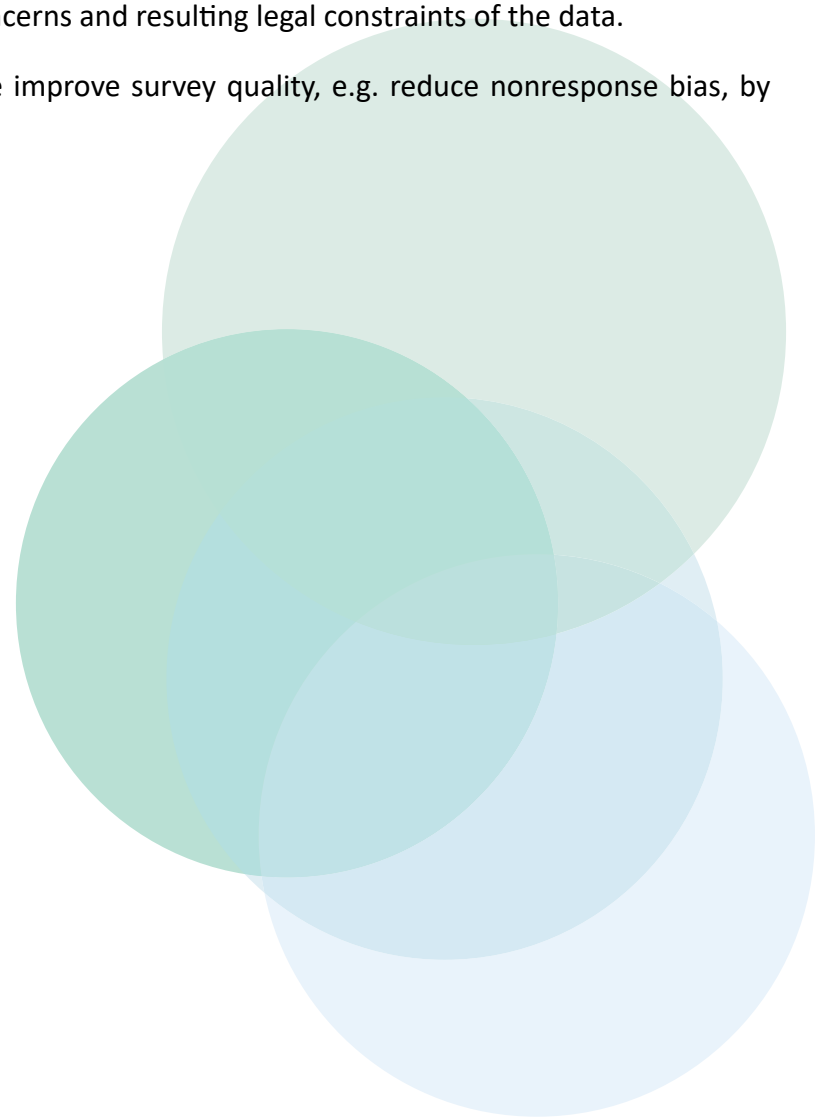
Mr Bengt Oscar Bengt Oscar[1], Mr Øyvin Kleven[1], Ms Marta Anna Krawczynska[1]

[1]Statistics Norway, Oslo, Norge

In general, surveys are about topics that are subjective in nature. This is also the case for election surveys. Several studies have paid attention to framing effect, that respondents may answer survey questions in ways that are systematically related to arbitrary features of the survey's design, such as the way questions are worded, the order in which answers are presented, the topics of the survey etc. One important feature is the information and communication strategies in the recruitment of the survey respondents.

In this study we conduct an experiment to examine how voting activities among non-respondents differs according to the context our questions are delivered to the non-respondents. Invitations are sent through the Norwegian internet portal for digital dialogue between businesses, private individuals, and public agencies – Altinn. Altinn will gives information about respondent behavior before responding or not responding that will give us useful knowledge about the decision to respond to the survey or not. We examine effects on response rate, bias for some relevant demographics and how this affect some of the key estimates from the voting survey 2023. We will also link survey and administrative data to study privacy concerns and resulting legal constraints of the data.

Our main research question is: Can we improve survey quality, e.g. reduce nonresponse bias, by framing the survey differently.



Can(non)probability online panels compensate national cross-sectional mixed-mode survey?

Ms Maruša Rehberger[1], Darja Lavtar[1], Nejc Berzelak

[1]National Institute Of Public Health Slovenia, , Slovenia

The increasing popularity of online panels can be attributed to their remarkable efficiency in terms of both time and cost, especially when compared to the resource-intensive nature of cross-sectional national surveys. Online panels are becoming increasingly used in research and, especially in marketing, are changing how we collect data and gain insights. However, can online panels be used as the main source of survey data for official statistics and how probability and non-probability panel samples perform in health survey context?

The Slovenian National Institute of Public Health (NIJZ), in the Slovenian statistical system holding a role of the Other National Authority (ONA) for health and health care statistics is data provider of illicit drug prevalence and other risky behaviours. This paper undertakes a comprehensive comparative analysis of data collected through three different survey sources: 1) a cross-sectional national survey on a probability sample, 2) a probability online panel, and 3) a non-probability online panel, all conducted in spring 2023, researching marihuana prevalence among 18 – 64 years' Slovenian adults. The primary aim is to examine the feasibility of (non)probability online panels as an alternative approach for cross-sectional mixed-mode surveys and assess their potential role in shaping the future of official statistics.

Challenges include potential response bias and resource-intensive survey processes. Probability online panels, leveraging internet platforms, offer an efficient and cost-effective approach while attempting to maintain representativeness through probability sampling for panel recruitment. However, concerns about panel attrition, non-response bias, and limited demographic coverage persist. Non-probability online panels are rapid and cost-efficient, but may lack representativeness due to non-probability panel recruitment. Despite this limitation, they can provide unique insights into specific subpopulations and dynamic trends often overlooked in traditional survey methods.

In the study, the cross-sectional national survey data serves as a benchmark, providing nationally representative data for a comprehensive understanding of health behaviours. By comparing national cross-sectional data and two online panels, the comparability of online panel data and currently used survey approach to obtain official estimates will be shown. Comparisons are performed on population of 18-64 years old Slovenian adults in general, by gender, age groups and educational attainment.

Remote access to microdata of the Italian National System

Ms Maria Assunta[1], Erika Lucarelli[1]

[1]Istat, Rome, Italy

Access to Istat Secure use files is regulated by art. 5-ter of Legislative Decree No. 33/2013 (introduced by Legislative Decree No. 97/2016), which adopted the Guidelines for access for scientific purposes to the elementary data of the Italian National System (Comstat Directive No. 11/2018).

According to current legislation, researchers may access to Istat Laboratory for the Analysis of Elementary Data to conduct statistical analyses on i) microdata collected and validated by Istat through surveys on individuals, households, enterprises and institutions; ii) databases integrating different sources.

While in the past this access to microdata could be only into physical laboratories of the data owners, now it can also be remotely with accredited research entities.

In 2022 Istat carried out a study, together with the Directorate for Economics, Statistics and Research at the Bank of Italy, to define an architectural prototype of a remote laboratory. This study sets the rules for research entities to require remote access, i.e.: a) the characteristics of the research project;

b) the adequacy of the technological and organizational system; c) the measures to protect and process microdata for guaranteeing their security and the confidentiality of statistical units.

This is a great result for the quality of Istat microdata in terms of improved accessibility and an important innovation for the research entities that may finally use their own laboratories to remotely access to microdata of the Italian National System. This allows them to achieve the same results of physical access by reducing costs and time.

The impact of innovation on statistical data quality challenges

Soad Yehia[1]

[1]Central Agnes For Public Mobilization And Statistics, Cairo, Egypt, [2]CentralAagency for Public Mobilization and statistics , Cairo, Egypt

This research study aims to analyze the impact of innovation on the challenges of the quality of statistical data production because data is the lifeblood of the decision-making process and is the raw material for developing plans to confront crises, as high-quality data, and statistics available at the right time help move at a faster pace in future decision-making processes. Policy formulation and focus on the type and quality of the statistical product are much more important than the quantity of statistical products. The challenges of the quality of statistical data are considered among the most important problems faced by statistical offices in the field of data collection and analysis. There are many challenges facing the quality of statistical data production, starting with collecting data from a variety of different sources and ensuring the accuracy and validity of the collected and recorded data, in addition to the continuity of data collection over a period. For a long period, especially in cases of continuous change in indicators or variables, analyzing and understanding the data, then keeping it securely and easily accessible, considering presenting statistics in an easy and effective way using technology while ensuring that the rights of individuals are not violated, and the confidentiality of the data is maintained.

Many changes and innovations have occurred in the process of collecting statistical data, such as the use of online communication technologies, reliance on automation programs, the use of new methods to increase the participation of respondents and ensure data accuracy, and the use of modern technology, open data, and artificial intelligence, which have contributed to improving the accuracy of error correction and providing new methods. To collect data, modern technology was used vigorously during the Corona period, and it was one of the biggest challenges at the level of all countries, as GPS geographical location was tracked, and China, South Korea, and Taiwan applied it. China developed the TenCent and Alibaba applications, South Korea developed Corona 100m, and Egypt implemented the Egypt Health application.

The descriptive approach was used in data analysis, and one of the most important results was realizing the importance of statistical data and information and making them in the form of electronic applications and investing in innovation, as necessity is the mother of invention and innovation contributes to improving the accuracy and completeness of the collected data,

Further studies in higher education: graduates of higher technical courses and degrees

Nuno Cabral[1], Artur Reguengo, Joana Duarte

[1]DGEEC, Lisboa, Portugal

In this poster, DGEEC analyzes the student's further studies in higher education - technical higher vocational courses (CTeSP) and degrees - using information gathered from official higher education statistics (Register of Students, Enrolments and Graduates in Higher Education - RAIDES).

These studies analyze the situation one year after completing the CTeSP and degrees, taking into account the following variables: a) higher education subsystem, b) higher education institution, c) final classification d) area of training and education, e) district, f) gender and h) age group.

These students are followed up in order to identify their situation after 1 year in the following categories: a) enrolled in a Bachelor's degree; b) enrolled in a Master's degree; c) enrolled in a PhD;

d) enrolled in a postgraduate specialization; e) enrolled in another level of higher education and f) not found in Portuguese higher education.

These studies, carried out on the basis of information collected by official statistics, allow for the construction of new statistical indicators on the pursuit of studies in higher education.

Using the information collected by official statistics to build these indicators allows new variables to be explored, and necessarily implies greater demands in reporting and validating the information. In addition, the fact that this information is made available to different types of audiences, such as the academic community, the general population and government bodies, requires continuous improvement and rigor in the reporting and validation of information and, consequently, contributes to improving the quality of official statistics.

InfoEscolas Portal: main indicators for primary and secondary education in Mainland Portugal

Ms Joana Netto Miranda Duarte[1], Mrs Patrícia Pereira[1]

[1]DGEEC, , Portugal

The InfoEscolas Portal was designed and built for the entire educational community (schools, school groups, teachers, parents, among others), in order to understand and analyze the results in a comparable and properly contextualized way, with a view to promoting the academic success of all students, at the various levels and educational and training offers that make up the 12 years of compulsory education in Portugal. The data is also used by other institutions under the Ministry of Education, such as the General Inspectorate of Education and Science, which is a regular user of this portal, using it as a tool for the external evaluation of schools.

It presents statistical data on pupils in mainland Portugal enrolled in public and private schools, covering around 1,175,000 pupils enrolled in more than 5,000 schools in all cycles of basic and secondary education.

It includes statistics for the 1st, 2nd and 3rd cycles of basic education and for scientific-humanistic and vocational courses in secondary education.

Starting with 9 indicators in the 1st edition launched in 2014, between 2016 and 2021 there are a total of 88 indicators/features, most of which are available for public consultation.

It presents statistics by school, school group, municipality, district, region (NUTS II and NUTS III) and for the whole Mainland Portugal.

The reference date for the latest edition, and for most of the data, is currently the 2020/21 school year, and for the indicators based on national tests and exams, the data is for 2021/22. By the time of the congress, the 2021/22 data for the context indicators and the 2022/23 data for the results indicators are expected to be available.

The InfoEscolas Portal consists of using the information collected by official statistics in an integrated way, while also tracking students in some of its indicators. The application of this statistical information to build this Portal requires the use of different variables to construct these indicators, which contributes to greater demands in the reporting and validation of this information. In addition, the fact that this data is available to different types of users, such as the educational community, the general population and government bodies, means that the accuracy of the reporting of this information is gradually improving, thus contributing to improving the quality of official statistics.

Classification of districts of Costa Rica using information from Google Maps

Luis Edo. Amaya-briceño[1]

[1]University Of Costa Rica, Liberia, Costa Rica

In our work, we share the computational experience developed, implementing the Google maps API, to collect, clean and quantify information on variables such as health, education, commerce, banking, among others for the entire country of Costa Rica.

The objective of the above is to be able to classify and characterize the country's districts based on said information.

The results obtained provide a very close proxy to what other existing quality indices provide, but at the cantonal level.

Data Power in the Issue of Combating Violence Against Egyptian Women

Nora Dalal[1]

[1]Egypt CAPMAS, cairo, Egypt

In this research, the power of data in strengthening efforts to combat violence against Egyptian women is demonstrated. This is by focusing on the importance of analyzing and using available data to contribute to a deeper understanding of the scope and dimensions of violence, which helps guide policies and programs more effectively.

The study focuses on the importance of data accuracy for its correct interpretation of the phenomenon of violence, with a focus on the age groups of women and girls who are exposed to violence, in addition to the geographical distribution of this phenomenon.

It will highlight the importance of data in shaping public opinion, enhancing awareness of issues of violence related to women, and directing awareness campaigns more effectively to achieve positive social transformation in favor of Egyptian women's rights.

This paper aims to understand the extent of violence against Egyptian women by analyzing available data and identify different types of violence such as physical, psychological, and economic violence also studding the distribution of violence at the level of Egyptian regions. Moreover, the paper analyze how violence affects women and society in general and the role of political makers.

- By Using descriptive approach to describe the nature of violence against women and Regression analysis of variables to determine related to violence against women.

The impact of clinical trials on Improving the quality of medical research and health outcomes among people living with HIV

Said Abdelrhman[1]

[1]Capmas, Nasr City, Egypt

In May 2022, the Seventy-fifth World Health Assembly endorsed resolution WHA75.8, which focuses on reinforcing clinical trials to furnish high-quality evidence regarding health interventions and enhancing the quality and coordination of research efforts. The integrity of clinical trials is contingent upon both data accuracy and the protection of subjects. The challenges of achieving global quality have intensified due to factors such as globalization, outsourcing, and the growing intricacies of clinical trials. This complexity is notably pronounced in populations affected by HIV and infectious diseases. The basis of HIV therapeutics lies in the evidence derived from randomized controlled trials. These trials have played a pivotal role in introducing numerous drugs and devices that extend survival, decrease morbidity, and prevent the use of interventions proven to be ineffective or unsafe. The landscape of HIV/AIDS research has undergone substantial evolution since the inception of the first randomized trials in this domain. To ensure the reliability of evidence concerning significant outcomes, HIV trials have grown in scale, often enrolling thousands of participants from numerous clinical sites across a multitude of countries.

This paper outlines and evaluates the efficacy of clinical trials that have contributed to reducing the prevalence of HIV in Miami-Dade County, recognized as the most heavily affected county by this virus in the United States in 2020. The design of the clinical trial has been structured to minimize bias in estimating the treatment effect. Furthermore, a well-conceived trial aims to possess adequate statistical power to identify a clinically significant effect, ideally the smallest meaningful effect achievable within the limitations of available research resources. Despite the ongoing challenge of developing an HIV vaccine, clinical trials play a crucial role by providing essential data to propel research efforts toward achieving this objective.

Building new process in statistical frame to improve quality in statistics, case study employment register data frame

Ms Daniela Avramoska[1]

[1]State Statistical Office Of North Macedonia, Skopje, Macedonia

The quality of statistics is a fundamental requirement for producing valuable statistical data that represents key elements for society's development. With the implementation of new statistical processes, innovation of new technologies, and modernization of the legal framework, there is a mark of continuous improvement in the quality of statistics. In this matter, administrative data emerges as a new data source for the production of official statistics. For this purpose, the State Statistical Office (SSO) developed the Employment registers based on the national and European methodology in the frame of SSO Labor Market Statistics. Creating and maintaining the administrative data into useful information for collecting, estimating, and disseminating the data for employees, wages, salaries, and working conditions within the domains of the Employment register improved the range and quality of statistical products, especially for creating social and economic indicators.

As a collecting method for statistical surveys in the field of labor statistics, SSO uses the Employment register. The Employment register is a data frame built from administrative data collected online daily exchange data between administrative sources, incorporated through the implementation of statistical processes for table integration, data transformation, and data validation into the final maintained data frame.

The ER was built into three phases. The first phase includes the determination of needs and uses in creating a core table of the register (the table of basic statistical units) with all possible representations (views) that can be finally derived from the data frame. The second phase was to build metadata tables from administrative data into an integrated table based on statistical purposes with the derived status of activities for each work relationship through a statistical model. The validation procedure is the most important part of the register created in the ER data frame after incorporating precondition logical validation used in administrative sources in the transformation of administrative data into statistical.

These are new challenges in statistical work based on administrative data as the main source for producing average net and gross wages, labor cost survey, structure of earnings, calculation of paid employees by NACE code, and have an impact on the future development plan based on administrative data use in a more rational and functional form.

Statistical disclosure control for the general public distribution of multidimensional cubes: an experiment at the french statistical service of agriculture

Levi Valensin Michael[1]

[1]SSP-Ministry Of Agriculture, Auzeville, France

As with all statistical sources, the publication of agricultural data on the Agreste website of the Ministry of Agriculture and Food, requires compliance to information diffusibility rules. The Department of Statistics and Forecasting (SSP), akin to other statistical services and INSEE, historically employs suppressive methods, refraining from disclosing sensitive values.

A distinctive feature of the SSP is the publication of multidimensional cubes in the form of interactive tables, allowing users to select variables at different levels of nomenclature . In such products, it becomes necessary to consider all possible crossovers and any induced secrets.

Historically, this task was performed heuristically in the SAS software using an internally designed function that applied the rules of induced secrecy. The SSP's strategy to transition from SAS led to a reevaluation of this process, prompting an examination of current techniques.

Suppressive methods for managing statistical secrecy have typically been implemented, for about two decades, using the τ-Argus software. This software allows the resolution of an optimization problem under constraints by minimizing the removal of information while ensuring compliance with secondary secrecy rules. More recently, R functions have been developed to apply similar algorithms, facilitating various steps in the anonymization of a file. For example, the sdcTable package, developed by Statistics Austria (INS of Austria), encompasses many τ-Argus functionalities R-based masking.

This contribution provides insights into the use of R packages for applying suppressive methods to the publication of agricultural data in the form of cubes. Several tests were conducted on files such as the annual survey "Forestry and Sawmills" (EXF-SRI) on wood sawing by species or Agricultural Land Areas (SAU) by municipality in the 2020 Agricultural Census.

Formatting tables and constructing hierarchical files in the appropriate format are preliminary but essential steps before any anonymization. Difficulties and technical problems arose during the processes, leading to the implementation of several automation rules to ease this delicate, binding, yet indispensable stage in the future.

These techniques have been integrated into a processing pipeline, culminating in the publication on the Agreste website in the form of dynamic tables.

The quality report of the Permanent Population and Housing Census in Italy: opportunities and perspectives for an evolving process

Dr Giancarlo Carbonetti[1], Gabriele Ascari[1], Andrea Bruni[1], Giorgia Simeoni[1], Fabrizio Solari[1], Donatella Zindato[1]

[1]Italian National Statistical Institute (Istat), Rome, Italy

In 2018, the Italian National Statistical Institute (Istat) launched the Permanent Population and Housing Census, based on the integration of register data and survey data. The decennial complete door-to-door enumeration has been replaced by two ad hoc annual sample surveys that, together with administrative data held in registers, allow the yearly release of data on the main characteristics of the usually resident population at national, regional and local level.

The field surveys involve a sample of municipalities each year (some participate every year, while the others every four years, according to a rotation scheme) from which representative samples of households are drawn.

This paradigm shift makes it possible to produce census data more frequently and timely than the decennial census, reducing both the costs of field operations and the statistical burden on households.

The new approach adopted by Istat is fully in line with the European legislation on population and housing censuses, that is output-oriented. Indeed, the authority and responsibility for developing census methods and technologies remain with the Member States, while a European programme of statistical data and metadata provision has been established in order to ensure the comparability of census data across EU. To this end, Member States have to produce census data according to the EU programme and at the same time provide the metadata and the quality report as required by the EU legislation on 2021 population and housing censuses.

In addition, Istat is defining its own quality reporting strategy for the population and housing census, to be at the same time aligned with European standards, tailored for users of the national dissemination and easily replicable every year.

In order to do this, the data production process has first to be analysed, taking into account its continuous evolution over the years. Furthermore, the national dissemination plan has to be evaluated: it foresees an annual dissemination for some results, and a dissemination every two, three or five years for others - depending on the classification and the territorial detail - according to a continuously evolving plan.

The aim is to define (i) a set of reference metadata agnostic to process changes that might occur from year to year and (ii) quality indicators that aim to represent the main quality characteristics in the various years of dissemination to accompany the annual release of the Permanent Census data.

Poster Session 2, June 6, 2024, 15:45-16:30

Marital status based on administrative records in the 2021 Population and Housing Census in the Basque Country

Ms Ana Maria Miranda Ligüérzana[1], Ms Elena Goni Rementeria[1]

[1]Eustat - Basque Statistics Office, Vitoria-Gasteiz, Spain

In Eustat, Basque Statistics Office, we have been using administrative sources in the preparation of population statistics for several years. This is the case of activity, education or housing annual statistics, in addition to our statistics of residents. The aim of the paper is to share the process to build the marital status variable, basically on administrative records, for the 2021 Population and Housing Census in the Basque Country.

In summary, the marital status variable for the 2021 Population and Housing Census is based on the linkage of the population register mainly to pension and taxes records, marriage and divorce files.

The incorporation of surveys, as the Demographic Survey, in the imputation process enhances the completeness and accuracy of the dataset, making it a robust resource for a deeper understanding of society dynamics.

The strategic use of administrative registers in the creation of census variables is a cornerstone of modern census methodologies. These records, derived from various official sources such as economic registers, marriage and divorce files, serve as a foundational framework for capturing essential demographic information. In Eustat, we use some of them to deliver an exhaustive and accurate marital status variable for the 2021 Population and Housing Census in the Basque Country.

Unlike traditional survey methods that rely on self-reported data, administrative records may offer an objective and standardized approach. Data collected through administrative processes are often thorough, reliable and regularly updated, reducing the likelihood of reporting biases and errors. This not only enhances the quality of data but also reduces the burden on respondents.

The categories differentiation within the marital status variable is linked to the information derived from these administrative files. For instance, widows are identified through the widow's pensions file, highlighting a unique aspect of marital status tied to the receipt of widow's pensions. Marriage and divorce files serve as crucial sources of information, contributing to categories like "married" and "divorced" within the census dataset.

Despite the richness of administrative records, we need to employ imputation processes to address missing information. Imputation techniques ensure that the census dataset is as complete and accurate as possible, especially when individuals are missing in the administrative files. In this context, surveys available in Eustat, i.e. Demographic Survey, play a crucial role in the imputation process. These surveys provide supplementary data to fill in gaps and refine the marital status variable, ensuring a more comprehensive representation of the population.

Peer learning in Africa: a partnership for statistical capacity building

Ms Ana Cánovas Zapata[1], Ms. Ana Carmen Saura Vinuesa, Ms. Dominique Francoz, Ms. Amalie Skovengaard, Ms. Janne Utkilen, Ms. Marika Pohjola, Ms. Dorota Paraluk

[1]National Statistics Institute of Spain (INE), , Spain

Building data capability in Africa is becoming a key objective for many international organizations, focusing efforts and resources in the region. The European Commission has become a key donor in the continent through the establishment of the Pan-African Statistics Programme II (PAS II). This programme aims to support the African integration process by strengthening the African Statistical System (AfSS), to ensure the use of quality statistical data in the African integration decision-making and policy monitoring, and to translate continental priorities at regional and national level.

The PAS II programme constitutes a great opportunity to contribute to sharing European standards and best practices in statistics with African Union countries and adapting them to the African context. The programme is being implemented via different budgetary mechanism, one of them is through grants awarded to National Statistics Institutes (NSIs) of the European Statistical System (ESS). It is a novelty in the context of statistical cooperation, since it is the first time that a project on statistical capacity building in Africa is implemented through EU grants to Member States. The European Commission has awarded two grants: one aimed at developing social statistics in African NSIs (SOCSTAF); another directed to developing economic and business statistics (ECOBUSAF). To cover both goals and enable transformation and progress, the creation of efficient partnerships is a key element to work across sectors and to address common needs.

In order to succeed in the implementation of this programme, two consortiums of European NSIs have been established which constitute an efficient partnership for capability and competence building, taking advantage of the key values of the different stakeholders. This partnership is a fundamental tool to increase the quality of the work, activities and practices of the institutions involved. Something that represents a great added value for the African counterparts is that the transmission of knowledge is implemented through a peer-to-peer approach, allowing for the sharing of good practices already implemented in similar organizations and collaborating to better adapt the procedures to the specific characteristics of the partner institutions. These good practices also include the way of working of the ESS as a supranational system, which could be of interest for the African statistical offices.

In this paper we analyse the characteristics of this kind of capability model, based on the case study of the ongoing grants implementing PAS II, and share some lessons learnt for the time being of the project.

Opportunity to improve national statistical systems in Africa by promoting quality

Paul-henri Nguema Meye[1]

[1]AFRISTAT, Bamako, Mali

The statistical systems of most African countries suffer from underdevelopment due to the inadequacy of resources mobilised for the training of human resources, equipment and the construction of appropriate working premises. However, the experience of recent years shows that political leaders are sensitive to speeches that refer to the quality of official statistics. On this basis, it would be wise to have a strategy which, on the one hand, makes it possible to take note of each message delivered in favour of statistical development by a political decision-maker and, on the other hand, to take immediate advantage of any facility offered to carry out activities which contribute to the improvement of statistical production.

Thus, as soon as reasonable means are available to engage in a quality statistical production process, it is recommended that a self-assessment of the national statistical system's compliance with internationally recognised standards be established. The diagnosis resulting from this assessment will make it possible to identify priority areas for action likely to contribute to improving statistical quality.

To achieve the desired results in terms of quality, it is well known that it will be necessary to:

- Examine the way in which the statistical system is coordinated. The fluidity of relations between stakeholders will also be assessed;

- Assess the state of the institutional environment and, in particular, whether statistical activity is governed by texts that facilitate the conduct of statistical activities;

- Measure the effectiveness of the processes used to produce statistics, in particular the availability of appropriate working tools and effective systems;

- Assess the extent to which the data produced are criticised or not.

It is clear that by pursuing this approach, the following recurring problems in the African context will be tackled:

- The availability of qualified human resources and its corollary, the existence of a credible training offer;

- The existence of a favourable working environment;

- The allocation of sufficient financial and material resources to statistical production structures. Other more specific concerns may also be highlighted, such as the distribution of work resources between the central body of the national statistical system and the sectoral statistical services.

Statistical Process Control and official statistics – The case of the Central Register of Driving Licenses in Germany

Dr Dirk Hillebrandt[1], Daniel Kopper[1], Lennart Duncker[1], Sebastian Rudolph[1]

[1]Kraftfahrt-Bundesamt (Federal Motor Transport Authority), Flensburg, Germany

The aim of this presentation is (a) to contribute to the establishment of a framework for process quality management in the production of official statistics and (b) to identify peculiarities in the data flow which might carry information of underlying processes relevant to the understanding and interpretation of resulting official statistics.

On January, 1, 2023 the German Central Register of Driving Licenses (ZFER) at the Kraftfahrt-Bundesamt (Federal Motor Transport Authority, KBA) holds 48.07 Million records of driving licenses issued since 1999. Each year approximately 1.7 Million records are added to the register.

Leveraging the wealth of data in the ZFER is crucial for the KBA tasked with efficiently producing official statistics on driving licenses.

Statistical Process Control (SPC), a methodology deeply rooted in statistical principles, provides powerful tools for ensuring the quality, reliability, and accuracy of statistical information through monitoring and controlling every stage of the production process.

The primary objective is to distinguish between common cause variation, inherent in any process, and special cause variation, which may indicate anomalies or errors in the data on driving licenses. These anomalies and errors can be a thread to the quality of official statistics if not assessed or ignored. Furthermore (and sometimes more importantly), they even bear information for a better understanding and interpretation of the resulting statistic.

During development and testing, we drew on over 15 years of data in the ZFER. Over the extended period under review, SPC helped to frequently detect trends and oscillating processes in the data flow (e.g. the effects of public holidays and weekends on the data flow at a high temporal resolution (unit: days), the effects of special administrative regulations during the COVID-19 pandemic).

In particular, the consideration of inherent data features (e.g. day of entry, conveyed reasons for changes in existing entries) during special short-term and long-term external influences and special events, as well as the use of various different temporal resolutions in the analysis facilitate the understanding of process characteristics. Some of the particularities found also make special reporting necessary to adequately inform all users of the official statistic on driving licenses.

Our conclusion is that SPC is suitable for monitoring data processing in an administrative register on a large scale and at the same time gaining valuable knowledge that facilitates interpretation for all users. The creation and purpose of automated and reproducible reports is discussed.

Use of international standards for the development of the National Quality Management Framework

Mrs. Aigul Zharkynbaeva1, Mrs. Zhazgul Beisheeva1

[1]National Statistical Committee Of The Kyrgyz Republic, , Kyrgyzstan

In Kyrgyz Republic there is a general commitment to ensure the quality of official statistics in the National Statistics Act, which also includes a dedicated chapter for quality. Measures to implement the requirements of the act are contained in the Strategy for 2022-2026. Additionally, the World Bank started a project for “Modernization of tax administration and statistical system”, in which the National Statistical Committee (NSC) of the Kyrgyz Republic has started to develop the National Quality Management Framework based on international standards.

From the organisational side, Department of Regional Statistics Development and Quality has started to work on quality related topics. Since 2022 there is also an internal Quality Working Group, that consists of the representatives of the top management and heads of units. In 2023, the Quality Working Group agreed on several important documents for the National Statistical System, such as quality declaration and quality policy. Both documents are in accordance with the National Statistics Act and United Nations Fundamental Principles in the Field of Official Statistics.

On a more detailed level, the United Nations National Quality Assurance Frameworks Manual for Official Statistics have been adopted. These guidelines contain recommendations for the development and implementation of a national framework for quality assurance of official statistics and fostering trust in it. To achieve this, the national basic principles for ensuring the quality of official statistics are established for the entire National Statistical System, thereby structurally covering interconnected main areas such as the coordination and management of the National Statistical System, the institutional environment, processes, and products.

Generic Statistical Business Processes Model (GSBPM) as well as Single Integrated Metadata Structure (SIMS) are under systematic implementation. Guidelines for GSBPM-based production are introduced and seminars have been carried out for heads of departments in the National Statistical Committee, but also for the heads of statistical bodies in regions across the country. SIMS-based metadata and quality reports are compiled and will be published on the new website of the NSC.

For this year and the following year, a quality road map has been drawn up under the leadership of the Quality Working Group, which includes a plan for the development and implementation of all elements still missing for the full implementation of the National Quality Management Framework. This presentation and the descriptive document provide a more detailed overview of the activities that have been implemented so far and future plans.

Impact of the Application of Total Quality Standards on the development of official statistics'

Dr Haidy Mahmoud[1]

[1]National Statistics Office (CAPMAS), Cairo, Egypt

This paper aims to improving the quality of statistical products in Egypt through examining the impact of applying the total quality standards on migration and mobility surveys, as one of the most important surveys carried out by statistical agencies worldwide.

The study uses descriptive analytical method, and Statistical methods to evaluation of the statistical status of Egypt, based on the European code of best practices, and Generic Statistical Business Process Model, in addition to the SWOT analysis, Statistical methods were used to identify the correlation between the various variables.

The study showed that the standard (accuracy) is the most affect by 38% correlated with the other quality standards; then the (availability) is 21%; that both the standard (accuracy) and (availability) affect 59% on the quality of the output of the statistical survey, It was also found that 99% of the sample frame design affects the quality of the statistical product, 61% of the response rates are due to the accuracy and clarity of the statistical form used in the data collection form, only 7% of the researcher's training on the data collection form affects the fieldwork method, it was found that 55% of the analysis of the survey results is due to the accuracy of published data.

Finally the study recommends the need to improve quality reporting through the use of measurable quality dimensions based on the GSBPM.

The usage of R programme for official statistics

Servete Muriqi[1], Ms. Drita Sylejmani[1], Mr. Hydai Morina[1]

[1]Kosovo Agency of Statistics, Prishtine, Kosovo

The Kosovo Agency of Statistics (KAS) has started to use R programme for different phases of survey which has helped overall the sample process, designing and extracting the sample, calculation of sampling errors, calculation of weights and the calibration process which allow us to improve the accuracy of the estimates.

The R contains different packages with the set of tools for selecting the sample and calibration of weights which is a technique that uses auxiliary information to weight the statistical units. Calibration allows us to reduce sampling variance or non-sampling errors such as nonresponse bias by taking advantage of the auxiliary information given by the auxiliary variables who's the total is known in the population. KAS is using R programme for estimation and calibration of weights in all household surveys.

The paper describes the sample design of EU SILC (Statistics on Income and Living Conditions), calculation of weights and the calibration process using R packages. For the first time the calibration process was applied at SILC survey and overall process of weights was made easier and has improved the accuracy of estimates of the data. Then the R has started to use also for other household surveys for different stages of survey.

Imputation and nowcast of highest educational attainment: Combining professional knowledge and machine learning techniques

Ms Christine Ning[1]

[1]Statistik Austria, Vienna, Austria

Missing data and the lack of timely data are problems that one may face while working with official statistics. For National Statistical Institutes it is important to provide timely data for policy makers to take data-driven decisions. In this paper the approach of Statistics Austria regarding imputation and nowcast of highest educational attainment of the Austrian population will be explained.

The information of highest educational attainment is contained in the educational attainment register, which is based on the 2001 census. It is updated annually with information from different sources such as Austrian school and universities, Austrian Economic Chambers, the Austrian Chamber of Agriculture, the ministry of health and the Public Employment Service Austria.

For a given reference date missing data can occur if a person was not registered in Austria when the 2001 census took place and no other information were reported from different sources till the reference date. Furthermore, due to the ongoing data collection and the plausibility checks there is a time lag of more than one year between the reference date of the last updated highest educational attainment and the present date.

In order to close this time lag and so to train a nowcast model, first an improvement of the current imputed data was sought. Logistic regressions trained on the non-missing data were implemented before. For our new approach the missing data is divided into two groups: The first group contains people where information after a certain reference date is available. For these people an imputation back in time incorporating professional knowledge of plausible values was implemented. The idea is to use specific information of a person in the future for the imputation of the past to reduce the estimation error. The second group contains people where no information neither in the future is available. For this group different machine learning techniques were trained on non-missing data.

Afterwards based on the imputed data different nowcast models were trained. The methods used and some of the challenges will be explained in detail.

A multivariate composite estimator for the Labour Force Survey

Håvard Hungnes[1]

[1]Statistics Norway, Oslo, Norway

This paper introduces a new multivariate composite estimator for the Labour Force Survey (LFS). Unlike the univariate composite estimators used in some countries, the multivariate estimator accounts for the different probabilities of moving between labour market categories, such as employed, unemployed, or inactive. This improves the accuracy of the population estimates for each category. The multivariate estimator also avoids the problem of residual determination, which occurs when one category is estimated as the difference between the total population and the sum of the other categories. Furthermore, the paper proposes a simple method to correct for time-varying biases in different parts of the LFS sample, depending on how long they have been in the survey.

A new prediction model for GDP using Granger lag causality and partial correlation

MS Vasiliki Sarantidou[1,2], Head of Business Statistics Division Christina Karamichalakou[1], Professor Dimitris Kugiumtzis[3]

[1]Hellenic Statistical Authority, Piraeus, Greece, [2]Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece, [3]Department of Electrical and Computer Engineering, University of Thessaloniki, Thessaloniki, Greece

The Gross Domestic Product (GDP) is one of the most well-established, well-known and relevant official statistical metrics. The exploration of the components that mainly affect the evolvement of GDP, aiming at concluding to the prediction of GDP, is placed high at the statistical and macroeconomic scientific agenda. Given quarterly time series measurements of GDP and its components, the objective is to predict GDP given the past data up to the current quarter. For many components and long horizon of past values, this is a high-dimensional regression problem, and dimension reduction has to be called in. There are different approaches in the literature for variable selection, such as the least absolute shrinkage and selection operator (LASSO), or variable extraction, such as the principal component regression (PCR). Here, a new prediction model is proposed applying a stepwise forward selection algorithm using as selection criterion the partial correlation for evaluating the conditional lag Granger causality of any of the candidate components (including past GDP) to the next quarter GDP. The termination criterion is a properly designed parametric hypothesis test, ensuring a balance between model complexity and predictive power. A simulation study is conducted to assess the reliability and consistency of the algorithm and compare it to other approaches, such as LASSO and PCR. The applicability of the proposed algorithm is demonstrated using time series data of the Greek GDP and its components, compiled by ELSTAT. By applying the algorithm to this data set, the variables that most influence GDP fluctuations are identified. The selected variables are then used to form a prediction model, contributing to accurate predictions.

This study is carried out in the framework of the EMOS programme of the Aristotle University of Thessaloniki in Greece.

Data Disaggregation on Sensitive Personal Data: Exploring Living Conditions and Discrimination among LGBTQ+ Individuals in Greece

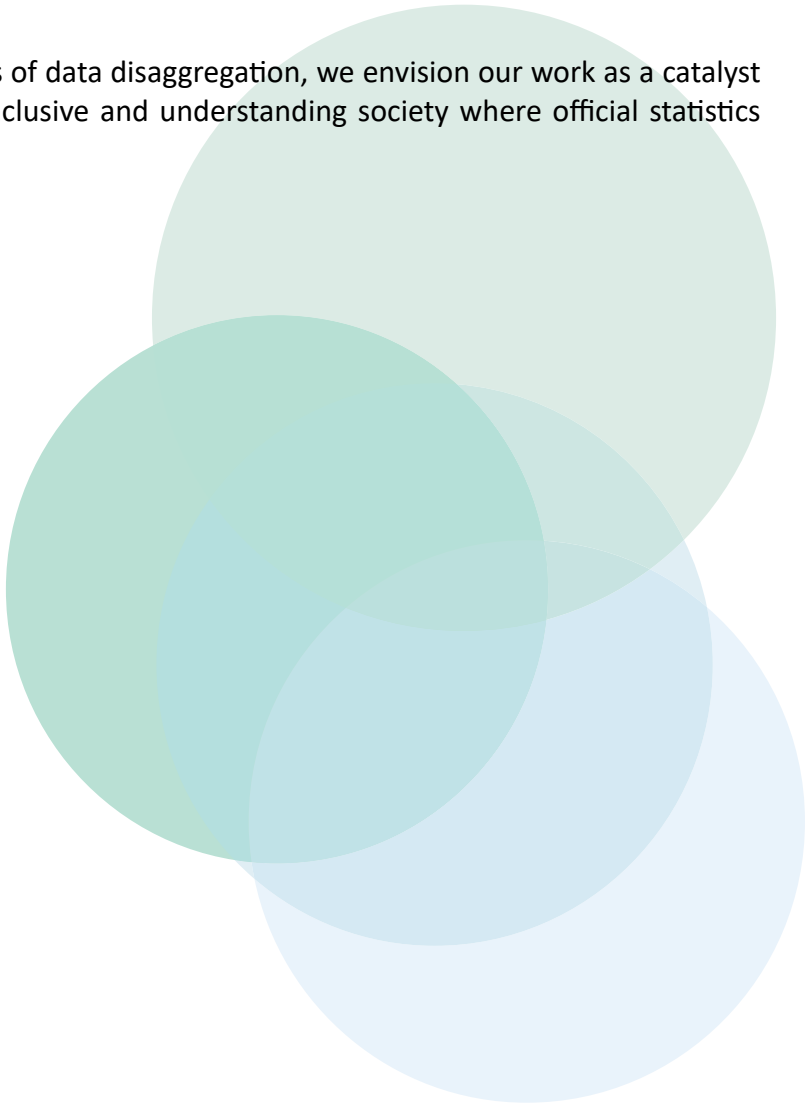
Mr Apostolos Kasapis[1], Vassiliki Benaki-Kyprioti[1]

[1]Hellenic Statistical Authority (ELSTAT), , Greece

This paper presents the challenges of collecting statistical disaggregated data on sensitive characteristics, shedding light on the challenges posed by respondents' potential reluctance to disclose personal details such as gender or sexual orientation. Utilizing the paradigm of the Pilot Survey conducted by the Hellenic Statistical Authority (ELSTAT), this paper meticulously examines the design of the questionnaire, the intricacies of the sampling methodology, and the careful evaluation of potential biases in individual responses.

In a groundbreaking effort, ELSTAT initiated the Survey in the beginning of 2024, marking the first comprehensive attempt to generate official statistics with data disaggregation based on sexual orientation or gender identity in Greece. The significance of this survey extends beyond its statistical implications since it aimed to provide valuable insights for policymakers, stakeholders and advocacy groups, towards the understanding of the experiences of LGBTQI+ individuals in Greece. This survey not only aimed to fill a crucial gap in existing knowledge but also advocated for the promotion of inclusivity and equality by providing data that aligns with the specific needs of diverse users of statistical data.

As we navigate through the complexities of data disaggregation, we envision our work as a catalyst for positive change, fostering a more inclusive and understanding society where official statistics leave no one behind.



Automated identification of potential interviewer-related errors in national mixed-mode surveys

Nejc Berzelak[1,2], Maruša Rehberger[1], Darja Lavtar[1]

[1]National Institute of Public Health, Ljubljana, Slovenia, [2]Social Protection Institute of the Republic of Slovenia, , Slovenia

Data quality in face-to-face surveys essentially depends on the performance of interviewers conducting the fieldwork. While interviewer-related variance and bias have been well-elaborated in survey methodology, the issue of timely identification and prevention of potential errors that may affect accuracy of estimates remains challenging. Relatively complex statistical methods are required to isolate interviewer effects, which is often not feasible for continuous monitoring during data collection.

This paper builds on two cases of surveys, conducted by the National Institute of Public Health that has the role of Other National Authority (ONA) for health and health care statistics in the Slovenian statistical system: European Health Interview Survey (EHIS) 2019 and National Survey on Tobacco, Alcohol, and Other Drugs (ATADD) 2023. Data collection for both surveys was conducted using a mixed-mode design, combining web and face-to-face modes. Data analysis revealed some anomalies that may indicate interviewer-related errors, despite non-specific patterns on standard indicators of response quality, interviewer monitoring and follow-up control interviews. Particularly concerning is that some of these anomalies significantly affected estimates at the level of NUTS3 statistical regions in which individual interviewers were conducting the fieldwork.

To enable early detection of such cases during the fieldwork of future surveys, this paper presents an approach established by the institute for automated flagging of anomalies in interviewer data that may indicate interviewer-related errors. The approach is based on a combination of metrics related to characteristics of the surveyed individuals, response quality indicators and benchmarking of survey estimates. The main objective is to establish a relatively simple and robust set of indicators that help identifying cases requiring further inspection and, if needed, appropriate fieldwork intervention.

Particular attention is devoted to assessing the impact of potential interviewer-related errors on regional-level data.

The performance of the proposed approach is evaluated against other model-based methods commonly used for assessing the interviewer-related errors. Strength and weaknesses are discussed along with recommended survey design decisions and fieldwork interventions to reduce such errors.

The implementation of a data culture, and the ethics of data use

Ms Natasha Natasha Cenova[1]

[1]State Statistical Office, Skopje, North Macedonia

This paper analyses the power of the official statistical data and its importance recognised by the users. One of the main goals of the official statistics is to serve the society and the public needs, in the same time ensuring that data are relevant. Today in the digital society statistical literacy is crucial for data usage. While producing data we have to focus on data usage, to implement data usage culture and see the ethics of data usage. Construction statistics produced in our SSO is very widely used by the public, private and government sector. In cooperation with the responsible institutions, the Association of the Units of Local Self-Government (ZELS) and the Ministry of Transport and Communication, the State Statistical Office signed an agreement of usage of administrative data source for issuing building permits, e-permits. The short-term indicators for issued building permits and production in construction provide important information on the economy and monetary policy, as well as for future investments in the construction sector. They are part of the PEEI main economic indicators for the economic developments in the EU and are collected according to the Eurostat Methodology of STS. The relevance and importance of data for the real estate market has increased a lot these days and is constantly increasing. Not only the future investors use them, also the eco activists are regularly following these indicators and raise the awareness in the public about the impact on the environment of the investments. In the most attractive municipality in the capital Skopje – the municipality of Centre the eco activists that are part of the local self-governmental body on their web site regularly post a map with the active building permits. Next, the Agency for Real Estate Cadastre (AREC) has established and maintains a Register of Prices and Leases of real estate.

They publish a quarterly report where they implement and link our statistical data on issued building permits, finished and unfinished dwellings. The National Bank on a quarterly level calculates indices of real estate prices with linking our statistical data. This proves that a culture of regular data usage and reuse of data is implemented and in the same time the ethics of data use is followed. All these justifies the main purpose of producing statistical data and their usage as a trusted source among other data in creating policies on all levels.

Measuring and monitoring the sustainability of tourism at the regional level: Catalonia's tourism sustainability indicators project

Mr Jordi Galter[1], Mrs Cristina Rovira Trepat, Mrs Carme Saborit Vidal, Mrs Mercè Escrichs Saez

[1]Idescat, Barcelona, Spain

In 2023, the Catalan tourism sector launched the National Commitment for Responsible Tourism, signed by all tourism stakeholders in Catalonia. Through a strategy of responsible growth, the Commitment wishes to promote a more sustainable level of prosperity in the tourism sector. To achieve this goal the strategy proposes to put sustainability at the core of all tourism activities, projects, and plans and to establish a balance of interests between tourists, entrepreneurs and investors, the local population, the natural and cultural environment.

Monitoring the sustainability of tourism policies implies major challenges for subnational territories. Until recently, regional stakeholders have had limited access to indicators that measure the economic, social, and environmental impact of tourism activities. A particular challenge has been the adoption of consistent methodologies across regions and linking data to the national statistical system, as well as the international frameworks. Regions need clearly defined, common sustainability indicators that are comparable and coherent with traditional tourism statistics. These can help regions to benchmark across time and geographic locations, providing a sound evidence base for decision-making.

Since 2022, the OECD and the EU are supporting four Spanish Autonomous Communities, Andalusia, Catalonia, Navarra, and Valencia, in developing a common set of indicators to measure and monitor the sustainability of tourism in each territory. Based on several technical workshops and through a participatory and iterative process with experts, the participants have agreed on a common set of indicators that has been evaluated in a pilot phase. The project considers existing international and national frameworks to measure the sustainability of tourism including those developed by the European Commission and the UNWTO.

This presentation introduces the common framework of indicators and the key phases and considerations that have led to their inclusion in the proposal. It summarizes the guiding principles, rationale and key considerations informing the identification of the common set of indicators. It also analyzes the difficulties that arise to provide quality information at the regional level. Finally, it presents a description of all the phases that have taken place to provides information on each of the indicators. The piloting phase has allowed to test the relevance and feasibility of the proposed set of indicators, it identifies the potential data and implementation challenges, and it may lead to revisions and adjustments in the selection of indicators, their specification and the metrices used to measure them.

Enhancing economic statistics quality by addressing large multinationals data through a Large Cases Unit (LCU)

Sixto Muriel De La Riva[1], Mr. Iván Pérez-Plaza Cuéllar[1], Mr. Juan José Cervigón Simo[1]

[1]Ine-spain, Madrid, España

The accelerating pace of economic globalization has introduced a multitude of challenges for statistical analysis, necessitating innovative frameworks to comprehend its multifaceted dynamics and captured them properly in statistics.

Firstly, large multinational corporations expand their operations internationally through an increasingly complex network of subsidiaries, which significantly complicates the representativeness and consistency of national business statistics. A precise profiling of the group and the 'enterprise' within the group as a statistical unit for analysis is crucial in statistics based on types and sizes of companies.

Secondly, the production processes of these large groups are often organized through global operations that are challenging or impossible to accurately reflect in statistical, administrative, and accounting sources. Furthermore, they are usually based on frequent utilization of intangible assets with diffuse economic ownerships.

The existence of a specialized and trained unit within national statistical offices dedicated to directly engaging with MNEs and analyzing their global corporate and operational structure is key. This approach stands out as the most effective and quite possibly the sole method to guarantee a definitive stride towards assuring the quality of national business and macroeconomic statistics (national accounts and balance of payments) in a globalized economy. This paper investigates the pivotal role of Large Cases Units (LCUs) as a strategic means to grapple with the intricate statistical implications of the economic globalization showing the first outcomes of the LCU in the Spanish statistical system.

Keywords: large cases unit, globalization, business statistics, profiling, national accounts, balance of payments

Improving the quality of seasonal adjustment process: an Italian case study based on per capita hours worked official indicator

Ms Annalisa Lucarelli[1], Cinzia Graziani[1], Maurizio Lucarelli[1], Emilia Matera[1], Andrea Spizzichino[1]

[1]Istat, Rome, Italy

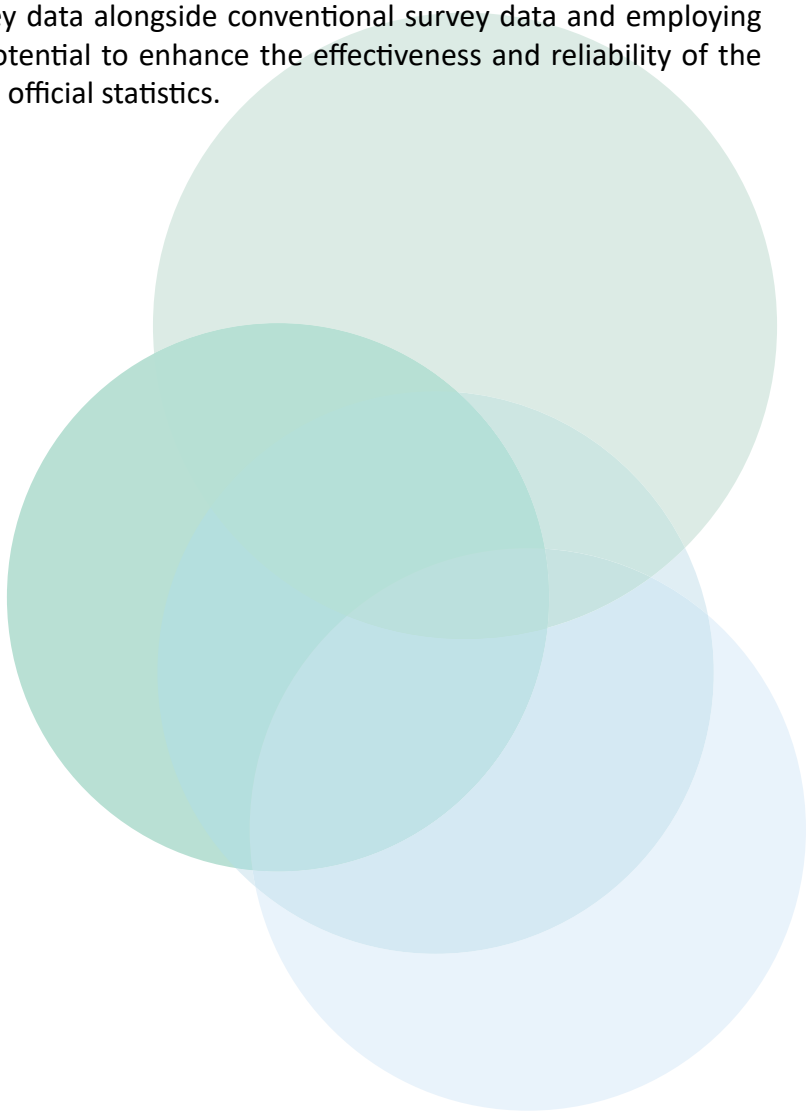
The quarterly Istat official survey on job vacancies and hours worked (VELA) produces two main indices of labour input: the number of hours worked and the number of hours worked per capita. In particular, this work focuses on hours worked per capita determined, for each section of economic activity, by dividing the total number of hours worked by the average number of job positions. In order to produce the seasonally adjusted (S.A.) indexes series a direct approach was in use, which consists in individually seasonally adjusting all series, whether they refer to a single economic activity section or to an aggregated sector. This method led to some not negligible issues, such as the presence of out-of-range values, especially in periods characterised by large fluctuations. Moreover, Eurostat guidelines suggest a direct approach only when dealing with component series with similar characteristics, which, according to our analyses, is not the case. Therefore, based on the similar positive experience in Istat for the Labour Force Survey (LFS), we have applied the indirect seasonal adjustment method to the VELA series. In the case of hours worked per capita, the application of the indirect approach has been twofold: by seasonally adjusting the numerator separately from the denominator and by obtaining the S.A. aggregated series by combining the single component S.A. series. As a matter of facts, applying this strategy has reduced the number of out-of-range data points to zero, while maintaining the general pattern of the series, as shown by the analysis of revisions, also ensuring greater consistency with international best practices. Moreover, the application of the same methodological approach in the seasonally adjustment process of similar series derived from different official surveys provides not only a gain in terms of comparability but also in terms of mutual validation opportunities. An attempt has been made by comparing the short- term dynamics resulting from the seasonally adjusted series of employees from the LFS and those of job positions from VELA. The two series show a quite similar pattern starting from the 2019, in the three main economic activity aggregates Industry, Construction and Services. This cross-validation process between the two series represents a step forward in improving the quality of the results obtained from the two surveys.

Machine Learning for Enhanced Estimation of Palm Oil Production: A Comparative Analysis of Random Forest and K-Nearest Neighbor

Ayu Paramudita[1]

[1]Statistics Indonesia, Jakarta, Indonesia

The conventional method of collecting data, especially from large and medium-sized enterprises, poses challenges for the National Statistics Office in generating official statistics. In this study, we explore the application of machine learning techniques to estimate data production, aiming to enhance the reliability of statistical outputs. Our focus is on estimating the production of Crude Palm Oil (CPO) and Palm Cooking Oil, given their pivotal role in Indonesia's economy and their significant impact on global markets. Beyond utilizing survey and administrative data, we incorporate environmental and climate conditions' variability into our regression model, recognizing their associations with the production process. The dataset is split into training and testing sets in an 80:20 ratio, and we compare the performance of the random forest and k-NN methods. The random forest model, after training, explains 89 percent of the variability in CPO and Palm Cooking Oil production, outperforming the k-NN model, which explains only 65 percent. Furthermore, the root mean square error of the random forest model is slightly lower than that of the k-NN model. In conclusion, our findings suggest that the random forest method exhibits superior performance in predicting the output values of crude palm oil and cooking oil compared to the k-NN method. In the end, integrating non-standard survey data alongside conventional survey data and employing machine learning techniques has the potential to enhance the effectiveness and reliability of the business processes utilized in generating official statistics.



Timeliness and punctuality in the State Statistical Office

Mr Ivan Spasovski[1]

[1]State Statistical Office, Skopje, Macedonia

The purpose of this paper is to present the timeliness and punctuality of the publication of output data, in news releases and in thematic publications, through the quality indicators related to Concept 14 of the SIMS v2.0 (Principle 13 of ESS Code of Practice, 2017).

This paper analyses the calculated quality indicators at individual and aggregate level and evaluates the quality of timeliness and punctuality of publication of output data. This analysis ascertains the situation and presents numerical assessments, descriptive explanations and interpretations of the quality of timeliness and punctuality of the publication of output data from 2016 to 2020. After the analysis, suggestions were made on how to improve the quality of publishing the output data in the State Statistical Office.

This paper also presents the inputs for the calculation of quality indicators such as: the statistical domain, the last day of the reference period, the year of the statistical survey and the planned and realised dates of publication of the news releases and thematic publications. Also, this paper provides explanations for standardisation of input and output data such as the last day of the reference period, correspondence of news releases and thematic publications with statistical surveys, and explanations of calculations of the quality indicators at individual and aggregate level.

Quality indicators TP1-Time lag - first results, TP2-Time lag - final results and TP3-Punctuality - delivery and publication are prepared at the level of statistical survey, statistical domain and periodicity of publication of the news releases and thematic publications for 2016, 2017, 2018, 2019 and 2020. The quality indicators are prepared according to: ESS Handbook for quality reports, 2014 (Eurostat), ESS Quality and performance indicators (QPI), 2014 (Eurostat).

Quality Improvement of Design Based Estimation by Different Administrative Records

Gülser Pınar Yilmaz Ekşi[1], Duygu Kılıç[2]

[1]Turkstat, Ankara , Turkey, [2]Turkstat, Ankara, Turkey

One of the indicator of quality in surveys is accuracy consisting of variance and bias components. This study concentrated on variance estimation in order to improve design estimates by taking into account different auxiliary information from administrative records and different calibration methods for Life Satisfaction Survey 2023. The quality of estimates are reviewed by variance estimation.

Survey errors are classified into two groups: “sampling error” and “non-sampling errors”. When evaluating sampling errors, the aim is to measure how much and how the randomly selected sample represents the population. Sampling errors are a measure of the variation between estimates obtained randomly from different samples and the amount of random error in the estimates is defined as precision. The most commonly used measure to represent the precision is expressed under variance estimation of complex surveys. According to the recommendations of EU regulations, precision requirements are concerned, can be identified in terms of variance by taking into account survey specificities such as indicators and regional disaggregation.

In recent years, administrative records have gained importance in terms of improving accuracy of design-based estimates in the statistical production process of administrative records. In order to reduce bias and increase the accuracy of estimates calibration methods by taking into account design based estimation auxiliary information from administrative records. In this context, the Life Satisfaction Survey 2023, which is aimed to produce estimates on a total basis in Turkey, design-based estimates using different auxiliary information from different administrative sources and sampling variance, coefficient of variation, confidence interval and etc. by different calibration techniques. The scope of this study is variance estimation by taking into account sampling design, type of indicators and suitable methods.

The aim of this study is to reduce bias and increase the accuracy of the estimates in Life Satisfaction Survey 2023 performed by Turkish Statistical Institute using the calibration estimators. For this scope, Integrated Calibration Method used in TURKSTAT, “Generalized Regression Estimator (GREG)” and raking calibration method are taken into account. It is aimed to estimate and compare with variance estimates. GREG, Raking and etc. are used as calibration techniques for the key indicators of the Life Satisfaction Survey 2023. Life Satisfaction Survey 2023 in Türkiye has multi- stage cluster sampling design. In this study, it is used ReGenesees R package developed by ISTAT is used to estimate variance of design based and model assisted analysis of complex sample surveys.


Sweden
2001
Stockholm, Sweden
The International Conference on Quality in Official Statistics (Q2001) Stockholm, Sweden, 14–15 May 2001
SCB
Statistics Sweden
eurostat

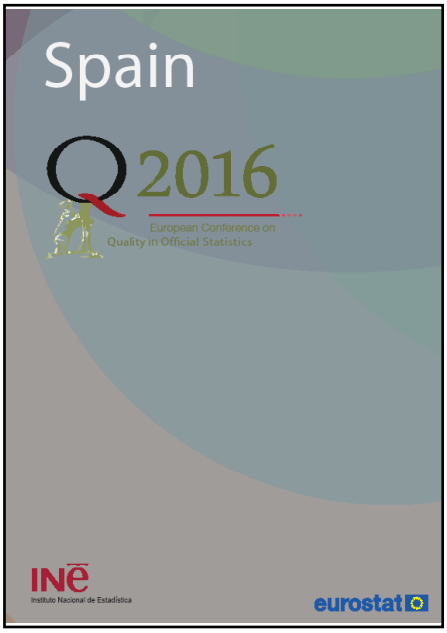

Germany
2004
European Conference on Quality and Methodology in Official Statistics (Q2004)
DESTATIS
Statistisches Bundesamt
eurostat


United Kingdom
2006
European Conference on Quality in Survey Statistics
Office for National Statistics
eurostat


Italy
2008
European Conference on Quality in Official Statistics
Istat
eurostat


Finland
Q 2010
Helsinki
European Conference on Quality in Official Statistics
Statistics Finland
eurostat


Austria
QM 2014
european conference on quality in official statistics
STATISTICS AUSTRIA
eurostat


Spain
Q 2016
European Conference on Quality in Official Statistics
INE
Instituto Nacional de Estadística
eurostat


Poland
2018
European Conference on Quality in Official Statistics
Statistics Poland
eurostat


Greece
athens
2012
European Conference on Quality
in Official Statistics (Q2012)
eurostat


Lithuania
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2022 VILNIUS
LRV.LT
eurostat


Portugal
EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL
INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

version 2024.05.29

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat