Web Intelligence
Network Conference
From Web to Data

WEB INTELLIGENCE NETWORK CONFERENCE FROM WEB TO DATA

February 4-5, 2025, Gdansk

BOOK OF ABSTRACTS

Table of contents

SESSION I. WEB SCRAPING AND INFRASTRUCTURE 4

URL finding: looking back, progress and plans for the future 4

Identifying official firm websites: a comparison of machine learning-based URL retrieval methods and AI-powered search engines 5

State of play of the Data Acquisition Service (DAS) of the Web Intelligence Hub (WIH) 5

Providing online based enterprise characteristics with the Web Intelligence Hub 6

Firms innovation capabilities and corporate websites: evidence on Italian SMEs 6

SESSION II. OJA USE CASE 8

Leveraging online job advertisements for green skills analysis in France 8

Development of a labour shortage indicator by occupation from OJA data 8

Combining online job advertisements with probability sample data for enhanced small area estimation of job vacancies 9

Assessment of classifiers using pre-defined data source 10

Online job advertisements classification using encoder-like large language model 10

Using language models for extracting regions of employment from online job vacancies 11

SESSION III. OBEC USE CASE 13

Evaluating the completeness of business databases: a comparison with official records using web scraping techniques 13

Use of dedicated business website to enhance the statistical business register in the Netherlands 13

Applying survey sampling theory to web-scraped data: an analysis of OBEC data using the IPW estimator 14

Online based enterprise characteristics (OBEC) in Statistics Poland 14

Trade links: estimating interregional trade using weblinks 15

SESSION IV. NEW USE CASES 16

New use-cases of web data for official statistics 16

Measuring construction activities using advertisements from real estate portals. ESSnet WIN Work Package 3, Use Case 2 16

Analysing housing market in Tricity Metropolitan Area in Poland 17

Constructing a Hedonic House Price Index for Poland using listings data from 1996-2024 17

Using web data for energy statistics: methodology and key lessons 18

SESSION V. QUALITY OF WEB DATA 20

Web content based statistics: the challenges ahead 20

Exploiting the web presence of enterprises to improve NACE code classification 20

Assessing the quality of enterprise characteristics and online job advertisement classifications derived from web data 21
Quality guidelines for acquiring and using web scraped data 21
A specialised architectural framework for web data: the BREAL extension and enhancement....... 22
SESSION VI. METHODOLOGY ON USING WEB DATA 23
Selective scraping, sampling and other methods to minimize known causes of biases of web data23
Online job advertisements deduplication using large language model 23
Finding the Goldilocks data collection frequency for the Consumer Price Index 24
Integrating big data and administrative sources for estimating vehicle mileage and analyzing road traffic accidents 24

List of abstracts

SESSION I. WEB SCRAPING AND INFRASTRUCTURE

URL finding: looking back, progress and plans for the future

Heidi Kühnemann , Hesse State Statistical Office

Abstract: Identifying correct enterprise websites is an important prerequisite both for deriving online-based enterprise characteristics and for enhancing the statistical business register (SBR) with website data. URL finding is one way to do so by sending automated search requests to a search engine, scraping and processing the results and applying machine learning or deterministic models to link websites to enterprises. In this presentation, I will look back on the work done on URL finding during the Web Intelligence Network (WIN). In a common report of Work Package 2 and Work Package 3 Use Case 5 of the WIN, we have summarized and compared the methodologies of different URL finders in the European Statistical System (ESS). We concluded that there is a need for a common open source URL finder software that can be used by the whole ESS, which would need to be adaptable to some country-specific particularities. Next to these methodological considerations, we also discussed challenges of URL finding, especially concerning legal and technological factors. One concern is that search engines might store search queries and are technically able to identify the search queries as business register data from a statistical institute. On the technological side, scraping search engine results is a computational expensive task that needs high bandwidth and relatively big storage sizes, especially when rendering dynamic website content. URL finding is still continuously being improved. I would like to give some insight into a small comparative experiment at Statistics Hesse on different scraping methods. We compared the performance of four browsers and HTTP get requests the quality of retrieved content and duration of scraping. As it turns out, Selenium (with Chrome) and Playwright performed similarly in terms of data quality, but Playwright retrieved results faster. What are some potential trajectories for URL finding in the future? During the WIN project, we investigated if URL finding could be offered centrally as a service, e.g. in the Web Intelligence Hub (WIH). Currently, this does not seem viable, since SBR data is the starting point of URL finding and this confidential data cannot be uploaded to the WIH. One promising idea is to use data of Open Web Search (OWS), an EU project which aims to create a European search engine. OWS crawls the whole web and could be able to deliver tailored web data information to statistical institutes. This data source could potentially be used instead of commercial search engines. Additionally, the OWS index could be filtered for candidate enterprise websites, e.g. by searching for all imprint pages. Website owners are legally required within the EU to list contact and identifying information on an imprint page. These imprints could therefore be used to link the website back to the enterprise. Finally, I will present how Statistics Hesse aims to fully automate URL finding and put it into production. Users will be able to start the scraping process independently without any scraping knowledge and retrieve the results.

Key words: URL finding, SBR, OBEC

Identifying official firm websites: a comparison of machine learning-based URL retrieval methods and AI-powered search engines

Donato Summa, Italian National Institute of Statistics

Abstract: The project aims to evaluate and enhance the retrieval of official company websites starting from administrative information, a critical task for integrating web data into official statistics and enhancing the quality of business registers. The study evaluates a machine learning (ML)-based pipeline alongside AI-powered search engines, such as Perplexity and Chatgpt, to assess their effectiveness in identifying correct URLs. The pipeline, which combines web scraping and pre-trained classification models, is tested against AI systems leveraging natural language processing and generative capabilities. In addition, an AI agent based on open-source large language models (LLMs) deployable on-premises was developed and employed to explore its potential and ensure data and processing confidentiality. The methodology involves standardized queries, incorporating company names and locations, tested across both systems. The results are analyzed using metrics such as precision and recall, alongside qualitative assessments of contextual relevance and noise. Post-processing steps include cross-referencing retrieved URLs with manually verified business records to ensure accuracy and reduce false positives. By benchmarking AI-driven tools against established machine learning methods, the study identifies complementary strengths, such as the speed of AI systems and the domain-specific accuracy of tailored ML models. The findings demonstrate the potential of hybrid approaches, offering significant benefits by combining ML pipelines with cutting-edge AI tools to enhance the accuracy and scalability of URL retrieval processes. For official statistics, this innovation provides advantages such as faster integration of web data, improved quality indicators, and broader coverage of business entities. These advancements support the modernization of statistical production processes, ensuring more reliable and timely data for policymakers and researchers. The use of open-source LLMs also promotes transparency and cost efficiency, facilitating their adoption by national statistical institutes. This study highlights a pathway to enhance the reliability and usability of web-based data in statistical production, fostering greater efficiency in the transition to modern data sources.

Key words: URL-Retrieval, Web-scraping, Machine-Learning, AI-Powered-Search-Engines, Web-Intelligence

State of play of the Data Acquisition Service (DAS) of the Web Intelligence Hub (WIH)

Mátyás Mészáros, Eurostat

Abstract: Since 2021, Eurostat has been developing a general Data Acquisition Service (DAS) within the Web Intelligence Hub (WIH). The DAS is designed to ingest information from the Web, significantly contributing to the creation of new statistics or the enhancement of existing ones. This development of the DAS was closely followed by the consortium partners during the Web Intelligence Network (WIN) project. In the upcoming Web Intelligence Network conference ”From Web to Data”, we would like to showcase the history of the service, the lessons learnt during the development and current state of play that the DAS achieved by the end of the WIN project. The presentation will discuss the evolution of features, the notable achievements made under the WIN project, and depending on the allowed time we might include a short demo presenting the newest capabilities. Through this presentation the participants will gain insights into the specific technologies and infrastructure underpinning the DAS and get a comprehensive understanding of how DAS operates within the WIH.

Key words: Data Acquisition Service, WIH

Providing online based enterprise characteristics with the Web Intelligence Hub

Jacek Maślankowski, Statistics Poland

Abstract: This paper outlines a comprehensive framework developed for the Online-Based Enterprise Characteristics (OBEC) use case within the Web Intelligence Hub, a component of the Web Intelligence Network consortium. This international initiative involved multiple countries collaborating on the methodological, technical, and quality-related aspects of deriving enterprise-level indicators from website-based data. The objective of the OBEC use case is to generate a set of indicators that characterize enterprises, with a focus on those employing 10 or more individuals, based on their online presence. The paper details various methodological approaches adopted in the OBEC use case, including the processes for compiling and validating a comprehensive URL list of enterprise websites. It addresses key phases of data collection and processing, with an emphasis on compliance with legal and ethical considerations when handling large-scale web data. Specific attributes of enterprise websites analyzed include social media integration, e-commerce functionalities, and multilingual support. Additionally, the role of the Web Intelligence Hub in facilitating the generation of these indicators is examined. The discussion encompasses the framework for utilizing the Hub, its functional limitations, and the methodological differences between web-derived indicators and those obtained via traditional survey methods. The presentation also introduces the proposed set of output tables generated through the Hub, highlighting their structure and practical applications. The conclusion reflects on the challenges associated with maintaining large-scale platforms like the Web Intelligence Hub and ensuring the validity of URL databases. Recommendations for sustaining such initiatives and improving indicator quality are also provided.

Key words: OBEC, web scraping, big data

Firms innovation capabilities and corporate websites: evidence on Italian SMEs

Caterina Liberati, University of Milano-Bicocca

Abstract: Assessing the presence and intensity of innovative activity within a firm presents significant challenges due to the multifaceted nature of innovation. This complexity is particularly pronounced when examining micro, small, and medium-sized enterprises (SMEs). Traditional approaches to measuring innovation—relying on financial statements, surveys, or patents—prove especially limited for SMEs, as they tend to provide insights that are both outdated and incomplete [1]. Recent studies have advocated for the use of textual features extracted from corporate websites to evaluate innovation [3-6]. Our research advances this discussion by emphasizing the analysis of HTML structures rather than textual content. Textual elements of websites are subject to frequent and rapid changes, rendering them less reliable for stable modelling. In contrast, HTML code exhibits greater stability and is less influenced by language and sector-specific semantics. Drawing on evidence from ICT research on SMEs, we hypothesized that the HTML structures of websites belonging to innovative SMEs would differ from those of their non-innovative counterparts. This hypothesis was empirically tested using data from Italian manufacturing SMEs. Our starting dataset, sourced from the AIDA-Moodys platform, included comprehensive structural (such as the industrial sector and geographic location) and accounting information about enterprises. These data were augmented with web-based indicators derived from the corporate website of each firm. The final analysis involved a matched sample of 680 Italian manufacturing SMEs, enabling robust comparisons between non-innovative and innovative firms, as identified by the Italian Startups and SMEs Acts. An examination of the HTML tags and dimensions of homepage structures revealed notable differences. Innovative SMEs were found to have larger, more content-rich, and better-organized websites, indicative of high levels of creativity and technical expertise. These findings suggest that a website's HTML structure can serve as a valuable proxy for assessing firm innovativeness. While this study offers novel insights, certain limitations should be noted. The focus on manufacturing firms and the exclusive analysis of homepage data restricts the generalizability of the findings. Future research could address these limitations by expanding the investigation to other sectors and incorporating additional elements of websites, which would further enhance the robustness of the results. However, our findings suggest future research to include HTML code-based features alongside text-based ones in building web-based firm-level innovation indicators.

References: [1] F. Gault (Ed.) (2013). Handbook of Innovation Indicators and Measurement. Edward Elgar Publishing. [3] A. Gök, A. Waterworth, * P. Shapira (2015). "Use of web mining in studying innovation." Scientometrics, 102, 653–671. [4] P.J.H. Daas * S. van der Doef (2020). "Detecting innovative companies via their website." Statistical Journal of the IAOS, 36(4): 1239-1251. [5] S. Ashouri et al. (2022). "Indicators on firm level innovation activities from web scraped data." Data in Brief, 42: 108246. [6] C. Rammer * N. Es-Sadki (2023). "Using big data for generating firm-level innovation indicators – a literature review." Technological Forecasting and Social Change, 197: 122874.

Key words: SMEs, Innovation, Corporate websites, HTML tag

SESSION II. OJA USE CASE

Leveraging online job advertisements for green skills analysis in France

Emiline Roger, French Ministry of Labor (DARES)

Abstract: Since 2019, France has been collecting a national database of online job offers using web scraping techniques. This database has opened new perspectives for analyzing labor market trends, particularly through the production of timely conjunctural series. Our presentation aims to share insights from this experience, focusing on the challenges and opportunities of integrating such data into official statistics. Notably, we explore the potential of combining online job advertisements with vacancy survey data to create a robust product that enhances both the volume and structure of job offers at detailed local and occupational levels. This approach could support more precise analyses of vacant positions across regions and professions. In addition, our experience highlights another key use case: analyzing emerging skill needs, particularly in the context of the ecological transition. We present a case study on green skills demand in France, analyzing trends from 2019 to 2023 across selected professions, such as solar panel installers, heat pump technicians, and electric vehicle mechanics. By showcasing our methodologies and findings, we aim to contribute to the broader European discussion on leveraging online data for labor market insights.

Key words: OJA, Green Skills, Ecological Transition

Development of a labour shortage indicator by occupation from OJA data

Annalisa Lucarelli, Italian National Institute of Statistics

Abstract: In the field of web data, online job advertisements (OJAs) represent a valuable and innovative source that can complement the official statistical production on labour demand. They enhance the understanding of labour demand, offering insights into a job market shaped by rapid economic changes. The use of OJA data is a well advanced use case, in fact, has been the subject of several Eurostat projects in recent years: the ESSnet Big Data I pilot project (2016-2018) and the Big Data II implementation project (2018-2020), followed in 2021 by the Web Intelligence Network (WIN) project, which focuses on the production of experimental statistics based on OJA data. OJAs offer more possibilities for analysing labour market trends and capturing new emerging needs of employers than traditional surveys. This is due to the granularity of the information contained in the advertisements and the high frequency with which this information is made available. Indeed, the OJA data provide a great deal of information on job characteristics (e.g. occupation, location, type of contract, working hours and pay), employer characteristics (e.g. economic activity), job requirements (e.g. education, skills and experience) and also on the advertisement itself (e.g. when the advertisement is published and when it expired from the website / job portal). The aim of this work is to utilize OJA data for an in-depth analysis of labour demand, focusing on emerging occupations and the skills sought by employers. Specifically, the objective is to define an experimental labour shortage indicator and contribute to the study of mismatch between labour supply and demand. This indicator, broken down by occupation and geographical area, enables the identification and ranking of occupations and geographical areas at risk of labour shortages. It is important to recognise that although OJA data provide a comprehensive and up-to-date perspective on labour demand, there may be discrepancies when compared with traditional survey data. These discrepancies could be due to potential biases within the OJAs, which need to be thoroughly analysed and understood before the resulting indicators are used to support official statistics and inform policy-making. It is therefore essential to ensure that these indicators are derived from accurate and reliable data and to carry out a thorough evaluation of their accuracy and robustness. To this end, this work proposes a quality assessment of the indicators, addressing some of the key quality issues associated with OJA data. First, the accuracy of the classification for the variable of interest, i.e. occupation, is taken into account. Next, the representativeness of the occupational distribution in OJA data is analysed by comparing it with Labor Force Survey (LFS) data. Finally, the consistency of the indicators with general economic trends is examined through a time series comparison with other macroeconomic indicators, (i.e. monthly employment rate, gross domestic product trends, industrial production index, construction production index, and services turnover index).

Key words: OJA, labour shortage, quality issues

Combining online job advertisements with probability sample data for enhanced small area estimation of job vacancies

Donatas Šlevinskas, State Data Agency, Statistics Lithuania

Abstract: Auxiliary information available in probability sample surveys is important in obtaining as accurate parameter estimates in the finite population and its domains as possible. Having auxiliary data related to the study variables at the unit or domain (area) level provides a range of models to choose from that can improve the direct design-based estimates. In this application, we combine the probability sample data on job vacancies with online job advertisements (OJA) information and administrative data to improve the estimates of job vacancy totals in small population domains. One of the ways of integrating big data samples such as OJA is to stratify the population into big data stratum and a missing data stratum (Kim and Tam, 2021). Then, apply the calibration method with different imposed conditions on artificial strata to exploit big data sample as complete auxiliary information. However, typical calibration methods rely on unit-level linear relationship between the variables. A model-calibration (MC) approach (Wu and Sitter, 2001) provides more flexibility by allowing general unit-level models. It improves probability sample-based estimates in small areas by predicting the study variable in the big data stratum using a measurement error model. These predictions are then used as constraints in the calibration process. Various modelling techniques can be used to obtain these predictions, including linear regression, nonlinear parametric models, or non-parametric methods such as k-nearest neighbours. The estimated totals from the MC and data integration approach are further refined using the Fay-Herriot area-level model (Fay and Herriot, 1979) to produce empirical best linear unbiased predictions (EBLUPs) for domain totals. This methodology can also be applied to other estimation problems where additional, incomplete data is available from administrative or web data sources.

Key words: data integration, small area estimation

Assessment of classifiers using pre-defined data source

Vladimir Kvetan, Cedefop

Abstract: This work builds on the cooperation between Cedefop, Eurostat and the European Labour Authority. It aims to evaluate data from the Web Intelligence Hub (WIH), online job advertisements (OJA), and the data production system (DPS), as well as data sourced from a single known source. The analysis covers data collected via API from April to June 2024. It benchmarks this known source job vacancy data against WIH DPS aggregates, focusing on job postings, sectors, occupations, and skills trends without establishing ground truth. The report identifies data gaps and similarities, proposing collaborative improvements among the involved organisations. The report outlines the data intake system, detailing the daily ingestion of approximately 110,000 job postings. This process identifies and eliminates duplicates, resulting in a deduplicated dataset of 6.7 million unique postings from 12.4 million ingested entries. Language pipelines process postings, predominantly in French, German, Dutch, Swedish, and Czech, ensuring cross-language comparability. Comparative analysis reveals challenges in data classification and completeness. For example, discrepancies exist between occupational classifications, with manual and machine learning-generated codes differing by large margins in specific cases. Similarly, the alignment of geolocation distribution is rather high at the country level but weakens with finer granularity due to differing classification methods. Economic activity and skills dimensions also exhibit varying completeness and consistency levels, highlighting potential improvement areas in both data sources. Key findings include:

- Data Similarities and Differences: The known source data relies on hard-coded classifications, whereas WIH DPS employs machine learning approaches, leading to distinct strengths and limitations in each dataset.
- Classification Challenges: Variability in occupational and sectoral classifications necessitates detailed investigation to harmonize methodologies.
- Data Completeness: Known source data lack coverage in key variables, such as skills and economic activity, and structural fields are often incomplete. However, WIH DPS partially compensates through textual analysis.
- Opportunities for Collaboration: Joint efforts between Cedefop, Eurostat, and EURES member states are essential to address data quality issues and optimise classification methods. The report emphasizes the importance of collaborative improvements, leveraging the strengths of both datasets.

Key words: Web Intelligence Hub, Online Job advertisements, job vacancies, data classification, skills analysis

Online job advertisements classification using encoder-like large language model

Mikołaj Tym, Poznan University of Economics and Business

Abstract: Online job advertisements are an accessible and abundant source of data for labour market statistics. However, these advertisements must be classified to analyze occupation demand and monitor changes in the labour market effectively. Inspired by the Web Intelligence Classification Challenge, we propose a scalable approach to classify online job advertisements into International Standard Classification of Occupations (ISCO) codes. We fine-tuned a lightweight Large Language Model (LLM) with 400 million parameters, employing an encoder-transformer architecture for this classification task.

The model was trained on rigorously preprocessed European Skills, Competences, and Occupations (ESCO) data, enabling classification across all 436 ISCO codes, including less common categories. This approach enhances the model`s applicability to unseen data. Leveraging a clean and synthetic dataset helped mitigate biases present in real-world data, such as associating company names with specific occupations. Consequently, our fine-tuned model predicts ISCO codes with human-interpretable probabilities. Scraped job advertisements often include irrelevant content, such as GDPR clauses, cookies, job benefits, or company descriptions. To address this, we trained the model not only for ISCO classification but also to determine whether a sentence contains job duty information. This step filters irrelevant content, allowing the model to focus on job duties and required skills, thereby improving classification accuracy. This methodology is applicable to other text sources containing extraneous content, broadening its potential impact on improving data quality across various domains. Given the hierarchical nature of ISCO codes, a reclassification mechanism was introduced to consider contextual cues such as required education levels and managerial responsibilities. For instance, a job posting for an AI engineer with team management duties is reclassified as Research * Development Manager rather than Specialist. Unlike approaches focused solely on training data quality, we emphasize preprocessing and filtering the inference data. Using a synthetic dataset avoided model biases, while the duty classifier improved input relevance. This novel methodology supports the processing of millions of advertisements using a lightweight LLM, even on modest hardware, thus enabling reliable and scalable labour market analytics.

Key words: ISCO codes, job advertisement classification

Using language models for extracting regions of employment from online job vacancies

Adam Tsakalidis, Cedefop

Abstract: A joint effort by Cedefop and Eurostat to aggregate and analyse hundreds of millions of online job advertisements (OJAs) in the context of the Web Intelligence Hub provides us with a rich source of data. This data contains valuable information about the demand of occupations and skills across different sectors and countries over time. Building natural language processing and machine learning models that can classify each OJA across such different variables with high accuracy can enable stakeholders, policy makers and scientists to monitor the evolution of the labour market, as expressed in OJAs, in almost real-time across the EU. Importantly, classifying the region of employment in OJAs with high accuracy, can enable such studies at a more fine-grained level. This work focuses on identifying the region of employment in OJAs, based on their description. Without loss of generalisation, we focus on the NUTS-2 region classification of OJAs written in the Greek language, though more fine-grained analysis is possible via our proposed approach. To achieve our goal, we firstly leverage the power of large language models (LLMs) to generate a sample of annotated data (i.e., mapping OJAs to their respective NUTS-2 regions) – a task that is typically performed by experts in a manual and thus expensive manner. Utilising these LLM-generated annotations, we then fine-tune a smaller Bidirectional Encoder Representations from Transformers (BERT) language model (Koutsikakis et al., 2020) – a process that can be completed with relatively low resources. We evaluate the performance of our fine-tuned BERT model on a manually (NUTS-2) labelled validation set of 528 OJAs written in Greek, with our results indicating that the

proposed approach can achieve very high accuracy (93%) on the NUTS-2 region classification task. Upon validation, the model is applied on a large-scale collection of 230,000 OJAs written in the Greek language during 2018-2023, extracting the NUTS-2 region of employment for each one of them. The resulting distribution of the number of OJAs per region shows very strong correlation against the official employment rates at the regional level in Greece. Overall, our main contributions include: (a) an automated, privacy-preserving and cost-effective way to generate NUTS-2 region labels for OJAs (via utilising state-of-the-art, yet resource-expensive, LLMs) to serve for model training purposes: (b) an approach to fine-tune a (BERT) language model on the data from (a), yielding very high accuracy on classifying regions from OJA descriptions: (c) demonstration of a solid correlation of the distribution of OJAs across regions against the employment rates at the regional level in Greece: and (d) an empirical investigation on the language used in the OJAs in specific regions, providing insights on the types of vacancies that are typically advertised (e.g., domination of tourism-relevant phrases in OJAs concerning the region of South Aegean).
References: Koutsikakis, John, et al. ”Greek-bert: The Greeks visiting sesame street.” 11th Hellenic conference on artificial intelligence. 2020.

Key words: Web Intelligence Hub, large language models, online job vacancies, natural language processing, region classification

SESSION III. OBEC USE CASE

Evaluating the completeness of business databases: a comparison with official records using web scraping techniques

Josep Domenech, Universitat Politècnica de València

Abstract: The Orbis database, a global corporate information repository managed by Bureau van Dijk, is a widely used resource for accessing company data across multiple countries. Similarly, one of its regional counterparts, the Sistema de Análisis de Balances Ibéricos (SABI), provides extensive company information for Spain and Portugal. This study evaluates the comprehensiveness of the SABI database by comparing it to the BORME, the official gazette for business registrations in Spain. Employing web scraping techniques, we identified companies listed in BORME that are absent from SABI. Our findings indicate that SABI covers only 38.3% of companies established between 2010 and 2023, with notable underrepresentation of younger firms, smaller enterprises, and certain sectors. Additionally, we observed a survivorship bias, with dissolved companies progressively less likely to be retained in the database over time. These results underscore critical biases in SABI, suggesting that researchers should exercise caution when using this database for economic and business research.

Key words: data quality, web scraping, reliability

Use of dedicated business website to enhance the statistical business register in the Netherlands

Arn*out van Delden, Statistics Netherlands*

Abstract: For many national statistical institutes (NSIs) a statistical business register (SBR) contains variables such size class, the economic activity (NACE code), and web site address (URL) for large number of legal units and statistical units. Keeping the SBR up to date is crucial for obtaining accurate business statistics. Manual editing of the variables in the SBR is cumbersome and time consuming. Methods that can help to reduce the amount of manual editing are very welcome and potentially interesting also for other NSIs. In this presentation we give an overview of the contribution by Statistics Netherlands to Use Case 5 of Work Package 3 of the Web Intelligence Network project on the use business websites to enhance the SBR and to try to reduce the manual editing work. Our contribution is threefold. Our first contribution concerns the process of finding URLs of legal units. We used quarterly data with scraped URLs that we have obtained from an external commercial provider called 'Data Provider'. We have studied how we could improve the linkage between the URLs in the commercial data set(s) and the legal units in the SBR, using non-unique identifiers. The second contribution concerns the use of texts scraped from those websites to predict NACE codes at the 5-digit level using machine learning (ML). Specifically we studied the impact of different types of feature sets on the performance of a number of ML models. These feature sets differed in the extent to which they were content-related to the classes that they aim to predict. We used the set that gave the best performance in our third contribution. That contribution concerns the development of a method that aims to identify for which of the legal units the currently registered NACE code is likely to be incorrect. This method combines two models: a classical ML model to predict NACE codes from website texts and a logistic model to predict the probability that a unit is misclassified as a function of background variables such as the size of the legal unit. Both models are re-fitted during a number of iterations until it convergences in a so-called estimation maximization (EM) algorithm. The algorithm is converged when the probabilities for each unit to be misclassified hardly change any more. The output of the model can, when implemented, be part of a score function that prioritizes which units manual editors should check for misclassifications. In the presentation I will highlight the main findings of each of the three contributions.

Key words: misclassified NACE codes, URL finding, feature selection, machine learning, website texts

Applying survey sampling theory to web-scraped data: an analysis of OBEC data using the IPW estimator

Vilma Nekrašaitė-Liegė, State Data Agency, Statistics Lithuania

Abstract: The growing availability of web-scraped data presents new opportunities for research but raises significant methodological challenges. The main challenge is that there are no possibilities to web-scrape all elements of the population (the population might be too large, or some URLs are not working, etc.). Thus, it is possible to treat web-scraped data as a nonprobability sample. Traditional survey sampling theory is built on the premise of probability samples, where each element’s inclusion probability can be estimated and is independent of the study variable. When these assumptions are violated, as is common with web-scraped data, applying traditional estimators can lead to biased results. This study explores the application of survey sampling theory to web-scraped data, specifically focusing on data collected for Online Based Enterprise Characteristics (OBEC). We investigate whether the Inverse Probability Weighting (IPW) estimator can be a viable solution to address bias in nonprobability samples derived from web scraping. By constructing weights based on auxiliary variables that are available for all elements of the population from business registers, IPW attempts to correct the bias compared with the naïve estimator, where the data are treated as a simple random sample. Our analysis shows that, when appropriate auxiliary information is incorporated, the IPW estimator can mitigate bias and yield more representative estimates. Our findings contribute to the growing field of research on the use of nonprobability samples and highlight the importance of adapting survey sampling theory for modern data sources. The IPW estimator provides a feasible method for improving the validity of web-scraped data analysis. This work underscores the need to incorporate survey sampling theory with web-scraped data to get more reliable results.

Key words: Web-scraped, nonprobability sample, IPW estimator

Online based enterprise characteristics (OBEC) in Statistics Poland

Ewelina Niewiadomska, Statistics Poland

Abstract: The methodology of the ICT usage survey in enterprises, conducted since 2005, is based on a model questionnaire developed by Eurostat in collaboration with experts from statistical offices across EU member states, the European Commission and the OECD. The survey targets enterprises in the manufacturing and services sectors employing more than 9 employee. The scope includes priority topics for the European Commission, such as artificial intelligence, data analytics, cloud services, cybersecurity, and ICT specialists. Additionally, the questionnaire contains questions about widely known technologies, such as websites and social media, more relevant to less digitally advanced enterprises. The growing use of automated methods for extracting content, such as web scraping, has created opportunities for public statistics to explore new data sources. Web scraping involves the automated collection of data from websites using software tools or scripts, enabling the extraction of structured information from both static and dynamic web pages. Analysis of the website content of enterprises participating in the survey demonstrated that the data obtained in this way could serve as a valuable resource, enhancing the quality of indicators derived from questionnaire-based surveys. The study on the use of ICT in Poland covers a sample of approximately 20,000 enterprises, which corresponds to a potential 20,000 websites. The primary goal was to develop a solution comprehensive enough to collect data from both dynamic and static websites. Data extraction from enterprises websites took place without download and store their content. The next steps focused on data processing and validation. The data obtained is currently used to verify the answers in the surveys. Experimental statistics in the ESSNet WIN project currently uses the Web Intelligence Platform, which enables the collection, storage and processing of data from websites.

Key words: web scraping, ICT, enterprise characteristics

Trade links: estimating interregional trade using weblinks

Juergen Amann, OECD

Abstract: Trade flows between regions within the same country are often larger than with their international counterparts. However, there is limited information on regional trade within national boundaries and its geographical and sectoral patterns, and therefore, little research has been done on its impact on regions. We address this issue by illustrating how information on links between websites (weblinks) can be used alongside Orbis firm-level data to produce proxies for interregional trade. We demonstrate the usefulness of our work by producing various benchmarking exercises of either micro-founded or model-based alternatives.

Key words: weblinks, subnational data, subnational trade, firm-level

SESSION IV. NEW USE CASES

New use-cases of web data for official statistics

Olav ten Bosch, Statistics Netherlands

Abstract: This presentation provides a high-level overview of the Web Intelligence Network (WIN) project`s third work package (WP3). WP3 aimed to explore the potential of new web data sources for official statistics, focusing on six specific use cases: 1. UC1 Characteristics of the real estate market 2. UC2 Construction activities 3. UC3 Online prices of household appliances and audio-visual, photographic and information processing equipment (and generalising the data collection to other activities) 4. UC4 Experimental indices in tourism statistics (hotel prices) 5. UC5 Business register quality enhancement 6. UC6 Faster Economic Indicators using new data sources For the first four use cases, a similar approach was adopted, involving exploring the new data source, developing scraping software, data acquisition and processing, modelling and interpretation to dissemination of the experimental results. However, each use case presented specific challenges related to data quality, availability, and the complexity of the underlying statistical concepts and had to use different techniques such as interpretation, standardization, image processing and deduplication techniques. The fifth use case, business register enhancement, took a different approach. Because the aim is to enhance a statistical business register (SBR) using any data source from the web in whatever way, it had to use a mix of techniques all combining information that the NSI already has in the SBR or other databases with readily available web data. The activities can be clustered into two main topics: (1) URL finding: improving or completing the business register with URLs or other web source identifiers and (2) using these web data to derive statistical variables. One example of such a variable is economic activity (NACE). Another example of the results is improving or deriving contact information contained in the SBR. The sixth use case, faster economic indicators, aimed to develop timely and accurate economic indicators using novel data sources. This involved exploring the potential of real-time camera data for a busyness indicator, which appeared to be a difficult task due to a changing privacy culture which causes limitations in data availability. This presentation provides a foundational understanding of the WP3 activities. It will delve deeper into specific use cases that lack dedicated presentations, while providing a concise overview of those that will be discussed in more detail.

Key words: webdata, webscraping, deduplication, interpretation, businessregisters

Measuring construction activities using advertisements from real estate portals. ESSnet WIN Work Package 3, Use Case 2

Tobias Gramlich, Hesse State Statistical Office

Abstract: A large share of the real estate market takes place online. Use Case 2 of Work Package 3 investigated the use of advertisements on online real estate platforms for producing official statistics in the context of newly constructed houses and apartments. Data from online real estate platforms are a valuable data source: They allow interesting and important insights into recent developments and may add to empirical foundations of political or societal discussions. The presentation gives an overview over the overall goal, techniques, data sources and results of Use Case 2, and discusses obstacles and shortcomings of data and methods used. Since there is a significant time gap between the completion or marketing of newly constructed buildings and apartments and the publication of official statistics on these, one of the tasks in Use Case 2 was to explore data and methods for producing early estimates of these numbers using current web data. While results derived from aggregated web data can be published much earlier than official statistics on construction activities, they exhibit notable discrepancies compared to the official figures. Additionally, the relationship between these numbers seems not to be stable over time. The presentation will present and discuss results mainly for Germany (Hesse and/or Berlin and Brandenburg).

Key words: webscraping, construction activities

Analysing housing market in Tricity Metropolitan Area in Poland

Olgun Aydin, Gdansk University of Technology

Abstract: There are dozens of real estate listing websites, with thousands of listings – not using it for the analysis of the changes happening in the market would be a huge miss. This research study focuses on gathering data from real estate listing portals using Web Scraping technology. Retrieved data was used to analyse the housing market of the Tricity Metropolitan area in Poland which covers Gdańsk, Gdynia and Sopot. Listings were gathered from two major portals, which are used by people selling, buying and renting flats in Gdańsk Metropolitan Area – olx.pl and trojmiasto.pl. After clean up and storage in the database, the data was made available through a user-friendly web application. The web application provides users to foresee changes in price trends in each district in a form of time-series graphs. Additionally, financial indices determined over gathered data, such as Price to rent ratio and Rental yield as well as their changes over time are provided through the web application. This allows users to find the most expensive districts in Tricity and those that are the most attractive for the real-estate investors.. The study presents real-world use cases in which Web Scraping technologies such as the open-source Selenium framework can be used in order to gather data in an automatic manner. Such approaches have various advantages such as low cost and ability to gather huge amounts of data in a reasonable time. Collected data can be later used to develop Machine Learning models or for contribution to official statistics.

Key words: Web Scraping, Selenium, Data Visualization

Constructing a Hedonic House Price Index for Poland using listings data from 1996-2024

Radosław Trojanek, Poznan University of Economics and Business

Abstract: House price indices (HPIs) are crucial in monitoring housing market trends and informing various stakeholders, including the general public, banks, the financial sector, and the government. While transaction prices are considered the most reliable indicator of a property`s market value, they often face delays in data availability due to the time required for public institutions to enter information into databases. Moreover, transaction prices may not reflect the housing market`s current state, as they are typically set several months before the transaction date. In light of these limitations, this study explores the use of listing prices as an alternative data source for constructing HPIs in Poland. Using a unique database of over 3 million housing offers from 27 cities in Poland between 1996 and 2024, hedonic HPIs were constructed for each city using the rolling time-dummy method. The data were obtained from archival advertisements and web-scraping procedures, and extensive cleaning and preprocessing were performed to ensure a homogeneous database with a uniform system of recording variables. The aggregate index for Poland was then constructed, considering the share of transactions in 2019, and compared with the official index published by the Central Statistical Office (CSO). Despite differences in data sources, methods, and subject coverage, the listings-based HPI showed similar price changes to the official index from 2010 to 2024. The listings-based HPI has provided information on the Polish housing market for longer than officially published indices, extending back to 1996. Furthermore, listings data offers the advantage of timely data availability without significant delays, unlike transaction-based indexes, which can lag by several months. The results of this study suggest that listings-based HPIs can serve as a good indicator of housing market changes, mainly when transaction data is unavailable or delayed. The ease of gathering listing data and the ability to provide HPIs without delay highlight the potential for more timely and comprehensive housing market monitoring. This study contributes to the growing body of research on the usefulness of listing data for constructing house price indices. It provides valuable insights for policymakers, financial institutions, and other stakeholders seeking to understand and respond to housing market dynamics in Poland. While the listings-based HPI has some limitations, such as the potential for offer prices to differ from actual transaction prices and the varying scope of information available across different data sources, the overall similarity with the official index suggests that these limitations do not significantly hinder its ability to capture housing market trends. Future research could explore methods to further refine the listings-based HPI, such as incorporating additional variables or adjusting for potential biases in offer prices. In conclusion, this study demonstrates the feasibility and value of constructing a hedonic house price index for Poland using listings data from 1996 to 2024. The resulting index provides a more comprehensive and timely picture of the Polish housing market, complementing official transaction-based indices and offering valuable insights for various stakeholders.

Key words: housing market, listings, hedonic methods

Using web data for energy statistics: methodology and key lessons

Herbeth Sandrine, Information Management S.A.

Abstract: The growing prominence of web data in official statistics presents unique opportunities to enhance timeliness, granularity, and comprehensiveness. Within the ”Tapping New Data Sources” project awarded by Eurostat, Artemis Information Management S.A. developed innovative methodologies to integrate web-based and alternative data sources into the statistical production process for energy statistics. This approach aimed to complement existing datasets while addressing the increasing demand for timely and relevant data. The methodology employed combined advanced web scraping, data mining, and validation techniques tailored to the specific requirements of energy statistics. Web data sources, including platforms from industry associations, agencies, and international organisations, were identified and assessed based on strict criteria such as timeliness, coverage, reliability, and compatibility with Eurostat`s

methodological standards. Significant emphasis was placed on harmonising data to ensure alignment with existing statistical frameworks and comparability across domains. Multiple challenges emerged during the process, resulting from the inherent nature of web-based data. One of the primary obstacles was the variability in data accessibility and formats. Many indicators were dispersed across diverse platforms and presented inconsistently. To address this, customised algorithms were designed to adapt to different data structures and publication schedules. Another major challenge was ensuring data quality and retrievability, as web data often requires extensive pre-processing to resolve issues such as incomplete records, ambiguous metadata, and unstructured formats. Tools were developed in Python using the Streamlit library to build interactive web applications and automate data extraction from platforms such as ENTSO-E (European Network of Transmission System Operators for Electricity), Energy Institute, GIE (Gas Infrastructure Europe), and ENTSO-G (European Network of Transmission System Operators for Gas). Each platform required specific solutions, such as handling API authentication, selecting relevant datasets, and managing diverse data formats. For ENTSO-E, additional challenges included implementing advanced API calls for specific indicator- requests and dynamic adjustment for country-specific parameters. GIE and ENTSO-G benefited from tailored API integration that supported flexible querying by date, region, and indicator type. Outputs were standardised in CSV format, ensuring usability and seamless integration into statistical workflows. The challenges encountered included navigating proxy restrictions, non-standardised data formats, managing API requirements, and obtaining web-scraping authorisation from data providers. These were addressed through adaptive programming techniques, dynamic error-handling mechanisms, and contact with data providers. While some platforms, such as Eurofuel, posed barriers due to persistent proxy issues and challenges in extracting text from images using Python libraries like BeautifulSoup and the image function from the PIL library, others demonstrated the potential of web data to effectively complement traditional energy statistics. This project highlights the promise and complexity of integrating web data into official statistics. Methodologically, it emphasised the importance of flexibility, innovation, and collaboration in overcoming technical and procedural challenges. Practically, it demonstrated the value of web data in providing timely and detailed insights, ultimately enriching traditional statistical systems. By leveraging automated tools and web intelligence, the project established a scalable model for using web data in official statistics, showcasing its transformative potential while maintaining high standards of quality and reliability.

Key words: Energy statistics, Web scraping, Python, Streamlit, Timeliness

SESSION V. QUALITY OF WEB DATA

Web content based statistics: the challenges ahead

Fernando Reis, Eurostat

Abstract: Web content offers opportunities for the production of new statistics and improving existing ones. However, it is accompanied by challenges that must be addressed to ensure the reliability, representativeness and sustainability of web content based statistics. This presentation explores the challenges of using web content as a statistical resource in the follow up of the lessons of ten years of its exploration in the context of European official statistics. Key challenges include the instability of the web, where websites appear, disappear, or change, hampering the continuity of web data production. The duplication of objects, that we may select as statistical units, in web content complicate proper estimation of statistical aggregates. The algorithms required to extract information automatically from web content introduce a source of measurement error that cannot be ignored. Further, misinformation, or fake information found in web content, poses significant risks as intentional distortions often target precisely the variables we aim to measure. The presentation concludes by emphasizing the need for developing new methodologies, engage in cross-disciplinary collaboration, and invest in infrastructure and expertise to harness the power of web content for producing official statistics.

Key words: Web, Statistics, Quality, Challenges, Methodology

Exploiting the web presence of enterprises to improve NACE code classification

Johannes Gussenbauer, Statistics Austria

Abstract: NACE is the European standard hierarchical classification method used to classify enterprises according to their economic activity and as such builds the foundation of various business statistics and indicators. Accordingly, it is imperative to mitigate NACE code misclassifications as best as possible to avoid biased statistical outputs. Hence, national statistical institutes carefully classify and edit NACE codes continuously causing a significant depletion of time resources. In order to assist and expedite the manual editing and classification processes, we propose to exploit the increasing web presence of enterprises to predict their NACE codes on the basis of their scraped webpages. In this paper we propose the current state of an automated 1) flat classification procedure to predict the economic activity for a fixed NACE level (level 2-4) and 2) hierarchical classification procedure to predict the economic activity in terms of all the NACE 1-5 levels. Clearly whether a proposed classification model has the ability to support the manual editing processes will depend on its quality. Thereby it is detrimental to use evaluation measures which take the structure of the classification models into account. While there is a general consensus regarding the quality measures to be used to assess flat classification models, hierarchical classification models do not enjoy the same benefit. Hence in this paper we also present evaluation measures, including a novel customized performance measure, which are more suitable to assess the quality of hierarchical models than the standard evaluation metrics.

Key words: NACE Classification, Machine Learning, Webscraping

Assessing the quality of enterprise characteristics and online job advertisement classifications derived from web data

Ville Auno, Statistics Finland

Abstract: One of the key tasks within Work Package 4 of the Trusted Smart Statistics – Web Intelligence Network (WIN) project was to assess the quality of web scraped data. The purpose of this quality assessment is to provide insight into the quality and usability of web scraped data in official statistics production. This presentation focuses on the quality assessment of web scraped Online Job Advertisement (OJA) data. The OJA data was assessed in a couple of ways. First, the quality and stability of data sources were evaluated using quality indicators formulated for this task. Second, the classification accuracy of the data was evaluated through two rounds of manual annotation exercises. The quality and stability of data sources were assessed using several quality indicators, including the inclusion of relevant job portals, changes in the number and ranking of sources over time, and data consistency across versions. These indicators were computed using a standardized RMarkdown script for ten countries within the WIN project. The findings suggest that the stability and relevance of sources remain critical challenges, emphasizing the need for more robust source selection and monitoring. Additionally, the accuracy of different classifications for the web scraped OJA data produced by a machine learning model was evaluated. Classifications included the International Standard Classification for Occupations (ISCO) and other key categories such as economic activity, location, education, and working time. These were assessed through two rounds of manual annotation exercises. These exercises raised concerns about the accuracy and, therefore, the reliable usability of the data for official statistics production. These results highlight the need for significant improvements in both source stability and classification accuracy before OJA data can be reliably integrated into official statistics. Furthermore, while web data remains a prominent alternative data source, this study underscores the importance of rigorous quality assessment practices as a foundation for the adoption of alternative data sources in statistics production.

Key words: quality, quality assessment, classification, accuracy

Quality guidelines for acquiring and using web scraped data

Magdalena Six, Statistics Austria

Abstract: In our paper, we focus on quality aspects related to the use of web information at various stages of the statistical production process. Our theoretical considerations are supported by practical examples from the use cases in Work Packages 2 and 3. We begin by presenting a framework for a transparent landscaping process for web data sources. Subsequently, we discuss quality guidelines for data acquisition and then elaborate on key quality aspects during the throughput phase, including linking, coverage, comparability over time, measurement errors, and process errors. Particular emphasis is placed on the classification of text data into existing categories, as this represents a common processing step for scraped data. Finally, we share our insights gained from using a centralized scraping platform and offer recommendations for how a centralized scraping infrastructure should be optimally designed in the future.

Key words: quality guidelines, web data, scrapingplatform

A specialised architectural framework for web data: the BREAL extension and enhancement

Giuseppina Ruocco, Italian National Institute of Statistics

Abstract: The development of a Web Intelligence Hub (WIH) was one of the main objectives of the four-year project ”Trusted Smart Statistics - Web Intelligence Network (WIN)” started in 2021. The WIH is a capability that promotes the use of web data in official statistics and provides services for collecting, storing and processing web data. Within the project activities, the focus of the architectural task was the “Extension and Enhancement (E*E)” of the BREAL framework (BREAL - Big Data REference Architecture and Layers) through the insights gained from the WIH use cases. BREAL is a reference architecture resulting from the ESSnet Big Data II project, designed to support NSIs in planning Big Data investments. BREAL provides a set of artefacts to model and build a statistical process based on Big Data. Starting with the definition of the business and functional requirements of the WIH in terms of BREAL Business Functions (BBFs), special attention was paid to the user experience in order to highlight the impact of technical choices from the user’s perspective. The lifecycle and maturity of the WIH use cases were considered as key elements for the E*E of the BREAL framework. The analysis of the process steps that can be centralised in a shared infrastructure and the tasks that need to be performed locally in the national environments was carried out through several user stories. In order to accelerate the implementation of the WIH and to strengthen the web scraping community within the European Statistical System (ESS), the specification of “Who can do what” and “Who is the owner of each task” is essential to manage multiple actors and roles. Concerning the enhancement of BREAL, the specialisation of BBFs for web data, based on the experience gained during the project, aimed at achieving a balance between: 1) process standardisation and the specificities of each use case and statistical domain: 2) national regulations, country-specific practices and a shared infrastructure. With regard to the extension of BREAL, a new BBF “Strategy and Process Management” was designed to deal with: 1) unexpected issues that prevent the integration of a use case into production: 2) the management of technical or organisational aspects, planned activities and related output. This additional BBF supports the definition of a statistical production model for each use case and its transition to production. Overall, the E*E of the BREAL framework for web data was designed to turn the challenges encountered during the development of the use cases into opportunities. The lessons learned during the project point to possible ways of building a common vision and implementing a holistic approach for the integration of web data into official statistics at EU level.

Key words: Web data capability, reference architecture

SESSION VI. METHODOLOGY ON USING WEB DATA

Selective scraping, sampling and other methods to minimize known causes of biases of web data

Alexander Kowarik, Statistics Austria

Abstract: The presentation will explore how sampling techniques specifically designed for web data can be used to tackle issues such as bias that may occur before or during data collection. In addition, it will cover specialized methods that cater to the distinct challenges posed by web-scraped datasets. The presentation also explores the processes and methodologies for leveraging webscraped data in statistical production with a focus on mitigating biases. Webscraping is increasingly valuable for enriching statistical registers and analyzing dynamic datasets, but it introduces new challenges that must be addressed to ensure data quality and reliability. The presentation is structured into two parts. The first examines the webscraping process, starting with defining a population frame. It then covers data acquisition via webscraping, feature extraction through data wrangling, and model-based estimation for interpretation. The second part highlights methods tailored to webscraped data, addressing issues like over- and under-coverage, deduplication of units, and concept drift detection. This work emphasizes a comprehensive approach to data acquisition, validation, and modelling, ensuring robust and actionable insights from webscraped datasets.

Key words: selective scraping, methods, process, web scraping

Online job advertisements deduplication using large language model

Jakub Żerebecki, Poznan University of Economics and Business

Abstract: Online job advertisements are widely available, but creating reliable statistics based on them involves several challenges. Job advertisements are often published in multiple languages, requiring solutions capable of processing and comparing text across diverse linguistic features and structures. Data is collected from various web portals with different formats, layouts, and conventions, complicating the task of standardizing and comparing job postings. Determining content similarity is often challenging. While some duplicates are exact matches, others may vary slightly in wording or presentation, making semantic similarity detection a critical yet complex task. Duplicate job postings may include updated information or slightly extended details. Deduplication is essential for producing high-quality statistics from the web job offers, as companies often post the same job multiple time using different sites. Identification and removal of duplicates is a key for efficient processing of data and avoiding double counting. Our methodology detects five categories of duplicates: Full Duplicates - Two job advertisements are considered full duplicates if they are identical in every respect. Semantic Duplicates - Two advertisements are considered semantic duplicates if they advertise the same job position but express the job characteristics differently, either in natural language or in different languages. Temporal Duplicates - Semantic duplicates with varying advertisement retrieval dates.

Partial Duplicates - Two advertisements are partial duplicates if they describe the same job position but differ in their characteristics. Non-Duplicates - Advertisements that do not fall into the above categories are considered non-duplicates. We combine text preprocessing and cleaning, advanced sentence embeddings generated by large language models, and approximate nearest-neighbour search for efficient similarity querying. These techniques enable categorization of duplicates based on textual overlap and semantic thresholds. Using a dataset of 112,000 job advertisements from 400 EU websites, including synthetic duplicates in multiple languages, our approach demonstrated robust performance and provides precise deduplication solution, forming a solid foundation for producing reliable European statistics.

Key words: deduplication, online job advertisements, multilinguality

Finding the Goldilocks data collection frequency for the Consumer Price Index

Luigi Palumbo, Bank of Italy

Abstract: So far there is no theoretical framework to assist National Statistical Organizations (NSOs) in determining the adequacy of their data collection for the Consumer Price Index – whether it is too sparse, too frequent, or just right. We propose a novel framework designed to achieve a balance between reducing uncertainty in price measurement and minimizing the expenses associated with data acquisition, processing, and storage. This cost-benefit analysis is particularly relevant with the emergence of big data and alternative data sources, alongside regulatory requirements for NSOs to archive their data over extended periods. An illustrative application is provided through an examination of electricity and gas utility prices in the Italian unregulated market during 2023, items that were notably affected by the energy price crisis stemming from Russia's invasion of Ukraine. Moreover, this application provides empirical insights into the uncertainty of CPI measurement, addressing a critical but underexplored issue.

Key words: consumer prices, measurement uncertainty, official statistics, data collection frequency

Integrating big data and administrative sources for estimating vehicle mileage and analyzing road traffic accidents

Marco Broccoli, Italian National Institute of Statistics

Abstract: The aim of this project is a comprehensive methodology for estimating vehicle kilometers travelled in Italy, focusing on data from vehicle registration and inspection records. It highlights measurement distortions due to the nationalization of vehicles registered abroad and classifies vehicles based on their inspection and mileage status. An algorithm is proposed to calculate average annual kilometers travelled, considering variables such as vehicle class, fuel type, and province of circulation. The importance of provincial analysis for more homogeneous and reliable estimates is emphasized. The research also discusses the use of iMacros for web scraping. An automated system for executing macros is described, used to collect data from sites like TruckScout24 and AutoScout24. The advantages and challenges of using proprietary commands and xPath for data extraction are highlighted, with a focus on the need for good decoupling between macros and websites to reduce maintenance costs. A ”divide and conquer” approach is proposed to simplify the scraping process, and modern alternatives like Selenium and Puppeteer are mentioned for web scraping automation. The project aims to estimate vehicle kilometers to build indicators on road accident rates, using data from sources like Open Street Map. The analysis utilizes big data and administrative databases to estimate kilometers travelled by vehicles in Italy, with a focus on heavy and light vehicles. A total of 319,895 ads for heavy vehicles and 778,931 for light vehicles were extracted. The methodology includes web scraping to gather data from online portals, using software like iMacros to automate the process. The collected data is compared with information from administrative sources, such as vehicle inspections, to validate the estimates. The analysis shows that average mileage estimates from big data and administrative sources are similar, but also highlights differences related to vehicle age. The project aims to improve understanding of traffic dynamics and road safety, emphasizing the importance of integrated approaches for more accurate results. The document discusses the estimation of vehicle kilometers in Italy using Big Data and administrative databases to analyze road accident rates. The need for more accurate performance indicators for road safety is highlighted, criticizing the use of the resident population as a denominator for accident statistics. The use of web scraping techniques and Big Data is proposed to obtain more representative data. The innovative approach to estimating the average mileage travelled by vehicles in Italy, using Big Data sources like Autoscout24 and TruckScout24, along with administrative databases such as the Public Vehicle Register (PRA) and the vehicle inspection archive of the Ministry of Infrastructure and Transport (MIT). The main objective is to build indicators on road accident rates, analyzing traffic flows and exposure to accident risk. The project integrates data from vehicle sales and administrative sources to provide a more accurate and detailed view of vehicle behaviour and road safety.

Key words: RoadAccidents, BigDataSource, VehicleMileage, VehicleSalesAds, RiskExposureRatio