Cena 15,00 zł
(VAT 8%)

Indeks 381306
e-ISSN 2543-8476
PL ISSN 0043-518X

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

BIG DATA I STATYSTYKI EKSPERYMENTALNE
BIG DATA AND EXPERIMENTAL STATISTICS

GRUDZIEŃ / DECEMBER
ROCZNIK / VOLUME 68

2023 | 12

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

POLSKIE TOWARZYSTWO STATYSTYCZNE
POLISH STATISTICAL ASSOCIATION

WIADOMOŚCI STATYSTYCZNE

THE POLISH STATISTICIAN

BIG DATA I STATYSTYKI EKSPERYMENTALNE
BIG DATA AND EXPERIMENTAL STATISTICS

GRUDZIEŃ / DECEMBER
ROCZNIK / VOLUME 68

2023 | 12 (751)

ADRES REDAKCJI / EDITORIAL OFFICE ADDRESS

Główny Urząd Statystyczny / Statistics Poland, al. Niepodległości 208, 00-925 Warszawa
tel./phone +48 22 608 32 25, e-mail: redakcja.ws@stat.gov.pl

Redakcja językowa: Wydział Czasopism Naukowych, Główny Urząd Statystyczny
Language editing: Scientific Journals Division, Statistics Poland

Redakcja techniczna, skład i łamanie, opracowanie materiałów graficznych, korekta, druk i oprawa: Zakład Wydawnictw Statystycznych – zespół pod kierunkiem Macieja Adamowicza
Technical editing, typesetting, preparation of graphic materials, proofreading, printing and binding: Statistical Publishing Establishment – team supervised by Maciej Adamowicz

Wersja elektroniczna, stanowiąca wersję pierwotną czasopisma, jest dostępna na ws.stat.gov.pl
The primary version of the journal, issued in electronic form, is available at ws.stat.gov.pl

Informacje w sprawie sprzedaży I prenumeraty czasopisma / Sales and subscription of the journal:
Zakład Wydawnictw Statystycznych / Statistical Publishing Establishment
zws.stat.gov.pl
tel./phone +48 22 608 32 10, +48 22 608 38 10

Zam. 372/2023 – nakład 285 egz.

SPIS TREŚCI
CONTENTS

OD REDAKTORKI PROWADZĄCEJ

Z wielką przyjemnością oddaję w ręce Czytelników numer tematyczny „Wiadomości Statystycznych. The Polish Statistician" poświęcony zagadnieniom z zakresu big data i statystyk eksperymentalnych. Niektóre artykuły zamieszczone w tym wydaniu powstały na podstawie referatów wygłoszonych podczas sesji specjalnej zorganizowanej przez redakcję „WS", która odbyła się 4 lipca 2023 r. w ramach międzynarodowej konferencji naukowej *Metodologia Badań Statystycznych MET2023* i której byłam moderatorką.

W artykule *TranStat: an intelligent system for producing road and maritime transport statistics using big data sources* dr Dominik Rozkrut, mgr Anna Bilska, mgr inż. Michał Bis i mgr Justyna Pawłowska omawiają innowacyjny system TranStat, opracowany przez Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnikę Morską w Szczecinie i Politechnikę Krakowską w ramach programu GOSPOSTRATEG. TranStat umożliwia produkcję statystyk transportu drogowego i morskiego z wykorzystaniem wielkich wolumenów danych i szybkie udostępnianie informacji wynikowych. Autorzy przedstawiają charakterystykę wykorzystanych źródeł danych i podsystemów funkcjonalnych, założenia opracowanych modeli oraz najważniejsze nowe statystyki. Podkreślają znaczenie informacji uzyskanych dzięki systemowi TranStat dla kształtowania polityki transportowej kraju.

Dr hab. Marek Cierpiał-Wolan, prof. UR, i dr Galya Stateva w pracy *The evaluation of (big) data integration methods in tourism* poruszają zagadnienie integracji danych z wielu źródeł, w tym dużych wolumenów danych, niezbędnej do uzyskania dobrej jakości informacji udostępnianych przez statystykę publiczną w czasie zbliżonym do rzeczywistego. W badaniu opierającym się na danych dotyczących Polski i Bułgarii, które zostały zaczerpnięte z trzech popularnych portali rezerwacyjnych, autorzy oceniają przydatność wybranych metod integracji danych w statystyce w dziedzinie turystyki: algorytmu *natural language processing* (NLP), algorytmu uczenia maszynowego, tj. *K*-najbliższych sąsiadów, z wykorzystaniem technik TF-IDF i N-gram, oraz parowania rozmytego (ang. *fuzzy matching*), wchodzących w skład metod probabilistycznych. Podkreślają, że na szczególną uwagę zasługują dane uzyskane za pomocą web scrapingu. Jako najskuteczniejszą metodę spośród wszystkich testowanych wskazują parowanie rozmyte oparte na algorytmie Levenshteina w połączeniu z formułą Vincenty'ego.

Current challenges and possible big data solutions for the use of web data as a source for official statistics to temat artykułu prof. Pieta Daasa i dr. Jacka Maślankowskiego. Autorzy akcentują znaczenie web scrapingu oraz rosnące zainteresowanie środowiska naukowego i administracyjnego tą techniką uzyskiwania informacji. Skupiają się na współczesnych problemach związanych z dostępnością, ekstrakcją i wykorzystywaniem informacji ze stron internetowych i proponują potencjalne metody ich rozwiązywania. Przedstawiają studium przypadku web scrapingu wykonanego w 2022 r. na próbie 503 700 stron internetowych. Stwierdzają, że jedno źródło adresów URL (baza danych) może nie wystarczyć do uzyskania wiarygodnych danych webowych – w przeprowadzonym badaniu prawie 20% adresów URL było niedostępnych, ponieważ strony internetowe przestały istnieć lub ich właściciele zablokowali dostęp do pliku robots.txt, co uniemożliwiło scrapowanie danych. Preferowane jest zatem korzystanie z wielu źródeł.

W pracy *Digital transformation and data ecosystem: implications for policy actions and competency frameworks* dr Monika Rozkrut przybliża politykę Unii Europejskiej w zakresie transformacji

cyfrowej, a także jej potencjał rozwojowy. Autorka dokonuje krytycznej analizy postępu w obszarze operacjonalizacji i wdrażania odpowiednich polityk. Identyfikuje szczególnie trudne zadania, które mogą mieć negatywny wpływ na osiągnięcie celów strategicznych. Zwraca uwagę, że istotnym problemem jest niedobór ekspertów przygotowanych do pełnienia nowych funkcji w dynamicznie rozwijającym się ekosystemie danych, wskazuje pożądane umiejętności umożliwiające efektywne zarządzanie danymi i podkreśla rolę data stewarda – kluczową dla wsparcia szybkiego rozwoju ekosystemu danych w UE.

Rozważania podejmowane w tym wydaniu „WS" zamyka opracowanie dr. hab. Jerzego Auksztola, prof. UG, pt. *Improving research on environmental noise pollution and its impact on the population in the context of sustainable development*. Autor odnosi się do kwestii poprawy statystyk umożliwiających badanie narażenia ludności na hałas w kontekście zrównoważonego rozwoju. Zaznacza, że badania nad natężeniem hałasu w środowisku mają długą tradycję, a wiedza o negatywnym wpływie tego czynnika na zdrowie oraz aspektach środowiskowych i ekonomicznych jest stale pogłębiana. Omawia narzędzia, jakie oferuje statystyka publiczna, umożliwiające kontynuację badań w tym zakresie.

Tradycyjnie numer zawiera też omówienie nowości wydawniczych GUS, przygotowane przez Joannę Sadowy.

Serdecznie zapraszam do lektury.

dr hab. Małgorzata Tarczyńska-Łuniewska, prof. US

FROM THE MANAGING EDITOR

It is my great pleasure to present to Readers the thematic issue of *Wiadomości Statystyczne. The Polish Statistician* devoted to topics related to big data and experimental statistics. Some of the articles were written on the basis of papers delivered at *MET2023 Statistical Research Methodology Conference*, during a special session organised by the *WS* editorial team and moderated by me, held on 4th July 2023.

In the TranStat: an intelligent system for producing road and maritime transport statistics using big data sources, Dominik Rozkrut, PhD, Anna Bilska, MSc, Michał Bis, BEng, MSc, and Justyna Pawłowska, MSc, discuss the innovative TranStat system, created by Statistics Poland, the Statistical Office in Szczecin, the Maritime University of Szczecin and the Cracow University of Technology in the framework of the *GOSPOSTRATEG* programme. TranStat enables the production of road and maritime transport statistics on the basis of large volumes of data, and allows their fast publication. The authors present the characteristics of the applied data sources and functional sub-systems, the assumptions for the created models and the most important new statistics. They emphasise the role of information obtained by means of TranStat for the shaping of the country's transport policy.

Marek Cierpiał-Wolan, PhD, DSc, Professor at the University of Rzeszów, and Galya Stateva, PhD, in the paper entitled *The evaluation of (big) data integration methods in tourism* deal with the issue of integration of data from several sources, including large volumes of information. According to the authors, this is a prerequisite to obtaining good-quality information that official statistics could disseminate in near-real time mode. Their study is based on data on Poland and Bulgaria, drawn from three popular booking portals. The authors asses the usefulness of the

following methods of data integration in tourism statistics: Natural Language Processing (NLP), machine learning algorithm, i.e. K-Nearest Neighbours (K-NN) using TF-IDF and N-gram techniques, and Fuzzy Matching, all belonging to probabilistic methods. The study shows that data obtained by means of web scraping deserve special attention. Fuzzy Matching based on the Levenshtein algorithm combined with Vincenty's formula turns out to be the most effective among all the tested methods.

The paper entitled *Current challenges and possible big data solutions for the use of web data as a source for official statistics* by Prof. Piet Daas and Jacek Maślankowski, PhD, underlines the importance of web scraping and points to the growing interest of the scientific and administrative circles in this technique of data extraction. The authors focus on the contemporary problems connected to the availability, extraction and application of information from websites, and propose potential solutions. Their research is based on the case study performed in 2022 on the sample consisting of 503,700 websites. The study demonstrates that one source of URLs (a database) might not be sufficient to obtain reliable web data. In the authors' case study, almost 20% of URLs turned out to be unavailable, because the websites did not exist anymore or their owners blocked the access to robots.txt file, which made web scraping impossible. Therefore, using several sources is preferable.

Digital transformation and data ecosystem: implications for policy actions and competency frameworks by Monika Rozkrut, PhD, presents the European Union's policy on digital transformation and its development potential. The author performs a critical analysis of the progress in the area of operationalisation and implementation of the relevant policies, and identifies especially difficult tasks which can negatively impact the achievement of strategic goals. The study draws attention to the serious problem of a shortage of experts able to perform new functions in a dynamically-developing ecosystem of data, defines desirable skills allowing the effective management of data, and emphasises the role of data steward, key to supporting the fast development of a data ecosystem in the EU.

The article entitled *Improving research on environmental noise pollution and its impact on the population in the context of sustainable development* by Jerzy Auksztol, PhD, DSc, Professor at the University of Gdańsk, deals with the question of improving statistics that enable research into the population's exposure to noise in the context of sustainable development. The author stresses that the research into the intensity of noise in the environment has a long tradition, and the negative influence of this factor on health and its environmental and economic aspects are constantly explored. The article discusses tools enabling further research into this topic offered by official statistics.

The issue concludes with Joanna Sadowy's presentation of Statistics Poland's new publications.

We wish you pleasant reading.

Małgorzata Tarczyńska-Łuniewska, PhD, DSc
Professor at the University of Szczecin

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 1–24 https://doi.org/10.59139/ws.2023.12.1
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 1.08.2023, poprawiony/revised: 10.11.2023, zaakceptowany/accepted: 20.11.2023

TranStat: an intelligent system for producing road and maritime transport statistics using big data sources[1]

Dominik Rozkrut,[a] Anna Bilska,[b] Michał Bis,[c] Justyna Pawłowska[d]

Abstract. The development of digital technologies, increasing the availability of big data and advanced processing techniques have enabled Statistics Poland to modernise the system for producing road and maritime transport statistics. As a result of the activities undertaken to adopt modern big data technologies and data from sensors, e.g. the Automatic Identification System (AIS) or the e-TOLL electronic toll collection system, new statistics have been obtained and data dissemination has accelerated. In addition, these activities ensure continuity in data production, especially in situations where collecting data from individuals may be difficult (e.g. the COVID-19 epidemic). The primary purpose of this article is to present the innovative TranStat system that enables the production of road and maritime transport statistics based on large volumes of data in order to shape the country's transport policy. The system was developed under the GOSPOSTRATEG programme and implemented by Statistics Poland, the Statistical Office in Szczecin, the Maritime University of Szczecin, and the Cracow University of Technology. The study presents the most important aspects of the TranStat system, i.e. the characteristics of data sources, the description of functional subsystems, assumptions of the developed models and result data for traffic statistics, transport performance and exhaust emissions calculations for both types of transport. This study also provides information on smart forms implemented by Polish official statistics, reducing the burden on respondents and the costs of surveys.

Keywords: road transport, maritime transport, traffic intensity, transport performance, exhaust emissions, smart forms, big data, TranStat, GOSPOSTRATEG, AIS, e-TOLL

JEL: C55, R41, G53, C80

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji *Metodologia Badań Statystycznych MET2023*, która odbyła się w dniach 3–5 lipca 2023 r. w Warszawie. / The article is based on a paper delivered at the *MET2023 Conference on Methodology of Statistical Research*, held on 3rd–5th July 2023 in Warsaw.

[a] Uniwersytet Szczeciński, Wydział Ekonomii, Finansów i Zarządzania, Instytut Ekonomii i Finansów; Główny Urząd Statystyczny, Polska / University of Szczecin, Faculty of Economics, Finance and Management, Institute of Economics and Finance; Statistics Poland, Poland.
ORCID: https://orcid.org/0000-0002-0949-8605. Corresponding author, e-mail: d.rozkrut@stat.gov.pl.

[b] Urząd Statystyczny w Szczecinie, Ośrodek Statystyki Morskiej, Polska / Statistical Office in Szczecin, Maritime Statistics Centre, Poland.

[c] Urząd Statystyczny w Szczecinie, Ośrodek Inżynierii Danych, Polska / Statistical Office in Szczecin, Data Engineering Centre, Poland. ORCID: https://orcid.org/0009-0007-0830-2889. E-mail: m.bis@stat.gov.pl.

[d] Urząd Statystyczny w Szczecinie, Ośrodek Statystyki Transportu i Łączności, Polska / Statistical Office in Szczecin, Transport and Communications Statistics Centre, Poland.
ORCID: https://orcid.org/0009-0009-7798-6457. E-mail: j.pawlowska2@stat.gov.pl.

TranStat – inteligentny system produkcji statystyk transportu drogowego i morskiego z wykorzystaniem big data

Streszczenie. Rozwój technologii cyfrowych, zwiększenie dostępności danych typu big data oraz zaawansowane techniki ich przetwarzania umożliwiły polskiej statystyce publicznej unowocześnienie systemu produkcji statystyk transportu drogowego i morskiego. Dzięki działaniom podjętym w celu adaptacji nowoczesnych technologii big data i danych sensorycznych, m.in. z systemu automatycznej identyfikacji statków AIS (ang. Automatic Identification System) czy elektronicznego systemu poboru opłat e-TOLL, uzyskano nowe statystyki oraz przyspieszono proces udostępniania danych. Ponadto działania te zapewniają ciągłość w obszarze produkcji danych, szczególnie w sytuacji, gdy zbieranie danych od respondentów może być utrudnione (np. podczas epidemii COVID-19). Głównym celem niniejszego artykułu jest zaprezentowanie innowacyjnego, opracowanego w ramach programu GOSPOSTRATEG systemu TranStat, który umożliwia produkcję statystyk transportu drogowego i morskiego z wykorzystaniem wielkich wolumenów danych i tym samym służy kształtowaniu polityki transportowej kraju. Projekt TranStat został zrealizowany przez Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnikę Morską w Szczecinie oraz Politechnikę Krakowską. W pracy przedstawiono najważniejsze cechy systemu TranStat – scharakteryzowano źródła danych, opisano podsystemy funkcjonalne i założenia opracowanych modeli oraz podano informacje wynikowe dla statystyk natężenia ruchu, pracy przewozowej i emisji zanieczyszczeń dla obu rodzajów transportu. Omówiono też inteligentne formularze wdrożone przez statystykę publiczną, mniej obciążające dla respondentów i umożliwiające redukcję kosztów badań.

Słowa kluczowe: transport drogowy, transport morski, natężenie ruchu, praca przewozowa, emisja zanieczyszczeń, inteligentne formularze, big data, TranStat, GOSPOSTRATEG, AIS, e-TOLL

1. Introduction

One of the most critical challenges in the era of the digital revolution is access to information in the shortest possible time after its collection and processing, resulting from the expectations of statistical data users. These data are used e.g. in analyses for monitoring policies and making decisions at all levels of public management. With the development of big data technology, increased availability of big data volumes, and the Internet of Things (IoT), Statistics Poland has an opportunity to modernise the system to produce road and maritime transport statistics. Many studies have recently explored the possibilities and challenges of using big data in official statistics, most of which point out the possible applications (Daas et al., 2015). New data sources are already being used in official statistics. This aligns with the tasks defined in the Fundamental Principles of Official Statistics. The use of new data sources helps in the implementation of these tasks (Rozkrut et al., 2021).

The response to the above-mentioned challenges and opportunities was the implementation of the TranStat project – an intelligent system to produce road and

maritime transport statistics using large volumes of data for making the country's transport policy as part of the GOSPOSTRATEG programme organised by the National Centre for Research and Development (Pol. Narodowe Centrum Badań i Rozwoju). The TranStat project was implemented in 2019–2021 by Statistics Poland, the Statistical Office in Szczecin, the Maritime University of Szczecin, and the Cracow University of Technology (Główny Urząd Statystyczny [GUS], Urząd Statystyczny w Szczecinie [US w Szczecinie], Politechnika Morska w Szczecinie & Politechnika Krakowska, 2019, 2020a, 2020b, 2020c, 2020d, 2020e, 2021).

The primary purpose of this article is to present the TranStat system that enables the production of road and maritime transport statistics in order to shape the country's transport policy. The study discusses the most important aspects of the TranStat system, i.e. the characteristics of data sources, the description of functional subsystems, the assumptions of the developed models and result information for traffic statistics, the transport performance, and the calculation of the exhaust emissions for both types of transport. The study also provides information on the smart forms implemented by Statistics Poland, which reduce the burden on respondents and the costs of surveys.

2. Characteristics of data sources used in the TranStat system

2.1. Automatic Identification System (AIS)

Applying big data in the area of transport can provide new insights beyond traditional transport datasets (Welch & Widita, 2019). AIS is an Automatic Identification System used on ships to exchange information electronically with nearby vessels, AIS base stations and satellites. The primary task of the AIS is to enhance navigation safety (anti-collision system) and to support marine traffic management for coastal vessel traffic services (VTS). According to the requirements of Chapter V of the SOLAS Convention developed by the International Maritime Organization (IMO), the AIS should be installed on:

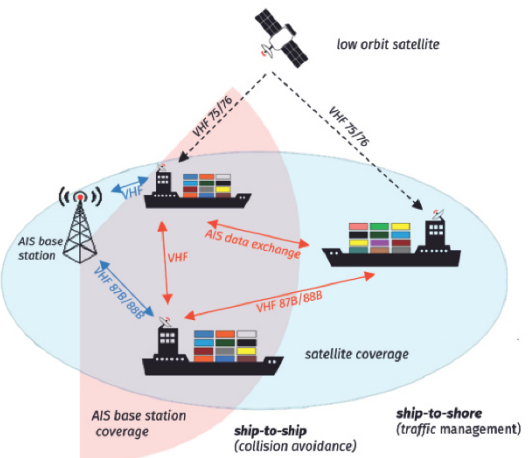
- all ships of a 300 gross tonnage and more – for international shipping;
- all vessels of a 500 gross tonnage and more not engaged in international shipping;
- all passenger ships, regardless of size.

Statistics Poland gained access to data from the AIS-PL system based on the Regulation of the Minister of Maritime Economy and Inland Navigation of 26th September 2018, on the National Ship Traffic Monitoring and Information Transmission System (Pol. Narodowy System Monitorowania Ruchu Statków i Przekazywania Informacji). The AIS's operation principle is based on the VHF radio frequency. Data are transmitted using Self Time Division Multiple Access

(STDMA). Data from the AIS-PL system come from 13 base stations along the Polish coast. GPS determines the ship's position.

There are four channels used for the AIS (Figure 1): AIS 1 (channel 87B), AIS 2 (channel 88B), and channels 75 and 76 for satellite communications.

Figure 1. Scheme of the AIS operation



Source: authors' work.

Within the AIS, there are 27 messages containing:

- dynamic data (related to the information about the ship's movement) from ship sensors (automatic data transmission). The transmission frequency depends on the speed and course change (2–10 s) when the ship is at anchor (3 min). Example attributes: Maritime Mobile Service Identity (MMSI) number – ship identification data, longitude, latitude, accuracy class indication, speed over the ground; course over the ground; angular velocity of turn, the vessel's navigational status, universal time coordinated (UTC);
- static data (related to the information about the ship's characteristics) is entered directly by the ship's crew (manual data transmission). Transmission frequency – 6 min. Example attributes: IMO number – ship identification data, MMSI number, ship name, call sign, ship dimensions, ship type, destination port, ship draught.

2.2. ViaTOLL/e-TOLL – electronic toll collection system

ViaTOLL is a toll collection system for toll road sections in Poland, based on radio technology, built by Kapsch. It operated until 30th September 2021, and was replaced by e-TOLL on 1st October 2021. Both systems worked simultaneously during the transition period from 24th June to 30th September 2021.

Map 1. National roads covered by the e-TOLL system

Source: e-TOLL (n.d.).

In Poland, the toll applies on toll motorways, expressways and selected national roads. The length of the paid sections is currently approximately 3,677 km. Revenues from the system contribute to the National Road Fund for further investments in expanding the road network in Poland and modernising the existing road infrastructure. The viaTOLL system (now e-TOLL) is a mandatory system for all motor vehicles and combinations of vehicles with a gross vehicle weight of over 3.5 tonnes, as well as for buses, regardless of their gross vehicle weight. The viaTOLL system consisted of 951 gates (Map 1) and on-board devices placed in vehicles. In addition, toll collection control vehicles were used. When driving under a gate, the recording device placed on it collected the toll from the individual user account. The e-TOLL system, the successor of the viaTOLL system, is a solution implemented and supervised by the Head of the National Tax Authority (Pol. Szef Krajowej Administracji Skarbowej). It is based on the technology of determining the user's position using satellite positioning with virtual gates. Each vehicle user obliged to pay the toll may choose one of the methods of transferring location data to the

system: using a free application installed on a mobile device, a GPS tracker factory installed in vehicles (Pol. Zewnętrzny System Lokalizacyjny – ZSL) or the On Board Unit (OBU).

3. TranStat system – assumptions, architecture, implementation

When developing the concept of the TranStat system, several assumptions were made based on the general requirements for modern IT systems, including: implementation of open standards, technological neutrality (vendor lock-in), compliance with applicable laws, modular construction, easy expansion with new system functionalities in the future, and ensuring an appropriate level of security. In addition, due to the specificity of sensor data and the need to process data in real time, the requirements for scalable big data solutions (volume, variety and velocity) were considered.

The TranStat IT system was developed and implemented in Statistics Poland's production environment.[2]

The following functional subsystems have been developed as part of the system:

- data collection and processing subsystem, responsible for the following processes: decoding AIS data; processing data from sensors; integration, validation, transformation and aggregation of data;
- the internal data presentation and analysis subsystem enables data exploration, visualisation, and statistical analyses using the RStudio and Apache Zeppelin tools;
- data presentation and analysis subsystem – external**,** intended for an external recipient, operating based on calculated aggregates and indicators.

Figure 2 shows the flow and processing of data in the TranStat system to obtain new statistics from large datasets from sensors, i.e. the AIS and the e-TOLL, and to contribute to smart forms.

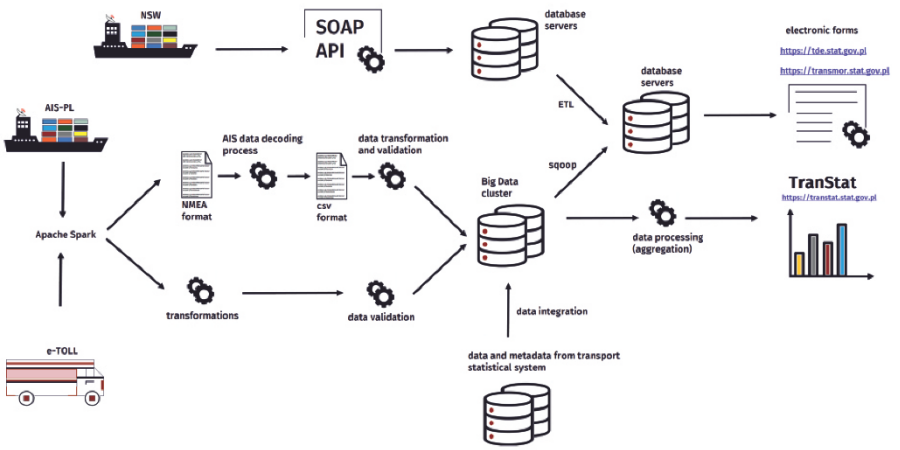
Due to the nature of the data, it was necessary to consider two types of data processing: batch processing (e-TOLL) and real-time processing of sensor data (AIS).

Data in the data collection and processing subsystem is stored in the Hadoop Distributed File System (HDFS) in the form of CSV files, and the process of storing sensor data is carried out by using dedicated tools for handling data streams, i.e. Spark Streaming. The data from the AIS-PL system are decoded from the NMEA format before being saved. Most planned sub-processes, i.e. validation, transformation, integration and aggregation as part of the data collection and

[2] Application link: https://transtat.stat.gov.pl (GUS, n.d.).

processing are implemented using the Scala programming language in the Apache Spark platform. To enable an advanced analysis and visualisation of spatial data for the internal subsystem of presentation and analysis, the RStudio Server and Apache Zeppelin tools have been implemented, through which it is possible to work directly on previously prepared data structures located on a data cluster in the HDFS.

Figure 2. Data flow and processing in the TranStat system



Source: authors' work.

As part of the external data presentation and analysis subsystem, an application and database server were implemented in a separate Demilitarized Zone to provide a dedicated application presenting statistical products (information on traffic volume, transport performance and emissions, metadata and charts). The designed web application was made in the ASP.NET MVC environment (Model, View, Controller) with the .NET Framework technology in the C# programming language and supported by front-end technology (XHTML, CSS, JavaScript, jQuery, Bootstrap).

4. Maritime traffic intensity statistics

4.1. Assumptions

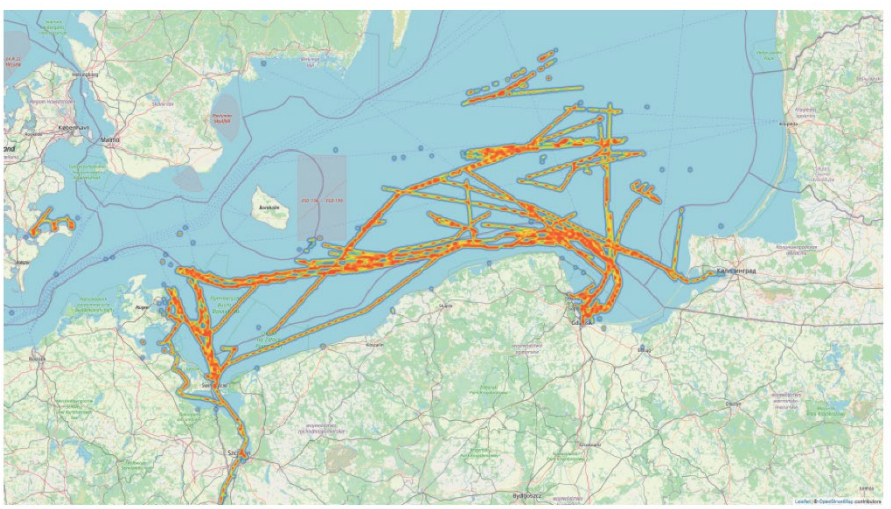
Many works indicate the need for an in-depth analysis of maritime traffic (Vasilev & Sulova, 2023) or, more broadly, multimodal transport flows (Zhang et al., 2018). To identify the phenomenon for four ports of primary importance to the national economy: Gdańsk, Gdynia, Szczecin and Świnoujście, points (containing geographic

coordinates: longitude and latitude) were determined, which form polygons that fall within the boundaries of the ports based on the regulation of the minister competent for maritime economy. These constituted areas for the study of ship traffic volume.

Traffic intensity is understood as flow intensity, defined as the number of transport units passing through the boundary line of an area in a specific time interval (e.g. Map 2). To develop a methodology for calculating the traffic intensity in a specified area and in a particular unit of time, depending on the method of calculating the intensity and the location of the calculation procedure on the time axis, the method of counting units based on notification times was used. As a result of the developed algorithms for traffic intensity in maritime transport, the following variables and breakdowns are obtained:

- variables: number of ships at a seaport; number of arrivals/departures by maritime vessels;
- breakdowns: time (day, month, quarter, year); spatial: ports located on the coast of Poland; means of maritime transport: by type, by country of flag.

Map 2. Traffic intensity for maritime transport as of 1st January 2023



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

4.2. Outcome information

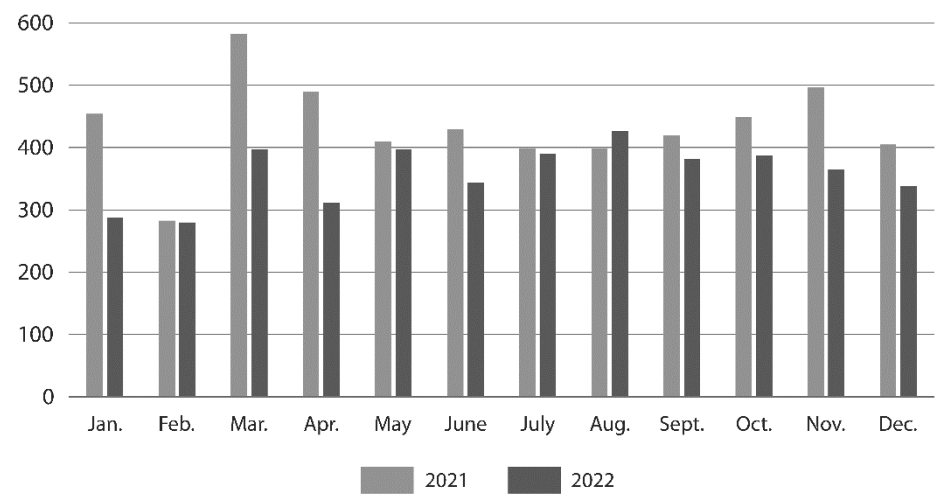
Years 2021 and 2022 were selected for the port of Szczecin to generate traffic statistics in maritime transport. The visualisation was made in months.

Figure 3 shows the number of ships arriving at the port of Szczecin by month in 2021 and 2022. The number of ship departures is very similar in a given month.

When analysing the graph, one can notice fluctuations – the traffic volume was not even in the analysed period. The largest number of vessel entries into the port in 2021 was recorded in March (583), while the largest number of entries into the port in 2022 was recorded in August (427). The total number of ship arrivals in 2022 was lower than in 2021, which was recordable almost every month. It is worth emphasising that the presented experimental statistics generated based on AIS data differ from the official statistics obtained based on the TransMor survey, implemented in 2022 as 'smart forms', where AIS data play a qualitative role. It can be expected that the experimental statistics will coincide with official statistics shortly.

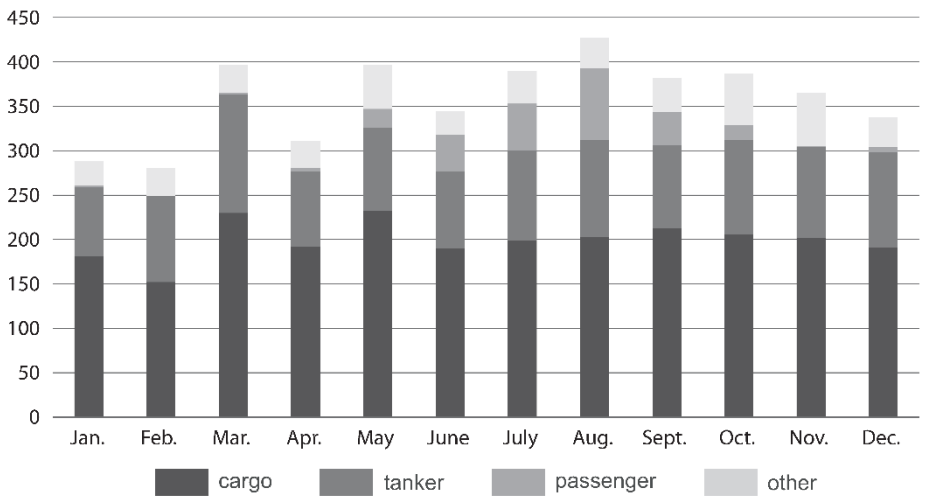
Figure 4 shows the number of ships entering the port of Szczecin by month and by ship type in 2022. The data presented relates only to cargo ships, passenger ships, tankers and ships classified as other (e.g. with an unknown ship type code). The analysis excludes such types of vessels as: fishing, service, tugboats, pushers, dredgers, research and scientific vessels, pilots, and rescue vessels – SAR. The dominant kind of ships arriving at the port of Szczecin were cargo ships, with the highest values in March (230) and May (232) 2022. The number of tankers entering the port of Szczecin in 2022 ranged from 78 to 133. The number of passenger ships is seasonal; the highest number of vessel arrivals was recorded in August (81) and July (53) 2022.

Figure 3. Number of vessel arrivals at the port of Szczecin by month



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

Figure 4. Number of vessel arrivals at the port of Szczecin by month and vessel type in 2022



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

5. Transport volume statistics in maritime

5.1. Assumptions

So far, maritime statistics have been presented as aggregated information obtained from respondents in a survey on carriages by maritime cargo-carrying and coastal transport fleets (on the T-08 form). The problem was that these data concerned transport carried out by Polish operators using their vessels or leased from foreign ship owners. In addition, the data were provided collectively for a given year and it was impossible to analyse e.g. the frequency with which ships travelled on specific routes and, thus, the variability of the transport volume over time. Gaining access to the AIS and the application of modern techniques in processing big data sets enabled receiving complete statistics on transport volume for goods and passengers carried on the routes with seaports located along the coastline of Poland. Transport volume is understood as the product of the transport performed by the given means of transport: the length of the road (number of kilometres) and the number of tonnes of transported goods (freight cargo). The unit of measurement is the tonne-kilometre (tkm) – one tonne-kilometre is the transport of 1 tonne of cargo over 1 km. In the case of the transport volume estimation model, it is planned to present possible ship routes in a directed (weighted) graph, where the graph's vertices are waypoints or quays and the edges are straight sections between them. Each edge contains the coordinates of the start and end points, and its weight is the length of

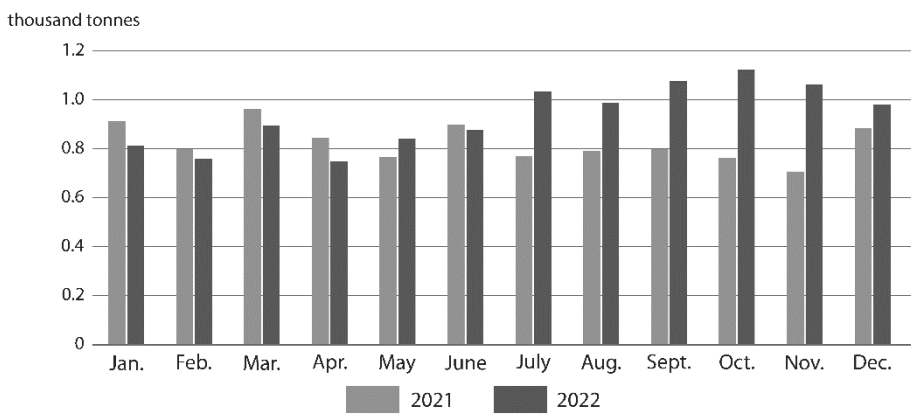
the segment between individual nodes, calculated by the Haversine formula (distance). As a result of the developed algorithms and combined data sources, the following was obtained:

- variables: transport volume for goods and passengers; total distance – the distance travelled by all vessels on arrival/departure relations when carrying goods or passengers;
- breakdowns: time (day, month, quarter, year); spatial: ports located on the coast of Poland, direction, country; means of maritime transport: by type, by flag, by gross tonnage; type of cargo: cargo group, commodity group.

5.2. Outcome information

Regarding transport volume, cargo transported by sea is delivered through Polish ports. The period of 2021–2022 for the port of Szczecin was analysed (Figure 5). Over the described time interval, the highest cargo throughput was recorded in October 2022 and amounted to 1.125 thousand tonnes. The type of cargo considered was dry bulk cargo, predominant in the port of Szczecin.

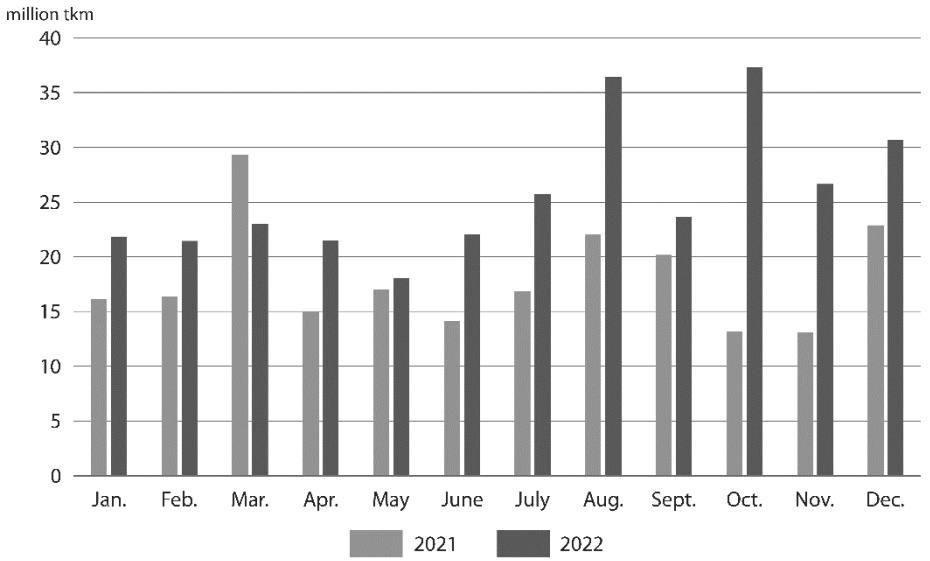
Figure 5. Cargo turnover in the port of Szczecin by month



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

The transport volume was obtained by combining information on the quantity of goods carried and the distance travelled. Based on the two-year data, it is easy to notice fluctuations in the transport volume, which consists of the weight of goods transported and the distances travelled. The highest value for maritime transport volumes on routes with the port of Szczecin was reached in October 2022 – over 37 million tkm, which in this case was associated with the highest annual throughput and the longest distance travelled (Figure 6).

Figure 6. Transport volume on the routes with the port of Szczecin



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

6. Emissions statistics generated based on maritime transport

6.1. Assumptions

Using big data to support low-carbon transport policies in Europe provides new opportunities for the analysis of real-world emissions, which is invaluable in this context (De Gennaro et al., 2016). Emission accounting has seen a major innovation in recent years. Big data, especially AIS data, has played a key role in this innovation (Yin et al., 2021). The emission of pollutants generated by ships significantly impacts the marine atmospheric environment, seaports and adjacent areas. Therefore, the issue of ship emissions as a local source of pollution for port cities is an essential aspect of air quality assessment. Transport ships with a gross tonnage of 100 GT and more were analysed to estimate the pollutants emitted by maritime transport.

To obtain information on the emissions of a given ship, a solution based on developed models has been implemented, i.e.:

- reference model: requiring the preparation of a matrix of characteristic technical parameters dedicated to the ship, enabling the determination of the value of individual emissions;
- specific model: using machine learning on a representative dataset from the reference model. The input parameters are the basic parameters of AIS messages, and the emission values are the output;

- generic model: used when a specific model obtains limit values or input data outside the acceptable range, e.g. ship length over 300 m. For such vessels, the boundary values of the pollutant estimation have been determined empirically. Exceeding the maximum values of any emission (CO_2, SO_x, NO_x, PM) of the specific model causes the estimation to be recalculated.

It was assumed that the input requirements for AIS messages will be as follows: ship type $t \in \langle 0; 39 \rangle \cup \langle 50; 99 \rangle$ (non-displacement units have been eliminated); speed SOG ≥ 0kt; length and width L > 0 m and B > 0 m; draft T > 0 m; position $\varphi \in \langle -90°; 90° \rangle$ and $\lambda \in \langle -180°; 180° \rangle$.

The statistics also consider and define an additional reference level of CO_2 emissions by MEPC.308(73) Resolution of 26th October, 2018.

As a result of the developed algorithms, the following variables and breakdowns are obtained:

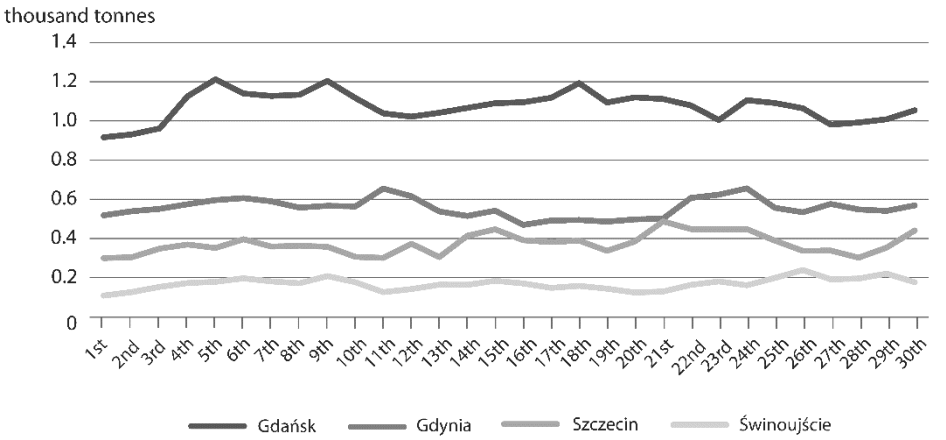
- variables, among others: NO_x emission (nitrates, nitrites); SO_x emission (sulphates, sulphites); CO_2 emission (carbon dioxide); PM emission (particulate matter);
- breakdowns: time (day, month, quarter, year); spatial: ports located on the coast of Poland; means of maritime transport: by type, by gross capacity.

6.2. Outcome information

The results of the work are statistics obtained from the TranStat system in the field of emissions generated by maritime transport, thanks to which it is possible to analyse the environmental impact of pollution from maritime vessels.

The largest share in the CO_2 emissions generated utilising maritime transport in November 2022 was by ships entering/leaving the port of Gdańsk (Figure 7). This port's highest amount of CO_2 pollution was recorded on 5th November 2022, and it was 1,212.5 tonnes. When interpreting the results, it is essential to note that it is a seaport with the country's highest annual throughput and many ship arrivals.

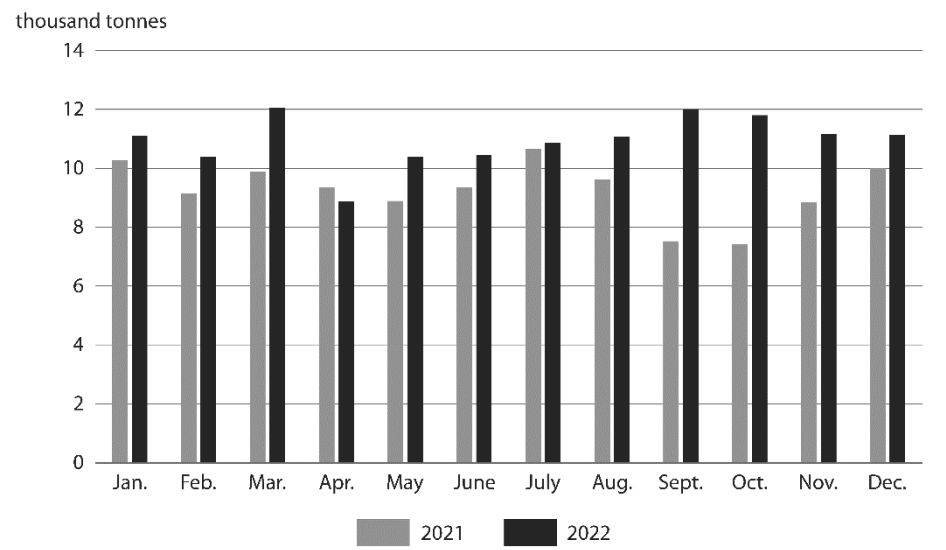
Figure 7. Daily CO_2 emission in November 2022 by seaport



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

Regarding the port of Szczecin, the analysis of CO_2 emissions was carried out for two years – 2021 and 2022 (Figure 8). Emissions for most months were higher in 2022 than in 2021. The highest emissions were recorded in March, September, October and November 2022, which is related to the high throughput observed in this period and more ship arrivals for bulk cargo transport, for which the emission volumes are the highest.

Figure 8. Monthly CO_2 emission in the port of Szczecin



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

7. Traffic statistics in road transport

7.1. Assumptions

The indicators presented in the TranStat system in the area of road transport are calculated based on parameters of all transactions generated in the e-TOLL system and the number of vehicles:

- number of transactions: the number of toll transactions for vehicles subject to toll, registered on the toll section;
- number of vehicles: unique number of vehicle occurrences at a toll collection point or section.

The analysed dataset is supplemented with an additional electronic set containing information on virtual gates of the e-TOLL system and toll collection stations of the system according to the following structure: unique name of the gate in the system and identifier of the toll collection station; longitude (GPS coordinate in decimal format); latitude (GPS coordinate in decimal format).

In total, there are 951 virtual gates on motorways, expressways and national roads covered by the e-TOLL system. Assuming that a journey is through at least two toll collection points, the following variables have been defined:

- number of trips: the vehicle completed a trip under the e-TOLL system if it was registered in at least two transactions from the analysed dataset;
- travel time.

To create statistics on traffic volume, it was assumed that a vehicle made a trip under the e-TOLL system if it was registered in at least two transactions from the analysed dataset. A journey lasting more than 0.15 hours was assumed to be long enough to be included. The elimination of 'zero-distance' journeys in the analysis allowed disturbances in the values of statistical indicators to be removed. After supplementing the toll collection points with the length of the section, the third variable was defined: several kilometres travelled – the number of vehicles that have travelled a given section multiplied by the length of the section.

The following dimensions were considered for the defined variables:

- time: day, week, month;
- spatial: vehicle registration country (Poland, abroad, unknown) and road number;
- categories of entities/vehicles according to payload groups (gross vehicle weight – GVW):
  - light vehicles: load group 13 (vehicles of a GVW of 3.5 tonnes or less); load group 14 (vehicles of a GVW of 3.5 tonnes or less, capable of towing a trailer and vehicles of a GVW exceeding 3.5 tonnes);
  - coaches, capacity group 30, with more than nine seats (including the driver);
  - heavy-duty vehicles: load group 41 (heavy-duty vehicles of a GVW above 3.5 tonnes and below 12 tonnes); load group 42 (heavy-duty vehicles of a GVW above 3.5 tonnes and below 12 tonnes with the physical ability to tow a trailer

and vehicles of a GVW above 12 tonnes); load group 50 (heavy-duty vehicles of a GVW of over 12 tonnes);

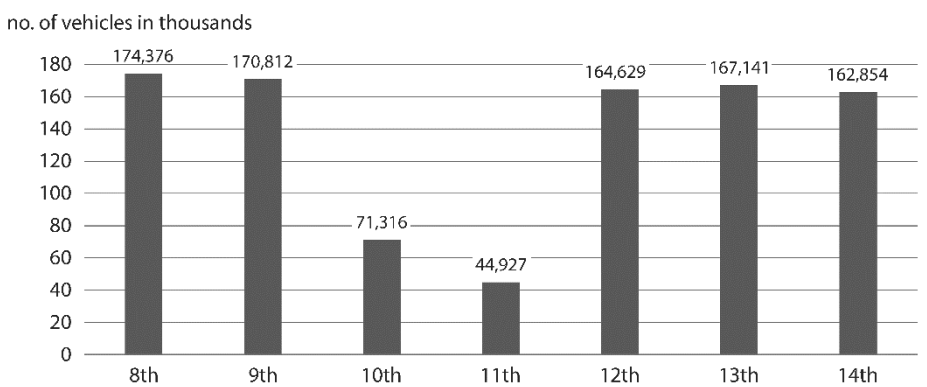
- categories of entities/vehicles according to the Euro emission class (0–6) – European emission standard specifying the permissible emissions in new vehicles sold in the EU and the European Economic Area.

In addition, the Enhanced Environmentally Friendly Vehicle (EEV) emission standard was included, assuming a reduced level of particulate emissions; compliance with this standard is voluntary.

7.2. Outcome information

The information obtained from the developed methodology for measuring traffic statistics in road transport makes it possible to characterise the fleet of vehicles, measure traffic on road sections covered by the e-TOLL system and present data on traffic volume. The specificity of the acquired data and the option of processing them in real time enables the presentation of indicators in breakdowns from daily, through monthly to annual. The results below show the traffic volume on road sections covered by the e-TOLL system by the number of vehicles for the exemplary period between 8th and 14th December 2022. The number of vehicles travelling on toll road sections in the analysed period was uneven for individual days of the week (Figure 9). On working days, the daily traffic volume ranged from approx. 160 to about 170 thousand vehicles, while on Saturday and Sunday, which fell on 10th and 11th December, the number of vehicles recorded in the e-TOLL system was significantly lower and amounted to approx. 71 and approx. 45 thousand vehicles, respectively.

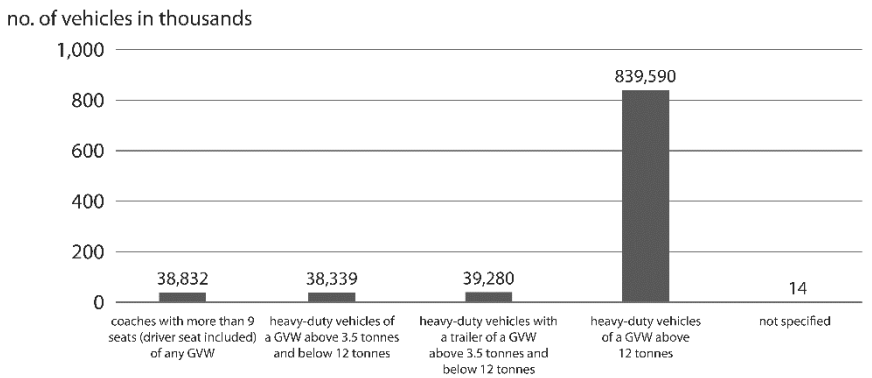
Figure 9. Daily traffic volume on the road network covered by the e-TOLL system by number of vehicles, 8th–14th December 2022



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

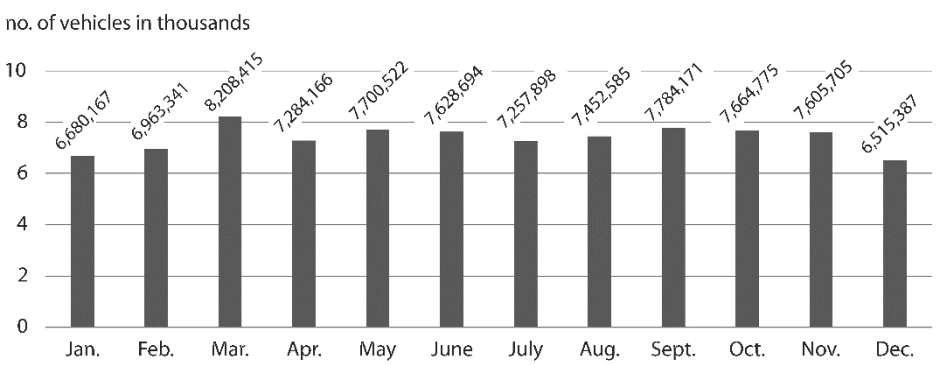
The daily transactions in the analysed week ranged from 3,924,247 (8th December 2022) to 996,293 (11th December 2022). The total number of records from transactions amounted to approx. 20 million. The e-TOLL system covered one million vehicles travelling on the toll road network in the presented week. The scope of data supplied with the bi-weekly frequency of the TranStat application allows the recipients to be presented with several detailed traffic indicators, including individual vehicle categories (Figure 10) and emission class or road section (Figure 11).

Figure 10. Weekly traffic volume on the road network covered by the e-TOLL system by vehicle category, 8th–14th December 2022



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

Figure 11. Monthly traffic volume on the A1 motorway for all vehicle categories in 2022



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

8. Emissions statistics generated based on road transport

A method using the COPERT programme – a standard calculator of vehicle emissions – was used to estimate the level of emissions generated by road transport. It uses vehicle population, mileage, speed and other data such as ambient temperature, and calculates the emissions and energy consumption for a specific country or region. The development of COPERT is coordinated by the European Environment Agency (EEA) as part of the activities of the European Topic Center on Air Pollution and Climate Change Mitigation. The Joint Research Center of the European Commission manages the scientific development of the model. COPERT has been developed to compile official inventories of road transport emissions in the European Economic Area member countries. However, it applies to all relevant scientific and academic research. Using a software tool for calculating emissions generated through transport allows for a transparent, standardised and thus consistent and comparable procedure for collecting data and reporting emissions.

8.1. Assumptions

The use of the COPERT programme makes it possible to estimate the amount of emissions generated by road transport based on the following input (supply) data:

- number of vehicles by type: lorries, road tractors, urban buses. The data source on the number of vehicles is the Central Vehicle Register (Pol. Centralna Ewidencja Pojazdów – CEP). The data set for the COPERT programme includes the number of cars for each category of vehicles, broken down by GVW and EURO emission class;
- data on vehicle mileage by type of vehicle from the CEP based on readings made by district vehicle inspection stations (Pol. okręgowe stacje kontroli pojazdów) and road inspections carried out by the Police: average annual mileage; average total mileage (since production);
- vehicle speed data by vehicle 1 type: on urban roads at peak, off-peak; on rural roads; on highways;
- share of specific types of vehicles on particular types of roads: on urban roads at peak, off-peak; on rural roads; on highways. Data on the share of vehicles on particular types of roads supplied to the COPERT programme are estimated values calculated based on the data of the General Directorate for National Roads and Motorways (Pol. Generalna Dyrekcja Dróg Krajowych i Autostrad – GDDKiA) after verification with the initial data of the COPERT programme;
- meteorological data of the Institute of Meteorology and Water Management (Pol. Instytut Meteorologii i Gospodarki Wodnej – IMGW): average monthly minimum and maximum temperature; average monthly air humidity.

As a result of the performed estimates, data on the volume of emissions are obtained, such as: NMVOC – non-methane volatile organic compounds; $PM_{2.5}$ –

particulate matter; NO_x – nitrogen oxides; CH_4 – methane; CO_2 – carbon dioxide; N_2O – nitrous oxide.

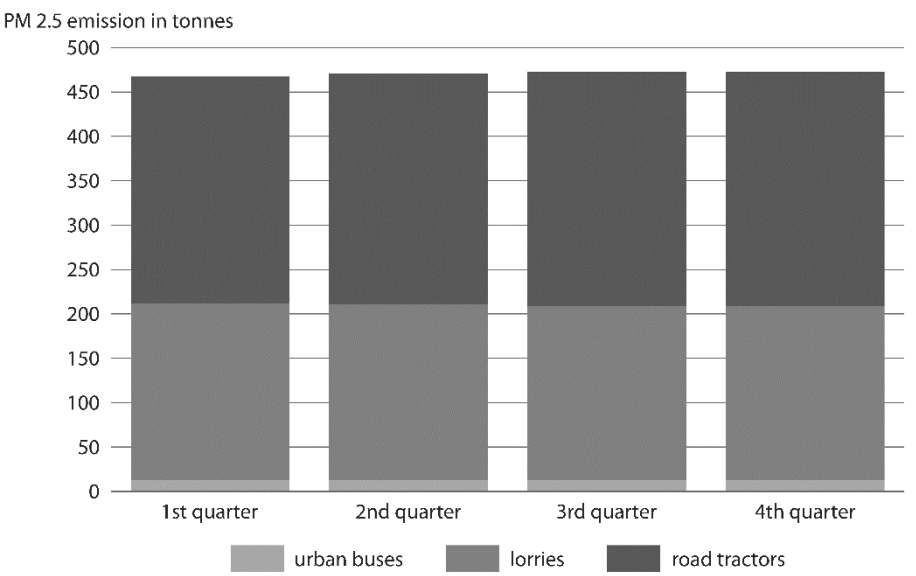
The following breakdowns are defined for the output variables:

- time: year, quarter;
- spatial: Poland, voivodships;
- vehicle category by type: lorries, road tractors, urban buses;
- category of the entity/vehicle according to the EURO emission class (2–6);
- category of the entity/vehicle by load group (GVW): heavy-duty vehicles (0–7.5 t; 7.5–12 t; 12–14 t; 14–20 t; 20–26 t; 26–28 t; 28–32 t; 32–36 t); heavy-duty vehicles with a trailer (14–20 t; 20–28 t; 28–34 t; 34–40 t; 40–50 t; 50–60 t); coaches (0–15 t; 15–18 t; 18+ t).

8.2. Outcome information

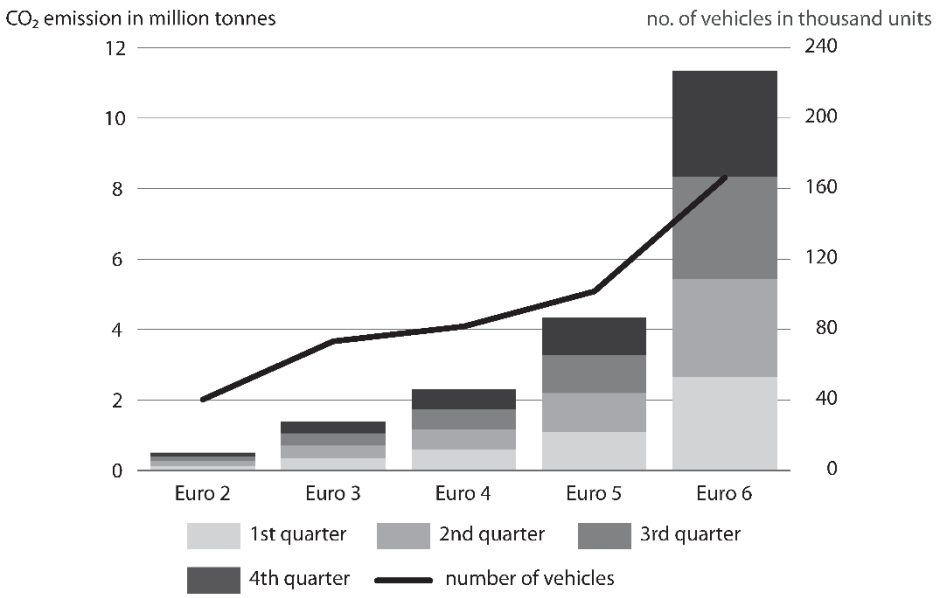
The data presented in the TranStat application show the volume of emissions generated by road transport on an annual and quarterly basis, broken down by type of pollution (Figures 12 and 13). These data are presented for individual categories of vehicles, detailing such vehicle characteristics as the gross vehicle weight and the Euro emission class. The data on vehicle emissions within individual emission classes also contain information on the number of registered vehicles for each Euro class. This additional variable allows for the correct interpretation of the results.

Figure 12. Particulate matter emission by vehicle type in 2021



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

Figure 13. Carbon dioxide emissions by vehicle emission class in 2021



Source: authors' work based on the results from the TranStat system (GUS, n.d.).

9. Electronic forms

As part of the TranStat project, the 1.48.02 Freight and passenger road transport (TD-E) and 1.50.01 Sea and coastal transport (TransMor) statistical survey forms were designed and adapted to the new requirements and needs of the respondents (intelligent electronic forms were provided to the respondents). It was carried out to:

- reduce the burden on respondents and the costs of statistical surveys;
- improve the method of collecting data in statistical surveys (TD-E, TransMor) by implementing mechanisms for autocomplete data (on vehicles and ships);
- improve the quality and completeness of statistical surveys.

Improving the method of collecting data in surveys allows respondents to fulfil the reporting obligation faster and easier while maintaining the security standards of the transmitted data, as well as high efficiency and ergonomics.[3]

The innovation of the solution consists in the implementation of rules for the automatic imputation of values from external sources, i.e.:

- for the TD-E survey – e-TOLL, the Database of Statistical Units (Pol. Baza Jednostek Statystycznych), and the Central Vehicle and Driver Register (Pol. Centralna Ewidencja Pojazdów i Kierowców);

[3] TransMor application – https://transmor.stat.gov.pl (GUS & US w Szczecinie, n.d. b). TD-E application – https://tde.stat.gov.pl (GUS & US w Szczecinie, n.d. a).

- for the TransMor survey, the AIS and the National Single Window (NSW) system – information on ship arrivals at seaports based on IMO FAL documents.

External data sources (outside official statistics) made it possible to obtain additional information on ships calling at ports (AIS-PL, NSW) and heavy-duty vehicles and coaches travelling on toll road sections covered by the e-TOLL system. In the case of sea transport, access to NSW gives an additional opportunity to view the transported cargo, significantly improving the data quality. To download information in real time from the NSW system (from the Maritime Office in Gdynia), a SOAP API server (Simple Object Access Protocol – communication protocol based on XML) was implemented in the DMZ of official statistics. The NSW system has a pervasive structure for data exchange directly between systems in the s2s mode (system to system). Implementing the environment in the Web Services technology was developed using the Apache HTTP server, PHP language and MS SQL Server database. Downloading data from sensors from the AIS-PL system and the e-TOLL is carried out by the data collection and processing subsystem of the TranStat system, e.g. for AIS-PL data by using a dedicated tool for handling data streams, i.e. Spark Streaming, which is a component of the Apache Spark platform. A detailed scheme of feeding forms from external sources is presented in Figure 2. Both forms were developed in the .NET Framework technology – ASP.NET MVC (Model, View, Controller) in the C# programming language, enabling data registration in a responsive web application (the interface is adjusted depending on the user's equipment, i.e. computer, tablet, mobile phone), thanks to which they retain complete functionality and ease of use.

10. Conclusions

The TranStat system has been implemented for statistical production and is an excellent enrichment of the Polish official statistics information system. Modernisation of the current approach for producing road and maritime transport statistics based on big data methods and tools was one of the overriding goals of the project.

The process was implemented through:

- obtaining access to big data streams for road (e-TOLL) and maritime (AIS, NSW) transport based on the completed legislative process, which guarantees the stability of data sources;
- cooperation with the scientific community as part of the methodology development for estimating traffic intensity, transport performance and emissions generated by road and maritime transport;

- creating a complete system production environment for all functional subsystems of the TranStat system;
- development of interfaces between individual system components;
- implementation of the necessary big data technology for data from sensors, enabling an automated data flow process, validation, processing and visualisation;
- development and implementation of algorithms (Apache Spark/Scala) enabling stream processing, data decoding (AIS) and necessary transformations for data from the AIS and e-TOLL systems;
- development and implementation of algorithms (Apache Spark/Scala) to generate new statistics for both types of transport (based on developed transport models);
- the design and implementation of intelligent electronic forms (containing data autocomplete mechanisms) that allow respondents to fulfil the reporting obligation faster and more efficiently while maintaining the security standards for the transferred data.

The benefits of using large data volumes (AIS, e-TOLL), downloaded in real time as part of the TranStat system, are:

- obtaining new information (before unavailable to recipients) on traffic intensity, transport performance and emissions generated by road and maritime transport that is necessary for making and monitoring transport policy at the national, regional and local levels;
- obtaining new knowledge trends regarding maritime and road transport statistics by using the correlation of multiple data sources;
- the ability to carry out in-depth analyses and evaluations of the communication system;
- providing high-quality data in a short time;
- reducing the burden on respondents fulfilling the reporting obligation through the use of smart forms; using methods and rules of automatic value imputation will strengthen the public's trust in public sector institutions;
- reducing survey costs by using non-statistical sources.

In addition, implementing the TranStat project strengthened the domain and analytical knowledge of the substantive employees of the Maritime Statistics Centre Transport and Communications Statistics Centre. It made it possible to build the competencies of the Data Engineering Centre employees at the Statistical Office in Szczecin in the area of big data, which guarantees the stability of the system's ongoing maintenance and the possibility of carrying out development works.

References

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, *31*(2), 249–262. https://doi.org/10.1515/jos-2015-0016.

De Gennaro, M., Paffumi, E., & Martini, G. (2016). Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities. *Big Data Research*, *6*, 11–25. https://doi.org/10.1016/j.bdr.2016.04.003.

e-TOLL. (n.d.). *Sieć dróg płatnych*. Retrieved June, 1, 2021, from https://etoll.gov.pl/ciezarowe/kalkulator-trasy/siec-drog/.

Główny Urząd Statystyczny. (n.d.). *TranStat*. https://transtat.stat.gov.pl.

Główny Urząd Statystyczny & Urząd Statystyczny w Szczecinie. (n.d. a). *TDE*. Retrieved May, 1, 2021, from https://tde.stat.gov.pl.

Główny Urząd Statystyczny & Urząd Statystyczny w Szczecinie. (n.d. b). *TransMor*. Retrieved June, 30, 2021, from https://transmor.stat.gov.pl.

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2019). *Periodic report No. 1 on implementing the TranStat project under the Social and economic development of Poland in the conditions of globalising markets GOSPOSTRATEG Program*.

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2020a). *Periodic report No. 2 on implementing the TranStat project under the program Social and economic development of Poland in the conditions of globalising markets GOSPOSTRATEG Program.*

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2020b). *Report on the methodology for estimating the volume of pollutants emitted using transport – road/maritime transport – task no. 4 as part of the research phase of the TranStat project.*

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2020c). *Report on the methodology for estimating the volume of transport performance – maritime transport – task no. 3 as part of the research phase of the TranStat project*.

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2020d). *Report on the methodology of measuring traffic statistics using large volumes of data – road/maritime transport – task no. 2 as part of the research phase of the TranStat project.*

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2020e). *Technical design of the system for measuring traffic intensity, transport performance and pollution generated by means of transport – task no. 5 as part of the research phase of the TranStat project.*

Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Politechnika Morska w Szczecinie & Politechnika Krakowska. (2021). *Final report on implementing the TranStat project under the Social and economic development of Poland in the conditions of globalising markets GOSPOSTRATEG Program*.

Rozkrut, D., Świerkot-Strużewska, O., & Van Halderen, G. (2021). Mapping the United Nations Fundamental Principles of Official Statistics against new and big data sources. *Statistical Journal of the IAOS*, *37*(1), 161–169. https://doi.org/10.3233/SJI-210789.

Vasilev, J., & Sulova, S. (2023). An Approach for the In-Depth Data Analysis of the Marine Traffic of Independent Nearby Ports. *Folia Oeconomica Stetinensia*, *23*(2), 402–426. https://doi.org/10.2478/foli-2023-0038.

Welch, T. F., & Widita, A. (2019). Big data in public transportation: A review of sources and methods. *Transport Reviews*, *39*(6), 795–818. https://doi.org/10.1080/01441647.2019.1616849.

Yin, Y., Lam, J. S. L., & Tran, N. K. (2021). Emission accounting of shipping activities in the era of big data. *International Journal of Shipping and Transport Logistics*, *13*(1–2), 156–184. https://doi.org/10.1504/IJSTL.2021.112922.

Zhang, G., Feng, S., & Wang, S. (2018). A Study on the Necessity of Statistical Index of Freight Multimodal Transport. *Management & Engineering*, (30), 3–9. https://doi.org/10.5503/J.ME.2018.30.001.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 25–48 https://doi.org/10.59139/ws.2023.12.2
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 18.07.2023, poprawiony/revised: 3.11.2023, zaakceptowany/accepted: 20.11.2023

The evaluation of (big) data integration methods in tourism[1]

Marek Cierpiał-Wolan,[a] Galya Stateva[b]

Abstract. In view of many dynamic changes taking place in the modern world due to the COVID-19 pandemic, the migration crisis, armed conflicts, and other, it is a major challenge for official statistics to provide high-quality information, which should be available almost in real time. In this context, the integration of data from multiple sources, in particular big data, is a prerequisite. The main aim of the study discussed in the article is to characterise and evaluate the following selected methods of data integration in tourism statistics: Natural Language Processing (NLP), machine learning algorithm, i.e. *K*-Nearest Neighbours (*K*-NN) using TF-IDF and *N*-gram techniques, and Fuzzy Matching, belonging to the group of probabilistic methods.

In tourism surveys, data acquired using web scraping deserve special attention. For this reason, the analysed methods were used to combine data from booking portals (Booking.com, Hotels.com and Airbnb.com) with a tourism survey frame. The study is based on data regarding Poland and Bulgaria, downloaded between April and July 2023. An attempt was also made to answer the question of how the data obtained from web scraping of tourism portals improved the quality of the frame.

The study showed that Fuzzy Matching based on the Levenshtein algorithm combined with Vincenty's formula was the most effective among all the tested methods. In addition, as a result of data integration, it was possible to significantly improve the quality of the tourism survey frame in 2023 (an increase was observed in the number of new accommodation establishments in Poland by 1.1% and in Bulgaria by 1.4%).

Keywords: data integration methods, tourism survey frame, web scraping
JEL: C1, C81, Z32

Ocena metod integracji danych dotyczących turystyki z uwzględnieniem big data

Streszczenie. W obliczu wielu dynamicznych zmian zachodzących we współczesnym świecie, spowodowanych m.in. pandemią COVID-19, kryzysem migracyjnym i konfliktami zbrojnymi,

[1] Artykuł został opracowany na podstawie referatu wygłoszonego na konferencji *Metodologia Badań Statystycznych MET2023*, która odbyła się w dniach 3–5 lipca 2023 r. w Warszawie. / The article is based on a paper delivered at the *MET2023 Conference on Methodology of Statistical Research*, held on 3rd–5th July 2023 in Warsaw.

[a] Uniwersytet Rzeszowski, Kolegium Nauk Społecznych, Instytut Ekonomii i Finansów; Urząd Statystyczny w Rzeszowie, Polska / University of Rzeszów, College of Social Sciences, Institute of Economics and Finance; Statistical Office in Rzeszów, Poland. ORCID: https://orcid.org/0000-0003-2672-3234. Autor korespon-dencyjny / Corresponding author, e-mail: m.cierpial-wolan@stat.gov.pl.
[b] National Statistical Institute, Bulgaria. ORCID: https://orcid.org/0009-0005-0755-6970.
E-mail: gstateva@nsi.bg.

ogromnym wyzwaniem dla statystyki publicznej jest dostarczanie informacji dobrej jakości, które powinny być dostępne niemalże w czasie rzeczywistym. W tym kontekście warunkiem koniecznym jest integracja danych, w szczególności big data, pochodzących z wielu źródeł. Głównym celem badania omawianego w artykule jest charakterystyka i ocena wybranych metod integracji danych w statystyce w dziedzinie turystyki: przetwarzania języka naturalnego (Natural Language Processing – NLP), algorytmu uczenia maszynowego, tj. *K*-najbliższych sąsiadów (*K*-Nearest Neighbours – *K*-NN), z wykorzystaniem technik TF-IDF i *N*-gramów, oraz parowania rozmytego (Fuzzy Matching), należących do grupy metod probabilistycznych.

W badaniach dotyczących turystyki na szczególną uwagę zasługują dane uzyskiwane za pomocą web scrapingu. Z tego powodu analizowane metody wykorzystano do łączenia danych pochodzących z portali rezerwacyjnych (Booking.com, Hotels.com i Airbnb.com) z operatem do badań turystyki. Posłużono się danymi dotyczącymi Polski i Bułgarii, pobranymi w okresie od kwietnia do lipca 2023 r. Podjęto także próbę odpowiedzi na pytanie, jak dane uzyskane z web scrapingu wpłynęły na poprawę jakości operatu.

Z przeprowadzonego badania wynika, że najbardziej przydatne spośród testowanych metod jest parowanie rozmyte oparte na algorytmach Levenshteina i Vincenty'ego. Ponadto w wyniku integracji danych udało się znacząco poprawić jakość operatu do badań turystyki w 2023 r. (wzrost liczby nowych obiektów w Polsce o 1,1%, a w Bułgarii – o 1,4%).

Słowa kluczowe: metody integracji danych, operat do badań turystyki, web scraping

1. Introduction

The modern world is determined by many threats, both global and local. World economies are facing increasing problems such as instability caused by numerous armed conflicts, energy crises, and the unprecedented scale of global migration. Additionally, since the beginning of 2020, the world has been struggling with the COVID-19 pandemic, which continues to affect many areas of our lives. All these circumstances are causing various effects of a socio-economic nature, which impact several economy sectors, and the tourism industry in particular. This leads to the emergence of huge demand for tourism-related data.

To provide high-quality, real-time information, it is necessary to integrate data from various sources, i.e. administrative registers, databases using information from censuses and sample surveys, and, most importantly, big data. Developing innovative methods of data integration has therefore become an imperative for academia and official statistics.

Data sources used so far by official statistics to create frames for tourism surveys have proven insufficient. This is due to both the specifics of the tourism market (short-term and sometimes incidental activity) and the fact that some part of the tourism-related activity is hidden in the shadow economy. In this context, big data is becoming an indispensable source of data.

The main aim of the study discussed in the article is to characterise and evaluate the following selected methods of data integration in tourism statistics: Natural Language Processing (NLP), machine learning algorithm, i.e. *K*-Nearest Neighbours

(*K*-NN) using TF-IDF and *N*-gram techniques, and Fuzzy Matching, belonging to probabilistic methods. In addition, we evaluated the quality of the tourism survey frame by obtaining data from web scraping of tourism portals. Selected methods were used to combine data from booking portals (Booking.com, Hotels.com and Airbnb.com) with the tourism survey frame in Poland and Bulgaria in 2023.

2. Big data in tourism statistics

Nowadays, the term 'big data' is used to describe a way of acquiring knowledge and learning about reality that is possible thanks to new technologies creating and processing large data sets. New data sources combined with innovative methods of processing them (especially machine learning, etc.) provide, in many cases, the possibility of publishing information on socio-economic phenomena and processes in a real-time mode, as well as better quality forecasts.

There are many classifications of big data. We assume that these are data from the following sources:

1. social interactions, especially social networks;
2. data-processing systems directly or indirectly related to business performance;
3. systems of electronic devices that automatically exchange data without human intervention – Internet of Things (United Nations Department of Economic and Social Affairs Statistics Division, 2015).

Another classification defines big data as information derived from sensors and any records of activity from electronic devices, social networks, business transactions, digital files (web pages, audio recordings, videos, PDF files, etc.) and real-time transmissions.[2]

As regards tourism surveys, useful information is that obtained by means of web scraping, as well as data from mobile network operators, traffic sensors or payment card operators.

Using web scraping, a technique for automatic extraction of data from websites, one can obtain valuable information on tourism phenomena and processes. In this context, online accommodation booking portals are particularly useful, and it is very important to select appropriate platforms from which data will be drawn. Hence, it is necessary to get to know the domestic market of online booking and identify the information resources of these portals. Both platforms with international coverage (e.g. Booking.com or Hotels.com) and local ones (e.g. Pochivka.bg for Bulgaria and Nocowanie.pl for Poland) should be taken into consideration. Data from traffic sensors, the Automatic Number Plate Recognition System (ANPRS), mobile

[2] Read more at: United Nations Economic Commission for Europe (UNECE), n.d.

network operators and payment cards are also very useful in monitoring tourist traffic.

It is worth noting that in European Union countries, traffic sensor data have been used for many years. Combined with the data from the ANPRS or the smart city more generally, they have been playing an increasingly important role in tourism statistics. Data held by mobile network operators, on the other hand, are an important source of information on the population movements. However, it is important to point out that this type of data should be subjected to detailed processing, especially in order to separate information on traffic related to daily activities from data relevant to tourism statistics. Still, gaining access to mobile network operators' data is very complicated, both legally and technologically (e.g. in the case of border areas, devices repeatedly log in outside the home network). As regards travel expenses of residents and foreigners, payment card data is the key source of information, because it takes into account spatial aspects. Companies operating payment terminals and ATMs have data on transactions made in individual countries. However, it has to be remembered that in countries with large migration, payment cards may be held by foreigners, which makes it difficult to accurately estimate the scale of spending. Therefore, the use of data obtained through payment cards should be accompanied by the analysis of additional sources of information to properly determine the status of the cardholder. A detailed breakdown of expenses (e.g. for lodging, food, transportation, goods) can be performed using the MCC (Merchant Category Code) classification.

In general, big data can be used in both census and sample surveys in various ways, e.g. only for data validation, as complementary sources, or to replace existing surveys entirely. When using these sources, especially in cyclical surveys, a pre-requisite is to secure the continuity of access to data, both formally and regularly. This could be done by diversifying data sources, as it is always possible to loose access to one or another source. In this context, it is important to constantly develop methods of data imputation and calibration. Data integration systems must be resistant to the loss of access to different sources of information. In practice, for example, simulators that integrate different sources of information can be an alternative to mobile data.[3] For traffic data at the EU's internal borders, an alternative to information from traffic sensors or the ANPRS is to use other sources such as smart city systems in nearby cities, parking meters, drones or satellite imagery.

[3] An interesting proposal for such a simulator was created in the framework of *ESSnet Big Data II* project (Oancea et al., 2019).

Another interesting way to provide constant access to big data is to develop mobile applications that would provide information and services related to users' needs, which, in return, collect diagnostic information from mobile devices (e.g. a travel planning application providing information on transportation, lodging, meal planning, preferred kinds of entertainment, etc.).

An obvious prerequisite for the use of big data is the positive assessment of their quality. In 2009, Daas provided some guidelines for the evaluation of the quality of various data sources, pointing out that each 'dimension' of quality can be defined by several indicators (Daas et al., 2009, pp. 6–9). Maślankowski (2015, pp. 173–174) asserted that the following 'dimensions' of the quality of big data were crucial: unambiguity, objectivity (mapping and reduction errors), inclusion of a timestamp, granularity or degree of detail, presence of duplicate data, completeness, accessibility, precision, interpretability, integrity, and consistency.

3. Selected methods of combining data

The process of combining datasets can be implemented by means of many methods. In the literature there are at least a dozen different methods, not to mention their variants and modifications. However, all of them belong to one of the following four groups of methods (Asher et al., 2020):

- Deterministic Record Linkage;
- Probabilistic Record Linkage;
- Data Fusion;
- Statistical Matching.

In the Deterministic Record Linkage and Probabilistic Record Linkage methods, the combined sets must be large enough for the probability of an object from one set beign also in the other set to be large as well.

In deterministic methods, the linkage process is based on simple rules of the logical exact matching of variables – keys. Any deficiencies in the data increase the risk of error of type I and type II, i.e. false matches and missing matches of records that should be linked. In probabilistic methods, the estimation of the probabilities of a random match between two values of a given variable, assuming that the paired records do not belong to the same unit, and the probabilities of a random mismatch between the values of a given variable, assuming that the paired records belong to the same unit, are incorporated into the process of combining data sets. The Data Fusion and Statistical Matching methods are dedicated to the process of combining relatively small sets, i.e. those for which the chance of an entity occurring in both sets is negligible. In the first case, linkage does not occur, but a concatenation of sets (union of sets) is created, taking into account common variables. Missing data are

imputed, for example, by interpolation. Statistical Matching is procedurally more complex and involves both the micro and macro approaches. In the micro approach, combining sets is done by either concatenation or matching based on similarity. Missing values are then imputed and record weights are calibrated. In the macro approach, a synthetic set is not created. The relevant parameters are estimated taking into account the special role of the covariance matrix of the variables present in both sets.

In this paper, we present the following data linkage and deduplication methods: Natural Language Processing machine learning algorithm, i.e. *K*-Nearest Neighbours using TF-IDF and *N*-gram techniques, and Fuzzy Matching belonging to probabilistic methods.

3.1. Natural Language Processing data linkage method

The NLP method with the Faiss library developed by Facebook AI Research for deduplication candidate generation and rapid comparison involves the following steps: tokenisation, vectorisation, comparison and similarity assessment, and deduplication decision.

The first step in this method is to transform the texts in character variables into tokens and then into semantic vectors using the SentenceTransformer library. This allows the meaning of the text to be expressed in a numerical form, which is necessary for further analysis. To find potentially duplicated accommodation establishments, we generated them on the basis of the input data. For that, we used the Faiss library, which enabled us to efficiently search the vector space to find similar items. With this search structure, we could quickly find similar accommodation establishments. The deduplication process also focuses on evaluating the coincidence between the two strings to determine the degree of their similarity and deciding whether they could be considered duplicates. To do this, we used Euclidean metrics to calculate the degree of difference between the two vectorised text strings.

In the context of the NLP processing, the selection of appropriate libraries plays a key role, especially when talking about differences between national languages. For instance, for English, there is a wide range of libraries and models ready to use, which facilitates research and application work. However, when we move on to inflectional languages, e.g. Polish or Bulgarian, the situation is different. For Polish, libraries such as spaCy offer dedicated models, allowing a more precise processing of the language, taking into account its grammatical and lexical peculiarities. For Bulgarian, on the other hand, the availability of tools and models is so far limited (libraries such as bgNLP and transformers allow basic text processing operations in this language).

3.2. Machine learning algorithm – *K*-Nearest Neighbours using Term Frequency-Inverse Document Frequency and *N*-gram techniques

Data linkage and deduplication procedures can be carried out using machine learning algorithms such as Random Forest, *K*-Nearest Neighbours (*K*-NN), clustering algorithms or neural networks (Quinlan, 1983, pp. 463–482).

The *K*-NN algorithm is a popular algorithm in machine learning that can be used to find similarity between data. The method used for data integration involves the use of Term Frequency-Inverse Document Frequency (TF-IDF) and *N*-gram techniques combined with the *K*-NN algorithm.

TF-IDF is a technique used to assess the validity of terms or tokens (string) in a text in the context of the entire dataset. It works by assigning weights to words based on their frequency in the document (TF) and the inverse frequency in the entire document set (IDF). A high TF-IDF value indicates that the term is valid in the document, allowing unique features of the data to be identified. *N*-grams are sequences of *N* consecutive tokens in a text. For example, bigrams are sequences of two words, and trigrams of three. The use of *N*-grams allows contextual information to be taken into account in text analysis. This is useful in the deduplication process, where the structure and layout of data are important.

Thus, the linkage and deduplication process begins with tokenisation, which consists in dividing the text into strings of characters. The use of the TF-IDF technique allows a more accurate detection of unique sequences of words or phrases in the text, which can be characteristic of duplicate data. The use of TF-IDF also helps reduce errors due to the similarity of single words. The next step is the use of N-gram creation, which makes it possible to take the context into account in text analysis. This is especially important in the deduplication process, as it helps detect similarities between data that differ not only in individual words, but also in their arrangement in sentences or phrases.

The next step involves using the *K*-NN algorithm, which determines similarities between different data. The *K*-NN algorithm operates on a feature space, where features are the TF-IDF values of *N*-grams that were previously calculated. An important part of this algorithm is the calculation of distances, using the appropriate distance metric. The choice of a particular distance metric depends on the type of data and deduplication goals. It is worth noting that *K*-NN is an unsupervised algorithm, which means that it does not require prior classification or data labels. Therefore, we use it as a tool to determine which data are similar to the largest extent. After calculating the distance between data instances, a similarity threshold is defined.

3.3. Fuzzy Matching

The last applied method is Fuzzy Matching along with Vincenty's formula, which is used to calculate the exact geodetic distances between accommodation establishments.

The Fuzzy Matching method allows the comparison of textual data, such as names of accommodation establishments, taking into account spelling errors, typos or differences in format. This makes it possible to find potential matches between records that are not identical, but may represent the same establishments. The method is based on algorithms such as the Levenshtein or Jaro-Winkler distance.

The Levenshtein algorithm is a text-editing algorithm that calculates the minimum number of operations (insertions, deletions or substitutions of characters) necessary to transform one text into another. With this algorithm, we can find potential matches between records that are not identical, but are similar to such a degree that they can represent the same accommodation establishments. The Jaro-Winkler algorithm, on the other hand, is an extension of the algorithm used to calculate the 'Jaro distance', and takes values from the [0, 1] interval, where 1 means the texts are identical, and 0 means there is no similarity between them at all. The Jaro-Winkler algorithm additionally has a prefix scale, which gives a higher similarity score when the strings share a common prefix (Cierpiał-Wolan et al., 2022).

Since the combined data may contain geographical coordinates, we can additionally use Vincenty's formula to calculate the exact distances, which might be important in the process of deduplicating accommodation establishments. Unlike simpler approximations such as the Haversine formula, Vincenty's formula takes into account the fact that the Earth is not perfectly spherical, but has the shape of an ellipsoid.

Finally, we consider two criteria, i.e. the distances between the strings for the selected variables and the geodetic distances. Thus, if the established thresholds for both criteria are met, we link the accommodation establishments.

The above-described methods were used to combine data from web scraping booking portals (Booking.com, Hotels.com and Airbnb.com) with the tourism survey frame. The procedure for combining this type of data is usually a multi-step process. Christen (2012) proposed five stages to it: standardisation, indexing, comparison, combining and the evaluation of results.

It is also worth noting that an interesting solution related to combining and deduplicating data regarding accommodation establishments on scraped portals is the use of algorithms that compare images. This method is based on analysing the visual characteristics of the images that are assigned to each establishment. There are several algorithms that can be useful in this kind of deduplication process, such as

comparing the similarity of colour histograms, comparing visuals using feature descriptors, and digital fingerprints.

4. Data sources

Information on accommodation establishments advertised in Poland and Bulgaria was downloaded between April and July 2023 from three booking portals: Booking.com, Airbnb.com and Hotels.com (a portal equivalent to Expedia). It is worth noting here that the number of scraped variables varied from a portal to portal. The information obtained from booking portals can be used both to supplement the tourism survey frame with new establishments, and to improve the quality of the results of the survey on the number of tourists and nights spent. Particularly important in this context is the information on the rental offers of accommodation establishments belonging to the NACE group 55.2 (Holiday and other short-stay accommodation), which are often difficult to identify primarily due to some part of them operating within the shadow economy.

Among the scraped variables, there are those that are crucial in the process of combining data obtained by web scraping with the tourism survey frame. These are: the name of the establishment, its address including longitude and latitude coordinates, and data on the services offered, to determine the size of the accommodation establishment and its type according to the NACE classification.

It is worth mentioning here that a major problem with linking and further processing data is classification differences regarding the types of accommodation facilities. In international comparisons, we usually use the NACE classification, but it is not used by booking portals. Hence, we need to make the data comparable – therefore, when assigning types of establishments to this classification, double verification was used. For this reason, we both used the classification of establishments declared on booking portals and we adopted a machine learning method (a classification tree). The learning dataset consisted of accommodation establishments from the booking portals linked with those found in the tourism survey frame. The application of this data processing procedure resulted in a relatively high accuracy of the units' assignment to accommodation-related NACE classification groups.

Web scraping was performed on the basis of simulated user interaction with the site using screen scraping. The adoption of this solution enabled full interaction with selected web pages, based on a dynamic modification of the Document Object Model (DOM) tree, cascading style sheet (CSS) components and JavaScript. The used web scraping system was prepared in the Python programming language with good practices to minimise the burden on the scraped portals.

4.1. Data scraped from Booking.com

As a result of web scraping from Booking.com, information on 82,965 booking ads, which in practice means 8,884 accommodation establishments in Poland, was obtained. In Bulgaria, the procedure yielded 3,873 establishments. 33 variables were scraped from Booking.com: the web address on Booking.com, the shortened web address on the portal, the name of the establishment, its address (street, number, postal code, city), the accommodation type, room name, maximum number of guests, rental price, facility area, the quality rating, number of nights spent, number of adults, number of children, number of rooms, number of double beds, number of single beds, number of couches, number of views, ability to communicate in English, availability of a car park, restaurant, bar, availability of onsite breakfast, availability of the Internet, TV, facility service, air conditioning, laundry, spa, fitness facilities, pool, and the availablity of facilities for the disabled.

It is worth mentioning that it was not always possible to obtain a full set of variables (e.g. about 67% of owners did not specify the area of the offered facility). This is very important in the process of classifying establishments in accordance with the NACE. It should also be noted that the classification of accommodation establishments adopted by Booking.com, as well as by other booking platforms, often differs from the classification used in official statistics (e.g. guesthouses are classified as hotel or agritourism accommodation). Therefore, a preliminary classification is usually made on the basis of the characteristics of the establishment, thanks to which we can unambiguously determine its type (Cierpiał-Wolan & WPJ Team, 2020). An important role in the process of classifying an accommodation establishment into a specific type is played by information on the rental price or the services offered, such as bed-making, serving breakfast, or the available catering facilities. The aforementioned amenities are characteristic for hotels and similar establishments classified as NACE 55.1. The pre-classification process is particularly important for new accommodation establishments which are not included in the tourism survey frame.

4.2. Data scraped from Hotels.com

Data were downloaded from Hotels.com on 4,620 accommodation establishments in Poland and 532 in Bulgaria. Twelve variables were extracted, namely: the web address on Hotels.com, the shortened web address on the portal, the name of the establishment, its address, type, number of rooms, the availability of a car park, restaurant, bar, the availability of onsite breakfast, and the availability of room and laundry services.

In contrast to the data obtained from Booking.com, only 11 accommodation establishments (0.2%) did not match any classification type. It is important to note

that the Hotels.com portal almost always provides information on the number of rooms in the accommodation establishment, which makes it possible to determine its size (e.g. Poland divides accommodation establishments into those with up to 9 or 10 and more beds). Out of the total number of the yielded accommodation establishments, 60% were hotels, and apartments accounted for 24%.

4.3. Data scraped from Airbnb.com

Using the web scraping method on the Airbnb.com portal, it was possible to obtain data on 12,556 accommodation establishments in Poland and 4,174 in Bulgaria. Eight variables were scraped from the Airbnb.com portal, i.e. the web address on Airbnb.com, the shortened web address on the portal, the name of the establishment, its address and type, the maximum number of guests, the rental price, and the number of beds.

A distinctive feature of this portal is that among the listed variables, the address of the accommodation establishment is not directly available. The address data on the site is limited to descriptive information like 'a cottage near the beach' or 'a sunny apartment at the foot of the mountains'. The great majority of facilities listed on the portal were those with fewer than 10 beds (97.2%).

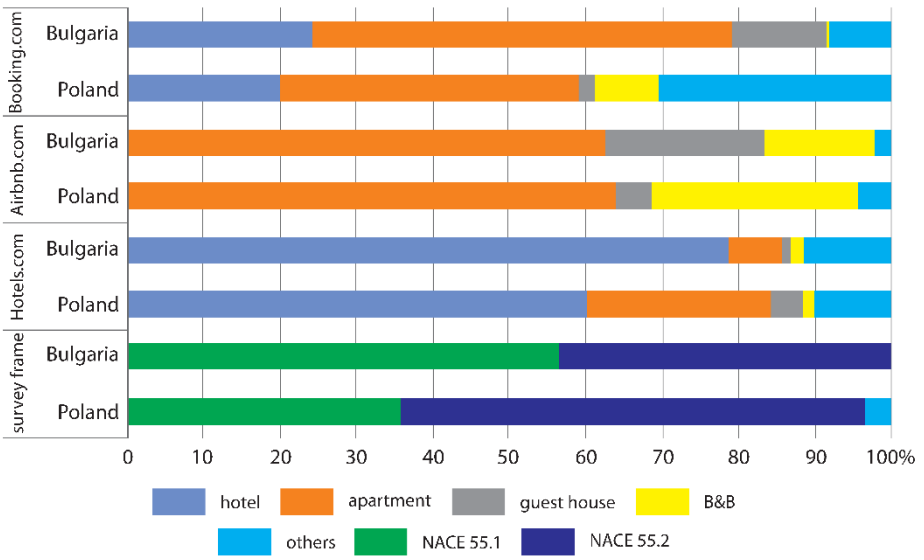
Compared to Booking.com, the vast majority of establishments on Airbnb.com (99.9%) were assigned a type. However, this classification contains generic names of the establishments that are not the same as those in the classification used in official statistics. Therefore, without additional information such as the area of the facility, the number of rooms or the available amenities (daily bed making, room cleaning and washing of sanitary facilities), it is not possible to classify an establishment into a particular type, which is labeled on the portal as e.g. a villa a loft or a condominium.

Figure 1 shows the structure of accommodation establishments on selected booking portals in Poland and Bulgaria by their generic name, as well as the classification of accommodation establishments in the tourism survey frame by the NACE classification in Poland and Bulgaria.

The structure of accommodation establishments scraped from Booking.com in Poland differs from such structure in Bulgaria. In the latter country, the vast majority of establishments are apartments (more than 50%), while the proportion of facilities classified as 'other' does not exceed 10%. In Poland, apartments also account for a considerable share of establishments (just under 40%), but the share of 'other' facilities is about 30%. This is due to the slightly different types accommodation on offer in these countries. In Poland, there are more establishments classified as other hotel establishments (which include, for example, B&Bs and

Aparthotels), while in Bulgaria, short-stay apartments rented by their owners prevail.

Figure 1. Structure of tourist accommodation in Poland and Bulgaria



Source: authors' work.

4.4. Tourism survey frame

The basis of our research was a tourism survey frame used in Poland and Bulgaria. In Poland it consisted of 13,804 establishments (with 10 and more beds), including 7,588 (55.0%) accommodation establishments classified as NACE 55.2. In Bulgaria, the tourism survey frame consisted of 4,031 establishments, of which 43.0% were classified as NACE 55.2.

From the tourism survey frame the following six variables were used: establishment name, establishment type, street, house number, postal code, city name.

5. Research results

The studies and analyses presented in this chapter were partially initiated in research projects[4] carried out by Statistics Poland and the National Statistical Institute in Bulgaria. Below, we present the results of the analyses for Poland.

[4] ESSnet Big Data II, Work Package WPJ (European Commission, n.d. b), ESSnet WIN, Work Package 3 – Use case 4 (European Commission, n.d. a).

In order to test the selected probabilistic methods of combining data, we used the tourism survey frame and the databases obtained by means of web scraping of the three booking portals mentiond before.

In addition, it was necessary to clean the web-scraped database (from typing errors, white spaces, HTML tags, etc.) in order to transform the unstructured data into a structured form corresponding to the data structure of the tourism survey frame. The cleaning process was performed in the Python programming language using the pandas, re, and the BeautifulSoup libraries. The cleaning process also used the web mapping and navigation service operated by HERE Technologies. Thanks to the aforementioned application, it was possible to carry out both the automatic parsing of address data into a common structure and the correction of language errors. The use of the HERE MAPS tool also made it possible to assign geographic coordinates to each accommodation establishment in the tourism survey frame and to do the same regarding the establishments in the web-scraped database (Cierpiał-Wolan et al., 2023).

All the above-described methods of data linkage and deduplication (NLP, *K*-NN, Fuzzy Matching) that we used generate linkage distances. To assess the usefulness of these methods, it was necessary to determine a distance threshold for the linkage. For this purpose, we used sensitivity (true positive rate) and specificity (true negative rate). In the case of data linkage, sensitivity is the probability of the correct match, while specificity is the probability of the correct non-match. Both terms are closely related to type I and type II errors. Since there is a trade-off between specificity and sensitivity, changing distance threshold always leads to improving one measure and worsening the latter. We evaluated various thresholds with the receiver operating characteristic (ROC) and found the optimal one by means of Youden's J statistic (Youden, 1950). With the optimal threshold, we generated a confusion matrix and derived a set of auxiliary statistics.

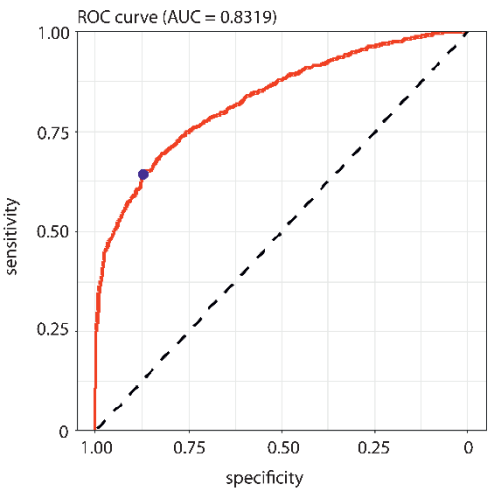
The ROC curve plots points of pairs – specificity and sensitivity – determined for a set of thresholds. When the curve is close to diagonal, this indicates that a given classifier is close to a random classifier. The better the classifier, the closer the curve to the top-left corner. Youden's J statistic is calculated as a sum of specificity and sensitivity minus one (Fawcett, 2006; Peirce, 1884; Powers, 2011).

5.1. Natural Language Processing

As a result of linking the establishments of the tourism survey frame with the establishments from web scraping by means of the NLP method (the main dataset was the tourism survey frame), 8,593 accommodation establishment connections were obtained. Then, based on the ROC curve, we determined the optimal threshold.

We examined a set of 2,314 thresholds ranging from 0.00 to 126.63. Figure 2 presents a ROC curve for the NLP method (solid red line), a ROC curve for the random classifier (dashed black line), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 2. The ROC curve for the NLP method



Source: authors' work using R package.

The optimal distance threshold amounted to 15.159. For this threshold, specificity and sensitivity amounted to 0.8730 and 0.6455, respectively, while Youden's J statistic reached 0.5184.

Table 1. Results of the NLP method

NACE	Number of matched establishments	Number of perfect matches	Distance		
			mean	minimum	maximum
55.1	1,440	90	17.8091	0	119.4789
55.2	717	3	24.2153	0	126.6318
55.3	49	0	33.8532	13.1397	85.4241
55.9	187	0	22.4814	4.2541	82.6862

Source: authors' work using Python package.

The distance at 0 consisted of 93 establishments (see Table 1). Differences at low distance values (0.68 to about 10) were mainly related to typos, e.g. missing the '-' symbol (in the names of accomodation establishments, cities and streets), incomplete names of establishments or streets, or missing Polish letters. However, it

could be detected that the link in this case concerned the same accommodation establishments. In the 10–20 range of distance values, there may have already been differences in the names of accomodation establishments, streets or their numbers. However, in most cases, the linked records concerned the same establishments. There were also mismatches of completely different records. Starting from the distance of 16, only half of the establishments were matched correctly. Above the value of approximately 30 (455 establishments), almost all the establishments were incorrectly linked, usually only by a few common characters, such as a fragment of the postal code or the phrase 'hotel, apartment' in the name of the establishment.

The quality of matching can be checked by the correctly and incorrectly matched and mismatched establishments, preferably using the confusion matrix. Four situations may arise after matching accommodation establishments:

1. the establishments have been correctly linked (true positive, TP);
2. two establishments have been mistakenly linked due to the short distance between them and similar names (distance smaller than threshold) (false positive, FP);
3. two establishments have not been linked (correctly) due to the large distance between them and low similarity between the names (true negative, TN);
4. two establishments have not been linked, but should have been, because there was only a small discrepancy in the establishment names (false negative, FN).

There are several approaches to building a confusion matrix for a data-matching problem. One of them assumes that we find the best match for all establishments in a smaller dataset with respect to a given distance or similarity score. Applying the optimal matching threshold, we obtain predictions: the match and mismatch. After a manual review of the linked data, we obtain the true state of the match and mismatch. Finally, we build a confusion matrix based on a set of true and predicted labels (match and mismatch).

Table 2 presents a confusion matrix for the data linkage result for the optimal distance threshold.

Table 2. Confusion matrix for the NLP method

Actual	Predicted	
	match	non-match
Match	0.31 (TP)	0.17 (FN)
Non-match	0.07 (FP)	0.45 (TN)

Source: authors' work using R package.

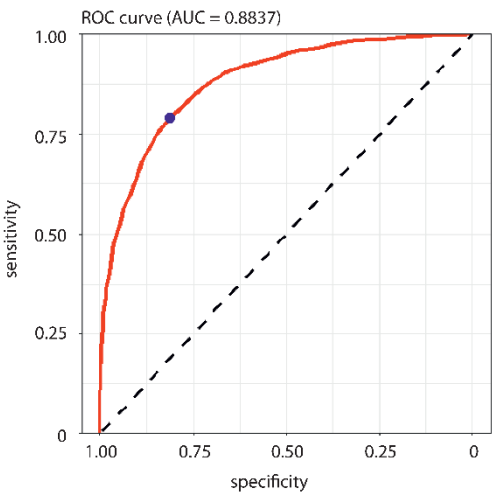
The accuracy amounted to 0.7622 with a 95% confidence interval (0.7446, 0.7792). The accuracy was tested against No Information Rate (NIR = 0.5132), and was significantly higher (p-value [Accuracy > NIR] < 0.0001).

5.2. Machine learning algorithm – *K*-Nearest Neighbours

Using the deduplication method based on machine learning algorithm, 3,157 establishments were combined. Similar to the NLP method, we set the optimal threshold.

We examined a set of 116 thresholds ranging from 0.00 to 1.29 (a large number of duplicated values of the metric). Figure 3, as in the case of NLP, presents ROC curves and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 3. ROC curve for the *K*-NN method



Source: authors' work using R package.

The optimal distance threshold amounted to 0.845. For this threshold, specificity and sensitivity reached 0.8121 and 0.7909, respectively, while Youden's J statistic totalled 0.6031.

Using the established value for the optimal distance threshold, data combining was performed.

The largest number of matches – 2,138 – was yielded for establishments belonging to the NACE group 55.1, and 54 of them linked perfectly (i.e. with zero distance). As regards accommodation establishments classified as NACE 55.2, 749 establishments were connected, of which three linked perfectly.

Table 3. Results of *K*-NN method

NACE	Number of matched establishments	Number of perfect matches	Distance		
			average	minimum	maximum
55.1	2,138	54	0.736908326	0	1.26
55.2	749	3	0.928958611	0	1.29
55.3	22	0	0.975454545	0.59	1.25
55.9	248	0	0.962540323	0.45	1.23

Source: authors' work using Python package.

A detailed analysis of the combined data showed that the distance measure ranged between 0 and 1.29. A distance of 0 covered 57 establishments. Differences at low distance values (0.14 to about 0.7) were related to incomplete names of accomodation establishments (e.g. lacking the owner's surname or abbreviation or company name), repetitions of keywords (e.g. 'hotel', 'apartment') and missing special characters. However, the links in almost all cases were true matches. Up to the distance level of 0.92, most establishment matches were correct; towards the end of this interval, differences in the names of accomodation establishments and numbers of buildings occured, but most street names and postal codes matched. Above the distance of 0.94, the vast majority of matches were incorrect, including street names or postal codes.

Table 4 presents a confusion matrix for data linkage result for the optimal distance threshold.

Table. 4. Confusion matrix for *K*-NN method

Actual	Predicted	
	match	non-match
Match	0.26 (TP)	0.07 (FN)
Non-match	0.13 (FP)	0.55 (TN)

Source: authors' work using R package.

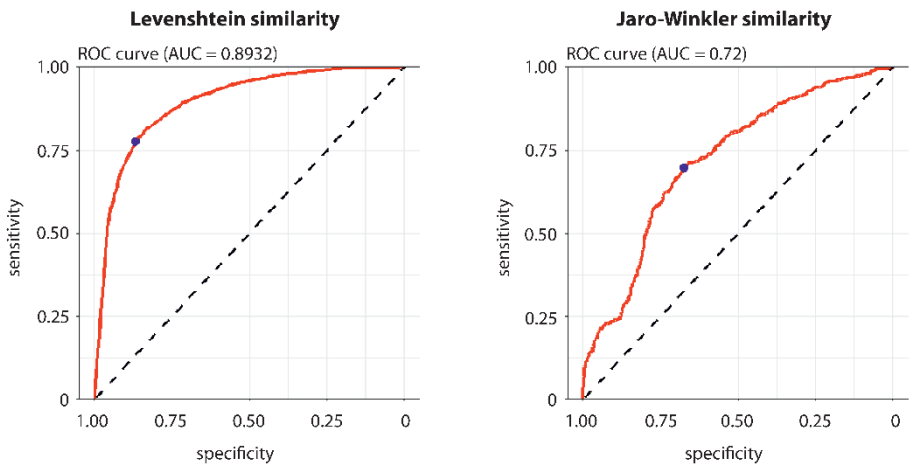
The accuracy amounted to 0.8052 with a 95% confidence interval (0.7969, 0.8134). The accuracy was tested against NIR (NIR = 0.6735). It was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001).

5.3. Fuzzy Matching and geolocation

The application of Fuzzy Matching and the use of geographic coordinates make it possible to obtain a set of linked accommodation establishments, containing two metrics for text strings (Levenshtein, Jaro-Winkler) and the geodetic distance between accommodation establishments in metres.

First, we checked which metric (edit distance) achieves better results. We examined a set of 1,977 thresholds ranging from 46 to 100 for the Levenshtein similarity and 1,843 thresholds ranging from 0.6 to 1 for the Jaro-Winkler similarity. Figure 4 presents the ROC curve for the applied method (solid red lines), the ROC curve for random classifier (dashed black lines), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue points).

Figure 4. ROC curve for the Levenshtein and Jaro-Winkler similarity



Source: authors' work using R package.

The optimal Levenshtein similarity threshold amounted to 83.5. For this threshold, specificity and sensitivity amounted to 0.8681 and 0.7757, respectively, while Youden's J statistic totalled 0.6438. The optimal Jaro-Winkler similarity threshold amounted to 0.73. For this threshold, specificity and sensitivity amounted to 0.6761 and 0.6975, respectively, while Youden's J statistic reached 0.3736. Table 5 presents a confusion matrix for the data linkage result for the optimal similarity threshold for the Levenshtein and Jaro-Winkler similarity.

Table 5. Confusion matrix for the Levenshtein and Jaro-Winkler similarity

Actual	Predicted	
	match	non-match
Levenshtein similarity		
Match	0.48 (TP)	0.07 (FN)
Non-match	0.04 (FP)	0.41 (TN)
Jaro-Winkler similarity		
Match	0.51 (TP)	0.25 (FN)
Non-match	0.07 (FP)	0.17 (TN)

Source: authors' work using R package.

For the two metrics tested, the best result was by far achieved by the Levenshtein algorithm, where approximately 48% of accomodation establishments were correctly matched. For 41% of the establishments from Booking.com, the accommodation establishments were not found in the tourism survey frame.

For the Levenshtein similarity, the accuracy amounted to 0.8912, with a 95%-confidence interval (0.8809, 0.9131). The accuracy was tested against NIR (NIR = 0.6355). It was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001). For the Jaro-Winkler similarity, the accuracy amounted to 0.6812, with a 95% confidence interval (0.6595, 0.7023). NIR (NIR = 0.7622) was higher than accuracy. It is worth noting that matching with the Jaro-Winkler similarity is near to a random classifier. The results of the analysis confirmed that combining Fuzzy Matching with the Levenshtein similarity is more effective than the Jaro-Winkler similarity.

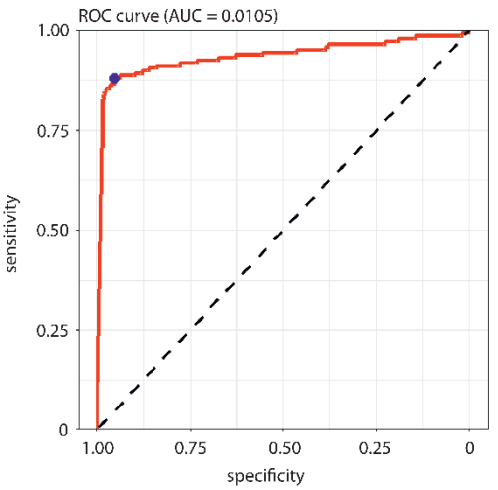
The Levenshtein similarity score ranged between 46 and 100. The similarity score of 100 covered 1,435 matches. Detailed analysis of the data showed that the differences in similarity scores in the range of 89–99 (924 establishments) were related to incomplete establishment names, repetition of keywords (e.g. 'hotel', 'apartment'), or lack of special characters. Linked records mostly involved the same establishments, but sometimes there were differences between the numbers of buildings or properties.

For the similarity score ranging from 85 to 88, the number of correctly and incorrectly matched establishments was similar. At the similarity score of 78, 80% of records were incorrectly matched. The reasons for these differences were the same as aforementioned. In addition, numerous discrepancies in the street and the name of establishments can be observed.

The detailed results of the algorithm using the Jaro-Winkler similarity were also analysed. The similarity score for addresseses using the Jaro-Winkler distance ranged between 60 and 100, and only five linked establishments reached the highest score (100). Similarity score values between 88 and 99 (41 links, including eight incorrectly linked) were the result of the incomplete name of the accommodation establishment (most often the lack of letters next to the building/apartment number, e.g. '9' where it should be '9a'), or the absence of individual special characters. Below the value of 88, correctly linked establishments totalled 404. However, their distribution was uneven.

To apply the geodesic distance between accommodation establishments, we also needed to determine the optimal threshold based on the ROC curve. For this purpose, we examined a set of 1,442 thresholds ranging from 0.006 to 697.3 km. Figure 5 presents a ROC curve for this method (solid red line), a ROC curve for a random classifier (dashed black line), and a pair of specificity and sensitivity for the optimal threshold derived from Youden's J statistic (blue point).

Figure 5. ROC curve for Vincenty's distance



Source: authors' work using R package.

The optimal Vincenty distance threshold amounted to 0.026 km. For this threshold, specificity and sensitivity totalled 0.9520 and 0.8795, respectively, while Youden's J statistic reached 0.832. Table 6 presents a confusion matrix for the data linkage result for the optimal similarity threshold.

Table 6. Confusion matrix for Vincenty distance

Actual	Predicted	
	match	non-match
Match	0.29 (TP)	0.02 (FN)
Non-match	0.02 (FP)	0.67 (TN)

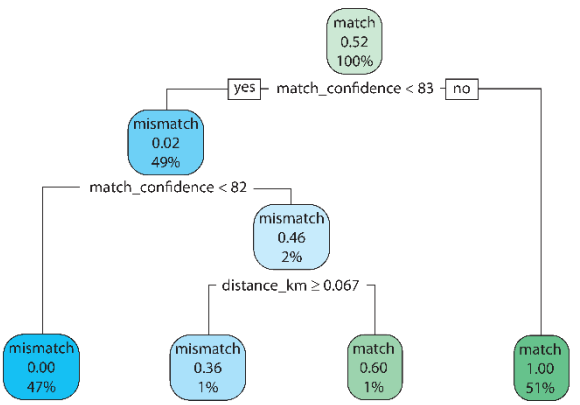
Source: authors' work using R package.

The accuracy amounted to 0.9612, with a 95% confidence interval (0.944, 0.9769). It was tested against NIR (NIR = 0.7622) and was significantly higher than NIR (p-value [Accuracy > NIR] < 0.0001).

As for the distance between accommodation establishments calculated using Vincenty's formula, all establishments within the distance up to 50 metres were linked correctly. In the distance range of 50–200 metres, the number of establishments linked correctly was comparable to the number of those linked incorrectly. A few establishments for which the distance was greater than 1 km were linked correctly, which was due to the specific notation of addresses obtained by web scraping.

Finally, we adopted a Fuzzy Matching method based on Levenshtein similarity and geolocation. Since we had two criteria to choose from, it was necessary to define decision rules for linking. The most intuitive solution seemed to be the conjunction of the two critical values. However, we did not want to rely on intuition, so we used a decision tree. The optimal complexity parameter value was determined at 0.016 by means of *k*-fold cross-validation. Figure 6 presents a decision tree using Levenshtein similarity and geolocation.

Figure 6. Decision tree for the Levenshtein similarity and geolocation



Source: authors' work using R package.

The model accuracy amounted to 0.9919, while specificity and sensitivity reached 0.9921 and 0.9917, respectively. Youden's J statistic totalled 0.9838.

Let us assume a match with a 100 similarity score and zero Vincenty distance is a perfect match. Table 7 presents the summary of the applied method.

Table 7. Results for the Levenshtein similarity and geolocation

NACE	Number of matched units	Number of perfect matches	Similarity score			Geodetic distance		
			mean	minimum	maximum	mean	minimum	maximum
Total	**3,170**	**1,400**	**94.13**	**82**	**100**	**0.63**	**0**	**711.86**
55.1	1,815	891	95.19	82	100	0.25	0	700.43
55.2	1,354	508	92.77	82	100	1.13	0	699.05
55.3	1	1	100.00	100	100	0	0	0

Source: authors' work using Python package.

For the NACE groups 55.1 and 55.2, the key criterion was a similarity score of at least 83 and less than 82. These two values separated 98% of cases. For values between 82 and 83, the geodetic distance was the deciding factor, which applied to about 2% of cases. The average value of the Vincenty distance was 0.250 metres for the NACE group 55.1 and over four times more, i.e. 1.13 meters, for the NACE group 55.2. Similarity score statistics were the same for both the above-mentioned NACE groups. In the case of the NACE group 55.3, only one establishment linked.

6. Conclusions

The growing demand for tourism information is caused both by external circumstances (the COVID-19 pandemic, large-scale migration, armed conflicts) and increasing expectations of tourists. In many countries, tourism is treated as a priority sector because of its role and the benefits it brings to the economy. The above circumstances as well as the dynamic development of new technologies and competition in information markets are forcing national statistical offices to carry out activities involving the continuous search for new sources of information and, above all, their integration into statistical databases and administrative records. Meeting these challenges is a highly complex phenomenon. Having reliable and real-time information on the tourist traffic, the average length of tourists' stay and the degree of their spending opens up new opportunities for effective tourism policies at the local, regional and international levels.

Data linkage and data deduplication from web scraping of tourism portals with the tourism survey frame are aimed at ensuring high-quality research results. Depending on the availability of data on websites, which also involves formal and legal considerations, different deduplication methods can be used. Each of these methods has its own strengths and weaknesses, which should be taken into account when choosing the right solution.

The article provides a detailed characterisation and evaluation of three data linkage and deduplication methods: NLP, *K*-NN and Fuzzy Matching. The use of NLP offers numerous benefits, such as scalability, flexibility and automation. Machine learning algorithms are also useful for data linkage and deduplication. When deciding which method to use, a combination of different algorithms for better results is worth considering (we, for example, considered the TF-IDF and *N*-gram techniques). A similar situation occurs with the Fuzzy Maching method, where, in addition, Vincenty's formula was used to calculate the exact geodetic distances between the establishments. It is also important to choose the appropriate distance metric, which affects the accuracy of the results.

The evaluation of the selected methods of linking and deduplicating the data was done using the confusion matrix, the ROC curve and Youden's J statistic. The best results were obtained by using the Fuzzy Matching method based on Levenshtein similarity combined with Vincenty's formula. It is worth noting that this method copes well with arbitrary notation of the names of establishments and can also be used to classify them.

The article analysed data from three booking portals, i.e. Booking.com, Hotels.com and Airbnb.com. It should be noted that these portals differ significantly in terms of the volume of information available. Therefore, Booking.com was chosen for the procedure of linking and deduplicating the data, due to its largest range of variables corresponding to the tourism survey frame. In this context, a very promising further research direction is the possibility of using algorithms that compare images. This way, it is possible to combine data from different portals more efficiently (photos become an additional key of correlation).

The presented research results are also important in the context of improving the quality of the tourism survey frame. It turns out that the use of web-scraped data resulted in an increase in the number of accommodation establishments classified as NACE 55.1 and NACE 55.2. In 2020, information was yielded on 151 new accommodation establishments in Poland and 56 in Bulgaria. These establishments accounted for 1.1% and 1.4% of the total number of accommodation establishments constituting the tourism survey frame in Poland and in Bulgaria, respectively. Most of them belonged to the NACE group 55.2 (64% of Poland-based units and 58% of the Bulgaria-based ones). Another large group were accommodation establishments belonging to the NACE group 55.1 (31% of Poland-based units and 26% of Bulgaria-based ones). They were mainly Aparthotels and B&B establishments, which, according to the adopted methodology, are classified as 'other hotel establishments'.

Currently in tourism statistics, information from booking portals is used for both data imputation and calibration. The process aimed at a full replacement of selected tourism surveys with information from online portals has already been launched. In this context, it is important to remember about prerequisites for the use of big data, namely a stable access to such data and a positive assessment of its quality. Combining new information with administrative registers and other sources of statistical data, by means of appropriate models, can lead to qualitatively new statistics at the micro-, meso- and macroscale.

References

Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An Introduction to Probabilistic Record Linkage with a Focus on Linkage Processing for WTC Registries. *International Journal of Environmental Research and Public Health*, *17*(18), 1–16. https://doi.org/10.3390/ijerph17186937.

Christen, P. (2012). *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2.

Cierpiał-Wolan, M., Truszyńska, A., Szlachta, P., Wnuk, Z., Sawicki, K., Oprych-Franków, D., Data, M., Ulma-Ciupak, B., Giełbaga, E., Wieczorek, G., Gumiński, M., & Mordan, P. (2022). *Feasibility project on digitalisation issues in national accounts*.

Cierpiał-Wolan, M., & WPJ Team. (2020). *Innovative Tourism Statistics Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used*. Eurostat, ESSnet Big Data II. https://ec.europa.eu/eurostat/cros/sites/default/files/WPJ_Deliverable_J2_Interim_technical_report_showing_the_preliminary_results_and_a_general_description_of_the_methods_used_2020_01_07.pdf.

Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Tóth, J. (2009). *Checklist for the Quality evaluation of Administrative Data Sources* (CBS Discussion Paper No. 09042). https://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60.

European Commission. (n.d. a). *Project Overview*. Retrieved July 8, 2023, from https://cros-legacy.ec.europa.eu/content/project-overview_en.

European Commission. (n.d. b). *WPJ Innovative tourism statistics*. Retrieved July 8, 2023, from https://cros-legacy.ec.europa.eu/content/WPJ_Innovative_tourism_statistics.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Maślankowski, J. (2015). Analiza jakości danych pozyskiwanych ze stron internetowych z wykorzystaniem rozwiązań Big Data. *Roczniki Kolegium Analiz Ekonomicznych SGH*, (38), 167–177. https://rocznikikae.sgh.waw.pl/p/roczniki_kae_z38_11.pdf.

Oancea, B., Necula, M., Salgado, D., Sanguiao, L., Barragán, S. (2019). *ESSnet Big Data II. Workpackage I: Mobile Network Data. Deliverable I.2 (Data Simulator). A simulator for network event data*. Eurostat.

Peirce, C. S. (1884). The Numerical Measure of the Success of Predictions. *Science*, *4*(93), 453–454. https://doi.org/10.1126/science.ns-4.93.453-a.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://bioinfopublication.org/pages/article.php?id=BIA0001114.

Quinlan, R. (1983). Learning efficient classification procedures. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.), *Machine Learning. An Artificial Intelligence Approach* (pp. 463–482). Springer-Verlag. https://doi.org/10.1007/978-3-662-12405-5.

United Nations Department of Economic and Social Affairs Statistics Division. (2015). *Classification of Types of Big Data*. https://unstats.un.org/unsd/classifications/expertgroup/egm2015/ac289-26.PDF.

United Nations Economic Commission for Europe. (n.d.). *Unece Statswiki*. Retrieved July 8, 2023, from https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 49–64 https://doi.org/10.59139/ws.2023.12.3
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 13.09.2022, poprawiony/revised: 28.08.2023, zaakceptowany/accepted: 20.09.2023

Current challenges and possible big data solutions for the use of web data as a source for official statistics[1]

Piet Daas,[a] Jacek Maślankowski[b]

Abstract. Web scraping has become popular in scientific research, especially in statistics. Preparing an appropriate IT environment for web scraping is currently not difficult and can be done relatively quickly. Extracting data in this way requires only basic IT skills. This has resulted in the increased use of this type of data, widely referred to as big data, in official statistics. Over the past decade, much work was done in this area both on the national level within the national statistical institutes, and on the international one by Eurostat. The aim of this paper is to present and discuss current problems related to accessing, extracting, and using information from websites, along with the suggested potential solutions.

For the sake of the analysis, a case study featuring large-scale web scraping performed in 2022 by means of big data tools is presented in the paper. The results of the case study, conducted on a total population of approximately 503,700 websites, demonstrate that it is not possible to provide reliable data on the basis of such a large sample, as typically up to 20% of the websites might not be accessible at the time of the survey. What is more, it is not possible to know the exact number of active websites in particular countries, due to the dynamic nature of the Internet, which causes websites to continuously change.

Keywords: big data, web data, websites, web scraping

JEL: C55, L86, M21

Współczesne wyzwania i możliwości w zakresie stosowania narzędzi big data do uzyskania danych webowych jako źródła dla statystyki publicznej

[1] Artykuł został zaprezentowany w postaci referatu na konferencji *Metodologia Badań Statystycznych MET2023*, która odbyła się w dniach 3–5 lipca 2023 r. w Warszawie. / The article was presented in the form of a lecture at the *MET2023 Conference on Methodology of Statistical Research*, held on 3rd–5th July 2023 in Warsaw.

[a] Eindhoven University of Technology, Department of Mathematics and Computer Science, the Netherlands. ORCID: https://orcid.org/0000-0002-1541-0315. E-mail: p.j.h.daas@tue.nl.

[b] Uniwersytet Gdański, Wydział Zarządzania; Urząd Statystyczny w Gdańsku, Ośrodek Inżynierii Danych, Polska. / University of Gdańsk, Faculty of Management; Statistical Office in Gdańsk, Centre for Data Engineering, Poland.
ORCID: https://orcid.org/0000-0003-0357-2736. Autor korespondencyjny / Corresponding author, e-mail: jacek.maslankowski@ug.edu.pl.

Streszczenie. Web scraping jest coraz popularniejszy w badaniach naukowych, zwłaszcza w dziedzinie statystyki. Przygotowanie środowiska do scrapowania danych nie przysparza obecnie trudności i może być wykonane relatywnie szybko, a uzyskiwanie informacji w ten sposób wymaga jedynie podstawowych umiejętności cyfrowych. Dzięki temu statystyka publiczna w coraz większym stopniu korzysta z dużych wolumenów danych, czyli big data. W drugiej dekadzie XXI w. zarówno krajowe urzędy statystyczne, jak i Eurostat włożyły dużo pracy w doskonalenie narzędzi big data. Nadal istnieją jednak trudności związane z dostępnością, ekstrakcją i wykorzystywaniem informacji pobranych ze stron internetowych. Tym problemom oraz potencjalnym sposobom ich rozwiązania został poświęcony niniejszy artykuł.

Omówiono studium przypadku masowego web scrapingu wykonanego w 2022 r. za pomocą narzędzi big data na próbie 503 700 stron internetowych. Z analizy wynika, że dostarczenie wiarygodnych danych na podstawie tak dużej próby jest niemożliwe, ponieważ w czasie badania zwykle do 20% stron internetowych może być niedostępnych. Co więcej, dokładna liczba aktywnych stron internetowych w poszczególnych krajach nie jest znana ze względu na dynamiczny charakter Internetu, skutkujący ciągłymi zmianami stron internetowych.

Słowa kluczowe: big data, dane webowe, strony internetowe, web scraping

1. Introduction

The use of web-scraped data for the production of official statistics encounters numerous methodological challenges. When the number of businesses maintaining websites is unknown, we can estimate it using web-scraped data. However, because this type of data is often biased, our estimate may not be accurate, i.e. some classes of enterprises could be over- or underestimated. Therefore a survey, understood as a questionnaire with questions to be answered, which is based on web data, may provide data aggregates that do not accurately represent the intended target population.

Web scraping refers to the process of using software for automatic extraction of data from websites (Khder, 2021). In this paper, we understand web scraping as a method to get the source of a website, i.e. the source file (mostly HTML), preceded by checking the robots.txt file and server headers. Web-scraped data are extracted from the website to get useful information.

The aim of this paper is to present and discuss current problems related to accessing, extracting, and using information from websites, along with the suggested potential solutions. The secondary aim is to provide an overview of methods and cases that can be used and replicated to extract statistical data from websites. This article is the result of the authors' long experience in working with this type of data.

The essential research question in this study is whether it is possible to collect internet data suitable for the production of official statistics using massive web scraping techniques. Along with the literature review, the research methods used in the paper comprised the authors' case studies of enterprise websites and case studies conducted at the European level, all focused on producing official statistics. In this

paper we demonstrate the results of a case study involving massive web scraping, performed in 2022 on a total population of 503,700 Polish websites. The results showed that such a large sample could not yield fully reliable data, as usually up to 20% of the studied websites are not accessible at the time of the survey. These findings are in line with other studies in this area, for example Oancea and Necula (2019) or Daas and van der Doef (2020). Web-scraped data has been regarded as a data source for official statistics for more than 10 years (Daas et al., 2015). Nowadays, web-scraping is a fundamental requirement during data scientist training (Dogucu & Çetinkaya-Rundel, 2020). This technique has not only been adopted in statistical research, but also in scientific papers or marketing reports, which includes, for example, marketing scholars using Application Programming Interface (APIs) to collect data from the Internet. In marketing research, the number of papers using online data increased from 1% in 2001 to 15% in 2020 (Boegershausen et al., 2022). Researchers relatively often collect price data from the web, for example, to calculate the value of the real estate market (Antonov & Laktionova, 2020), to produce price indices of real estate (Pegueroles et al., 2021) and used cars (Nasiboglu & Akdogan, 2020), to compile an experimental consumer price index (Oancea & Necula, 2019), or to produce consumer electronic products (goods) and airfares (services) price indices in order to improve the Harmonised Index of Consumer Prices (Polidoro et al., 2015). Another well-known example is the Billion Prices Project at the Massachusetts Institute of Technology (MIT), which scrapes massive amounts of prices from the web to produce daily online price indices for the USA and several other countries (Cavallo & Rigobon, 2016). Financial and other types of information can easily be extracted from web data with a variety of supporting packages, which provide basic tools used for pre-processing web data (Krotov & Tennyson, 2018).

More advanced examples of research into web-scraped data include the use of text mining and Natural Language Processing techniques (NLP) to study local policies (Anglin, 2019) or to identify particular types of enterprises (Daas & van der Doef, 2020). In the latter case, machine learning methods are used to find words that correlate with a particular type of enterprise, e.g. innovative companies. NLP has been applied, for example, in analyses of the labour market by studying online job advertisements. Usually, a large portion of information, such as skills required for a certain job advertised online, is extracted from unstructured descriptive texts (Schedlbauer et al., 2021). Online job advertisement data can also provide input for different indicators on labour market statistics, such as the Labour Market Concentration index (Ascheri et al., 2022).

Web scraping for official statistics has been particularly well studied in a number of ESSnet projects: *Big Data I* (European Commission, n.d. a) and *Big Data II* (European Commission, n.d. b). They yielded some experimental statistics using

web data on online job vacancies, enterprise characteristics, and innovative tourism statistics (European Commission, n.d. c). The varied and complex use of web data in official statistics necessitated creating a web-scraping policy, which was formulated at the European or the NSI level (European Commission, n.d. d). It features the principles of web scraping according to good practice, such as delaying accessing pages on the same domain or adding idle time between requests (Office for National Statistics, n.d.). Currently, most of the work regarding the use of web data at the European level is done in the Web Intelligence Hub (WIH; Wirthmann & Reis, 2021) project, which is supported by an international community of statisticians within the WIN. The latter is a centralised repository offering services used to scrape data, store them in a repository, and provide data processing and analysis tools. One of the goals of the WIN is to improve knowledge and strengthen web intelligence competencies of statisticians in the use of the WIH services across the ESS and beyond (European Commission, n.d. e).

2. Research method

There are different approaches to defining statistical populations while using web data. These can be divided into three groups, as presented in Table 1.

Table 1. Web scraping examples by population size

Population size	Examples
P1: One website	Satellite data Search engine results
P2: Selected websites (Purposive sampling)	Online job advertisements Real estate prices Price statistics
P3: All websites	Enterprise characteristics Innovative company detection

Source: authors' work.

Often a single website (P1) is scraped (single-site web scraping) – for example, search engine results are collected for an analysis. In such a case, the web scraping software is collecting data from an individual website represented by one URL, i.e. a website address.

Collecting data from a set of websites selected by researchers (P2) is called purposive sampling (Palys, 2008). It involves the selection of a sample on the basis of the researcher's judgement as to which subjects fit the criteria of the study best (Purposive sampling, n.d.). Price statistics, for example, are based on a number of specific e-commerce portals. A collection of job advertisements compiled from

several websites is another example of P2. Since the sample of the scraped websites is selected before the research is performed, it may not be statistically representative of the target population the researcher had in mind, which is a potential source of bias that might seriously affect the outcome.

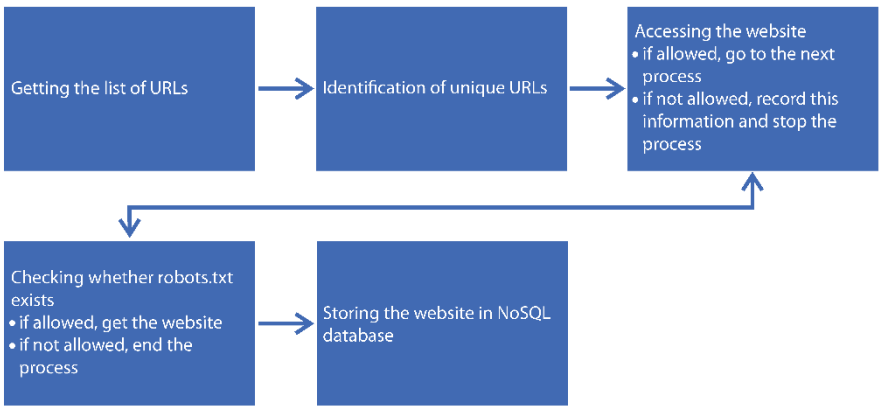
Collecting data on the entire population (P3) is the most technically-challenging approach, for three reasons. Firstly, a publicly available database of all active websites is not available in every country. There are domain registration authorities headed by the Internet Assigned Numbers Authority (IANA), but their databases are not publicly available. Secondly, creating a complete database by combining all URLs from various sources is a good starting point, but most probably, some websites will still be missing. Hence additional search and crawl process is required. Crawling refers to the process of finding additional URLs on websites already accessed. This step can also be used to check the quality of the link between the statistical units, e.g. enterprises, and websites found, and may also provide information on the units without a website. Thirdly, the composition of active websites is very dynamic; data yielded may already be outdated by the time the process is finished.

For all these reasons, obtaining an almost full overview of all websites in a country requires considerable effort. What is more, the total population of websites is often not exactly known. The size of the population is therefore often just estimated, and usually only popular websites are selected, potentially introducing biases. This is because some less popular websites, yet important from the point of view of the study, are skipped. How this can be avoided, for instance, was described in Daas and van der Doef (2020).

3. Case study on web scraping of the selected collection of websites

This section presents the case study of collecting data on a selection of enterprises. The study is based on the data scraped from the websites of Polish enterprises (URLs gathered from the Orbis database), but we will also be referring to the Dutch experience. The 'Polish enterprise' is understood as a company present in the Orbis database, located in Poland. The URL data for the Dutch study were obtained from DataProvider (a Dutch company). These data were subsequently linked to the corresponding businesses in the Business Register of Statistics Netherlands at the most detailed level possible. The linking procedure compared the Chamber of Commerce number and address from the website with those in the Business Register; see Daas and van der Doef (2020) for more details. The process used for web scraping in this case study is presented in Figure.

Figure. Web-scraping process used in the case study



Source: authors' work.

First, we decided to create a list of URLs based on the Orbis entities database managed by Bureau van Dijk. It is a commercial data provider service which maintains the database. Orbis is the repository for entity data. It consists of information on nearly 450 million companies and entities around the world (Orbis, n.d.). We took from this database all URLs linked with companies in Poland. As presented in Table 2, the collection of URLs for Poland was comparatively large (503,700 enterprises) and linked to enterprise business IDs. The linking process was based on the attributes such as an ID, name, address, etc., included in the Orbis database. Next, the set of URLs was limited to those with unique domain names, to prevent the same domain from being scraped multiple times. There were duplicates in the dataset, i.e. we found thousands of enterprises using the same domain name. This usually occurs when there is a consortium of enterprises with many branches, or when the local enterprise is based on franchise. Then, it was checked whether it was possible to access the website. Some of the websites were inaccessible, generally due to three reasons:

- website rejected by the server due to suspicious internet traffic (robot detected);
- the server was out of service/web domain was not active or
- a time-out of the request.

Sometimes it helps to re-visit an 'inaccessible' website at a later time. For instance, in the study performed by Daas and van der Doef (2020), inaccessible websites were re-visited at four different times. The next step was to check whether the robots.txt file, if available, allowed the website to be scraped. We found that for more than 95%

of the websites in which the robots.txt file denied scraping (at a certain level), it was still possible to collect the data, when ignoring this file for testing purposes. If this was not possible, the process was stopped and the appropriate flag was recorded in a NoSQL database. If the robots.txt file allowed access to the homepage, the homepage was stored in the database for further processing.

Even though we decided to use the URLs from the Orbis database, it is important to note that there are other ways of obtaining an URL list. One of the options is to actively search for websites with the URL retrieval software (European Commission, n.d. a, n.d. b). This is described in more detail in the subsequent paragraphs. The second option is to get various databases of URLs and merge them to have the most complete database. This involves using Wikipedia, the Whois database or publicly available business registers storing URLs, e.g. for Poland it could be the National Court Register.

Table 2. Results of the case study of web scraping of a selected collection of websites

Specification	Websites	
	number	percent
Population size	503,700	100
Unique domain names	459,700	91
Accepted connections	340,700	74

Source: authors' work.

The list of URLs taken from the Orbis database contained only websites of enterprises. The study was conducted in the first quarter of 2022. Duplicates were identified for about 44,000 enterprises, i.e. 8.7% of websites. In the unique population of URLs, nearly 26% of websites did not respond correctly. Servers failed to connect because of the three already-mentioned reasons. This shows that the URL database must be maintained and updated on a regular basis.

During the URL collecting-and-scraping process, we identified some problems and proposed solutions to them. They are presented in Table 3.

Table 3. Problems identified during the URL sampling and web scraping

No.	Issue	Methods of mitigating
1	Incomplete URL list	Use URL search to find additional URLs
2	Non-updated data on the list of URLs	Use URL search script to verify if URLs have changed
3	Outdated information on websites	Regularly scrape websites
4	Website is blocking robots	Try to use an alternative approach, i.e. use a different web browser engine to scrape the data and inform the website owner of the issue

Table 3. Problems identified during the URL sampling and web scraping (cont.)

No.	Issue	Methods of mitigating
5	Robots.txt rejection	Inform website owner of the intention to scrape data (scrape anyway)
6	Temporary unavailability	Scrape the website at another time/date
7	No time stamps	Regularly scrape the website and monitor changes by comparing stored data in NoSQL database
8	Duplicates of websites	Apply de-duplication mechanisms and URL-forward checks
9	Only partial information obtained	Check if the website is still active and if yes, check the script to extract more data
10	The quality of the link between an enterprise and the URL	Check whether the website refers to the enterprise in the population by verifying that the company's details, like the name or address, exist on the website
11	Information on enterprises without a website (if relevant)	Check whether there are other sources of information available, such as a survey, or contact a small sample to obtain an indication of the number of enterprises and type(s) of data missing

Source: authors' work.

With regard to incomplete URL lists (Table 3 issue 1), it is possible to get more information by using additional sources. According to the report from the ESSnet Web Intelligence Network group responsible for obtaining data from the web, the number of URLs in the official statistical databases in selected ESS countries ranges between 2,000 and 20,000. These are the URLs of enterprises, governmental institutions, and other types of entities, like NGOs. It is important to note that URLs in official statistics are usually not collected on a regular basis, and this process is often supported by external software or third-party databases. In Poland, for example, there are several such databases, e.g. the CEiDG (the Central Register and Information on Economic Activity) and KRS (National Court Register), which are publicly available and used to support official statistics. However, some countries are only sourcing URLs from third-party institutions. An example of a fairly complete set of URLs is the Orbis database (European Commission, 2022a). The Orbis database is probably the most comprehensive set of URLs, however it requires paid subscription, according to the purchasing plans of Orbis (n.d.).

Another option is to create a list of URLs by using search terms such as the company's name, address, etc. in relevant search engines. A tool that uses search engines and evaluates and validates the resulting URLs is called a URL retrieval software. In this approach, URLs are directly obtained from web search engines and checked by visiting websites. Such software is using Google, Bing, Yahoo, or DuckDuckGo search engines to obtain a set of URLs based on the enterprise characteristics, e.g. the name, address, business ID. Usually, multiple URLs are

yielded by the search engine that needs to be checked to either reduce the number of possible websites or select the appropriate one (European Commission, n.d. a, n.d. b). Finding a correct URL can be especially challenging for small companies. This is even more the case if company names resemble each other or are too generic. For this reason, the URL retrieval is usually followed by a machine learning-based classification which classifies an URL as correct or incorrect along with the confidence rate, to increase the chance that the correct URL is found.

Another possibility to expand the URL list is to extract domain names from company e-mail addresses (European Commission, 2022b). This approach is particularly interesting for official statistics, as most of the surveys are conducted online, and the contact between the respondent and the NSI is via e-mail. While e-mail address domains may not directly relate to the enterprise in all the cases (e.g. some companies use gmail.com, outlook.com, etc.), they can nonetheless help to increase the total number of URLs found. The downside of any URL retrieval approach is the fact that they might be found for enterprises that actually do not have a website. It may happen when there are companies with the same name located in the same city or area. Even nowadays, this can still be the case for (some) small companies active in specific branches, such as farms. To sum up these issues, even though the use of third-party databases and URL retrieval software can support official statistics' URL databases, it takes much of work to obtain a complete and reliable set of enterprise URLs.

As mentioned before, official statistics requires a regularly updated list of URLs. Outdated URL lists (Table 3 issue 2) can create a scenario where a large part of the listed websites may not respond. A solution is to use software to check the availability of websites. However, in some cases, we experienced a situation where a website that did not respond at the beginning of the data collection period was active the following week. During our experiments, we established that when a website that does not respond after the first visit, another scraping should be done after a few days, up to a maximum of four attempts. If none of these attempts are successful, the website is assumed to be inactive. Usually, massive web scraping of thousands of websites may last up to 2 weeks to assure that all scrapable websites responded. Another important issue is how to deal with enterprises that have changed their URL. This can be done by performing an URL search for enterprises that were found to have an inactive URL.

Avoiding out-of-date information on websites (Table 3 issue 3) requires regular visiting and scraping websites. This is the key to obtaining high-quality web data. However, there are several enterprises that provide very limited information on what they actually do. In addition, some websites remain publicly accessible even if the enterprise is no longer operating.

Blocking robots (web scrapers) by the website owner (Table 3 issue 4) is a very important potential obstacle to consider. One solution is informing the website owner about the intention to scrape their website and asking their permission for that, but this is not convenient while scraping thousands of websites. The alternative is to use another supplementary scraping approach, such as a web-browser-based engine, e.g. a headless browser from Chromium or Mozilla Firefox, that might mitigate the criteria used by the website owner and prevent being identified as a robot. The use of these web browser engines should be the same as the typical use of web browsers, i.e. there should be delays between requests and not all the attachment links (e.g. PDFs) should be scraped. Using such an approach will certainly increase the number of the collected websites. In our case study, we observed that adopting this option increased the positive response of about 10% of the websites. It is very important to indicate the robot properly in the user agent variable by including a 'web scraper for official statistics'.

Robots.txt rejection (Table 3 issue 5) is challenging. On the one hand, it is possible to access the data even if the robots.txt denies this type of traffic on the website; we found it was possible in 95% of such cases. On the other hand, we should respect the rules laid down by the website owner in the robot.txt file. However, since the data is used for the production of official statistics, our suggestion is to scrape the data and inform the website owner about this unexpected traffic via the appropriate channels used by the NSI of the country.

Website temporary unavailability (Table 3 issue 6) is also related to the second item on the list of possible reasons for the unavailability of a website, i.e. outdated URLs. However, in this case, we are focusing on the temporary unavailability of a website. This issue can be solved by repeating the requests at another date and time (as mentioned before). If the requests repeatedly fail for a large numbers of websites, it might be helpful to change the IP address, use a VPN or delete cookies.

In many cases, the time stamp is very important when collecting data (Table 3 issue 7). For example, when collecting job advertisements, it is important to have information on the date and time when a specific job advertisement was published. The seventh issue shown in Table 3 illustrates a situation where it is not possible to extract the time stamp from the website. If that is the case, one option is to perform web scraping at a regular basis for a specific period, e.g. daily or weekly, according to the requirements of the research, and to compare the results to see what has been added or removed. However, an easier soultion is to simply use the date of web scraping, withouth worrying when the ad was first published. If the advertisement is on a website, most probably it is valid.

Website duplicates (Table 3 issue 8) come in two different forms. One occurs when a selected number of websites is scraped, in order to, for example, obtain real

estate advertisements, and the same or a very similar advertisement is discovered in the data collected. The second relates to a URL list in which one URL is used by several (different) enterprises or when different URLs redirect users to the same website. In the first of the above-mentioned cases, it is necessary to include the detection of similar items in the data-processing phase. Very similar advertisements should be treated as the same record. In the second case, where a URL is linked to more than one enterprise, it is important to check whether the enterprises with the same link are connected (e.g. branches, etc.). If this is actually so, the results of the website analysis should also be linked to each of these enterprises. When different URLs refer to the same webpage, this usually indicates that an enterprise wants to increase the traffic to its webpages by increasing the chance that the website is found. In this case, it is important to verify if the original URLs are all correctly associated with the enterprise and not with others.

Limited amount of information provided by websites (Table 3 issue 9) can negatively affect the results of web-scraping. Manual check is required to assure that the website is still owned by the company and that the information extracted is all that is available on the webpage. We found that this situation is quite often caused by websites reporting that the domain is either 'for sale' or 'under construction'. If this is the case, these websites are actually inactive and need to be excluded. However, when the website is owned by a company and some (relevant) information is provided, this needs to be included in the subsequent processing steps. One way of dealing with these kinds of websites is to include them in the final estimation process as a separate group (Daas & van der Doef, 2020).

The quality of the relation between an enterprise (sample unit) and its website (Table 3 issue 10) needs consideration when, in some cases, there are errors in the databases resulting in a website not being linked to the appropriate company. It is also possible that the domain has expired due to non-payment and requests are redirected to the website of the service provider. In all these cases, the solution is to check the content of the website and compare it with the data in the business register, i.e. the company's name, address, etc.

Collecting data on enterprises which do not have a website but are included in the sample population (Table 3 issue 11) is only a problem if the data from these enterprises is required. Obtaining all the required information from enterprises without a website may be difficult when a large sample is studied. However, one can attempt to study a smaller population, e.g. a selected type of companies, for which data is available in another source, such as a survey. If this is possible, one could attempt to estimate the total number of companies with no website to get an idea of the size of this part of the population. Such an approach is briefly explained in the study on innovative Dutch companies (Daas & van der Doef, 2020). In this study,

the number of innovative companies without a website was estimated via the units included in the Community Innovation Survey and by contacting a small sample of potential innovative small companies directly. The final estimate of innovative Dutch companies without a website was 0.1%.

4. Discussion and examples

The use of website data in statistical production cannot be overestimated. First of all, all kinds of data on the demand on the labour market, real estate advertisements or price statistics can be downloaded this way at a minimum cost. The cost is predominantly the work of people collecting and processing the data. The essential issue is obtaining a representative (part of a) population to be used in the study. Knowledge of the market and the largest players in the studied field might be helpful. During our extensive work in the area of web scraping, we formulated some helpful recommendations. Firstly, when looking for the most relevant websites to be scraped, do not assume that the biggest are the best. For example, there are numerous websites with job ads, but only a few of them have been stable and reliable over time. This is essential, as time series might be significantly disturbed by including websites with volatile job advertisements in the population. As shown in Table 1, it affects P1 and P2 population sizes.

If voluminous web scraping is necessary, the authors of this study prefer to scrape as many websites as possible. A typical example of massive web scraping is the Online-Based Enterprise Characteristics (OBEC) survey. It has been repeatedly conducted for a selected number of EU countries and the results can be found in the experimental statistical website (European Commission, n.d. c). The difference between the case study described in this paper and the OBEC survey is the population size. In our case study, we used all the websites of Polish enterprises available in the Orbis database. In the OBEC survey, on the other hand, which is conducted for several EU countries, only those websites of the OBEC population were scraped that have been traditionally used in the survey on the ESS enterprises' application of the ICT. The expectation of similar results for both studies independent of the data collection mode is the main motivation for using website data to extract enterprise characteristics.

However, the voluminous web-scraping case study described in this paper demonstrates that when this technique is used, researchers should make allowances for the issues described in Table 3. Here, it actually helps to collect as much data as possible. Additional advantage of massive scraping is that the findings for websites of smaller enterprises (i.e. those with fewer than 10 employees) are included, which is not the case in traditionally-conducted surveys. On the other hand, excluding

enterprises with fewer than 10 employees makes the mitigation of problems much easier. This is why there are some suggestions to limit the population to the most reliable URLs. The same arguments hold for the study of innovative companies (Daas & van der Doef, 2020). In this Dutch study it was shown that i) the traditional survey-based estimate of the number of large innovative companies could be done with web-scraped data only, and that ii) the number of small innovative companies could be determined for the first time.

Slightly less complicated is a study that uses a selected number of websites, like job advertisements. Our experience of working with such data shows that these can be easily included and fulfill the requirements of a traditional survey. Additional advantage of small-scale web scraping, i.e. based on a smaller number of websites, is the fact that there is the possibility to contact all website owners and inform them about scraping their domains. However, the occurrence of duplicates in web-scraped data is a much more complicated problem. For instance, in the case of job advertisements, many providers might add (a set of) the same advertisements, some of which may even be repeated at different locations within the same domain. According to our experience with the OJA data for Poland from the four largest OJA portals, up to 10% of the job advertisements collected were found to be duplicates. Duplicates can be very difficult to detect, because the enterprise to which the job applies sometimes cannot be accurately identified.

5. Conclusions

This paper presents an overview of issues that affect the use of website data for official statistics. Some of these problems are of technical nature, and can be solved relatively easily. However, as our case study, the analysis of the literature and our personal experiences demonstrate, the most challenging problem is related to the selection of a set of websites, as well as the quality of the link between the units and the websites used in the study. The larger the number of websites used, the more serious these issues become. The difficulty is that a certain number of pages and objects needs to be collected to represent the target population. This population may not be known in advance; what is more, it is often determined no sooner than the data has been collected. This particularly concerns enterprises where, based on the Eurostat data, it should be possible to estimate the percentage of firms having a website, but it may not be possible to indicate exactly which of them actually have one. The previously mentioned URL-search approach can be used here to mitigate this problem.

The application of internet robots using search engines provides an opportunity to increase the number of URLs. This increases the sample size, which is very helpful

when conducting research based on websites. External sources provided by third-party companies may also prove helpful for public statistics. The synergistic effect of different URL retrieval methods and additional sources will certainly contribute to creating a list of URLs as complete as possible, and are also likely to enhance the relationship between an enterprise and the accompanying website. However, it should not be forgotten that the Internet is constantly changing and data collected today may differ significantly from the those collected tomorrow. This is relevant for all studies that use web data. Time stamps are essential here.

This enables us to answer the research question posed at the beginning of this paper, namely whether it is possible to collect internet data suitable for the production of official statistics using massive web scraping techniques. The tentative answer is that one source of URLs (a database) may not be enough to conduct reliable massive web scraping surveys. Multiple sources, which need to be maintained, managed and updated, and supplemented by third-party providers (whenever possible), are generally preferable. The case study conducted on one database revealed that nearly 20% of URLs were not accessible, due to non-existing websites or website owners blocking access of the robots.txt file to prevent scraping. Therefore, we suggest using the methods and solutions listed in Table 3 to deal with these issues. When all information is available and the data has been processed, subsequent bias correction methods need to be applied to produce the best possible estimate.

References

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576.

Antonov, O., & Laktionova, O. (2020). Evaluation of Real Estate Market Value in Ukraine Using Web-Scraping. *Galician Economic Journal*, *63*(2), 35–44. https://doi.org/10.33108/galicianvisnyk_tntu2020.02.035.

Ascheri, A., Marconi, G., Meszaros, M., & Reis, F. (2022). Online Job Advertisements for Labour Market Statistics using R. *Romanian Statistical Review*, (1), 3–26. https://www.revistadestatistica.ro/2022/03/online-job-advertisements-for-labour-market-statistics-using-r/.

Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of Gold: Scraping Web Data for Marketing Insights. *Journal of Marketing*, *86*(5), 1–20. https://doi.org/10.1177/00222429221100750.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Inflation Measurement and Research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151.

Daas, P. J. H., & van der Doef, S. (2020). Detecting Innovative Companies via their Website. *Statistical Journal of IAOS*, *36*(4), 1239–1251. https://doi.org/10.3233/SJI-200627.

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, *31*(2), 249–262. https://doi.org/10.1515/jos-2015-0016.

Dogucu, M., & Çetinkaya-Rundel, M. (2020). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education*, *29*(sup1), 112–122. https://doi.org/10.1080/10691898.2020.1787116.

European Commission. (n.d. a). *ESSNet Big Data I*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en.

European Commission. (n.d. b). *ESSNet Big Data II*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/essnet-big-data-2_en.

European Commission. (n.d. c). *Experimental big data statistics*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/content/Experimental_big_data_statistics_en.

European Commission (n.d. d). *Web scraping policy*. Retrieved April 21, 2023, from https://cros-legacy.ec.europa.eu/content/item-04-web-scraping-policy_en.

European Commission. (n.d. e). *Trusted Smart Statistics – Web Intelligence Network*. Retrieved August 17, 2022, from https://ec.europa.eu/eurostat/cros/WIN_en.

European Commission. (2022a). *Deliverable 2.1: WP2 1st Interim Progress Report.* https://cros.ec.europa.eu/system/files/2023-12/wp2_deliverable_2_1_wp2_1st_interim_progress_report_20220331_revision_2.pdf.

European Commission. (2022b). *Report: URL finding methodology*. https://cros-legacy.ec.europa.eu/system/files/20220131_url_finding_methodology.pdf.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, *13*(3), 144–168. https://doi.org/10.15849/ijasca.211128.11.

Krotov, V., & Tennyson, M. (2018). Research Note: Scraping Financial Data from the Web Using the R Language. *Journal of Emerging Technologies in Accounting*, *15*(1), 169–181. https://doi.org/10.2308/jeta-52063.

Nasiboglu, R., & Akdogan, A. (2020). Estimation of the Second Hand Car Prices from Data Extracted via Web Scraping Techniques. *Journal of Modern Technology & Engineering*, *5*(2), 157–166. http://jomardpublishing.com/UploadFiles/Files/journals/JTME/V5N2/NasibogluR.pdf.

Oancea, B., & Necula, M. (2019). Web scraping techniques for price statistics – the Romanian experience. *Statistical Journal of the IAOS*, *35*(4), 657–667. https://doi.org/10.3233/SJI-190529.

Office for National Statistics. (n.d.). *Web Scraping Policy*. Retrieved August 17, 2022, from https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/webscrapingpolicy.

Orbis. (n.d.). *Overview* [Data set]. Retrieved April 28, 2023, from https://www.bvdinfo.com/en-gb/our-products/data/international/orbis.

Palys, T. (2008). Purposive sampling. In L. M. Given (Ed.), *The Sage Encyclopedia of Qualitative Research Methods, Vol. 2* (pp. 697–698). Sage. https://doi.org/10.4135/9781412963909.

Pegueroles, P., Guerrero, R., Fernández, A., & López, D. (2021). Price's Index through of Web Scraping. *Revista Chilena de Economía y Sociedad*, *15*(1), 32–54. https://rches.utem.cl/wp-content/uploads/sites/8/2022/01/revista-chilena-de-economia-y-sociedad-vol15-n1-2021-Pegueroles-Guerrero-Fernandez-Lopez.pdf.

Polidoro, F., Giannini, R., Lo Conte, R., Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, *31*(2), 165–176. https://doi.org/10.3233/SJI-150901.

Purposive sampling. (n.d.). In *Oxford Dictionary*. Retrieved April 28, 2023, from https://www.oxfordreference.com/display/10.1093/oi/authority.20110810105658510.

Schedlbauer, J., Raptis, G., & Ludwig, B. (2021). Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, *150*, 1–9. https://doi.org/10.1016/j.ijmedinf.2021.104453.

Wirthmann, A., & Reis, F. (2021). *The Web Intelligence Hub – A tool for integrating web data in Official Statistics*. 63rd ISI World Statistics Congress, Online. https://cros-legacy.ec.europa.eu/sites/default/files/isi_-_web_intelligence_hub_eurostat_paper.pdf.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 65–82 https://doi.org/10.59139/ws.2023.12.4
Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges
Zgłoszony/submitted: 2.10.2023, poprawiony/revised: 1.12.2023, zaakceptowany/accepted: 4.12.2023

Digital transformation and data ecosystem: implications for policy actions and competency frameworks

Monika Rozkrut[a]

Abstract. The article discusses EU policy for digital transformation and the associated development potential. The article aims to critically analyse the current progress of the operationalisation and implementation of the relevant policies. It is followed by the recognition of the challenges that can contribute negatively to the necessary strategic objectives and obstacles that may hinder reaching the policy goals. In particular, a significant obstacle may be a major deficiency of adequately prepared experts ready to work in new roles in a dynamically developing data ecosystem. A remarkable example is the role of the Data Steward. This role is essential for fostering the rapid development of the data ecosystem in the EU. We propose creating a universal competence framework for Data Stewards to streamline human resource allocation. The article proposes a basic outline of the necessary skills and competencies ensuring effective data stewardship.

Keywords: digital transformation, data ecosystem, Data Steward, data stewardship, data sharing, competence framework

JEL: C80, O30, J24, O15

Transformacja cyfrowa i ekosystem danych – implikacje dla tworzenia polityk i wymagań kompetencyjnych

Streszczenie. W artykule omówiono politykę Unii Europejskiej dotyczącą transformacji cyfrowej i związany z nią potencjał rozwojowy. Celem pracy jest krytyczna analiza postępu, jaki dokonuje się obecnie w zakresie operacjonalizacji i wdrażania odpowiednich polityk. Ponadto zidentyfikowano wyzwania, które mogą mieć negatywny wpływ na osiągnięcie koniecznych celów strategicznych, oraz przeszkody mogące utrudniać realizację przyjętych polityk. Jako szczególne utrudnienie postrzega się znaczny niedobór ekspertów przygotowanych do pełnienia nowych funkcji w dynamicznie rozwijającym się ekosystemie danych. Znaczącym przykładem jest rola data stewarda, kluczowa dla wsparcia szybkiego rozwoju ekosystemu danych w UE. Zaproponowano stworzenie uniwersalnych ram kompetencji dla data stewardów w celu usprawnienia zarządzania zasobami ludzkimi. W artykule przedstawiono podstawowy zarys niezbędnych umiejętności i kompetencji zapewniających efektywne zarządzanie danymi.

[a] Uniwersytet Szczeciński, Wydział Ekonomii, Finansów i Zarządzania, Instytut Ekonomii i Finansów, Polska / University of Szczecin, Faculty of Economics, Finance and Management, Institute of Economics and Finance, Poland. ORCID: https://orcid.org/0000-0002-9701-6388. E-mail: monika.rozkrut@usz.edu.pl.

Słowa kluczowe: transformacja cyfrowa, ekosystem danych, data steward, zarządzanie danymi, wymiana danych, ramy kompetencji

1. Introduction

Over the last few years, modern digital technologies have significantly influenced changes in economies and societies, affecting individual enterprises, entire sectors of activity, public administration, and citizens' everyday lives. Digital transformation is changing the way people live, work and communicate. Data are at the heart of this transformation. The amount of information generated by citizens, businesses and public authorities is constantly growing. Whole new data ecosystems are created, with enterprises, public institutions, and even households collecting data. These can be treated as a new resource and a new factor of production. The growing amounts of industrial and public data, combined with technological changes in how they are stored and processed, represent a potential source of growth and innovation that should be harnessed. The wide availability of data and their safe and open exchange offer the possibility to respond to many vital economic and societal challenges. The proper use will enable enterprises, the public sector and citizens to make more informed and beneficial decisions positively influencing their development. Fast and secure data exchange will ensure the further growth of a knowledge-based economy and enhance the quality of life. Therefore, there is a growing need to tackle data governance, better organise and improve the functioning of the data ecosystem, and make it cohesive and coherent, at least in the basic scope, to facilitate data exchange and application between all interested parties. New regulations are necessary to increase trust in sharing data and ensure the security of their processing; guidance over data quality is likewise crucial. All these needs, in turn, translate into a growing demand for qualified staff with competencies in specific areas related to the broadly understood digital transformation. The article aims to critically analyse the strategic activities of the European Union aimed at stimulating the progress of the digital transformation and creating a data space. The analysis points out that a significant obstacle to reaching the policy goals is the shortage of adequately trained specialists to perform the role of Data Stewards. The emergence and rapidly increased interest in jobs related to data stewardship seem particularly important. It seems that fostering this newly reinvented role determines the development of properly set up and functional data ecosystems adequate to its potential and needs.

2. Digital transformation

As general-purpose technologies, information and communication technologies (ICT) saturate all aspects of socio-economic life. On the one hand, they change the ways of social interactions, and on the other hand, they more and more often enable direct communication between devices and objects that create what is called the Internet of Things. Companies are changing their internal structures to use ICT effectively. Employees acquire new skills to use them, process information and learn. Public administration adjusts the ways of interacting with citizens and businesses accordingly. New means of communication lead to the creation of new behaviour models and, as a result, consumption patterns are changing as well. Undoubtedly, new technologies are currently one of the fastest-growing sectors of the economy in developed countries, characterised by an increasing share in the creation of GDP.

The progress in modern computer techniques and technologies has led to a revolution in collecting, processing, storing and transmitting information (Bell, 1973). Modern digital technologies influence and support business activity. The digitalisation process concerns communication and business management and often affects production or service processes. This results in IT systems in enterprises that collect, send, store and process larger and larger data sets.

These technologies fundamentally change how various entities operate: producers, service providers, customers and eventually the whole economy (Horton, 2007; MacKay & Vogh, 2012). Hence, we discuss how the private and public sectors are transforming, including the transformation of government and local government authorities and administration. The use of modern technologies in providing services by public administration bodies is described as e-services or e-government, synonymous with electronic administration, e-administration, and e-governance. Contacts between a given organisation and its service providers and recipients occur electronically, thanks to ICT.

The digitalisation of the economy produces more and more data. Data can, therefore, be determined as a new resource, a new production factor. Data collected by companies, especially on what are called platforms (platform economy), allow them to assess the demand, profile the advertising content and personalise the product. Digitalisation technology, i.e. everything that allows data to be collected, processed and analysed leads to new business models that affect user behaviour and business response, distorting companies' established operations in numerous sectors of the economy. The emergence of Airbnb led to a shock in the hotel market; the appearance of Uber revolutionised public transport. The platform economy expands to other sectors of the economy in the first place, where information is at the heart of the business model.

The transformation processes permeate the entire economy, gradually changing it into a digital economy founded on electronic data exchange. It is the crucial axis of a modern economy that also develops in virtual space. A digital economy results from technological progress, new means of communication, the accumulation of knowledge and data processing methods. It is a consequence of the development and convergence of data processing techniques (dynamic growth of computing power, data storage capacity), telecommunications (data transfer speed, advancement of protocols, infrastructure investments) and knowledge accumulation (algorithms, data science).

3. Data ecosystems – a strategic perspective

Technological changes in storing and processing the growing amount of data are a major source of innovation. New digital data ecosystems are created that can be understood as systems for generating processes, collecting and storing data resources, and continuous exchange between individual entities of this ecosystem participating in social and economic life. They are the effect of the possibilities that digital technologies jointly offer. Data enable ecosystem entities to make better decisions, resulting in higher performance, competitiveness and more efficient management. Here, the issue of entities entering the data ecosystem and their potential roles appears. We deal with an outburst of possibilities in this case as well. With advancing digital transformation processes, the generation and acquisition of data cease to be the domain of only a narrow group of specialised entities. These are not only IT enterprises anymore, because digital technologies are increasingly often used across almost all industries and beyond that, not only in businesses but also in the public sector, NGOs or simply among citizens who can passively share their data using modern devices and services, or actively participate in citizen-generated data projects.

Data is at the core of the digital economy. The increasing volume of the generated data and the number of actors involved lead to a growing need to undertake necessary governance-related policies to improve the functioning of data ecosystems. Firstly, it is important to facilitate the exchange of the collected data between individual entities of socio-economic life. The gathered data should be available to everyone regardless of the nature of the entities (public, NGOs or private), size or maturity. Thanks to this proceeding, the benefits that data bring will be maximised.

Decomposing the EU strategic perspective may be a helpful tool to contextualise further considerations in this article. The European data strategy aims to enable the EU to make better decisions and solve current political issues, such as resource and climate problems, leading to a data-agile economy driving overall innovation. In

February 2020, the European Commission published two documents key to the data-based economy: 'Shaping Europe's Digital Future' (European Commission [EC], 2020b) and 'A European Strategy for Data' (EC, 2020a). In the first one, three essential elements were indicated for the correct shaping of the digital future of Europe: technology benefits people, a fair and competitive economy, and an open, democratic and sustainable society. The document on the digital future of Europe contains essential elements of the EU's strategy for a data-based economy. It is an introduction to the content laid out in the second regulation. In the data strategy, the European Commission described the vision of European Data Spaces, a common, single data market on which they could be used regardless of the country of origin. It provides a list of the legal activities and investments that will be made over the next few years. The goal is to exploit data and demand for goods and services based on data. The strategy for data undertaken by the European Commission activities is based on four pillars. Firstly, there is a plan to create a legal framework for data management, tackling the availability of public sector data and actions used to exchange data efficiently within sectors should also be intensified. The second strategy pillar applies to the provision of support for technology development and digital infrastructure. The third pillar of the strategy involves investment in competencies and general skills to use data. The European Commission's activities aim to reduce the gap in extensive data acquisitions and data analysis capacities. The fourth pillar of the strategy for building a data-based economy concerns the development of a common European data space in strategic sectors and fields of public interest (e.g. relating to mobility, green governance, industrial data, energy, agriculture and health).

As indicated in the document, the possibilities of using data for innovative purposes are essential in the data-based economy. Data availability is a crucial problem. Even in the cases where one may expect an abundance of data, a lack of sharing mechanisms may severely hamper the innovation potential. Therefore, the data value is best assessed through the prism of its reuse. Hence, the need to unblock the flow channels is one of the critical elements of the European strategy. A typical data ecosystem for the development of collaboration in the field of data sharing may be decomposed into four main spheres – B2B (Business-to-Business), B2G (Business-to-Government), G2B (Government-to-Business) and finally, G2G (Government-to-Government).

In the case of B2B data sharing, the consensus is that it is still not dynamic enough and needs enhancement in the EU. In the coming years, it should be made more viable economically and, thus, more attractive thanks to the new policy initiatives. Fears against the loss of competitive advantage, the lack of mutual trust, concerns about data appropriation and exploitation by third parties are among the factors that

affect the current state. There are plans to run new policies and programmes to financially support this type of data sharing while offering guidelines to facilitate agreements guaranteeing equal negotiation positions of the parties and the security of the information provided.

Next, B2G data sharing is associated with the use of private data by public authorities. The European Commission's documents state that using private-sector data is insufficient. At the same time, the public sector can significantly improve the policy-shaping process and delivery of public services based on the evidence available through data. One prominent example of the above is the potential to considerably increase official statistics' scope, granularity and timeliness. Strengthening data sharing in this regard would accelerate the development of a data-based society and improve decision-making processes in the public sector. The European Commission recommends developing appropriate incentives to build data sharing. It proposes to analyse the legitimacy of introducing an EU framework regulating the reuse of private data in the public interest. There is a chance that the individual member states will initiate national programmes facilitating access to privately-held data by the public sector.

The primary assumption in the G2B model is to provide data to businesses by public administration. Data generated thanks to public funds should benefit the society to the broadest possible extent. It applies particularly to high-value data sets, including different forms of protected data, which are only sometimes made available for research due to the need for mechanisms consistent with the provisions on personal data protection.

A crucial role in building a data-based economy is G2G data sharing, i.e. data exchange between public authorities. In this respect, the preparation and implementation of the relevant regulations remain a competency of member states' administration at the national level. Cooperation on this platform will significantly improve the policy and provision of public services and reduce administrative burdens for market-operating enterprises.

The EU should become a place where data enable better decisions. However, this goal must be founded on a solid legal framework dealing with fundamental rights, data protection and security. Suppose the EU is to play a leading role in a data-based economy. In that case, it must take action now and in a coordinated manner to deal with access and data storage, computational and cyber security. In addition, the EU will have to improve its data processing structures and increase the pool of high-quality data available for reuse. Issues related to cyber security tackle many fields – public administration, financial institutions, international corporations, small and medium-sized companies, and individual users. A balance must be kept between

a vast flow and use of data and a high level of privacy, security and ethical standards to unleash the potential of data to the fullest.

4. EU policy actions

Undoubtedly, effective policies need the support of appropriate regulatory frameworks and governance. The primary goal is to create regulations that raise confidence in the sharing process and ensure the security of processed data. In 2014, the European Commission published the 'Towards a thriving data-driven economy' (EC, 2014b) communication, which relied on a coordinated action plan involving the member states and the EU. The European Commission proposed policy initiatives to address bottlenecks in the data 'reuse potential' by creating a common European data space. Consequently, the adopted 'Towards a common European data space' (EC, 2014a) European Commission communication proposed a package of measures pointing to the four primary modes of data sharing referred to above, i.e. B2B, B2G, G2B, and G2G. As one of the results, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, also known as the 'Open Data Directive', entered into force on 16 July 2019, providing a common legal framework for a European market for government-held data (public sector information). The Directive was to be transposed into national legislation by June 2021.

In its European Data Strategy put forward in February 2020 (EC, 2020a), the Commission declared subsequent initiatives: an implementing act on high-value data sets and two major legislative proposals: a governance framework for common European data spaces (Data Governance Act; EC, 2020d) and a Data Act. Implementing the regulation on high-value data sets will include a list of data with high commercial potential, speeding up the emergence of value-added EU-wide information products.

The new regulations are to become the basis for the European Data Policy. Many regulatory initiatives are in preparation (see below), but two are fundamental. The main policy implications of the EU data strategy are prepared as the Data Governance Act and the Data Act. The Data Governance Act provides an overarching governance framework for establishing and functioning common European data spaces, constituting the core part of the European Commission's broader data and digital strategies. It is designed to encourage investment in new infrastructure for sharing data, increasing data availability, strengthening the mechanisms for their sharing and ensuring the safety of processed data, helping to build a single digital market for data across EU member states. Legal solutions apply mainly to ensuring a high level of privacy and data security, which are subject to

sharing based on a new data management framework. The task of the Data Governance Act is primarily to increase market trust in the sharing process. The proposed provisions build the confidence of enterprises, government authorities and citizens to share information on the European data market.

The Data Governance Act has four pillars:

- granting access to public sector data for reuse in situations where these data may be copyrighted;
- sharing data between enterprises in exchange for remuneration in any form;
- enabling the use of personal data with the aid of an 'intermediary' helping individuals to follow Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation [GDPR]);
- enabling the use of data from altruistic motives.

The document contains information on how public and private sector data (usually unavailable due to intellectual property rights, trade secrets or privacy rights) could be made available using mechanisms guaranteeing anonymity and confidentiality under the applicable legal acts. These shall also include personal data protected by the General Regulation on Data Protection and within its legal standards. The Data Governance Act outlines how the data in possession of the public sector, charged with third-party rights (e.g. trade secrets and intellectual property rights) may be used. It contains provisions enabling data intermediaries (data brokers), defined as trustworthy actors, and sets the registration rules of entities that collect and provide data for charitable purposes. The new law aims to improve information protection and create a safe environment for obtaining, using and controlling data. The entire process of transferring and managing data is based on the neutrality and transparency of 'data intermediaries' responsible for data collection. These are to be trustworthy entities to which data are made available. Data intermediaries will be required to maintain neutrality and observance of the strict requirements, including the ban on using data for their interest. A certification or labelling framework has been proposed together with a notification obligation and then the monitoring of the compliance with the requirements for designated competent authorities in the member states. These proposals do not apply to data-sharing initiatives in closed groups.

In conclusion, the regulation is primarily to help build trust and raise process security, thus strengthening the practice of using data and creating innovative solutions. Government agencies should consider what data can be made available to guarantee their quality and what policy (and perhaps a price list) of their sharing should be adopted. The analysis of which data may be helpful and how to ensure

their safety and compliance with the requirements of other regulations will be necessary on the part of companies. In turn, potential data intermediaries should consider the relevant ideas for business and ways to facilitate data access and sharing. In each of the above, one has to be able to 'reign over data' to know what data one has and how they are processed and used. In practice, applying the proposed regulation will involve the inventory of the data held, classifying them according to the regulations they are subject to and ensuring their safety. The Data Governance Act also sets out a framework for the voluntary registration of entities that collect and process data made available for charitable or altruistic purposes.

The Act aims to enable data-driven innovation by setting a governance framework to promote confidence in data sharing and incentivise the expansion of EU data spaces while ensuring that natural persons and legal entities are in control of the data they generate. We are currently facing many forms of barriers to sharing data, like restrictive intellectual property rights, concerns about compliance with GDPR, fears of breaches of confidentiality, or fears of others deriving value from shared data when those who actually shared them were unable to do so. Data governance aims to promote data assets, going beyond some simple set of protection rules, but aiming more broadly at breaking down barriers to sharing. Thus, it has the potential to benefit businesses and the society more generally. The Data Governance Act establishes a framework to promote confidence in data exchange between organisations. It creates the basis for managing data spaces that comply with the values and laws of the EU, such as personal data protection, consumer protection and competition rules.

The last of the main proposed acts was planned to be adopted at the end of 2023. The Data Act will explicitly support B2B data sharing and B2G data sharing for public interest purposes, fostering access to data held by private sector entities when these data are of public interest. The European Commission hints that the right to data portability could be enhanced to give individuals more control over who can access and use their data; changes in the EU's intellectual property rights framework may be introduced, particularly in database rights and trade secrets.

5. The increasing complexity of data ecosystems

Europe's digital strategy also highlights many other legislative initiatives the European Commission plans to introduce. These include: laws on crypto assets and digital operational and cyber resilience in the financial sector; online platforms (Digital Services Act; EC, 2020c), data centres, cybersecurity, a review of the Network and Information Security Directive, EU's existing eIDAS Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on

electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC (digital identities), action plans on '5G' and '6G'; new digital sector inquiry; or new strategies for payments, the industry, blockchains and quantum computing. Add the Open Data Directive, supporting the use of public sector data collections, the Database Directive, and the well-known General Data Protection Regulation, and we will receive an increasingly complex EU legal system related to data. Naturally, this brings in substantial issues, some still unclear. For example, the Digital Governance Act provides an oversight of data sharing by competent authorities. However, there is a potential for uncertainty of responsibilities if providers are subject to regulatory oversight by several different authorities in the countries in which they are located.

The next issue to mention relates to cross-border transfers of data. According to the same Governance Act, non-personal data that is subject to the rights of others may be transferred from an EU country to a third country only if proper safeguards are in place. The Act should ensure the protection of the fundamental rights of data holders. At the same time, third countries providing some equivalent (to the EU) level of protection should be allowed to transfer data across boarders. How the European Commission is going to operationalise it needs to be made clear. One solution might be offering model contract clauses to gain reassurance that the non-personal data transferred outside the EU is protected. Next, the Digital Governance Act does not require data-sharing service providers to have an EU establishment, though the provider must appoint a legal representative in the EU.

Prospective data-sharing service providers with multiple establishments in the EU will be deemed to have their main establishment where their central unit is located. The GDPR, however, allows choosing the main establishment where the most important decisions about personal data processing are taken, not necessarily the central administration unit. Some other risk emerges in data altruism, specifically concerning forms proposed to gather consent from individuals to use their data, which is very broadly described in the regulation as consent to specific purposes or data processing in certain areas of research or parts of research projects when it is initially difficult to precisely identify the purpose at the time of data collection.

The above-mentioned issues naturally impact all modes of possible data sharing and may become extensively problematic in more complex cases like data collaboration and collaboratives. In these cases, the distributed nature of the data supply is matched with the distributed nature of demand for data. These can bring in a value-added insight that can potentially be generated only thanks to such initiatives. Naturally, the policy's intention is to instigate such collaboration in a more agile and instantaneous manner. The pandemic is clear evidence of such a need; it also clearly exposed a lack of preparedness for deep and swift data

collaboration. Data innovations may happen, yet many concerns still need to be addressed regarding incentives, limitations, obstacles, the lack of regulatory solutions or governance framework. New regulations are expected to fill this gap, but not without problems. However, the purpose of this article is not a detailed analysis of the potential problems, difficulties or issues requiring solutions. The intention is to demonstrate that the entities or persons who want to take advantage of the digital transformation and exponentially growing data ecosystems will need help with this increasingly complicated matter. Inevitably, a high degree of complexity can result in a significant burden, cost and lack of trained staff.

6. The role of the Data Steward

Although permeated with significant social and economic goals and undoubtedly necessary for the digital development of the EU, numerous legislative proposals create an increasingly complex system of relations, connections, rules and limitations. Undoubtedly, as often in such cases, attempts to solve particular problems create new ones. In such a complicated system of mutual dependencies and numerous regulations, high competencies are necessary to enable efficient navigation between individual elements of the system according to the prescribed set of recommendations and rules. Although the perspective of broader use of data by analysts, resulting from the adopted measures seems very attractive, the analysts themselves will need to constantly acquire new competencies and extensive contextual knowledge in such a complex arrangement.

The solution is already available in the form of Data Stewards, a concept that has long been developed in the business and scientific community (Peng, 2018). Undoubtedly, the role of specialised Data Stewards will become more significant in an increasingly complex and exponentially growing data ecosystem. At this point, however, we want to propose the following side theses: changes caused by the dynamic progress of digital transformation mean that the current group of specialists in this field is insufficient for today's needs and that the growing strategic, political and regulatory interest causes that the usual extent of knowledge, competence and skills of an exemplary Data Steward must be significantly developed. Let us agree that these theses are valid. It subsequently leads to the following conclusions: the need for a notable increase of Data Stewards supply while ensuring an appropriate profile of knowledge, skills and competencies. Thus, we put forward the central thesis that the most appropriate tool to respond to the formulated conclusions is to create a proper competency framework for Data Stewards tailored to a rapidly changing environment, conditioned by the ongoing digital transformation processes and extensive new legal regulations. It is the fundamental conclusion of this paper.

The growing volume of data and effective data management processes require a proper blend of technologies and people. Demand for digital competencies goes far beyond the extent and quality needed before, even just a few years ago (Organisation for Economic Co-operation and Development [OECD], 2016a, 2016b, 2018, 2020). The skills necessary to manage and supervise a whole data life-cycle, assess data value, keep adequate data quality, their efficient sharing and reuse are highly important. The expectations towards adequately trained staff are growing. New roles in the data ecosystems emerge, further extending the scope of training. One particular example supported by the regulations is that of an intermediary. Their task will ensure the availability, quality and reuse of the existing data. By organising pooling and sharing data, they will not process it on their account. However, experts (European Commission & Directorate-General for Research and Innovation, 2016) point out substantial deficits in people and skills. Good Data Stewards are exceedingly rare. Some claim that one Data Steward is needed for every twenty data analysts, so a quick estimate suggests that many Data Stewards will be needed over the coming years to fill in the gap (Versweyveld, 2016).

Data stewardship concerns those who work with, protect and use data. A Data Steward is at the heart of any data governance programme. In short, data stewardship aims to design, build, implement and manage data, enabling users to make consistent decisions based on information. Typically, the tasks of Data Stewards include:

- helping define and implement data definitions, shared metadata, standardised, controlled dictionaries and standards;
- setting data quality guidelines that face requirements for what is considered to be good data quality;
- consistent management of data resources throughout the entire life-cycle in order to preserve their quality, integrity and consistency, and avoid redundancies and integrity-related failures;
- facilitating the reuse of and providing access to data resources (for internal or external purposes);
- maintaining high-quality metadata;
- cooperating with others engaged in creating, collecting, accessing, using, sharing and maintaining data;
- collecting information on the needs and feedback on the quality of data resources for which they are responsible, ensuring that data is fit for the purpose;
- being aware of the data protection policies, intellectual property and information security; ensuring data is protected and security procedures are enforced;
- organising and contributing to communication and promotional activities to increase awareness and use of data resources for which they are responsible;

- building, supporting and sharing knowledge; helping users understand the data better and recommending improvements.

Undoubtedly, the above list does not fully exhaust the scope of activities of Data Stewards; depending on the context, not all of these tasks need to be performed or can only be performed to a certain extent. Nevertheless, such a synthetic list of essential items constitutes a complex set of competencies, requiring years of training and gathering experience. While carrying out his or her activities, a Data Steward must not only demonstrate knowledge and deep understanding of the issues that remain in a strong relationship with his or her area of competence, but also of those going beyond it. For example, when assisting data analysts, a Data Steward should have some experience in carrying out such research, which can help to efficiently prepare an appropriate quantitative analysis development environment, allowing analysts to concentrate on the task at hand. The evolution of the role of the Data Steward requires an appropriate skill set. Data Steward duties cover all aspects of data governance; they must understand all levels of the business and are expected to demonstrate cooperation skills and the ability to collaborate and effectively communicate with other internal and external stakeholders, both in the language of business processes and technology. Data Stewards must be able to build relationships with others, promote best practices and demonstrate the proper use of data, following appropriate guidelines, rules and regulations, whether internal or external (Sen, 2018). Data stewardship aims to use information as an asset by defining strategies, standards, policies, models, processes, tools and methodologies to identify opportunities. Undoubtedly, the proper set of competencies is challenging to achieve.

7. Competency framework for Data Stewards

Technological progress inevitably raises the issue of digital competencies. The discussion is not easy; the first obstacle is the very concept of competencies. They are interdisciplinary and broad, referring primarily to the practical ability to use modern technologies' wealth of tools and methods. Discussions on digital competencies can be found mainly in national and international strategic documents and scientific literature. The variety of definitions and approaches is undoubtedly associated with the dynamic evolution of technology and its proliferation in the economy and society. The concept of digital competencies is related to technology-oriented skills. Initially, digital skills were associated predominantly with access to equipment and the Internet and extended to more complex usage skills to achieve various life goals. 'Digital competencies' may be defined as a harmonious set of knowledge, skills and attitudes that allow the effective use of digital technologies in various areas of life

(Erstad, 2010). The concept of digital competencies covers an extensive set of skills that determine the efficient and conscious use of new technologies and the active participation in the life of the information society.

In this article, we argue for the need to eliminate the deep deficit of Data Stewards. A natural proposition to solve this problem is to invest in training appropriate personnel. A quick review of the existing literature in this area and observing what is happening in the education and labour markets indicate two crucial aspects to consider when planning further activities. Firstly, the role of Data Stewards has long been present in business, and there are also career paths that prepare for this role. Secondly, the analysis of the existing education programmes indicates a fundamental need for more opportunities at the university level.

The first of the indicated problems may mean that in the dynamic progress of the digital transformation, accompanied by an active response in the legislative sphere, we will face a deficit of specialists in the government and non-governmental sectors. The second problem means that there is still no consensus as to what competencies should characterise a Data Steward, at least one with a general profile of competencies, ready to assume the duties, regardless of the specificity and context of the tasks performed (after all, it is already possible to indicate a few potential specialisations for Data Stewards).

Therefore, we propose to develop an appropriate framework for the competencies of Data Stewards. A basic set of ready-made requirements would undoubtedly increase the awareness of the concept of Data Stewards, facilitating the preparation of appropriate fields of study, vocational training and training materials. There is no doubt that the expectations of the labour market in this respect will increase. Employers wish to know the potential employees' knowledge, skills and personal and social competencies. In turn, the candidates want to be aware of the content and level of learning outcomes, within and outside higher education, in line with lifelong learning. Competency frameworks also balance the different paths to achieve the intended qualifications (Punie et al., 2013). When developing the desired competencies of Data Stewards, it is possible to use the existing schemas based on the standard defined in the European Qualifications Framework (EQF) – the structure of qualifications levels adopted in the EU, constituting a reference system of the national qualifications framework, enabling the comparison of qualifications obtained in different countries. The qualifications framework defines the qualifications obtained in the education system and their mutual relations, defining e.g. learning pathways, making it easier to compare qualifications acquired at different times, places and forms. The qualifications framework aims to adapt competencies to the needs of the labour market and increase employee mobility, promoting and facilitating lifelong learning.

The primary element of the framework is the learning outcomes, defining what the learner knows, understands and can do after completing the learning process. In other words, it is a language for the description of competencies to be understood by all stakeholders in the learning process. The scheme for describing the levels of qualifications is based on identifying three categories of learning outcomes: knowledge, skills and competencies, according to which specific characteristics of learning outcomes are composed. Competencies in this tripartite are understood as the proven ability to apply knowledge, skills, personal, social or methodological abilities demonstrated at work or study and professional and personal careers. In the EQF, competencies are defined in terms of responsibility and autonomy (in the Polish competence framework, the term 'social competencies' is used). The first category describes the expected range of theoretical and factual knowledge. Skills are described as cognitive (involving logical, intuitive and creative thinking) and practical (relating to manual dexterity and using methods, materials, tools and instruments). Responsibility and autonomy are described as the learner's ability to apply knowledge and skills independently and responsibly (Europass, n.d.).

A detailed elaboration of the possible variants of the competency framework is quite a significant endeavour. Here, we limit ourselves to indicating essential areas and issues to consider when developing such a framework. A Data Steward, in a nutshell, is someone who understands the value of data as a resource, who knows about the problems of information systems in the organisation, who wants to question the status quo and introduce changes, who has the necessary communication skills, and is a person with high credibility, above all thanks to the unique competencies in the area of data governance.

The scope of the expected competencies can be divided into three main groups: technical competencies, non-technical (business) competencies and soft competencies. The scope of technical competencies should primarily cover the following areas: knowledge management, data processing, data archiving, infrastructure and network, and services. This list can be further elaborated. For example, these should include a good understanding of data and information concepts, data modelling skills (conceptual, logical and physical), knowledge of the organisation's data management technologies and systems, and the related tools. Non-technical skills refer to strategic planning and change management, project management, compliance and legal environment, and a deep understanding of the organisation's foundations, functions, goals and environment. While often underestimated, soft traits and skills are particularly important in the case of Data Steward. Here, one can indicate creativity, independence, teamwork, listening skills, openness and diplomacy, objectivity, communication and persuasion skills, networking skills,

patience, calmness, and composure. To sum up, the competence of a Data Steward is about the careful and responsible management of the data ecosystem.

8. Conclusions

Currently, businesses are confronted with a fairly complex interplay between the existing and forthcoming EU data laws. From a business perspective, practical guidance is needed to help enterprises navigate through this complexity and meet their legal obligations while harnessing the power of data. A recognised Data Steward function should be promoted in both the public and private sectors, along with developing digital skills and capacity building to ultimately create a data-sharing culture based on the principle of reciprocity (EC, 2020e). Trained Data Stewards are needed to increase capacity, both on the supply and demand side of data, to democratise access to the data and facilitate the growth of data-driven solutions within functional data ecosystems. Data Stewards are at the heart of data governance efforts.

Technology and regulation can only solve some problems of the lack of adequate human resources, expertise or poorly defined functions. Data Stewards must take on diverse roles, tasks and responsibilities, aligning data processes and applications in developing, and enforcing data governance in compliance with regulations and data ethics. Implemented and planned legislation requires support through designing education programmes in the field of new roles necessary for the efficient functioning of data ecosystems, particularly the role of a Data Steward. Unfortunately, the current state of data stewardship education and training, especially in the public sector, is in its infancy. Future education in data stewardship should be built on a solid collaboration between the private and public sectors and the academia to ensure relevant competencies in documenting, curating and structuring data across public and private organisations, facilitating data sharing and use in an increasingly regulated environment.

We argue that democratic countries with highly effective state authorities should conduct an active policy on creating the state's information order and provide citizens and business entities with adequate information as a public good, regardless of whether the information is needed. Thanks to this, information also reaches these citizens or entities who do not realise what information they need. In this way, it effectively minimises a social information gap and increases social knowledge resources (Oleński, 2000). The right policy should focus on developing infrastructure and the knowledge, skills and attitudes necessary for the smooth functioning of data ecosystems or (more generally) public information governance. Skills and habits for using information and stock of derived knowledge which

a citizen, an enterprise or the public administration can use constitute an essential aspect of the country's socio-economic development. With the exponential increase in the potentialities of generating, collecting and processing information in all areas of socio-economic life, this puzzle's critical element becomes the Data Steward's role.

References

Bell, D. (1973). *The Coming of Post-Industrial Society*. Basic Books.

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (OJ EU L 172, 26.6.2019). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024.

Erstad, O. (2010). Educating the Digital Generation. *Nordic Journal of Digital Literacy*, *5*(1), 56–71. https://doi.org/10.18261/ISSN1891-943X-2010-01-05.

Europass. (n.d.). *Description of the eight EQF levels*. Retrieved April 3, 2021, from https://europa.eu/europass/en/description-eight-eqf-levels.

European Commission. (2014a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Towards a common European data space*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0232.

European Commission. (2014b). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Towards a thriving data-driven economy*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52014DC0442.

European Commission. (2020a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European strategy for data*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066.

European Commission. (2020b). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Shaping Europe's digital future*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0067.

European Commission. (2020c). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825.

European Commission. (2020d). *Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=COM:2020:767:FIN.

European Commission. (2020e). *Towards a European strategy on business-to-government data sharing for the public interest. Final report prepared by the High Level Expert Group on Business to Government Data Sharing*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/731415.

European Commission & Directorate-General for Research and Innovation. (2016). *Management Plan 2016*. https://commission.europa.eu/system/files/2016-05/management-plan-2016-dg-rtd-march2016_en.pdf.

Horton, F. W. (2007). *Understanding information literacy: a primer*. United Nations Educational, Scientific and Cultural Organization. https://www.ifla.org/publications/understanding-information-literacy-a-primer/.

MacKay, K., & Vogt, Ch. (2012). Information Technology in Everyday and Vacation Contexts. *Annals of Tourism Research*, *39*(3), 1380–1401. https://doi.org/10.1016/j.annals.2012.02.001.

Oleński, J. (2000). *Elementy ekonomiki informacji. Podstawy ekonomiczne informatyki gospodarczej*. Uniwersytet Warszawski.

Organisation for Economic Co-operation and Development. (2016a). *Skills Matter. Further Results from the Survey of Adult Skills*. OECD Publishing. http://dx.doi.org/10.1787/9789264258051-en.

Organisation for Economic Co-operation and Development. (2016b). *G20/OECD INFE Core Competencies Framework on Financial Literacy for Adults*. https://www.oecd.org/finance/Core-Competencies-Framework-Adults.pdf.

Organisation for Economic Co-operation and Development. (2018). *Financial Markets, Insurance and Pensions. Digitalisation and Finance*. https://www.oecd.org/finance/Financial-markets-insurance-pensions-digitalisation-and-finance.pdf.

Organisation for Economic Co-operation and Development. (2020). *OECD Digital Economy Outlook 2020*. OECD Publishing. https://doi.org/10.1787/bb167041-en.

Peng, G. (2018). The State of Assessing Data Stewardship Maturity – An Overview. *Data Science Journal*, *17*, 1–12. http://doi.org/10.5334/dsj-2018-007.

Punie, Y., Brečko, B., & Ferrari, A. (2013). *DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe*. Publications Office of the European Union. https://publications.jrc.ec.europa.eu/repository/handle/JRC83167.

Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC (OJ EU L 257, 28.8.2014).

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ EU L 119, 4.5.2016).

Sen, H. (2018). *Data Governance. Perspectives and Practices* (1st edition). Technics Publications.

Versweyveld, L. (2016). *We need 500,000 respected data stewards to operate the European Open Science Cloud*. http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud.

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 83–92 https://doi.org/10.59139/ws.2023.12.5
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

Improving research on environmental noise pollution and its impact on the population in the context of sustainable development

Doskonalenie badań nad zanieczyszczaniem środowiska hałasem i jego oddziaływaniem na ludność w kontekście zrównoważonego rozwoju

1. Introduction

In the early decades of the 21st century, environmental statistics became one of the main pillars of information resources which describe the world we live in, prepared by official statistics services for individual, corporate and institutional users. Data sources are also being enhanced, which not only widens the scope of the presented information and statistical analyses, but also reduces the labour intensity experienced by producers of official statistics. One of these extensions involves the study of population exposure to noise in cities with more than 100,000 inhabitants in Poland, planned to be carried out in 2024 by Statistics Poland. The undertaking evolved from an experimental project into a target solution. Strategic noise maps prepared by cities with a population exceeding 100,000 constitute the source of data, as does the number of people living in a specific building to which a spatial location can be assigned based on data collected in the 2021 Census. These two administrative records enable the preparation, conduct and publication of survey results without the additional involvement of respondents, which is consistent with the direction of development that official statistics is set to follow (Allin, 2021).

In the context of the study of the population's exposure to noise, the direction of development of official statistics related to the spatial reference of the observed phenomena is significant, as it enables the recorded measures to be combined with the precise geographical location of the surveyed units, e.g. the population. The hitherto method of referring the measures to spatial location has been implemented by apportioning the population among the administrative divisions. The current state of research in the field of census taking and noise assessment offers the possibility of combining them and relating them to a wide range of population, ensuring precise mutual reference. This article explores the new directions of development in this area. Its purpose is to discuss the process of improving statistical research on population exposure to noise pollution in the context of sustainable development with the application of new trends in data acquisition.

2. The progress of research on noise pollution in the EU

The discourse on the need to harmonise various activities aiming to reduce population exposure to noise could be traced back to as early as 1993 in an official document of the European Union, titled *Towards sustainability: A European Community programme of policy and action in relation to the environment and sustainable development*). 'To maintain the overall quality of life' was one of the five actions described in the environmental programme, which included a commitment to measures reducing population exposure to noise. It was formulated as follows: 'No person should be exposed to noise level which endangers health and quality of life' (*Towards sustainability*, p. 56). It can be therefore clearly stated that the documents defining the concept of sustainable development included a clear goal: to reduce noise pollution, which poses a threat to health and quality of life until its complete elimination.

A comprehensive programme for limiting the exposure of the population to excessive noise emissions in the EU was presented in a follow-up document, published in 1996, referring entirely to the issues of noise pollution (Commission of the European Communities, 1996). It quoted the results of the conducted surveys which indicated that 20% of the population of Western Europe experienced negative effects of noise (Commission of the European Communities, 1996, p. 1). The standardisation of the assessment of noise levels relating to its classification, methods of measurement and territorial distribution was soon introduced, as was the obligation of the member states to take clearly defined measures to reduce the level of noise affecting the population. Based on this document, the Environmental Noise Directive (END; Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise, further referred to as Directive 2002/49/EC) was prepared and adopted in 2002. The following expectations were formulated: 'Achieving a high level of health and environmental protection is part of the Community policy, and protection against noise is among the adopted goals'. In addition to indicating the need to take measures to minimise the impact of noise, the END also imposed an obligation on member states to prepare strategic noise maps, also referred to as acoustic maps, every five years. By 2022, four editions were prepared: in 2007, 2012, 2017 and 2021. Progress in the measurement methodology can be seen in each successive edition, eliminating any emerging discrepancies in the interpretation of a complex and self-perceived phenomenon. During these two decades, the mandatory territorial scope of strategic noise maps was extended. The minimum population ceiling for towns was lowered from 250,000 to 100,000, thus increasing the population covered by the study and (i) for major roads, the level of over

3 million vehicle passages per year was defined, (ii) for railway lines above 300,000 train passages per year, and (iii) for major airports over 50,000 take-offs and landings per year (Directive 2002/49/EC). The methods of noise level assessment were successively clarified in the subsequent editions of the END. In 2015, a directive was passed specifying the method of measuring and preparing strategic noise maps (Commission Directive (EU) 2015/996 of 19 May 2015 establishing common noise assessment methods according to Directive 2002/49/EC of the European Parliament and of the Council, further referred to as Commission Directive (EU) 2015/996). The document consists of over 800 pages, which is indicative of the complexity of the subject matter and the difficulties in coordinating activities enabling the harmonisation and comparability of the presented results. Moreover, this complexity is multidimensional, where, apart from the technical area related to the measurement methods, the estimation of the number of people exposed to noise is the problematic issue. This topic will be discussed in the next chapter.

3. Contribution of official statistics to research on noise pollution

The key issue in preparing strategic noise maps is the identification of the population affected by excessive noise emissions. In the directives (Directive 2002/49/EC; Commission Directive (EU) 2015/996; Commission Directive (EU) 2020/367 of 4 March 2020 amending Annex III to Directive 2002/49/EC of the European Parliament and of the Council as regards the establishment of assessment methods for harmful effects of environmental noise, further referred to as Commission Directive (EU) 2020/367) and guidelines relating to the methods of developing noise maps (Główny Inspektorat Ochrony Środowiska [GIOŚ], 2021), the focus is laid on acoustic issues that enable the determination of accurate immission maps of road, rail and industrial noise and showing the spatial distribution of pollution in the studied area. However, a reasonable confidence estimation of the population exposed to precisely defined noise levels requires an equally precise allocation of the number of people living in dwellings subject to spatial analysis. This task was entrusted to local governments obliged under the END regulation to prepare strategic noise maps. However, they do not have datasets to accurately determine the number of people living in particular buildings. The information available to local governments on registered residents, based on which the number of residents is estimated, or resulting from detailed records (e.g. for the purposes of waste management) shows significant biases.

The solution proposed in the recommendations (GIOŚ, 2021, pp. 208–210) refers to the information published by Statistics Poland, which in annual cycles provides

the size and structure of the population in territorial breakdown, at the level of gminas (Polish equivalent to communes – NUTS 5). This number should be then distributed in proportion to the number of residential units, considering single-unit, two-unit and other buildings. The estimates prepared in this way are used to determine the number of people exposed to noise in buildings using one of the two methods specified in the END. The first one considers the location of the building together with the directions of exposure of the façade towards the source of the noise, while the other one assigns the population to an address point located within the building.

The above-mentioned methods of estimating the number of people living in specific buildings reveal significant limitations which should be considered when assessing the phenomenon of noise pollution.

To reduce the identified barriers, it becomes necessary to include information provided by official statistics in the process of estimating the number of people exposed to noise according to the location of buildings. The new methods of the 2021 population and housing census, taking into account the location of the residential building, make meeting the requirements for publishing data in a 1 km^2 grid possible, and help develop strategic noise maps with greater accuracy. Another advantageous phenomenon is the synchronisation of the dates of the preparation of both datasets, i.e. five-year periods for the preparation of strategic noise maps (e.g. 2021, 2026, 2031...), with ten-year deadlines for the implementation of censuses (e.g. 2021, 2031...). Acoustic maps can be therefore fed with population data from the most up-to-date census data.

The set of 244 indicators illustrating the 17 Sustainable Development Goals (SDGs) does not directly refer to noise reduction, which can be recognised as a shortcoming. The need to eliminate this type of pollution is clearly indicated in the documents constituting the sustainable development initiative (Towards sustainability), which encourages a precise reference to it in the implementation process. References to noise pollution can also be found in research works presenting composite indicators of sustainable development (Lafortune et al., 2022; Sachs et al., 2022), but only as a supplement to the categories of pollution listed in the objectives (3.9, 6.3, 14.1), i.e. air and water, or housing conditions, e.g. urban sprawl and overcrowded settlements (Lafortune et al., 2022, pp. 53, 96). The lack of noise indicators in UN documents can be explained by the complexity of the issue of measuring this phenomenon and the difficulties in implementing them on a global scale.

However, the services of official statistics have taken steps to fill the existing gap. As part of the project entitled *On the 2030 Agenda and SDGs*, they present an

indicator called *Population living in households considering that they suffer from noise*. This indicator shows that the share of people experiencing inconvenience due to this reason was constantly decreasing: from 16.2% in 2010 to 12.5% in 2019 (Główny Urząd Statystyczny, 2022). The source of these data is a survey coordinated within the EU called EU-SILC (European Union statistics on income and living conditions; Eurostat, n.d.). The selection of the research sample is representative, so we can legitimately apply its results to the entire population. It should be noted here that they reflect the subjective perception of the surveyed community whose representatives are selected irrespectively of the noise indicators. The scope of this study covers many social areas defined by the European Pillar of Social Rights initiative. Therefore, the postulate of transposing the subjective assessment of noise pollution to the entire population becomes of significant importance in this study. It is also worth adding that strategic noise maps encompass only 27.8% of the population of Poland covered by the fifth edition of the survey carried out in cities with more than 100,000 inhabitants in 2021.

The next chapter discusses the prospects of integrating the trends in the development of official statistics and research on noise pollution, thus enabling a wider exploration of its impact on the population from the health, economic and environmental perspectives.

4. Research project on exposure of the population to noise in cities exceeding 100,000 inhabitants, carried out by Statistics Poland

The guiding idea behind the research project entitled *Exposure of the population to noise in cities of more than 100,000 inhabitants*, launched by Statistics Poland was to extend the scope of reliable statistics describing the impact of selected pollutants on humans using sources of data that do not impose much burden on the respondents. The aim was to indicate the number of people exposed to selected categories of noise in cities with more than 100,000 inhabitants in a uniform and comparable for the country way. The study combined strategic noise maps based on a well-established methodology, presenting the excess levels of permissible road, tram, railway, aviation and industrial noise with reliable data on the number of people living in individual buildings with assigned geographical locations (X, Y).

The study was carried out in accordance with the assumptions formulated for research projects: from experimental statistics to statistical production. In 2020, an experimental project was launched in cooperation with three cities – Gdańsk, Gdynia and Bydgoszcz, which provided a strategic noise map from the 2017 edition for analytical purposes. These maps were compared with the results of an

experimental project entitled *Development of a method for estimating the size and structure of the population* according to the actual place of stay, considering the criterion of staying and absence of 12 months or more according to the territorial division and a 1 km^2 grid. It identified sources of data that could be used in estimating the population size in relation to a 1 km^2 grid in accordance with the requirements set out by the European Commission (Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km^2 grid, further referred to as Commission Implementing Regulation (EU) 2018/1799). The results of the conducted analyses were presented to the Methodological Commission of Statistics Poland, who acknowledged the need to include the proposed experimental study in statistical production starting from 2024. The administrative collections will be used as the source material. Qualitatively verified acoustic maps prepared as of 31 December 2021 will be provided to Chief Inspectorate of Environmental Protection by local governments and infrastructure operators obliged to prepare strategic noise maps. The database of address points with assigned geographic coordinates (X, Y) and the estimated population, calculated using the 2021 census data, will be obtained from official statistics resources. Such a list will be used to prepare many comparative analyses in a unified and coherent manner. Breakdowns by noise pollution category and by territory will be available.

The experience gained will be subject to further research in this area and new solutions will be proposed or existing ones will be strengthened with the use of new data sources, which is considered in the next chapter.

5. Further research development

The integration of two datasets, i.e. the size and structure of the population related to address points with acoustic measurements from strategic noise maps, conducted by official statistics services, creates a synergistic potential for obtaining additional results otherwise unattainable from separate surveys. It should be emphasised here that the current state of their maturity resulted from the development of methodological works as well as from technological progress. The methodological development of strategic noise maps, manifested by a precise description of noise intensity assessment methods (Commission Directive (EU) 2015/996) enables the comparison of the results provided by different producers of noise maps. Statistics Poland, following the EU requirement to map the population in a 1 km^2 grid in the 2021 census (Commission Implementing Regulation (EU) 2018/1799), used spatially

located references to buildings. These two trends combined into one research bundle create new cognitive opportunities absolutely necessary to reduce environmental noise pollution. This issue will be dicussed from an environmental, health and economic perspective.

The environmental perspective of the development of noise research refers to two matters: the precise addressing of the studied phenomena, in which population estimates can be related to the spatial location of residence with greater accuracy, and to the expansion of the scope of research. Local governments in cities with more than 100,000 inhabitants used official statistics estimates showing the number of people assigned to a broad administrative division. Moreover, the administrators of the main roads, railway lines and airports, who are obliged to prepare strategic noise maps, encounter difficulties in preparing precise estimates for the number of people exposed to noise generated by the respective means of transport. The new formula of integrated research overcomes these limitations. The precise assessment of the effectiveness and efficiency of the measures taken to reduce the impact of noise is also an interesting research topic. Another area of application of the new indicators may be the unification of the method of measuring noise pollution in terms of two sustainable development goals: 3 – *Ensure healthy lives and promote well-being for all at all ages* and 11 – *Make cities and human settlements inclusive, safe, resilient, and sustainable*.

The health perspective in the development of research on noise is related to the possibility of launching many studies examining the impact of noise on the health of a population throughout the whole country. The initiated studies pertaining to selected locations where data on the degree of noise intensity were available (Argys et al., 2020) can be extended to a wider sample of the population. An opportunity arises to precisely study the incidences of the already recognised disorders (Basner et al., 2014), such as ischemic heart disease (IDH), high annoyance (HA) or significant sleep disorders, and the resulting consequences. Discovering correlations between morbidity and residential locations exposed to nuisance noise of various categories (road, rail, tram, air and industrial) will also allow the identification of the potential hidden associations between noise and morbidity (Hammer et al., 2014).

The economic perspective of the development of noise research is also promising. The above-mentioned possibilities of carrying out research on the effectiveness of the preventive measures aimed at minimising or eliminating the impact of noise on the environment make it possible to select the most effective ones. Another aspect of economics-based research is the attempt to determine the costs of the effects of noise pollution and compare them with the measure of the value of goods and services produced by a specific community of a country or region, i.e. with the Gross

Domestic Product (GDP). Such attempts have already been made and certain approximations were even quoted (King et al., 2011, p. 756). However, after reaching the indicated sources (Cvetković & Praščević, 2006, p. 22), it turns out that the calculation of the share of 'the costs caused by noise pollution' in GDP was difficult to find. This issue is cognitively interesting and important for determining further directions for the undertaken activities. Nevertheless, it requires data collection from numerous sources, which should be available to official statistics. Based on such calculations, an attempt may be made to approximate the health costs, the value reduction of real estate due to the exposure of residents to noise, or the expenses incurred to reduce the effects of noise. Thus, the analysis of the economic aspect of noise pollution may become an important direction in further research.

6. Conclusions

Research on environmental noise pollution has a long tradition. The knowledge about its negative impact on the health and the environmental and economic aspects of human life is constantly being extended. The subject of noise reduction has even become a component of the framework constituting the Sustainable Development Goals (*Towards sustainability*). However, a question arises whether the previously expressed expectations regarding environmental noise have been forgotten or ignored (King & Murphy, 2016). Such conclusions may be prompted by the lack of a reference to noise pollution on the list of indicators of the United Nations Sustainable Development Goals (United Nations, n.d.). In the European circle, on the other hand, the consistent development of this study is visible, manifested by the adoption of the END and the preparation of five editions of strategic noise maps for selected areas. Each of them was accompanied by an extension of the territorial scope, as well as a specification of the noise measurement methodology (Commission Directive (EU) 2015/996). The last edition was supplemented with the requirement to prepare an assessment of the harmful effects of environmental noise (Commission Directive (EU) 2020/367). The European experience can therefore be used to include the aspect of the area affected by noise pollution in the existing sustainable development goals and indicators.

However, the European way is characterised by serious barriers associated with the costly process of the cyclical development of strategic noise maps, which may prove unbearable for many countries. Official statistics presented a medium option, applicable to those countries that could not initially undertake the task of a unified method of noise measurement for key urban (cities) and linear (roads, railways) spaces. It is a sample survey where one of the research areas would involve the

subjective perception of noise. Official statistics studies also offer tools for the further exploration of noise impact by using sets of the number of people living in specific buildings with an assigned XY spatial location. In the area of health, there is research capacity for population studies on a representative sample of people exposed to specific categories of noise and its intensity, compared to a control group. In the environmental area, an interesting issue would be a comparative analysis of the activities undertaken by individual local government units to eliminate noise pollution. In the economic field of study, there are many research directions relating to the effectiveness and efficiency of the implemented protection from noise. All these threads should support the system of monitoring the achievement of sustainable development goals.

References

Allin, P. (2021). Opportunities and challenges for official statistics in a digital society. *Contemporary Social Science*, *16*(2), 156–169. https://doi.org/10.1080/21582041.2019.1687931.

Argys, L. M., Averett, S. L., & Yang, M. (2020). Residential noise exposure and health: Evidence from aviation noise and birth outcomes. *Journal of Environmental Economics and Management*, *103*. https://doi.org/10.1016/j.jeem.2020.102343.

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., & Stansfeld, S. (2014). Auditory and non-auditory effects of noise on health. *The Lancet*, *383*(9925), 1325–1332. https://doi.org/10.1016/S0140-6736(13)61613-X.

Commission Directive (EU) 2015/996 of 19 May 2015 establishing common noise assessment methods according to Directive 2002/49/EC of the European Parliament and of the Council (O.J. EU L 168 2015.7.1).

Commission Directive (EU) 2020/367 of 4 March 2020 amending Annex III to Directive 2002/49/EC of the European Parliament and of the Council as regards the establishment of assessment methods for harmful effects of environmental noise (Text with EEA relevance) (O.J. EU L 67 2020.3.5).

Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km^2 grid (Text with EEA relevance) (O.J. EU L 296 2018.11.22).

Commission of the European Communities. (1996). *Future noise policy: European Commission Green Paper*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:51996DC0540&from=PT.

Cvetković, D., & Praščević, M. (2006). Strategic directions in implementation of environmental noise directive in international and national legislation. *Facta universitatis. Series: Physics, Chemistry and Technology*, *4*(1), 21–34. http://facta.junis.ni.ac.rs/phat/pcat2006/pcat2006-03.pdf.

Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise (O.J. EU L 189 2002.7.18).

European Commission. (2000a). *Position Paper on EU Noise Indicators*. https://op.europa.eu/en/publication-detail/-/publication/10d75ba4-7279-4df2-aa50-3ed7fdf656a8.

European Commission. (2000b). *The Noise Policy of the European Union*. https://www.moa.gov.cy/moa/environment/environmentnew.nsf/69D658BCCDE2FDB5C2258041002D61F5/$file/%CE%9Doise%20brochure.pdf.

Eurostat. (n.d.). *EU statistics on income and living conditions*. https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Główny Inspektorat Ochrony Środowiska. (2021). *Dobre praktyki wykonywania strategicznych map hałasu. Wytyczne Głównego Inspektora Ochrony Środowiska*. https://www.gov.pl/web/gios/opracowania.

Główny Urząd Statystyczny. (2022). *Goal 11 – Sustainable cities & communities. Indicator 11.1.a – Noise from neighbours or from the street*. https://sdg.gov.pl/en/statistics_nat/11-1-a/.

Hammer, M. S., Swinburn, T. K., & Neitzel, R. L. (2014). Environmental Noise Pollution in the United States: Developing an Effective Public Health Response. *Environmental Health Perspectives*, *122*(2), 115–119. https://doi.org/10.1289/ehp.1307272.

King, E. A., & Murphy, E. (2016). Environmental noise – 'Forgotten' or 'Ignored' pollutant?. *Applied Acoustics*, *112*, 211–215. https://doi.org/10.1016/j.apacoust.2016.05.023.

King, E. A., Murphy, E., & Rice, H. J. (2011). Implementation of the EU environmental noise directive: Lessons from the first phase of strategic noise mapping and action planning in Ireland. *Journal of environmental management*, *92*(3), 756–764. https://doi.org/10.1016/j.jenvman.2010.10.034.

Lafortune, G., Fuller, G., Bermont Diaz, L., Kloke-Lesch, A., Koundouri, P., & Riccaboni, A. (2022). *Europe Sustainable Development Report 2022. Achieving the SDGs: Europe's Compass in a Multipolar World*. Cambridge University Press. https://s3.amazonaws.com/sustainabledevelopment.report/2022/europe-sustainable-development-report-2022.pdf.

Sachs, J. D., Lafortune, G., Kroll, C., Fuller, G., & Woelm, F. (2022). *Sustainable Development Report 2022*. Cambridge University Press. https://doi.org/10.1017/9781009210058.

Towards sustainability: A European Community programme of policy and action in relation to the environment and sustainable development (O.J. EU C 138 1993.5.17).

United Nations. (n.d.). *SDG Indicators. Metadata repository*. https://unstats.un.org/sdgs/metadata/.

World Health Organization. (2018). *Environmental noise guidelines for the European region*. https://iris.who.int/bitstream/handle/10665/279952/9789289053563-eng.pdf?sequence=1.

Jerzy Auksztol
Uniwersytet Gdański, Wydział Zarządzania; Urząd Statystyczny w Gdańsku; Główny Urząd Statystyczny, Polska / University of Gdańsk, Faculty of Management; Statistical Office in Gdańsk; Statistics Poland, Poland

Wiadomości Statystyczne. The Polish Statistician, 2023, *68*(12), 93–94
Dyskusje. Recenzje. Informacje / Discussions. Reviews. Information

WYDAWNICTWA GUS. LISTOPAD 2023
PUBLICATIONS OF STATISTICS POLAND. NOVEMBER 2023

W ofercie wydawniczej Głównego Urzędu Statystycznego z ubiegłego miesiąca warto zwrócić uwagę na następującą publikację:
Among Statistics Poland's publications from the previous month, we would like to recommend:

Narodowy Spis Powszechny Ludności i Mieszkań 2021. Ludność. Stan i struktura demograficzno-społeczna w świetle wyników NSP 2021
National Population and Housing Census 2021. Population. Size and demographic-social structure in the light of the 2021 Census results

Jedno z opracowań tematycznych przedstawiających wyniki Narodowego Spisu Powszechnego Ludności i Mieszkań 2021.

Język: polski, angielski
Language: Polish, English
Seria: Spisy powszechne
Series: Censuses
Dostępne wersje: drukowana i elektroniczna z tablicami w formacie Excel
Available in: printed and electronic form with Excel tables

W publikacji podano informacje dotyczące stanu i struktury demograficznej ludności (według płci i wieku oraz według stanu cywilnego), rozszerzone o takie elementy charakterystyki demograficzno-społecznej, jak: kraj urodzenia, obywatelstwo, przynależność narodowo-etniczna i wyznaniowa, wykształcenie, migracje zagraniczne na pobyt czasowy i migracje wewnętrzne. Zawarto w niej także dane o osobach z niepełnosprawnościami, osobach przebywających w obiektach zbiorowego zakwaterowania oraz osobach bezdomnych.

Część analityczną publikacji wzbogacono o mapy i wykresy. Część tabelaryczna składa się z tablic przeglądowych zawierających porównanie wyników spisów z lat 2011 i 2021 oraz tablic wynikowych z danymi pochodzącymi ze spisu przeprowadzonego w 2021 r.

W listopadzie br. ukazały się ponadto:
- „Biuletyn statystyczny" nr 10/2023;
- *Ceny robót budowlano-montażowych i obiektów budowlanych (wrzesień 2023 r.)*;

- *Emerytury i renty w 2022 r.*;
- *Gospodarka finansowa jednostek samorządu terytorialnego 2022*;
- *Gospodarka materiałowa w 2022 r.*;
- *Gospodarka mieszkaniowa i infrastruktura komunalna w 2022 r.*;
- *Gospodarka morska w Polsce w latach 2021 i 2022*;
- *Gospodarka paliwowo-energetyczna w latach 2021 i 2022*;
- *Koniunktura w przetwórstwie przemysłowym, budownictwie, handlu i usługach 2000–2023 (listopad 2023)*;
- *Kultura fizyczna w latach 2021 i 2022*;
- *Obrót nieruchomościami w 2022 r.*;
- *Ochrona środowiska 2023*;
- *Produkcja ważniejszych wyrobów przemysłowych w październiku 2023 r.*;
- *Produkcja wyrobów przemysłowych w latach 2018–2022*;
- „Przegląd Statystyczny. Statistical Review" nr 2/2023;
- *Rocznik Demograficzny 2023*;
- *Rocznik Statystyczny Leśnictwa 2023*;
- *Rolnictwo w 2022 r.*;
- *Sytuacja społeczno-gospodarcza kraju w październiku 2023 r.*;
- *Transport intermodalny w latach 2020–2022*;
- „Wiadomości Statystyczne. The Polish Statistician" nr 11/2023;
- *Wybrane aspekty rynku pracy w Polsce w 2022 r.*;
- *Wypadki przy pracy w 2022 r.*

Joanna Sadowy
Główny Urząd Statystyczny, Departament Opracowań Statystycznych, Polska
Statistics Poland, Statistical Products Department, Poland

Wszystkie publikacje GUS w wersji elektronicznej są dostępne na stronie stat.gov.pl/publikacje/publikacje-a-z. Wersje drukowane (jeśli zostały wydane) można zamawiać pod adresem: zws-sprzedaz@stat.gov.pl.
All the publications of Statistics Poland available in electronic form can be accessed at stat.gov.pl/en/publications. Printed versions (if available) may be ordered at: zws-sprzedaz@stat.gov.pl.

Wiadomości Statystyczne. The Polish Statistician, 2023, vol. 68, 12, 95–100
Spis treści numerów 1–12/2023 / Contents of the issues 1–12, 2023

SPIS TREŚCI NUMERÓW 1–12/2023
CONTENTS OF THE ISSUES 1–12, 2023

DLA AUTORÓW
FOR THE AUTHORS
(for the English translation of the information given below, please visit ws.stat.gov.pl/ForAuthors)

W „Wiadomościach Statystycznych. The Polish Statistician" („WS") zamieszczane są artykuły o charakterze naukowym poświęcone teorii i praktyce statystycznej, które prezentują wyniki oryginalnych badań teoretycznych lub analitycznych wykorzystujących metody statystyki matematycznej, opisowej bądź ekonometrii. Ukazują się również artykuły przeglądowe, recenzje publikacji naukowych oraz inne opracowania informacyjne. W czasopiśmie publikowane są prace w języku polskim i angielskim.

Od 2007 r. „WS" znajdują się na liście czasopism naukowych MEiN. Zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 1 grudnia 2021 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych „WS" otrzymały 70 punktów.

„Wiadomości Statystyczne. The Polish Statistician" są udostępniane w następujących bazach, repozytoriach, katalogach i wyszukiwarkach: Agro, BazEkon, Biblioteka Nauki, Central and Eastern European Academic Source (CEEAS), Central and Eastern European Online Library (CEEOL), Central European Journal of Social Sciences and Humanities (CEJSH), Directory of Open Access Journals (DOAJ), EBSCO Discovery Service, European Reference Index for the Humanities and Social Sciences (ERIH Plus), Exlibris Primo, Google Scholar, ICI Journals Master List, ICI World of Journals, Norwegian Register for Scientific Journals and Publishers (The Nordic List) oraz Summon.

Za publikację artykułów na łamach „WS" autorzy nie otrzymują honorariów ani nie wnoszą opłat.

1. Zgłaszanie artykułów

Prace przeznaczone do opublikowania w „WS" należy przesyłać za pośrednictwem platformy Editorial System: www.editorialsystem.com/ws.

Zgłaszany artykuł powinien być zanonimizowany, tj. pozbawiony informacji o autorze/autorach (również we właściwościach pliku), podziękowań i informacji o źródłach finansowania, a także innych informacji wskazujących na afiliację lub umożliwiających zidentyfikowanie autora. Jeżeli w pracy występują tablice, wykresy lub mapy, powinny być umieszczone w treści artykułu. Materiały graficzne, razem z danymi do nich, należy ponadto załączyć jako osobny plik / osobne pliki, najlepiej w formacie Excel. **Prosimy o niestosowanie stylów i ograniczenie formatowania do wymogów redakcyjnych.** Więcej informacji w pkt 4 *Wymogi redakcyjne*.

Razem z artykułem należy przesłać skan/zdjęcie oświadczenia o oryginalności pracy i niezłożeniu jej w innym wydawnictwie. **Załączenie oświadczenia jest warunkiem poddania pracy ocenie wstępnej i skierowania do recenzji.**

Zgłoszenie artykułu do opublikowania w „WS" oznacza zgodę na jego udostępnienie na licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0).

Autorzy mają prawo do samodzielnego umieszczania w wybranych przez siebie repozytoriach artykułu w wersji zarówno zgłoszonej do „WS", jak i zaakceptowanej do opublikowania

oraz opublikowanej, z zastrzeżeniem wymogu niezwłocznego podania w repozytorium informacji o numerze „WS", w którym praca się ukazała, wraz z linkiem do niej (DOI).

2. Przebieg prac redakcyjnych

Zgłoszony artykuł jest oceniany i opracowywany w czteroetapowym procesie:

1. **Ocena wstępna**, dokonywana przez redakcję. Polega na weryfikacji naukowego charakteru artykułu oraz jego struktury i zawartości pod kątem wymogów redakcyjnych, a także zgodności tematyki z profilem czasopisma. Autor uzupełnia i poprawia artykuł stosownie do uwag redakcji, a w przypadku nieuwzględnienia danej uwagi uzasadnia swoje stanowisko. Warunkiem skierowania pracy do recenzji jest potwierdzenie oryginalności tekstu uzyskane za pomocą systemu antyplagiatowego. W przypadku wykrycia znacznego podobieństwa do innych prac artykuł zostanie odrzucony.
2. **Ocena recenzentów**, dokonywana przez specjalistów w danej dziedzinie. Artykuł oceniają dwaj recenzenci spoza jednostki naukowej, przy której afiliowany jest autor; w przypadku pracy w języku angielskim co najmniej jeden recenzent jest afiliowany przy jednostce zagranicznej. W razie sprzecznych opinii dwóch recenzentów powoływany jest trzeci recenzent. Recenzenci kierują się kryteriami oryginalności i jakości opracowania zarówno w odniesieniu do treści, jak i formy artykułu.

   Autorzy artykułów, które otrzymały pozytywne oceny, wprowadzają poprawki zalecane przez recenzentów i przesyłają zmodyfikowaną wersję pracy. Jeśli pojawi się różnica zdań dotycząca zasadności proponowanych zmian, autorzy są zobligowani do uzasadnienia swojego stanowiska.
3. **Ocena Kolegium Redakcyjnego** (KR), decydująca o przyjęciu pracy do publikacji. Jest dokonywana na podstawie recenzji, z uwzględnieniem opinii redaktorów tematycznego i merytorycznego. Polega m.in. na weryfikacji dokonania przez autora zmian w artykule stosownie do uwag recenzentów. KR ocenia artykuł pod względem poprawności i spójności merytorycznej oraz zaleca autorowi wprowadzenie poprawek, jeśli są one konieczne, aby praca spełniała wymogi czasopisma. Autorowi przysługuje prawo do odwołania od decyzji o niepublikowaniu artykułu. W takim przypadku powinien on skontaktować się z redakcją „WS" i przedstawić uzasadnienie. Ostateczna decyzja w tej sprawie należy do redaktora naczelnego.

   W „WS" publikowane są wyłącznie te artykuły, które otrzymają pozytywną ocenę na każdym z wymienionych etapów i zostaną poprawione przez autora zgodnie z otrzymanymi uwagami (chyba że autor przedstawi argumenty uzasadniające nieuwzględnienie danej uwagi).

   Artykuły przyjęte przez KR do publikacji są zamieszczane na stronie internetowej czasopisma w zakładce Early View, gdzie znajdują się do czasu opublikowania w konkretnym wydaniu „WS".
4. **Opracowanie redakcyjne, autoryzacja i korekta.** Artykuł zakwalifikowany do druku jest poddawany opracowaniu merytorycznemu i językowemu. Redakcja zastrzega sobie prawo do zmiany tytułu i śródtytułów, modyfikowania tablic, wykresów i innych elementów graficznych oraz przeredagowania treści bez naruszenia zasadniczej myśli autora.

   Po opracowaniu redakcyjnym artykuł jest przesyłany do autoryzacji. Tekst zatwierdzony przez autora, po składzie i łamaniu, jest poddawany korekcie i rewizji (II korekcie).

Autor dokonuje korekty autorskiej tekstu na etapie rewizji. Wykresy i inne materiały graficzne są opracowywane na podstawie plików i danych przekazanych przez autora i poddawane korekcie i rewizji. Autor dokonuje ich akceptacji na etapie rewizji.

W przypadku odkrycia błędów w opublikowanym artykule zamieszcza się na łamach „WS" sprostowanie, a artykuł w wersji elektronicznej jest poprawiany i umieszczany na stronie internetowej „WS" ze stosownym wyjaśnieniem.

3. Zasady etyki publikacyjnej COPE

Redakcja „WS" podejmuje wszelkie starania w celu utrzymania najwyższych standardów etycznych zgodnie z wytycznymi Komitetu ds. Etyki Publikacyjnej (COPE), dostępnymi na stronie internetowej www.publicationethics.org, oraz wykorzystuje wszystkie możliwe środki mające na celu zapobieżenie nadużyciom i nierzetelności autorskiej. Przyjęte zasady postępowania obowiązują autorów, Radę Naukową, Kolegium Redakcyjne, redakcję, pracowników Wydziału Czasopism Naukowych GUS, recenzentów i wydawcę.

3.1. Odpowiedzialność autorów

1. Artykuły naukowe kierowane do opublikowania w „WS" powinny zawierać precyzyjny opis badanych zjawisk i stosowanych metod oraz autorskie wnioski i sugestie dotyczące rozwoju badań i analiz statystycznych. Autorzy powinni wyraźnie określić cel artykułu oraz jasno przedstawić wyniki przeprowadzonej analizy. Prezentacja efektów badań statystycznych zaprojektowanych i przeprowadzonych przez autorów wymaga opisania zastosowanej w nich metodologii. W przypadku nowatorskich metod analizy pożądane jest podanie przykładu ilustrującego ich zastosowanie w praktyce statystycznej. Autorzy ponoszą odpowiedzialność za treści prezentowane w artykułach. W razie zgłaszania przez czytelników zastrzeżeń odnoszących się do tych treści autorzy są zobligowani do udzielenia odpowiedzi za pośrednictwem redakcji.
2. Na autorach spoczywa obowiązek zapewnienia pełnej oryginalności przedłożonych prac. Redakcja nie toleruje przejawów nierzetelności naukowej autorów, takich jak:
   - duplikowanie publikacji – ponowne publikowanie własnego utworu lub jego części;
   - plagiat – przywłaszczenie cudzego utworu lub jego fragmentu bez podania informacji o źródle;
   - fabrykowanie danych – oparcie pracy naukowej na nieprawdziwych wynikach badań;
   - autorstwo widmo (*ghost authorship*) – nieujawnianie współautorów, mimo że wnieśli oni istotny wkład w powstanie artykułu;
   - autorstwo gościnne (*guest authorship*) – podawanie jako współautorów osób o znikomym udziale lub niebiorących udziału w opracowywaniu artykułu;
   - autorstwo grzecznościowe (*gift authorship*) – podawanie jako współautorów osób, których wkład jest oparty jedynie na słabym powiązaniu z badaniem.

   Autorzy deklarują w stosownym oświadczeniu, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i nie został złożony w innym wydawnictwie oraz że jest ich oryginalnym dziełem, i określają swój wkład w opracowanie artykułu. Jeżeli doszło do zaprezentowania podobnych materiałów podczas konferencji lub

sympozjum naukowego, to podczas składania tekstu do publikacji w „WS" autorzy są zobowiązani poinformować o tym redakcję.
3. Autorzy są zobowiązani do podania w treści artykułu wszelkich źródeł finansowania badań będących podstawą pracy.
4. Główną odpowiedzialność za rzetelność przekazanych informacji, łącznie z informacją na temat wkładu poszczególnych współautorów w powstanie artykułu, ponosi zgłaszający artykuł.
5. Autorzy zgłaszający artykuły do publikacji w „WS" biorą udział w procesie recenzji double-blind peer review, dokonywanej przez co najmniej dwóch niezależnych ekspertów z danej dziedziny. Po otrzymaniu pozytywnych recenzji autorzy wprowadzają zalecane przez recenzentów poprawki i dostarczają redakcji zaktualizowaną wersję opracowania wraz z pisemnym poświadczeniem uwzględnienia poprawek. Jeśli pojawi się różnica zdań co do zasadności proponowanych zmian, należy wyjaśnić, które poprawki zostały uwzględnione, a w przypadku ich nieuwzględnienia – uzasadnić swoje stanowisko.
6. Jeżeli autorzy odkryją w swoim maszynopisie lub tekście już opublikowanym błędy, nieścisłości bądź niewłaściwe dane, powinni niezwłocznie poinformować o tym redakcję w celu dokonania korekty, wycofania tekstu lub zamieszczenia sprostowania. W przypadku korekty artykułu już opublikowanego jego nowa wersja jest zamieszczana na stronie internetowej „WS" wraz ze stosownym wyjaśnieniem.

3.2. Odpowiedzialność Rady Naukowej, Kolegium Redakcyjnego i Wydziału Czasopism Naukowych GUS

1. Rada Naukowa (RN) kształtuje profil programowy czasopisma, określa kierunki jego rozwoju i konsultuje jego zakres merytoryczny.
2. Kolegium Redakcyjne (KR) podejmuje decyzję o publikacji danego artykułu z uwzględnieniem ocen recenzentów oraz opinii zespołu redakcyjnego. W swoich rozstrzygnięciach członkowie KR kierują się kryteriami merytorycznej oceny wartości artykułu, jego oryginalności i jasności przekazu, a także ścisłego związku z celem i zakresem tematycznym „WS". Oceniają artykuły niezależnie od płci, rasy, pochodzenia etnicznego, narodowości, religii, wyznania, światopoglądu, niepełnosprawności, wieku lub orientacji seksualnej ich autorów.
3. Zespół redakcyjny, wyodrębniony z KR, tworzą redaktor naczelny i jego zastępca, redaktorzy tematyczni i redaktor merytoryczny. Członkowie zespołu redakcyjnego weryfikują nadsyłane artykuły pod względem merytorycznym, oceniają ich zgodność z celem i zakresem tematycznym „WS" oraz sprawdzają spełnienie wymogów redakcyjnych i przestrzeganie zasad rzetelności naukowej. Ponadto wybierają recenzentów w taki sposób, aby nie wystąpił konflikt interesów, i dbają o zapewnienie uczciwego, bezstronnego i terminowego procesu recenzowania.
4. Za sprawny przebieg procesu wydawniczego, poinformowanie wszystkich jego uczestników o konieczności przestrzegania obowiązujących zasad i przygotowanie artykułów do publikacji odpowiadają pracownicy Wydziału Czasopism Naukowych (WCN) GUS. W celu uzyskania obiektywnej oceny oryginalności nadsyłanych artykułów przed skierowaniem ich do recenzji WCN wykorzystuje system antyplagiatowy. Informacje dotyczące

artykułu mogą być przekazywane przez WCN wyłącznie autorom, recenzentom, członkom RN i KR oraz wydawcy.

5. Zmiany dokonane w tekście na etapie przygotowania artykułu do publikacji nie mogą naruszać zasadniczej myśli autorów. Wszelkie modyfikacje o charakterze merytorycznym są z nimi konsultowane.
6. W przypadku podjęcia decyzji o niepublikowaniu artykułu nie może on zostać w żaden sposób wykorzystany przez wydawcę lub uczestników procesu wydawniczego bez pisemnej zgody autorów. Autorzy mogą się odwołać od decyzji o niepublikowaniu artykułu. W tym celu powinni się skontaktować z redaktorem naczelnym lub sekretarzem redakcji „WS" i przedstawić stosowną argumentację. Odwołania autorów są rozpatrywane przez redaktora naczelnego.
7. Członkowie RN i KR ani pracownicy WCN nie mogą pozostawać w jakimkolwiek konflikcie interesów w odniesieniu do artykułów zgłaszanych do publikacji. Przez konflikt interesów należy rozumieć sytuację, w której jakiekolwiek interesy lub zależności (służbowe, finansowe lub inne) mogą mieć wpływ na ocenę artykułu lub decyzję o jego publikacji.
8. W celu przeciwdziałania nierzetelności naukowej wymagane jest złożenie przez autorów oświadczenia, w którym deklarują, że zgłaszany artykuł nie narusza praw autorskich osób trzecich, nie był dotychczas publikowany i jest ich oryginalnym dziełem, a także określają swój wkład w opracowanie artykułu.
9. W celu zapewnienia wysokiej jakości recenzji wymagane jest złożenie przez recenzentów oświadczenia o przestrzeganiu zasad etyki recenzowania COPE i niewystępowaniu konfliktu interesów.
10. W przypadku uzasadnionego podejrzenia na jakimkolwiek etapie procesu wydawniczego, że autorzy dopuścili się nierzetelności naukowej (zob. pkt 3.1. Odpowiedzialność autorów), zespół redakcyjny skrupulatnie zbada sprawę ewentualnego nadużycia. Jeśli nierzetelność autorów zostanie udowodniona, to zgłoszony przez nich artykuł zostanie odrzucony przez KR, a autorzy otrzymają informację o podjętej decyzji wraz z jej uzasadnieniem.
11. Czytelnicy, którzy mają wobec autorów opublikowanego artykułu uzasadnione podejrzenia o nierzetelność naukową, powinni powiadomić o tym redaktora naczelnego lub sekretarza redakcji. Po zbadaniu sprawy ewentualnego nadużycia czytelnicy zostaną poinformowani o rezultacie przeprowadzonego postępowania. W przypadku potwierdzenia nadużycia, na łamach czasopisma zostanie zamieszczona stosowna informacja.

3.3. Odpowiedzialność recenzentów

1. Recenzenci przyjmują artykuł do recenzji tylko wtedy, gdy uznają, że:
   - posiadają odpowiednią wiedzę w określonej dziedzinie, aby rzetelnie ocenić pracę;
   - zgodnie z ich stanem wiedzy nie istnieje konflikt interesów w odniesieniu do autorów, przedstawionych w artykule badań i instytucji je finansujących, co potwierdzają w oświadczeniu;
   - mogą wywiązać się z terminu ustalonego przez redakcję, aby nie opóźniać publikacji.
2. Recenzenci są zobligowani do zachowania obiektywności i poufności oraz powstrzymania się od osobistej krytyki. Zawsze powinni uzasadnić swoją ocenę, przedstawiając stosowną argumentację.

3. Recenzenci powinni wskazać ważne dla wyników badań opublikowane prace, które w ich ocenie powinny zostać przywołane w ocenianym artykule.
4. W razie stwierdzenia wysokiego poziomu zbieżności treści recenzowanej pracy z innymi opublikowanymi materiałami lub podejrzenia innych przejawów nierzetelności naukowej recenzenci są zobowiązani poinformować o tym redakcję.
5. Po ukończeniu recenzji przechowywanie przesłanych przez redakcję materiałów (w jakiejkolwiek formie) oraz posługiwanie się nimi przez recenzentów jest niedozwolone.

3.4. Odpowiedzialność wydawcy

1. Materiały opublikowane w „WS" są chronione prawem autorskim.
2. Wydawca udostępnia pełną treść wszystkich artykułów w internecie w trybie otwartego dostępu, tj. bezpłatnie i bez technicznych ograniczeń, od 1 stycznia 2022 r. na licencji Creative Commons Uznanie autorstwa – Na tych samych warunkach 4.0 (CC BY-SA 4.0). W przypadku artykułów zgłoszonych do „WS" od 2022 r. dozwolone jest dzielenie się artykułem (kopiowanie i rozpowszechnianie go w dowolnym medium i formacie) oraz adaptowanie go (w dowolnym celu, także komercyjnym) na warunkach określonych w tej licencji. Z pozostałych artykułów zamieszczonych w czasopiśmie można korzystać w ramach otwartego dostępu, zgodnie z ustawą o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego.
3. Wydawca deklaruje gotowość do opublikowania poprawek, wyjaśnień oraz przeprosin.

4. Wymogi redakcyjne

Zgodnie z wymogami czasopisma omawiany w artykule problem badawczy powinien być jednoznacznie zdefiniowany oraz istotny dla oceny zjawisk społecznych lub gospodarczych. Artykuł powinien zawierać wyraźnie określony cel badania, precyzyjny opis badanych zjawisk i stosowanych metod, uzyskane wyniki przeprowadzonej analizy oraz autorskie wnioski.

4.1. Struktura i zawartość artykułu

Wymagane elementy artykułu:
1. Tytuł.
2. Dane autora: imię/imiona i nazwisko, afiliacja w języku polskim i angielskim, ORCID, wkład w powstanie artykułu, adres e-mail. Wśród autorów artykułu wieloautorskiego należy wskazać autora korespondencyjnego.
3. Streszczenie (zalecana objętość – do 1200 znaków ze spacjami, forma bezosobowa). W przypadku artykułu opisującego badanie empiryczne powinno zawierać: cel, przedmiot, okres i metodę badania, źródła danych i najważniejsze wnioski z badania. W przypadku artykułów o innym charakterze należy podać co najmniej cel pracy, jej przedmiot i najważniejsze wnioski.

   Streszczenie to podstawowe źródło informacji o artykule, warunkujące też decyzję czytelnika o zapoznaniu się z całą pracą. Dlatego powinno być przygotowane ze szczególną starannością i dbałością o umieszczenie w nim wszystkich wymaganych elementów.

4. Słowa kluczowe – najistotniejsze pojęcia lub wyrażenia użyte w pracy (nie mniej niż trzy). Powinny być zawarte w streszczeniu i/lub tytule.
5. Kod/kody z klasyfikacji Journal of Economic Literature (JEL).
6. Tłumaczenie tytułu, streszczenia i słów kluczowych (na język angielski w przypadku artykułu napisanego w języku polskim, a na język polski w przypadku artykułu napisanego w języku angielskim).
7. W artykule opisującym badanie empiryczne wymagane są następujące części:
   - wprowadzenie, zawierające syntetyczne przedstawienie zagadnień teoretycznych, uzasadnienie podjęcia danego problemu badawczego, cel badania i krytyczne odniesienie do literatury przedmiotu. W wyjątkowych przypadkach, kiedy istotne dla podjętego tematu jest obszerniejsze przedstawienie dyskusji toczącej się w literaturze, przegląd literatury może stanowić odrębną część artykułu;
   - metoda badania, uwzględniająca przedmiot i okres badania, źródła danych i zastosowane metody badawcze, w tym uzasadnienie ich wyboru;
   - wyniki badania – analiza danych oraz interpretacja wyników i odniesienie ich do rezultatów wcześniejszych badań (dyskusja). W uzasadnionych przypadkach dyskusja może stanowić odrębną część artykułu;
   - podsumowanie, które powinno być zwięzłe i odzwierciedlać istotę problemu badawczego przedstawionego w artykule, bez podawania danych liczbowych; końcowe wnioski powinny odnosić się do treści artykułu, a w szczególności do celu badania.

   Wszystkie części powinny być opatrzone numerami.
8. Bibliografia, zawierająca pełny wykaz prac i materiałów przywołanych w artykule, przygotowana zgodnie z wymogami czasopisma.

4.2. Przygotowanie artykułu

1. Artykuł powinien być utrzymany w formie bezosobowej.
2. Tekst należy zapisać alfabetem łacińskim. Nazwy własne, tytuły itp. oryginalnie zapisane innym alfabetem powinny być poddane transliteracji.
3. Nie należy stosować stylów; formatowanie należy ograniczyć do wymogów redakcyjnych.
4. Objętość artykułu łącznie ze streszczeniem, słowami kluczowymi, bibliografią, tablicami, wykresami i innymi materiałami graficznymi nie powinna być mniejsza niż 10 stron maszynopisu ani przekraczać 20 stron.
5. Edytor tekstu: Microsoft Word, format *.doc lub *.docx.
6. Krój czcionki:
   - Arial – tytuł, autor, streszczenie, słowa kluczowe, kody JEL, śródtytuły, elementy graficzne (tablice, zestawienia, wykresy, schematy), przypisy;
   - Times New Roman – tekst główny, bibliografia.
7. Wielkość czcionki:
   - 14 pkt – tytuł, autor, śródtytuły wyższego rzędu;
   - 12 pkt – tekst główny, śródtytuły niższego rzędu;
   - 10 pkt – pozostałe elementy.
8. Marginesy – 2,5 cm z każdej strony.

9. Interlinia – 1,5 wiersza; tablice i przypisy – 1 wiersz; przed tytułami rozdziałów i podrozdziałów oraz po nich – pusty wiersz.
10. Wcięcie akapitowe – 0,4 cm; bibliografia – bez wcięcia, wysunięcie 0,4 cm.
11. Przy wyliczeniach należy posłużyć się listą punktowaną z punktorami w postaci kropek (wysunięcie 0,4 cm, wcięcie 0 cm); wiersze (oprócz ostatniego) zakończone średnikiem.
12. Strony ponumerowane automatycznie.
13. Tablice i elementy graficzne (wykresy, mapy, schematy) muszą być przywołane w tekście.
14. Wykresy, mapy i schematy należy zamieścić w tekście głównym. Wykresy powinny być edytowalne (optymalnie wykonane w programie Excel; w przypadku wykonania w programie graficznym powinny mieć postać wektorową). Wykresy i inne materiały graficzne należy przekazać osobno, najlepiej w pliku programu Excel lub innym edytowalnym w pakiecie Microsoft Office.
15. Tablice muszą być edytowalne. Nie należy stosować rastrów, cieniowania, pogrubiania czy też podwójnych linii itp.
16. Wskazówki dotyczące opracowywania map znajdują się w publikacji *Mapy statystyczne. Opracowanie i prezentacja danych*, dostępnej na stronie internetowej GUS.
17. Pod tablicami i każdym elementem graficznym należy podać źródło opracowania, a także objaśnić użyte w nich skróty i symbole.
18. Literowe symbole liczb i innych wielkości niezłożonych należy zapisywać małą lub dużą literą i pismem pochyłym (np. a, A, $y(x)$, a_i); wektorów – pismem pochyłym i pogrubionym (np. $\boldsymbol{a}$, $\boldsymbol{A}$, $\boldsymbol{w}$, $\boldsymbol{y}(x)$, $\boldsymbol{w}_i$); macierzy – pismem prostym i pogrubionym (np. $\mathbf{A}$, $\mathbf{a}$, $\mathbf{M}$, $\mathbf{Y}(x)$, $\mathbf{M}_i$).
19. Objaśnienia znaków umownych i zapisów w tablicach: kreska (–) – zjawisko nie wystąpiło; zero (0) – zjawisko istniało w wielkości mniejszej od 0,5; (0,0) – zjawisko istniało w wielkości mniejszej od 0,05; kropka (.) – brak informacji, konieczność zachowania tajemnicy statystycznej, wypełnienie pozycji jest niemożliwe lub niecelowe; „w tym" – oznacza, że nie podaje się wszystkich składników sumy.
20. Stosowane są następujące skróty: tablica – tabl., wykres – wykr.
21. Wszystkie zawarte w artykule informacje, dane i stwierdzenia wykraczające poza wiedzę powszechną – np. wyniki badań innych autorów, zarówno o charakterze empirycznym, jak i koncepcyjnym – muszą być opatrzone przypisem bibliograficznym. Przez wiedzę powszechną należy rozumieć informacje ogólnie znane i niebudzące wątpliwości ani kontrowersji w danej grupie społecznej, np. utworzenie GUS w 1918 r. lub powstanie UE w 1993 r. na podstawie traktatu z Maastricht. Natomiast dane statystyczne udostępniane lub publikowane np. przez GUS lub Eurostat nie należą do takich informacji. Charakteru wiedzy powszechnej nie mają również stwierdzenia odnoszące się do idei, zjawisk i procesów społecznych, politycznych czy gospodarczych. Nawet pozornie zdroworozsądkowe idee zmieniają bowiem swój sens w zależności od kultury, języka lub dyscypliny naukowej, a także bywają w rozmaity sposób konceptualizowane, jak np. pojęcie poznania w naukach społecznych.

    Podanie źródła jest konieczne niezależnie od tego, czy informacje lub stwierdzenia są ujęte w ramy cytatu, czy przedstawione bez dosłownego przytoczenia, np. w formie parafrazy. Jeżeli stwierdzenie może budzić jakiekolwiek wątpliwości odbiorców, autor powinien wskazać stosowne źródło podawanej informacji.

22. Przypisy rzeczowe, słownikowe lub informacyjne należy umieszczać na dole strony. Przypisy bibliograficzne, zgodnie ze standardem APA (American Psychological Association), należy podawać w tekście głównym.
23. Bibliografię należy przygotować zgodnie ze standardem APA.

4.3. Zasady przywoływania publikacji w treści artykułu

Wyszczególnienie	Przykład przywołania	
	w odsyłaczu	w treści zdania
Autor indywidualny		
Jeden autor	(Iksiński, 2001)	Iksiński (2001)
Dwóch autorów	(Iksiński i Nowak, 1999)	Iksiński i Nowak (1999)
Trzech autorów lub więcej	(Jankiewicz i in., 2003)	Jankiewicz i in. (2003)
Autor instytucjonalny		
Nazwa funkcjonuje jako powszechnie znany skrótowiec: pierwsze przywołanie w tekście	(International Labour Organization [ILO], 2020)	International Labour Organization (ILO, 2020)
kolejne przywołanie	(ILO, 2020)	ILO (2020)
Pełna nazwa	(Stanford University, 1995)	Stanford University (1995)
Typ publikacji		
Publikacja bez ustalonego autorstwa	(*Skrócony tytuł* …, 2015)	*Pełny tytuł* (2015)
Publikacja bez roku wydania	(Iksiński, b.r.)	Iksiński (b.r.)
Akt prawny	(Pełny tytuł)	Pełny tytuł
Strona internetowa / Zbiór danych: znana data publikacji	(Iksiński, 2020) / (Nazwa instytucji, 2020)	Iksiński (2020) / Nazwa instytucji (2020)
nieznana data publikacji	(Iksiński, b.r.) / (Nazwa instytucji, b.r.)	Iksiński (b.r.) / Nazwa instytucji (b.r.)
Rodzaj przywołania		
Przywoływanie kilku prac (porządek prac – chronologiczny, porządek autorów – alfabetyczny)	(Iksiński, 1997, 1999, 2004a, 2004b; Nowak, 2002)	Iksiński (1997, 1999, 2004a, 2004b) i Nowak (2002)
Przywoływanie publikacji za innym autorem (uwaga: w bibliografii należy wymienić tylko pracę czytaną)	(Nowakowski, 1990, za: Zieniecka, 2007)	Nowakowski (1990, za: Zieniecka, 2007)

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

4.4. Przykłady opisu bibliograficznego

Bibliografia powinna być zamieszczona na końcu opracowania. Prace należy uszeregować alfabetycznie według nazwiska pierwszego autora. W przypadku dwóch lub więcej prac tego samego autora / tych samych autorów trzeba je uporządkować chronologicznie według roku publikacji. Jeśli kilka prac tego samego autora / tych samych autorów zostało opublikowanych w tym samym roku, należy podać je w kolejności alfabetycznej według tytułu i odpowiednio oznaczyć literami a, b, c itd.

Typ publikacji	Przykład opisu bibliograficznego
Artykuł w czasopiśmie	
W wersji drukowanej	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik* (zeszyt), strona początku–strona końca.
Dostępny w internecie, z DOI	Nazwisko, X., Nazwisko 2, Y. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://doi.org/xxx.
Dostępny w internecie, bez DOI	Nazwisko, X., Nazwisko 2, Y., Nazwisko 3, Z. (rok). Tytuł artykułu. *Tytuł czasopisma*, *rocznik*(zeszyt), strona początku–strona końca. https://xxx.
Maszynopis	
Niepublikowany / przygotowywany przez autora / zgłoszony do publikacji, ale jeszcze niezaakceptowany	Nazwisko, X. (rok). *Tytuł* [maszynopis niepublikowany / w przygotowaniu / zgłoszony do publikacji].
Zaakceptowany do publikacji	Nazwisko, X. (w druku). Tytuł artykułu. *Tytuł czasopisma.*
Opublikowany nieformalnie (np. na stronie internetowej autora)	Nazwisko, X., Nazwisko 2, Y. (rok). *Tytuł artykułu*. https://xxx.
Opublikowany w trybie online first (przed włączeniem do zeszytu)	Nazwisko, X. (rok). Tytuł artykułu. *Tytuł czasopisma.* Online first. https://xxx.
Książka	
W wersji drukowanej	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo.
Dostępna w internecie, z DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://doi.org/xxx.
Dostępna w internecie, bez DOI	Nazwisko, X. (rok). *Tytuł książki*. Wydawnictwo. https://xxx.
W przekładzie	Nazwisko, X. (rok). *Tytuł książki* (tłum. Y. Nazwisko). Wydawnictwo.
Wydanie wielotomowe: tom zatytułowany	Nazwisko, X. (rok). *Tytuł książki: nr tomu. Tytuł tomu*. Wydawnictwo.
tom niezatytułowany	Nazwisko, X. (rok). *Tytuł książki* (nr tomu). Wydawnictwo.
Kolejne wydanie	Nazwisko, X. (rok). *Tytuł książki* (nr wydania). Wydawnictwo.
Pod redakcją: w języku polskim	Nazwisko, X. (red.). (rok). *Tytuł książki*. Wydawnictwo.
w języku angielskim	Nazwisko, X. (Ed.). (rok). *Tytuł książki*. Wydawnictwo.
Rozdział w pracy zbiorowej	Nazwisko, X. (rok). Tytuł rozdziału. W: Y. Nazwisko, Z. Nazwisko 2 (red.), *Tytuł książki* (s. strona początku–strona końca). Wydawnictwo. https://doi.org/xxx lub https://xxx.
Inne prace	
Raport: autor indywidualny	Nazwisko, X. (rok). *Tytuł raportu*. Wydawnictwo.
autor instytucjonalny	Nazwa instytucji. (rok). *Tytuł raportu*. Wydawnictwo.
Working Papers	Nazwisko, X. (rok). *Tytuł pracy* (nazwa serii i numer). https://doi.org/xxx lub https://xxx.
Sesja konferencyjna / prezentacja / referat	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł pracy* [typ wystąpienia, np. referat]. Nazwa konferencji, miejsce konferencji.
Rozprawa doktorska: nieopublikowana	Nazwisko, X. (rok). *Tytuł pracy* [niepublikowana rozprawa doktorska]. Nazwa instytucji nadającej tytuł doktorski.
opublikowana	Nazwisko, X. (rok). *Tytuł pracy* [rozprawa doktorska, nazwa instytucji nadającej tytuł doktorski]. https://xxx.
Akt prawny	Pełny tytuł aktu prawnego wraz z datą publikacji w dzienniku urzędowym.

Typ publikacji	Przykład opisu bibliograficznego
Strona internetowa	
Znana data publikacji, zawartość strony się nie zmienia	Nazwisko, X. (rok, dzień i miesiąc). *Tytuł*. https://xxx.
Nieznana data publikacji, zawartość strony się zmienia	Nazwa instytucji. (b.r.). *Tytuł*. Pobrane dzień, miesiąc i rok pobrania z https://xxx.
Zbiór danych	
Surowe dane nieopublikowane	Nazwisko, X. (rok wydania pracy, w której dane są wykorzystywane) [opis danych, np. surowe dane nieopublikowane dotyczące…]. Źródło danych (np. nazwa uniwersytetu).
Dane opublikowane: znana data publikacji, zawartość zbioru się nie zmienia	Nazwisko, X. (rok). *Nazwa zbioru danych* [zbiór danych]. Wydawca. https://xxx.
nieznana data publikacji, zawartość zbioru się zmienia	Nazwa instytucji. (b.r.). *Nazwa zbioru danych* [zbiór danych]. Wydawca. Pobrane dzień, miesiąc i rok pobrania z https://xxx.

Źródło: opracowanie na podstawie: American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th edition). https://doi.org/10.1037/0000165-000.

Praca przygotowana w sposób niezgodny z powyższymi wskazówkami będzie odesłana do autora z prośbą o dostosowanie formy artykułu do wymogów redakcyjnych.

DZIAŁY „WS" – TEMATYKA ARTYKUŁÓW
WS SECTIONS – TOPICS OF THE ARTICLES

Pełny opis zakresu tematycznego działów: ws.stat.gov.pl/AimScope
Description of the topics covered in each section: ws.stat.gov.pl/AimScope

Studia metodologiczne / Methodological studies

- Oryginalne teoretyczne rozwiązania metodologiczne ze wskazaniem ich praktycznej użyteczności
- Prace przeglądowe i porównawcze oraz dotyczące etyki w statystyce, które wnoszą pionierski wkład poznawczy do obecnego stanu wiedzy

Statystyka w praktyce / Statistics in practice

- Nowatorskie zastosowania narzędzi i modeli statystycznych oraz analiza i ocena statystyczna zjawisk społeczno-ekonomicznych i innych, prowadzona w szczególności na danych z zasobów statystyki publicznej
- Wykorzystanie narzędzi informatycznych do uzyskiwania i przetwarzania informacji statystycznych, naliczania danych wynikowych, ich prezentacji i rozpowszechniania
- Projektowanie badań statystycznych, uzyskiwanie, integracja i przetwarzanie danych oraz generowanie wynikowych informacji statystycznych i kontrola ich ujawniania

Studia interdyscyplinarne. Wyzwania badawcze / Interdisciplinary studies. Research challenges

- Wyzwania badawcze wynikające z rosnących potrzeb użytkowników danych statystycznych i wymagające zaangażowania znacznych środków oraz rozwiązań z różnych dziedzin nauki i techniki
- Wykorzystanie technologii informacyjnych i komunikacyjnych, innowacyjność, przetwarzanie i analiza zagadnień związanych z data science i big data
- Wyniki badań prowadzonych przez przedstawicieli dyscyplin innych niż statystyka z wykorzystaniem metod statystycznych

Spisy powszechne – problemy i wyzwania / Issues and challenges in census taking

- Propozycje rozwiązań – zarówno organizacyjnych, jak i metodologicznych – możliwych do zastosowania w spisach oraz rezultaty analiz danych spisowych
- Praktyczne aspekty związane z gromadzeniem i udostępnianiem danych ze spisów, w tym dotyczące obciążenia odpowiedzi i ochrony tajemnicy statystycznej

Edukacja statystyczna / Statistical education

- Metody i efekty nauczania statystyki oraz popularyzacja myślenia statystycznego i rzetelnego posługiwania się informacjami statystycznymi
- Problemy związane z kształceniem w zakresie umiejętności stosowania statystyki na wszystkich poziomach edukacji, a także dotyczące wykorzystywania nowoczesnych koncepcji i metod dydaktycznych oraz pomocy naukowych w nauczaniu statystyki

Z dziejów statystyki / From the history of statistics

- Historia prowadzenia obserwacji statystycznych oraz rozwoju ich metodologii i narzędzi
- Życie i osiągnięcia zawodowe wybitnych statystyków, jak również działalność najważniejszych instytucji i organizacji statystycznych w Polsce i za granicą

In memoriam

- Nekrologi i artykuły wspomnieniowe

Informacje. Recenzje. Dyskusje / Discussions. Reviews. Information

- Teksty nierecenzowane i niemające charakteru artykułów naukowych: sprawozdania z konferencji naukowych i innych wydarzeń dotyczących statystyki, recenzje książek, omówienia nowości wydawniczych GUS, polemiki i dyskusje