Statistical Journal of the IAOS 40 (2024) 421–434
DOI 10.3233/SJI-230100
IOS Press

Optimising port arrival statistics: Enhancing timeliness through Automatic Identification System (AIS) data

Nele van der Wielen*, Justin McGurk and Labhaoise Barrett
Central Statistics Office, Mahon, Cork, Ireland

Abstract. Today, there is a greater demand to produce more timely official statistics at a more granular level. National Statistical Institutes (NSIs) are more and more looking to novel data sources to meet this demand. This paper focuses on the use of one such source to compile more timely and detailed official statistics on port visits. The data source used is sourced from the Automatic Identification System (AIS) used by ships to transmit their position at sea. The primary purpose of AIS is maritime safety. While some experimental statistics have been compiled using this data, this paper evaluates the potential of AIS as a data source to compile official statistics with respect to port visits. The paper presents a novel method called "Stationary Marine Broadcast Method" (SMBM) to estimate the number of port visits using AIS data. The paper also describes how the H3 Index, a spatial index originally developed by Uber, is added to each transmission in the data source. While the paper concludes that the AIS based estimates won't immediately replace the official statistics, it does recommend a pathway to using AIS-based estimates as the basis for official port statistics in the future.

Keywords: Official statistics, maritime, ports, AIS, H3 spatial index

1. Introduction

National Statistical Institutions (NSIs) are trusted to produce high quality statistics to generate independent insight for all. Traditionally NSIs have relied on surveys to produce official statistics but more recently the trend is to leverage administrative data in Government databases to meet the ever increasing demand for statistical information on a broader range of subjects, with increased timeliness and at greater granularity [1]. Today, in a world where global data creation is projected to surpass 175 zettabytes by 2025 [2], novel data sources are providing new opportunities for NSIs to meet this rising demand with constrained budgets [1,3,4].

To exploit the value of these novel data sources (including big data), NSIs, and international statistical organisations such as Eurostat and the United Nations Statistical Division (UNSD) are working together and collaborating to explore how the potential of these new data sources for official statistics can be exploited [5,6]. Studies to date, including mobile network data, smart city data [7] and web scraping for labour market statistics [8] have shown the potential of these new data sources in official statistics.

Ocean shipping is the main mode of transport for international trade activities [9] making up 80% of global trade in terms of volume [10]. The production of maritime indicators is vital for the understanding of trade activity. Automatic Identification System (AIS) is the international system for tracking ship movements which was originally developed by the International Maritime Organization (IMO) [11]. Ships of a certain size and type are required to carry an AIS transponder that transmit location and other key information in real-time as signals. This paper focuses on near real-time shipping data generated by AIS that can enable the faster compilation of economic and maritime indicators [4].

Data generated by AIS is an example of big data due to the large amount of signal data generated by ships

*Corresponding author: E-mail: Nele.vanderwielen@cso.ie.

globally. It offers potential for statistical analysis going beyond its original goal of maritime safety [12]. The data can be used to generate more timely maritime statistics on port calls, berth calls, carbon emissions, trade routes or port congestion. The UNSD has set up an AIS task team to evaluate the potential of this data [6]. The European Statistical System also has a similar initiative as part of its ESSnet Big data project [5].

This paper focuses on calculating port visits in Ireland using AIS data. The article resulted from a collaboration between the Central Statistics Office (CSO) and UNSD. This partnership was formed to explore whether AIS data could be used to generate official maritime statistics. This is the first time the CSO has used AIS data to produce faster maritime indicators. Ireland, as an island relying heavily on sea transportation, is ideally placed to capitalise on the potential offered by faster AIS-based maritime indicators.

Several NSIs have published experimental statistics with this data source in order to reduce the time lag between the reference period and the publication of official maritime statistics. The Office for National Statistics in the UK has published more timely shipping indicators [13,14]. Statistics Denmark publishes regular experimental statistics on port calls to Danish Sea ports [15]. Statistics Netherlands considers the data source as having significant potential [16]. The UNSD has also produced a handbook providing a snapshot of case studies from around the world including Ireland [12]. The case studies include port calls, tracking of fishing fleets and CO2 shipping emissions which were carried out by the CSO, other NSIs, the United Nations and Eurostat. However, overall, the usage of this data source by NSIs is still limited and the abovementioned projects remain classified as experimental in nature.

There is a lack of studies that have compared outputs from this data source with official statistics to assess its suitability for official production [4]. This paper addresses this gap by comparing port call statistics compiled from AIS sourced data with CSO's official statistics compiled using administrative data sources.

This paper is laid out as follows. The next section is a literature review on the usage of big data in official statistics. Section 3 covers a description of the data sources and the deployment of the H3 spatial index on AIS records and the data processing environment and includes an overview of the United Nations Global Platform. Section 4 presents a new method, Stationary Marine Broadcast Method (SMBM), that can be used to efficiently identify stopped ships that can subsequently be used to identify port visits. In Section 5 the results are presented. The paper is then concluded with a discussion of results and some final remarks.

2. Literature review

In recent years, the usage of big data in official statistics has been debated across NSIs and has presented both challenges and opportunities for the field of official statistics. NSIs acknowledge that big data cannot be ignored in this modern world. Ignoring the potential of big data can question the relevance of NSIs and risks pushing them out of the information market [17].

A common view among researchers is that traditional surveys and big data can and should be used together to maximise the value of each [21]. For example, big data can complement existing surveys and can fill gaps in coverage, providing insights into hard-to-reach populations. For example, NSIs have started using big data to complement traditional statistics, including on migration [22]. One key benefit is the ability to enhance the timeliness of data, providing more up-to-date information. Traditional statistical methods often involve time-consuming data collection and processing. The usage of big data, like AIS, offer real-time or near-real-time data to produce official statistics. Harchaoui and Janssen [18] have shown how the U.S consumer price index could be produced in a timelier manner using prices available online, combined with web scraping technology.

Big data sources enable statisticians to generate analysis quickly to provide more up-to-date information for decision-makers. Gallego and Font [19] successfully developed a methodology for the early detection of reactivation of tourist markets to help mitigate the effects of the COVID-19 crisis, using Skyscanner data. With its increasing availability and completeness, AIS data can provide more detailed and timely port statistics compared with administrative data sources [20]. The richness and variety of big data sources enable NSIs to gain a more comprehensive understanding of complex phenomena, leading to more detailed and granular analyses [23]. Analysis of big data can produce new findings about variables that have not been previously captured by official sources [21].

However, despite the potential benefits, the integration of big data into official statistics is not without challenges. Methodological issues, such as ensuring representativeness and addressing selection biases, pose significant hurdles [24]. Understanding the methods be-

hind the sampling, measuring, and assembling of big data is critical to extracting value from big data [25]. Many big data sources, including AIS data, are not collected with statistical purposes in mind, leading to issues of reliability and comparability.

The quality of big data sources always needs to be critically assessed. For example, Bähr and his colleagues [26] have developed a framework to identify the possible sources of error when dealing with sensor data.

Privacy concerns are another critical consideration. Big data often contains personally identifiable information, raising ethical questions about the responsible use of such data [27]. Striking a balance between harnessing the power of big data and protecting individual privacy is a complex challenge that NSIs must navigate.

This paper presents a novel methodology for producing official statistics based on geospatial big data and proves it reliability and accuracy by comparing it against administrative data sources in Ireland.

3. Data sources and processing

In this section, we describe the AIS system, the data it generates and how it is collected and enhanced for statistical purposes. We then describe the UNGP before describing the H3 spatial index and how it is deployed on the data records to enable fast and efficient searching and sub setting.

The official statistics used for comparison purposes are the 'Official Statistics of Port Traffic' published by the CSO, Ireland. These statistics are compiled quarterly based on returns made by harbour authorities [28].

The reference years for the analysis are 2021 and 2022.

3.1. About AIS

AIS is a receiver and transmitter system used by ships to transmit their position. It is used as a safety tool that functions with a combined satellite and ground receiver system. AIS allows other ships, coast guards, and emergency services to be aware of the ship's current position in both coastal and ocean traffic.

The AIS data source consists of the collection of AIS messages from ships with information including three categories: 1) *static data* (information on vessel characteristics like IMO number), 2) *dynamic data* (information on ship movements like latitude and longitude position, speed, heading, and time in terms of UTC when these were observed) and 3) *voyage-related data* (information on current voyage e.g. destination, draught) [12].

Despite the relative newness of this data source, there is already extensive evidence that it can be used to produce accurate shipping statistics [4,12]. This is helped by the rigour associated with AIS as a legal requirement for maritime safety of larger vessels. The legal framework for AIS operation in Irish waters is specified in legislation, including "S.I. No. 573/2010 – European Communities (Vessel Traffic Monitoring and Information System) Regulations 2010" [29]. In addition, there is a detailed outline on international obligations for AIS use in the International Maritime Organisation (IMO) guideline document [30]. While this legal framework ensures that AIS records are available, extracting information from these big datasets to produce official statistics is still a challenging task [31]. Hence, for NSIs a key question remains how the quality of official statistics can be guaranteed when using novel data sources. As Kitchin and Stehle [7, p. 123] summarise, "maintaining data quality; gaining and validating methodological transparency and ensuring long-term stability in the measurement and production processes" are key challenges when using big data.

AIS provides for the collection of homogenous, near real-time, shipping data that can be used to produce maritime indicators much more frequently than traditional methods. This a real advantage when producing timely statistics. Compared with other administrative datasets, AIS has a broader vessel coverage, and these datasets are widely available and well documented [12, 21]. Furthermore, the automated, centralised nature of the process adds no response burden to ports authorities or shipping companies.

There are limitations and sources of error contained in AIS data. For example, the data can contain noise which can lead to errors and the occurrence of such errors may differ depending on the ship type [33]. Poor quality of AIS equipment can lead to corrupted AIS messages and the location or time data attributes may not be accurate if the coverage of the AIS receiving stations near the relevant port is weak or if AIS transponders are turned off when ships are in port. This all can lead to AIS data showing unrealistic tracks [34]. In addition, manually entered data is prone to human error [35], hence manually recorded information like vessel type can often be incorrect. However, to overcome this issue, it is possible to link AIS data to ship registers which hold official vessel related information.

3.2. The AIS data source on the UN Global Platform (UNGP)

For this project, the AIS sourced data was accessed through the United Nation Global Platform (UNGP) which holds a global repository of live and archived AIS sourced data [36]. The data is provided by ExactEarth who combine their own satellite data with terrestrial data from FleetMon. In addition to location, bearing and navigation status, the data includes unique identifier information on International Maritime Organisation (IMO) number and Maritime Mobile Service Identity (MMSI) number.

The UNGP data source is also enriched with the IHS Shipping Registry. Incorporating SeaWeb and Lloyd's Register of Ships (published since 1764), the IHS Shipping Registry provides detailed information on all self-propelled and seagoing merchant ships. Among the information included in the registry are IMO number, MMSI number, ship name, ship type, cargo type, ownership, registration, tonnage, dimensions, and propulsion. Based on the IMO and MMSI number, the AIS data source can be linked with the IHS ship register data.

3.3. The H3 spatial index

The data is further enriched by adding a spatial index to every AIS message record to enable detailed and efficient geographical analysis. The spatial index system used is H3, a system originally developed by Uber, containing sixteen resolution levels, each level comprising cells that cover the earth's surface with a corresponding H3 Index [37,38].

H3 Index, a hierarchical geospatial index, is based on cells shaped as hexagons. Every hexagonal cell, up to the maximum resolution supported by H3, has seven children cells below it in this hierarchy. The hexagons have the property of expanding rings of neighbours approximating circles (see Fig. 1).

By way of illustration, Brazil is about the size of two H3 cells of resolution zero. The indicative cell size of H3 hexagons is shown in Table 1.

On the UNGP the H3 Index is added to each AIS message for all sixteen resolution levels. The UNGP uses this index to store the data in parquet files. This allows for efficient querying and updating of AIS data within the platform.

The main benefit of the H3 Indexing systems is that any position on the earth, and by extension any AIS message, can be assigned to a H3 cell at a given resolution via the cell's corresponding H3 Index. Using the H3 Index allows analysts to go beyond the ships actual GPS location to a more generalised location. The H3 Index also allows areas, such as ports and ship neighbourhoods, to be defined in terms of H3 indices.

Table 1
Indicative average H3 cell area by resolution

H3 index resolution	Average hexagon cell area hectare (Ha) (1 Ha = 10,000 SqM)
0	435,745,000
1	60,979,000
2	8,680,000
3	1,239,000
4	177,000
5	25,300
6	3,610
7	516
8	73.7
9	10.5
10	1.5
11	0.2
12	0.0307
13	0.0044
14	0.0006
15	0.0001

4. Stationary Marine Broadcast Method (SMBM)

The project developed a novel method, "Stationary Marine Broadcast Method" (SMBM), to estimate the number of port visits in Ireland using AIS data. The Python code and pseudo code for the newly developed method is available on the United Nations Global Platform Git Hub which can be accessed on request [12].

The basic idea behind the SMBM is that a ship needs to be stationary for an extended period of time in a port in order to load and unload. If a ship is stationary within a port area for long enough and has multiple AIS messages around the same zone, then it is highly likely to be a port visit. In a nutshell, under the SMBM a port visit was estimated from stopped ship events that occurred within a port area or polygon. A stopped ship event was based on a ship's speed being reduced to zero within the port or area of interest followed by multiple AIS messages within a neighbourhood of the initial zero speed observation.

This method takes advantage of the H3 Index system to define an area or 'neighbourhood' around a stationary ship that can be used to identify when the ship leaves the neighbourhood.

Once a ship was identified as stationary or stopped (the triggering event) subsequent AIS signals were assessed to determine if the ship remains within the

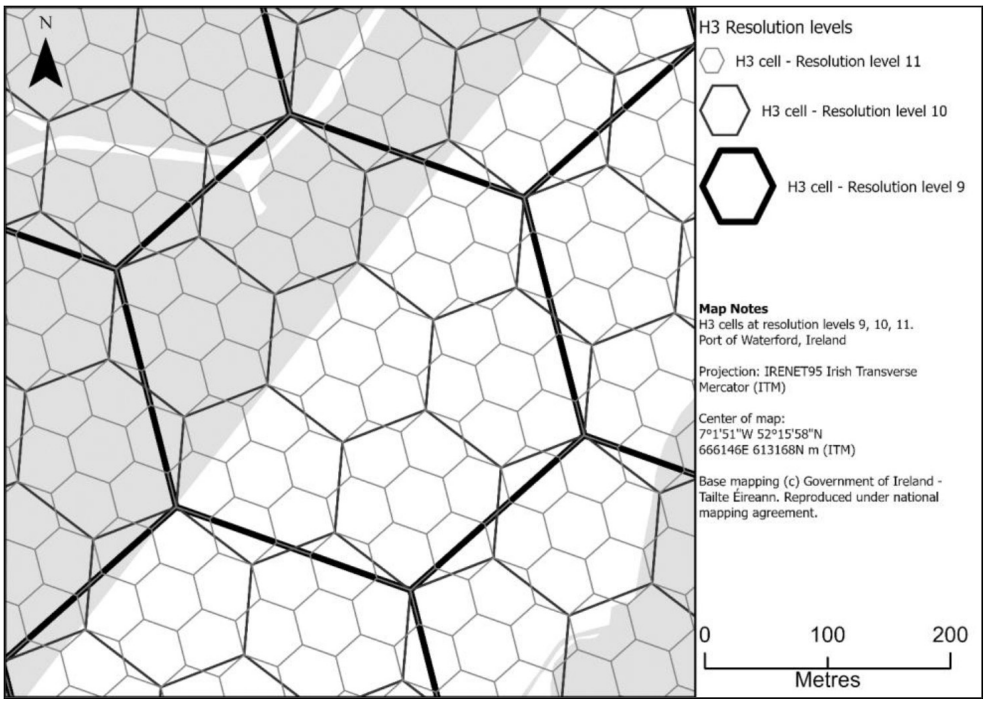

Fig. 1. H3 – Parent H3 Cell with Seven Children.

Table 2
K-depth and number of cells for k-ring

K depth	Number cells in k-ring
0	1
1	7
2	19
3	37
4	61
5	91
6	127

'neighbourhood' or has left the 'neighbourhood' defined by the triggering event (the escape event).

A ship was determined to no longer be in the neighbourhood (i.e. the escape condition was met), if its H3 Index was not within the set of H3 Indexes that defined the neighbourhood of the original triggering event (stopping). To define the neighbourhood of a stopped ship, we used a H3 k-ring (griddisk) function that returns a set of H3 Indexes that surround a ship's current H3 Index location. The term "k-depth" refers to the number of rings used to define the neighbourhood for testing the escape condition.

Depth 0 was defined as the origin index, k-ring 1 was defined as k-ring 0 and all neighbouring indices, and so on, an arithmetic series (see Table 2).

This function introduced two hyper-parameters: *k-depth* of the k-ring and the *H3 resolution* to use for the k-ring.

For the SMBM a H3 Index resolution of level 10 was used where the width of a cell is approximately 150 m. This width also approximates the size of ships being studied in this article. A K-depth of three was used as this takes account of errors in reporting position, minor changes in position due to tide and current, and small operational movements of the ship (see Fig. 2).

The choice of k-depth and H3 index was based on a sensitivity analysis alongside visual assessments of the spatial mapping. The choice represents a balanced trade-off between k-depth and H3 index, reflecting typical ship and port characteristics.

All in all, the methodology and data processing for the SMBM can be broken down into eight steps which are summarised below in detail.

4.1. Step 1: Generation of port polygons

For this release we created port polygons for six main ports in Ireland, namely: Bantry Bay, Cork, Drogheda, Dublin, Rosslare, and Waterford. The port polygons defined in this article served several purposes. Firstly, port polygons define the area in which a port visit can take place and assign a visit to a particular port. Secondly, the polygons are used to create a study area that acts as spatial filter to reduce the amount of AIS data to be

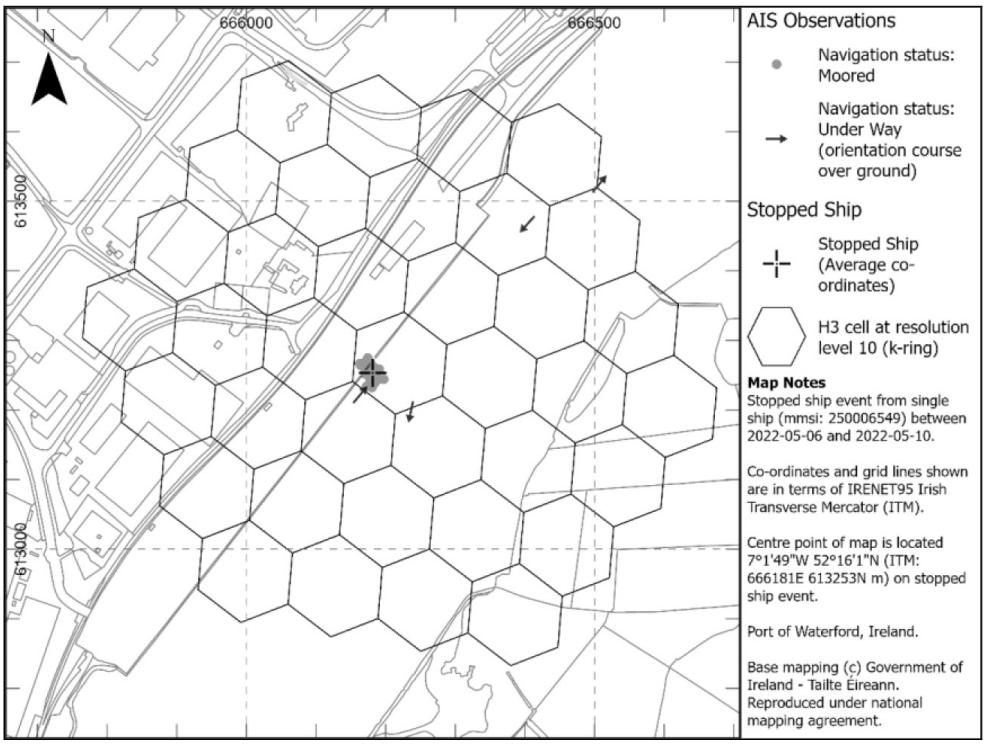

Fig. 2. H3 Neighbourhood: K-depth 3 and H3 Resolution 10 (example for Waterford Port).

processed, as AIS messages which fall outside the study area are irrelevant.

The port polygons in Ireland do not overlap and were digitised with GIS software in terms of Irish Traverse Mercator (EPSG 2157). When digitising the port polygons, care was given to defining an area whereby an ocean-going ship upon entering would be expected to be visiting the port. Ships are confined to water areas, and the amount of land covered by these polygons is irrelevant as any onshore AIS report is most likely an error arising from GPS signal error, a transmission error or rounding. As a result, the port polygons did not represent an area corresponding to port activities but rather an area that contains the port.

It is worth noting applicability of the SMBM method outlined in this paper is not restricted to Ireland, it can be used for any region or country in the world.

4.2. Step 2: Data reduction

The data source contains a large volume of information not relevant for analysis. Therefore, it is useful to discard this information. To reduce the volume of data to be analysed, a study area was first defined based on a 10 km box of port areas (see Fig. 3). The AIS data was filtered to only include signals within the study area, reducing the amount of data to be processed by about 80% from the initial fetched AIS data.

4.3. Step 3: Adding sorting variable

Working with the reduced dataset, as a next step, the timestamp value for each AIS observation was used to create an integer variable using UNIX timestamp. The UNIX timestamp is the integer number of seconds from January 1^{st} 1970. This allows the data to be sorted in a natural time order and is also used for the calculation of event duration.

4.4. Step 4: Creating ship list in study area

Next, a set of ship MMSI values were created within the study area.

$$S = \{s_1, \ldots, s_n | s_i \in \textit{MMSI used in study area}\} \quad (1)$$

4.5. Step 5: Identifying stopped ships

For each element of ship MMSI set S, we got the set of observations associated with this ship. Each observation was a list of data items consisting of at least the following:

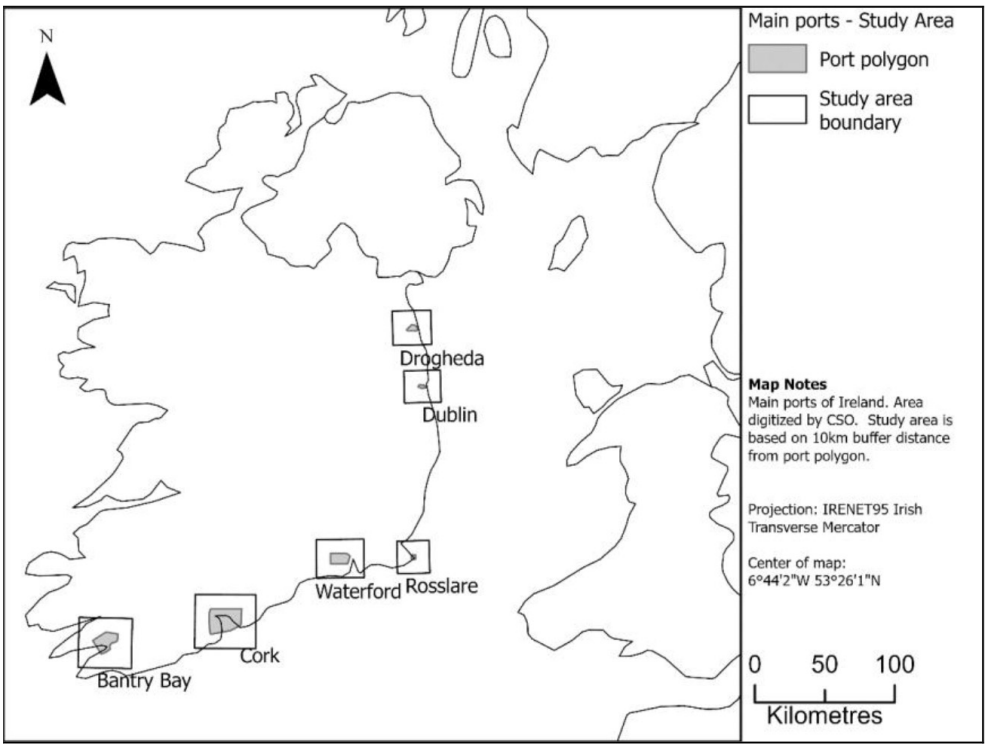

Fig. 3. Study Area: 10 km Bounding Box Buffer.

- MMSI (mmsi)
- UNIX time (unix)
- Timestamp in terms of UTC (ts)
- Navigation Status (nav)
- Speed Over Ground (sog)
- Latitude (lat)
- Longitude (lng)
- H3 Index at a given resolution (h3)

$$A = \{a_1, \ldots, a_n | a\,[mmsi]_i = s_i \in S\} \tag{2}$$

The set of observations A was converted to a list by sorting on UNIX time. A trigger object (T) was a list object that holds data on the triggering event MMSI, IMO, timestamp, H3 Index, and lists for each observation associated with the triggering event for latitude, longitude, UNIX time, and navigation status.

The triggering event T was defined as the case when an observation speed over ground (sog) was zero and the length of T was zero.

When this condition was met an observation was said to be the triggering observation. Subsequently, data from the triggering observation were put into the trigger object. The observation before the triggering event P was additionally saved.

$$(a\,[sog]_i = 0 \wedge length\,(T) = 0) \begin{cases} Populate\,(T) \\ P = a_{i-1} \end{cases} \tag{3}$$

By using the value of the H3 Index of the triggering event we created a set of H3 Indexes that define a neighbourhood by using the K-ring function.

$$K = \{h_i | h_i \in KRing\,(T\,[h3])\} \tag{4}$$

This enabled subsequent observations to be tested on this neighbourhood via:

$$a\,[h3]_i \in K \tag{5}$$

If this returns as true, then the observation was still associated with the trigger event and the dynamic data of location, navigation status, UNIX time was updated by appending the current observation to the trigger list object associated with that data item.

$$a\,[h3]_i \in K\,\{Append\ data\,(T) \tag{6}$$

If this returns as false this observation is said to be an *escape event* as the ship is no longer within the neighbourhood and the escape condition is met. Once the escape condition is met, a stopped ship event (E) was created from the trigger object (T), the escaping observation (a_i), and the prior observation (P), and saved. The trigger object was then flushed of data and

returned to an empty state

$$a\,[h3]_i \notin K \begin{cases} Save\,(E) \\ Flush\,(T) \end{cases} \tag{7}$$

Because we were iterating through observations of a ship sorted on time, it meant the list data stored in the trigger object for UNIX time, latitude, longitude, observation status was also sorted on time as the data was appended as each observation was tested against the k-ring of triggering observation. This meant when we reached an escape condition, we could access the observation data corresponding to the triggering event, as it will be the first list item in the list data *(T[item][0])*. Likewise the last list item on the list corresponded to the final observation within the neighbourhood *(T[item][-1])*. The current observation, which is an escape event, is outside the neighbourhood.

We then used these list items to create statistical aggregates for a stopped ship event. In the case of latitude and longitude, the arithmetic mean allows for an average co-ordinate value as well as standard deviation for the latitude and longitude values for the stopped ship event. In the case of navigation status, we used the mode of the list item *(T[nav])* to identify the most used value.

Note, if the study area were to cross the 180°E/W line in the Pacific Ocean, arithmetic mean would not produce sound results and more sophisticated means of calculating an average co-ordinate would need to be used. This was not necessary in our case, as the ports of interest do not cross this line.

4.6. Step 6: Creating the stopped ship event

Using UNIX time data from the trigger object T, prior observation P and the current observation a_i we estimated upper and lower duration times for a stopped ship as:

$$\begin{aligned} &Upper\ time\ estimate = a\,[unix]_i - P\,[unix] \\ &Lower\ time\ estimate = \\ &\quad (T\,[unix]\,[-1]) - (T\,[unix]\,[0]) \end{aligned} \tag{8}$$

A stopped ship event was created from data in the trigger object. Here a stopped ship event with 10 or less observations was flagged as invalid and standard deviations were not calculated.

4.7. Step 7: Linking stopped ships to ship register

The next step was to link the AIS data to the IHS ship register based on the IMO and MMSI number. Firstly, a link was made via the IMO number, secondly the remaining unmatched recodes were attempted to be matched via the MMSI number.

4.8. Step 8: Linking to port areas

The data was then linked to port area by doing a spatial operation known as a spatial join. By using the average coordinates of a stopped ship, a point geometry was created. Any point that was within a port area polygon gained the attributes of the port area thus completing the process by linking stopped ship events to a port area.

4.9. Step 9: Filtering on relevant ship types

Finally, only activities of cargo vessels, car ferries and other passenger vessels over 250 gross tonnage were included in this paper to align with Eurostat Maritime Transport [39] and to allow comparison to CSO's official statistics [28]. The following vessels were excluded:

- Fish-catching vessels
- Fish-processing vessels
- Vessels for drilling and exploration
- Tugs
- Pusher craft
- Research and survey vessels
- Dredgers
- Naval vessels
- Vessels used solely for non-commercial purposes.

5. Results

In this section, the AIS based estimates are presented and compared with the official statistics compiled from administrative data sources for six Irish ports.

In 2021, a total of 11,552 vessels arrived in the six ports according to CSO's official statistics. In comparison, the AIS based results estimate 11,889 vessels called to the same six ports, a difference of 2.9%. In 2022 11,399 vessels arrived in Ireland based on CSO's official statistics compared to 12,652 based on AIS data. Figures 4 and 5 compare the two sets of estimates over 4 quarters in 2021 and 2022 and indicate strong alignment in the trend across the quarters. They show that the AIS-based estimates follow the same trend as the official statistics, but the AIS-based port call figure exceeds official statistics in most quarters. This is in line with findings from Arslanalp et al. [4].

Two plausible explanations for this overestimate could be i) AIS based estimates offer better coverage than the official data compiled from administrative data and/or ii) AIS based estimates may include an element of overcounting. The Stationary Marine Broadcast

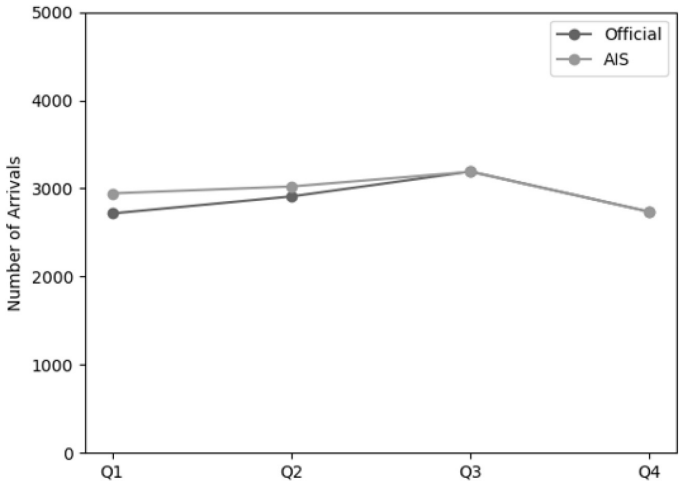

Fig. 4. Comparison of official versus AIS-based Port Calls, Q1 2021 to Q4 2021.

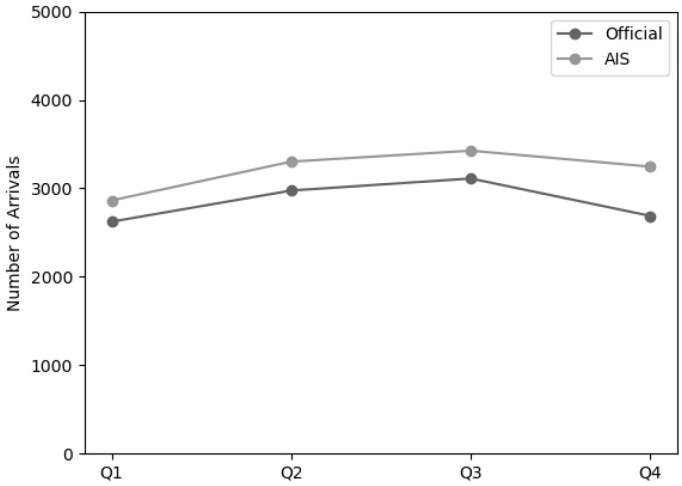

Fig. 5. Comparison of official versus AIS-based Port Calls, Q1 2022 to Q4 2022.

Method risks over-counting arrivals when a ship stops more than once within a port area. A vessel that arrives and stays for several days within a port may visit several locations within the port polygon and a count will occur for each stopping event, for example, a vessel arriving in port, unloading its cargo, and then moving to another location to take on fuel or stores. Another scenario may be navigation within the port area and waiting for a berthing location to become available. Deduplicating overcounted observations using IMO/MMSI poses challenges due to the diversity in ship journey patterns and schedules, as well as differences in port characteristics. Implementing blanket deduplication rules might exclude valid visits, especially considering variations among ship types, such as high-frequency ferries vs container ships. In this context, we opted to retain the overcount to ensure a more accurate representation. While there is a risk of redundancy, maintaining the data allows for more comprehensive analysis.

The overall trends are very similar when looking at individual ports. However, the range of differences vary between the AIS based estimates and official statistics when looking at individual ports, see Table 3. Only small differences are observed for Bantry Bay. However, larger, and possibly more volatile differences are observed for the busiest port, Dublin.

Figure 6 provides a comparison by type of vessel, showing broadly similar breakdowns as would be expected. In 2022 cargo ships accounted for 76% of the total port calls in Ireland according to the administrative data, while both AIS-based port call methods show that cargo ships accounted for nearly 77% of visits. The

Table 3
Comparison of official versus AIS-based port calls by ports, 2021

	2021		2022	
	Official	AIS	Official	AIS
Main Irish Ports	*11,552*	*11,889*	*11,399*	*12,652*
Bantry Bay	10	7	8	8
Drogheda	411	239	252	296
Dublin	7219	7451	7402	8352
Cork	1635	1644	1386	1809
Rosslare	1843	2098	1915	1766
Waterford	434	450	436	421

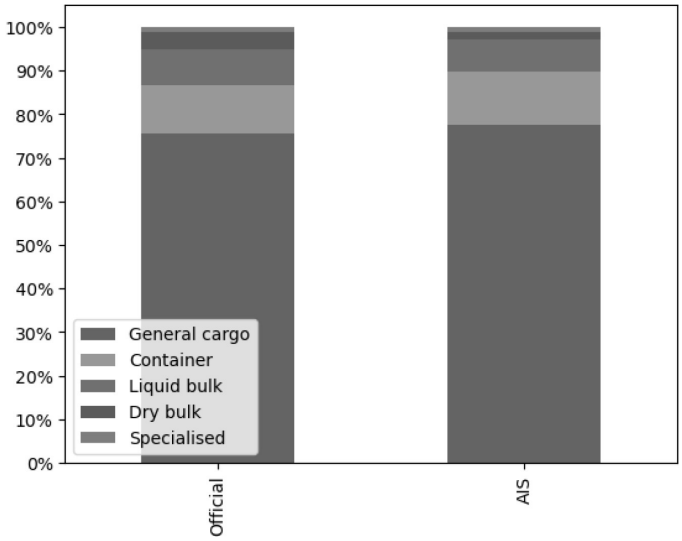

Fig. 6. Port calls by type of Cargo, 2022.

smallest share of arrivals were dry bulk and specialised ships, and Fig. 6 highlights that AIS data also captures this breakdown correctly. The same was found for 2021. The key advantage of these AIS-based statistics is that they track the overall trend well and can be produced in real-time to provide early maritime estimates.

Figures 7 and 8 show the duration of stopped ships by vessel type. For comparison purposes we split general cargo into Roll-on Roll-off (RoRo) and not RoRo. The duration of port visits is not available in the administrative data sources used in the compilation of official statistics and Figs 6 and 7 have been included to show the potential of AIS data.

The assumption was that 264 hours (11 days) is a reasonable upper limit for a port visit, and as such, any port stops above this limit and those related to stopped ship events located near a dry dock/ship repair yard were excluded from the compilation of the AIS based estimates. Based on investigation of ships with long durations in port, a stay past 11 days indicates a reason other than economic activity, such as detainment. Figures 7 and 8 show the short turnaround time for RoRo vessels, while container and dry bulk vessels stay in port for a longer period.

6. Discussion and concluding remarks

The use of big data in official statistics is still relatively recent and can provide challenges. However, novel data sources provide new opportunities.

The AIS based data source is fairly new. A significant benefit of this data source is that there is one global standard, providing for the opportunity of easily compiled coherent and standardised maritime statistics. International standards are currently being developed based on experiences using this data in the UNGP and other projects.

The analysis in this paper does produce differences between the AIS based estimates and official statistics for the two years investigated. As noted earlier, official port statistics in Ireland are compiled from summary

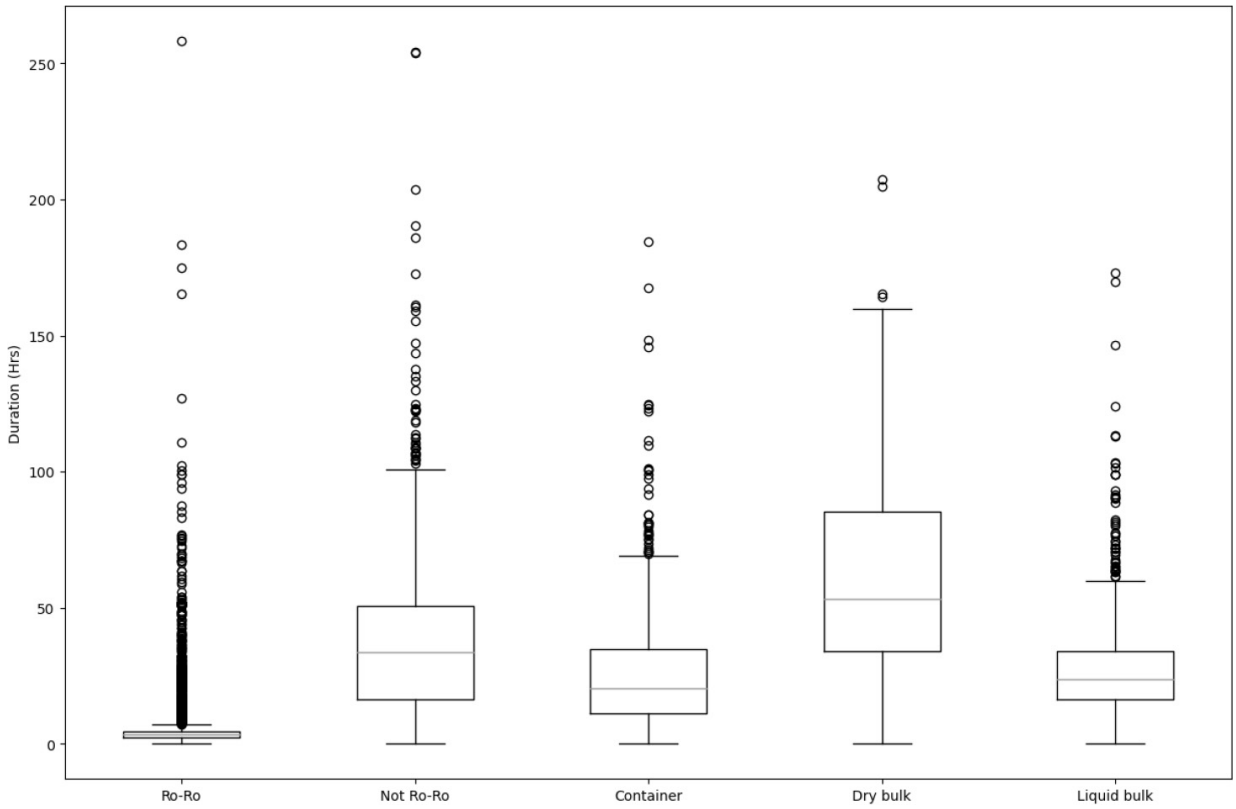

Fig. 7. Duration of Stopped Ships under 264 hours, 2021.

data provided by each port authority on a quarterly basis, and as such, detailed disaggregation beyond what is received is not possible without significant burden on the part of the port authorities. There are two possible reasons (or a combination there of) for the higher estimate in AIS based statistics compared to the administrative data. The first reason is that the difference could be explained by a higher coverage in the AIS based estimates – indicating that the current official statistics are underestimating the number of port calls. The second reason is that there may be an element of double counting included in the AIS based estimation procedure indicating over-coverage errors with the AIS based estimates. In all likelihood the differences are probably explained by some combination of these error types.

The AIS based source, like many novel data sources, will enable greater temporal granularity (daily and hourly statistics are possible if required). The AIS based data source also provides for other statistical opportunities including:

- Trade Routes: generating statistics on international trade routes,
- Port Congestion: calculating the waiting times of a specific vessel at the port before entering the berth,
- Other informative statistics can be compiled based on the mobility behaviour of ships (e.g., fishing boats will probably exhibit certain behaviours when actively engaged in fishing),
- CO2 emissions: calculating the impact of emissions of CO2 on the environment.

There is significant incentive to exploit the associated benefits for official statistics. However, directly replacing the existing system of official statistics with the AIS based estimates will involve a discontinuity as there is some incoherence between them.

The traditional approach, summary reporting by port authorities, has the advantage that different attribute values can be agreed over time for collection by port authorities. The traditional approach can adapt to changing needs. There is significant value for an NSI to have access to the underlying microdata used to compile reports and reduce the burden on the port authorities. Access to underlying microdata would also provide for more in depth analysis in the differences between the two sets of estimates. This paper recommends that

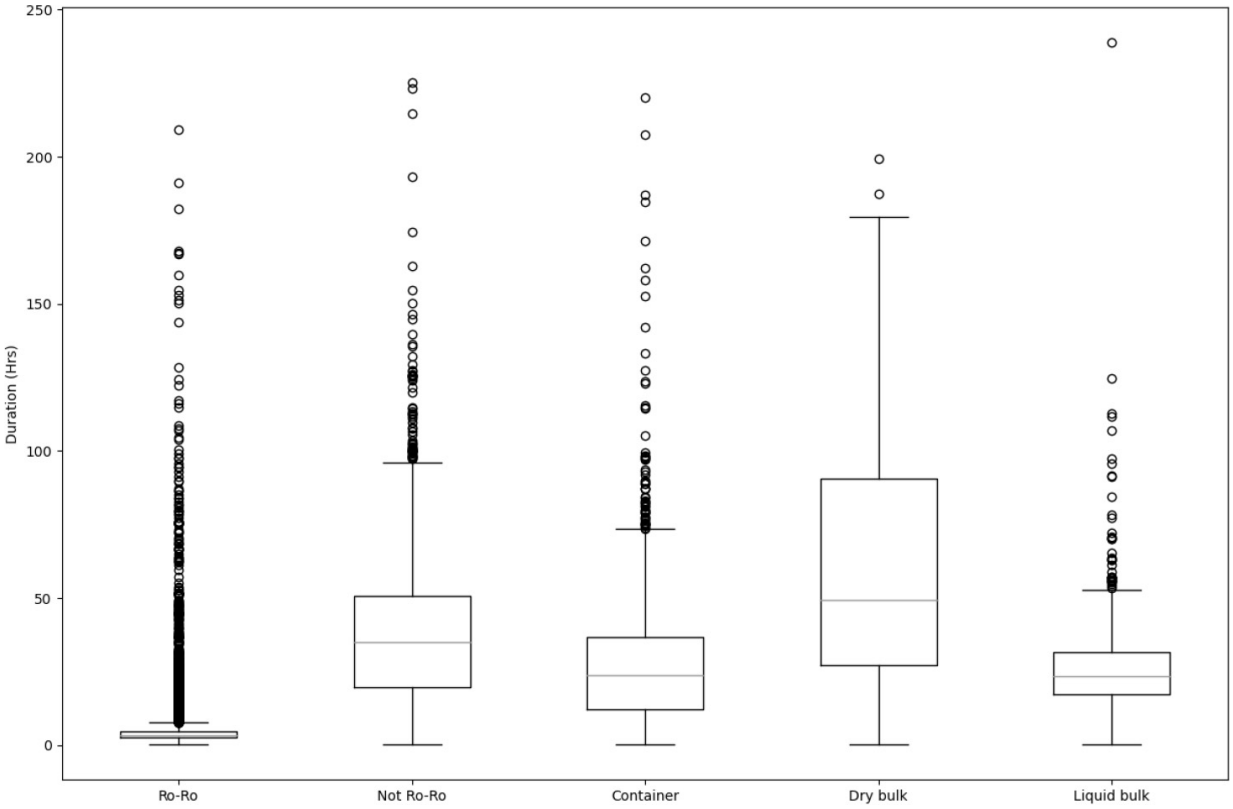

Fig. 8. Duration of Stopped Ships under 264 hours, 2022.

where NSIs do not have access to underlying micro-data to compile statistics that they should seek it. This would be in the spirit of the UN Fundamental Principles of Statistics [40] where NSIs need to be open and transparent in how the compile their statistics.

The methods and approach described in this paper have been designed to be applied in any geographical setting – they should easily translate to different country settings. The H3 Index greatly enhances processing efficiency. The SMBM allows for events to be quickly and efficiently identified from the message data. The paper considers that SMBM may overestimate port calls, and as such, the authors recommend further evaluation to investigate how big a problem this might be, and if it is indeed a problem, if the method itself can be adjusted to mitigate it.

Given the experiences in this project, the authors suggest the following course of action as a migration path to using AIS based data for compilation of official statistics.

The CSO should continue to develop and produce experimental statistics for port traffic in parallel with the official statistics. This will allow monitoring of how both sets of statistics behave over a greater time scale. The experimental statistics can be compiled at a greater granularity (say monthly) and exploit the attributes and possibilities available in the AIS based data sources for new statistical indicators of value. It should promote benefits of experimental statistics to users and seek feedback. Investigate how the traditional reporting used by the port authorities might best be deployed to enhance the new AIS based estimates – particularly where there are information gaps. In the meantime, timely publication of these experimental statistics will serve as early indicators for the official statistics.

Adoption of AIS based estimates worldwide holds the promise of higher quality statistics when we consider the UN Fundamental Principles. Such statistics would be more timely with higher granularity on many dimensions and use a standard method on a standard data source with global coverage that can easily be audited. The whole process can also be easily audited by external parties.

The methods developed and used in this paper are designed to be applicable in different geographical settings, and as such this paper should make a contribution

in signposting the pathway to using the AIS data source for the compilation of official statistics.

Acknowledgments

All authors thank the United Nations Statistics Division for their support, especially Markie Muryawan and Cherryl Chico. Their support, along with help of CSO staff Timothy Linehan, John Flanagan, Paul M. Crowley, Francesca Kay, John Dunne, and Donal Kelly, who made the collaboration between the authors possible.

References

[1] Kitchin R. The opportunities, challenges, and risks of big data for official statistics. Statistical Journal of the IAOS. 2015; 31(3): 471-81. doi: 10.3233/SJI-150906.

[2] Rydning DR, Reinsel J, Gantz J. The digitization of the world from edge to core. Framingham: International Data Corporation. 2018; 16: 1-28. Available from: https//www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. [accessed 2022 December].

[3] Japec L, Lyberg L. Big data initiatives in official statistics. In: Hill CA, Biemer PP, Buskirk TD, Japec L, Kirchner A, Kolenikov S, Lyberg LE, editors. Big data meets survey science: A collection of innovative methods. New York: John Wiley & Sons. 2020; 273-302. doi: 10.1002/9781118976357.ch9.

[4] Arslanalp S, Marini M, Tumbarello P. Big data on vessel traffic: Nowcasting trade flows in real time. IMF Working Paper No. 2019/275. [Internet]. International Monetary Fund; 2019. Available from: https//www.imf.org/en/Publications/WP/Issues/2019/12/13/Big-Data-on-Vessel-Traffic-Nowcasting-Trade-Flows-in-Real-Time-48837. [accessed 2023 May].

[5] Eurostat 2020. ESSnet Big data [Internet]. CROS – European Commission. 2020. Available from: https//cros-legacy.ec.europa.eu/content/essnet-big-data-1_en. [Accessed 2023 March].

[6] United Nations 2023 UN-CEBD. UN Big data. [Internet]. 2023. Available from: https//unstats.un.org/bigdata/task-teams/index.cshtml. [Accessed 2023 January].

[7] Kitchin R, Stehle S. Can smart city data be used to create new official statistics? Journal of Official Statistics. 2021; 37(1): 121-47. doi: 10.2478/JOS-2021-0006.

[8] de Pedraza P, Visintin S, Tijdens K, Kismihók G. Survey vs scraped data: comparing time series properties of web and survey vacancy data. IZA Journal of Labor Economics. 2019; 8(1): 2-23. doi10.2478/izajole-2019-0004.

[9] Kaluza P, Kölzsch A, Gastner MT, Blasius B. The complex network of global cargo ship movements. Journal of The Royal Society Interface. 2010; 7(48): 1093-103. doi: 10.1098/rsif.2009.0495.

[10] UNCTAD. Review of maritime transport 2022. Geneva: United Nations; 2022.

[11] What is the Automatic Identification System (AIS)? [Internet]. MarineTraffic Help. 2021. Available from: https//help.marinetraffic.com/hc/en-us/articles/204581828-What-is-the-Automatic-Identification-System-AIS-. [Accessed October 2022].

[12] Task Team on AIS Data. AIS Handbook [Internet]. UN Statistics Wiki; Updated 2022. Available from: https//unstats.un.org/wiki/display/AIS/AIS+Handbook. [Accessed 2023 January].

[13] Faster indicators of UK economic activity: more timely and relevant shipping indicators. [Internet]. Data Science Campus; 2019. Available from: https://datasciencecampus.ons.gov.uk/projects/faster-indicators-of-uk-economic-activity-improving-the-shipping-indicators/. [Accessed 2022 October].

[14] Bonham C, Noyvirt A, Tsalamanis I, Williams S. Analysing port and shipping operations using big data. Data Science Campus, ONS. 2018. Available from: https//datasciencecampus.ons.gov.uk/project/analysing-port-and-shipping-operations-using-big-data/. [Accessed May 2023].

[15] Port calls in Danish Sea ports (experimental statistics) [Internet]. Denmark. Short Terms Statistics, Business Statistics, Statistics Denmark; 2020. Available from: https://www.dst.dk/en/Statistik/dokumentation/documentationofstatistics/port-calls-in-danish-sea-ports–experimental-statistics-. [Accessed 2023 March].

[16] Maritime data: big data source with great potential [Internet]. Statistics Netherlands; 2023. Available from: https://www.cbs.nl/en-gb/about-us/innovation/nieuwsberichten/big-data/maritime-data-big-data-source-with-great-potential/. [Accessed 2022 December].

[17] Demunter C. Tourism statistics: Early adopters of big data. Statistical Working Paper, Eurostat, 2017.

[18] Harchaoui TM, Janssen RV. How can big data enhance the timeliness of official statistics? The case of the US consumer price index. International Journal of Forecasting. 2018; 34(2): 225-34. doi: 10.1016/j.ijforecast.2017.12.002.

[19] Gallego I, Font X. Changes in air passenger demand as a result of the COVID-19 crisis: using Big data to inform tourism policy. Journal of Sustainable Tourism. 2021; 29(9): 1470-89. doi: 10.1080/09669582.2020.1773476.

[20] Yang D, Wu L, Wang S, Jia H, Li KX. How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. Transport Reviews. 2019; 39(6): 755-73. doi: 10.1080/01441647.2019.1649315.

[21] Callegaro M, Yang Y. The role of surveys in the era of "big data". The Palgrave Handbook of Survey Research. 2018; 175-92.

[22] Ahmad Yar AW, Bircan T. Big data for official migration statistics: Evidence from 29 national statistical institutions. Big data & Society. 2023; 10(2). doi: 10.1177/20539517231210244.

[23] Struijs P, Braaksma B, Daas PJ. Official statistics and big data. Big data & Society. 2014; 1(1). doi: 10.1177/2053951714538417.

[24] Tam SM, Kim JK. Big data ethics and selection-bias: An official statistician's perspective. Statistical Journal of the IAOS. 2018; 34(4): 577-88. doi: 10.3233/SJI-170395.

[25] Brave SA, Butters RA, Fogarty M. The perils of working with big data, and a SMALL checklist you can use to recognize them. Business Horizons. 2022; 65(4): 481-92. doi: 10.1016/j.bushor.2021.06.004.

[26] Bähr S, Haas GC, Keusch F, Kreuter F, Trappmann M. Missing data and other measurement quality issues in mobile geolocation sensor data. Social Science Computer Review. 2022; 40(1): 212-35. doi: 10.1177/0894439320944118.

[27] Sagiroglu S, Sinanc D. Big Data: A Review [Internet]. 2013 International Conference on Collaboration Technologies and Systems (CTS). 2013; 42-7. Available from: https://ieeexplore.ieee.org/abstract/document/6567202. doi: 10.1109/cts.2013.6567202.

[28] Statistics of Port Traffic [Internet]. Central Statistics Office; 2023. Available from: https://www.cso.ie/en/statistics/transport/statisticsofporttraffic/. [Accessed 2023 April].
[29] Government of Ireland. S.I. No. 573/2010 – European Communities (Vessel Traffic Monitoring and Information System) Regulations. 2010. Ireland: Government of Ireland; 2010.
[30] International Maritime Organisation; 2015. Revised Guidelines for the onboard operational use of shipborne Automatic Identification Systems (AIS). London: International Maritime Organisation. 2015.
[31] Sofie De Broe, Olav ten Bosch, Daas P, Gert B, Laevens B, Kroese B. The need for timely official statistics. The COVID-19 pandemic as a driver for innovation. Statistical Journal of the IAOS. 2021; 37(4): 1221-7. doi: 10.3233/SJI-210825.
[32] Lapinski ALS, Isenor AW. Estimating reception coverage characteristics of AIS. Journal of Navigation. 2011; 64(4): 609-23. doi: 10.1017/S0373463311000282.
[33] Greidanus H, Alvarez M, Eriksen T, Vincenzo G. Completeness and accuracy of a wide-area maritime situational picture based on automatic ship reporting systems. Journal of Navigation. 2015; 69(1): 156-68. doi: 10.1017/S0373463315000582.
[34] Emmens T, Amrit C, Abdi A, Ghosh M. The promises and perils of Automatic Identification System data. Expert Systems with Applications. 2021; 178: 114975. doi: 10.1016/j.eswa.2021.114975.
[35] Harati-Mokhtari A, Wall A, Brooks P, Wang J. Automatic Identification System (AIS): Data reliability and human error implications. Journal of Navigation. 2007; 60(3): 373-89. doi: 10.1017/S0373463307004298.
[36] AIS Data: Task Team of the UN Committee of Experts on Big data and Data Science for Official Statistics. [Internet]. United Nations CEBD; 2023. Available from: https://unstats.un.org/bigdata/task-teams/ais/index.cshtml. [Accessed January 2023].
[37] H3: Uber's Hexagonal Hierarchical Spatial Index. [Internet]. Uber Technologies Inc; 2018. Available from: https//www.uber.com/en-IE/blog/h3/. [Accessed 2022 December].
[38] Grid traversal functions [Internet]. Uber Technologies Inc; 2018. Available from: https//h3geo.org/docs/api/traversal/#griddisk. [Accessed 2023 January].
[39] Reference Manual on Maritime Transport Statistics [Internet]. Luxembourg: Eurostat; 2019.
[40] Fundamental Principles of Official Statistics [Internet]. United Nations Statistics Division; 2014. Available from: https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx. [accessed 2023 January].