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:

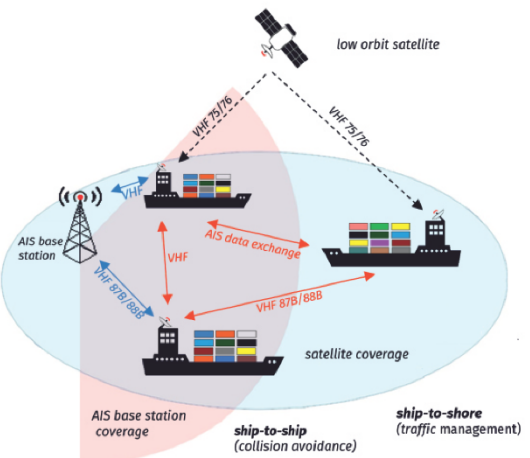
- all ships of a 300 gross tonnage and more – for international shipping;
- all vessels of a 500 gross tonnage and more not engaged in international shipping;
- all passenger ships, regardless of size.

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.

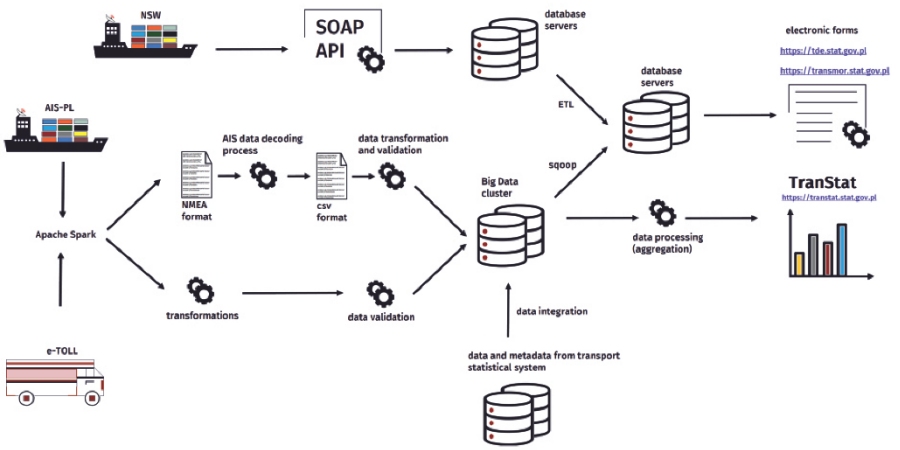
Due to the nature of the data, it was necessary to consider two types of data processing: batch processing (e-TOLL) and real-time processing of sensor data (AIS).

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

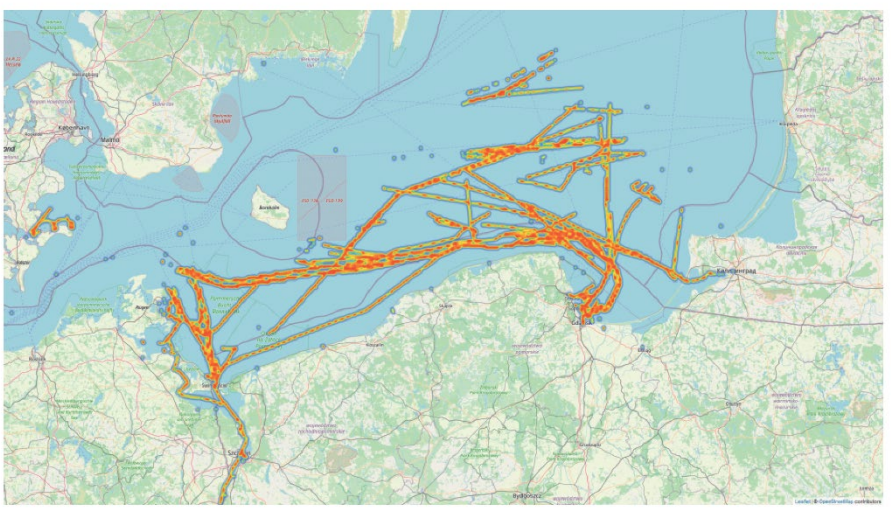
Many works indicate the need for an in-depth analysis of maritime traffic (Vasilev & Sulova, 2023) or, more broadly, multimodal transport flows (Zhang et al., 2018). To identify the phenomenon for four ports of primary importance to the national economy: Gdańsk, Gdynia, Szczecin and Świnoujście, points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

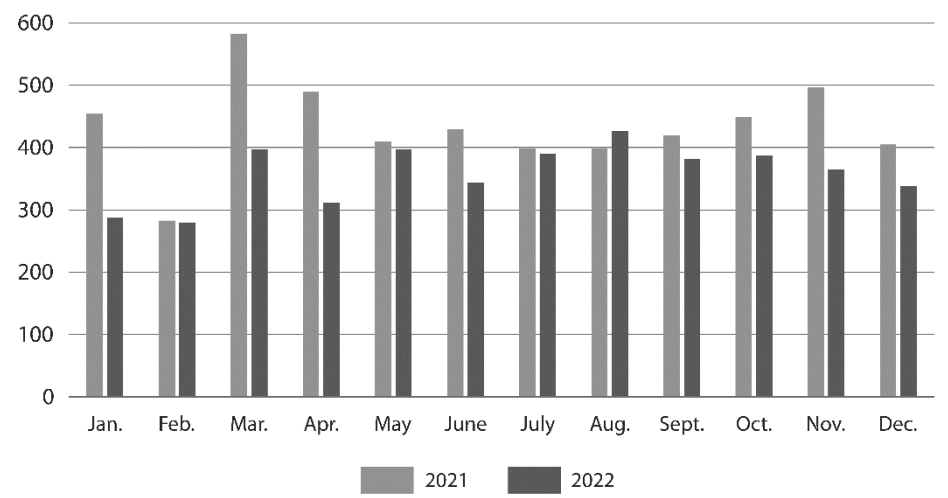
Years 2021 and 2022 were selected for the port of Szczecin to generate traffic statistics in maritime transport. The visualisation was made in months.

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.

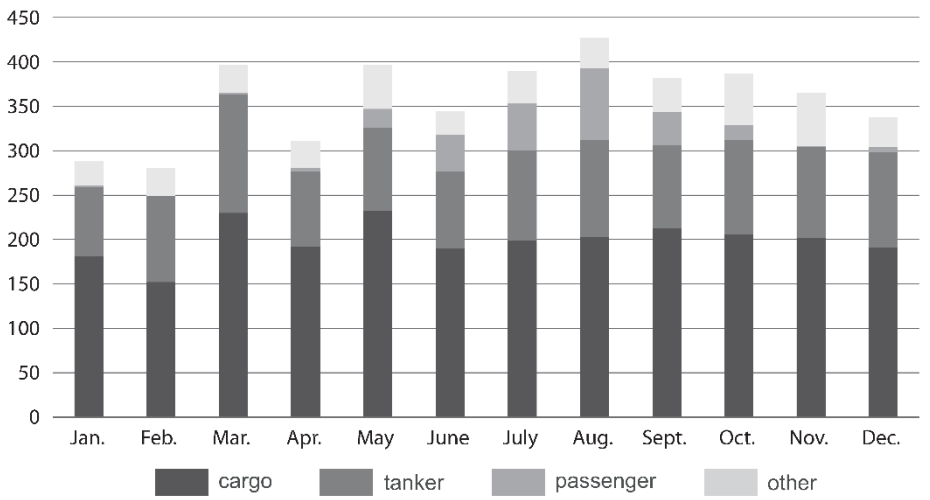
Figure 4 shows the number of ships entering the port of Szczecin by month and by ship type in 2022. The data presented relates only to cargo ships, passenger ships, tankers and ships classified as other (e.g. with an unknown ship type code). The analysis excludes such types of vessels as: fishing, service, 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

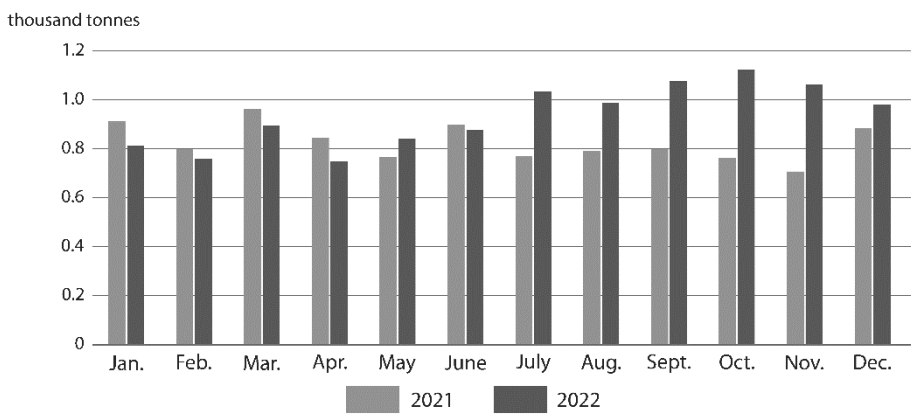
the segment between individual nodes, calculated by the Haversine formula (distance). As a result of the developed algorithms and combined data sources, the following was obtained:

- 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.

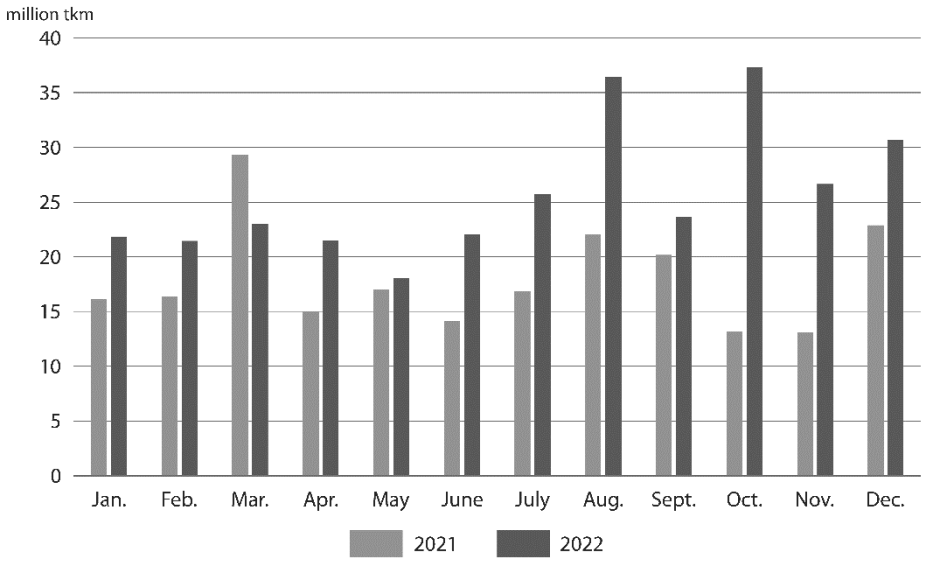
Figure 5. Cargo turnover in the port of Szczecin by month



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:

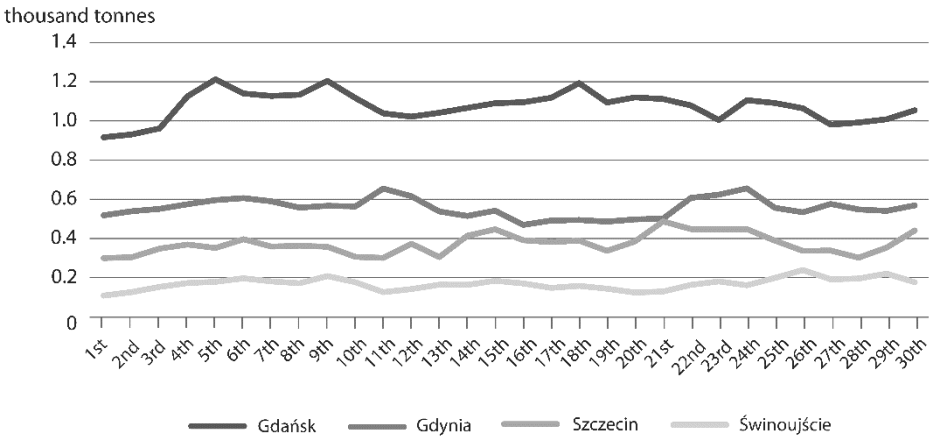
- variables, among others: NO_x emission (nitrates, nitrites); SO_x emission (sulphates, sulphites); CO_2 emission (carbon dioxide); PM emission (particulate matter);
- breakdowns: time (day, month, quarter, year); spatial: ports located on the coast of Poland; means of maritime transport: by type, by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

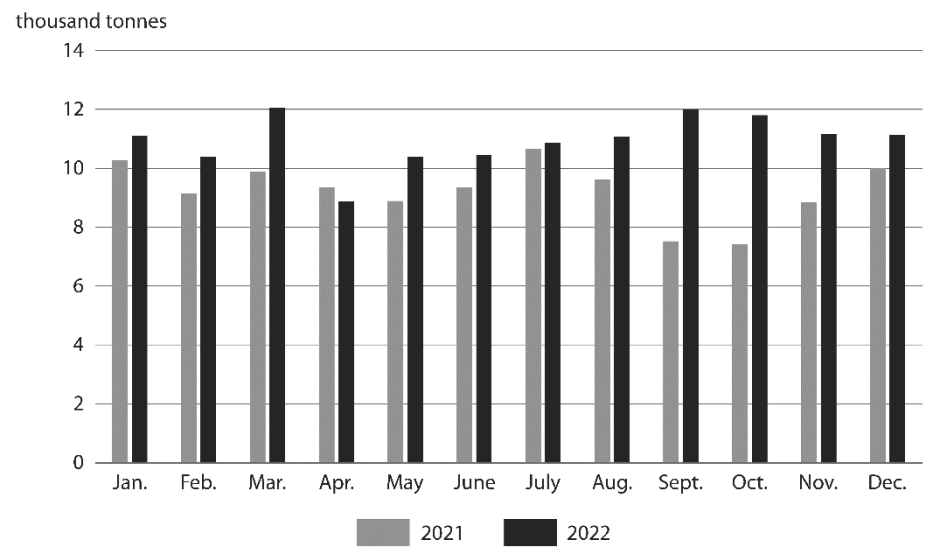
Figure 7. Daily CO_2 emission in November 2022 by seaport



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);

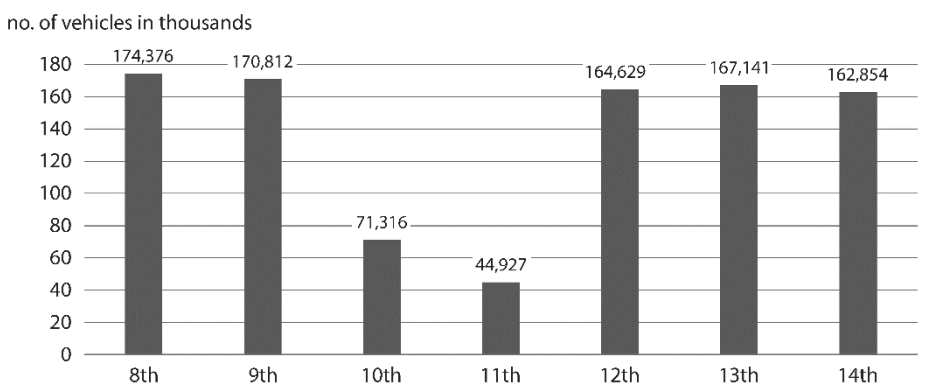
- categories of entities/vehicles according to the Euro emission class (0–6) – European emission standard specifying the permissible emissions in new vehicles sold in the EU and the European Economic Area.

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.

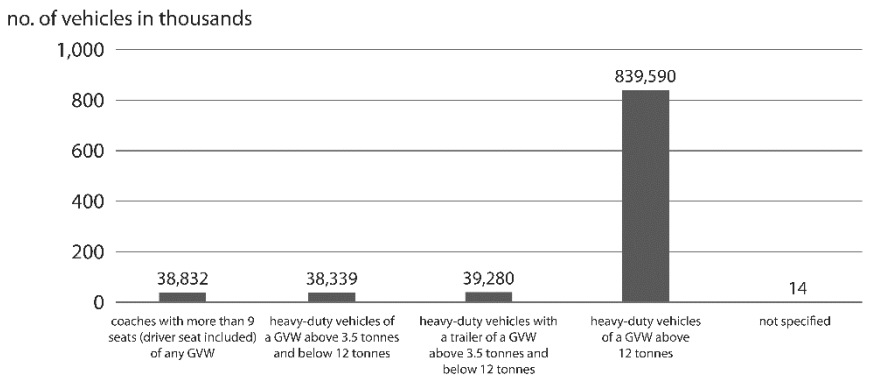
Figure 9. Daily traffic volume on the road network covered by the e-TOLL system by number of vehicles, 8th–14th December 2022



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

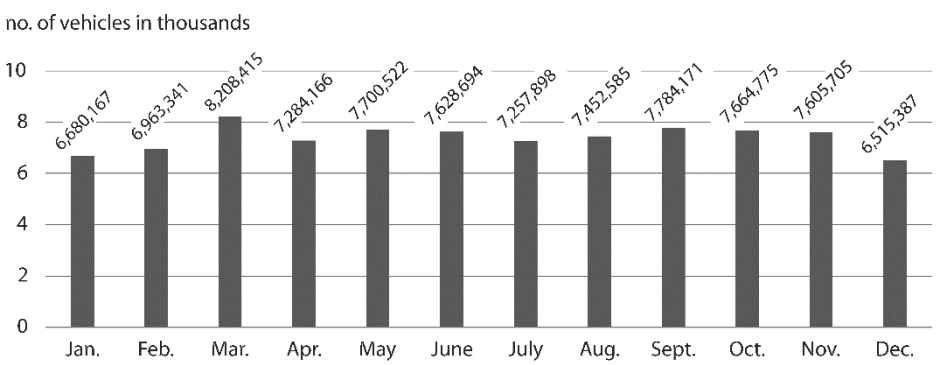
The daily transactions in the analysed week ranged from 3,924,247 (8th December 2022) to 996,293 (11th December 2022). The total number of records from transactions amounted to approx. 20 million. The e-TOLL system covered one million vehicles travelling on the toll road network in the presented week. 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.

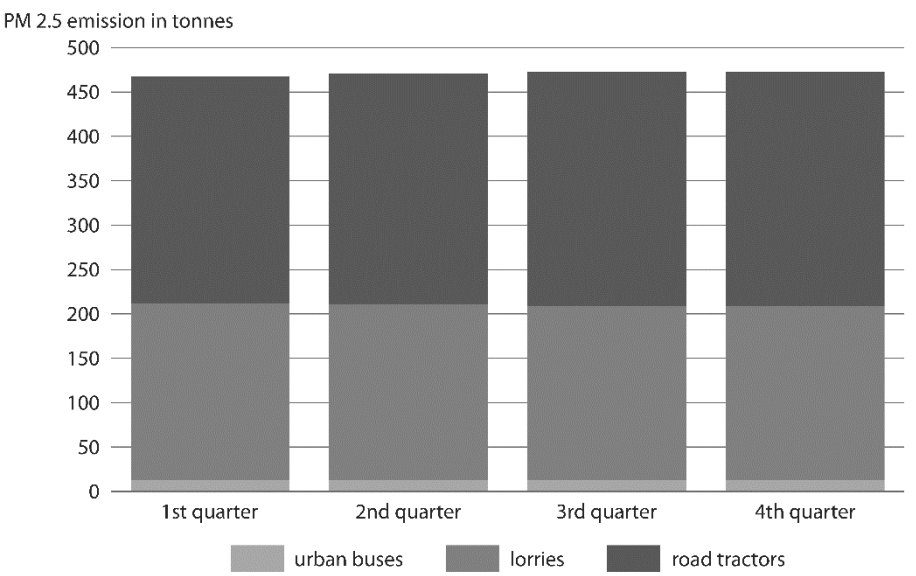
The following breakdowns are defined for the output variables:

- 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

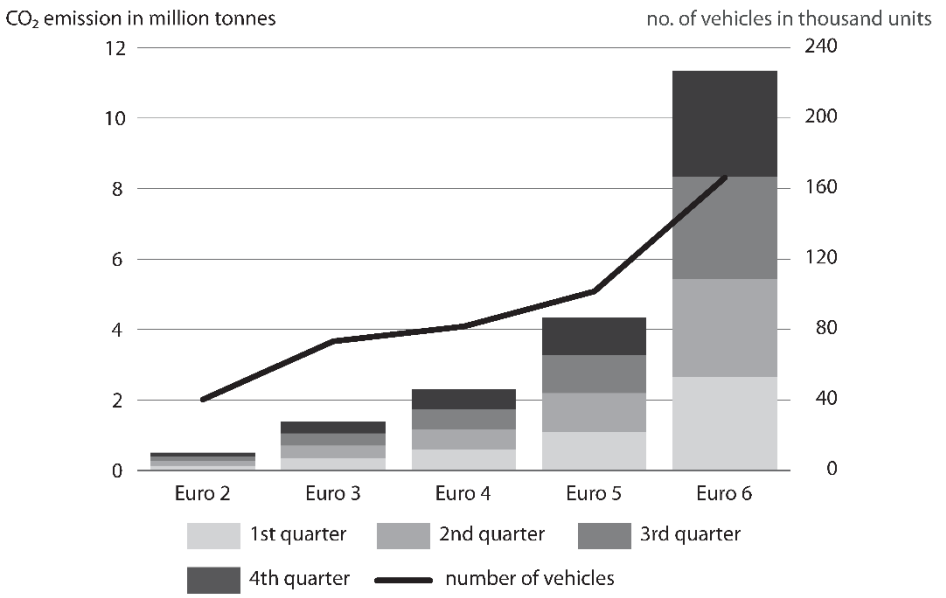
The data presented in the TranStat application show the volume of emissions generated by road transport on an annual and quarterly basis, broken down by type of pollution (Figures 12 and 13). These data are presented for individual categories of vehicles, detailing such vehicle characteristics as the 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.