AIS-Driven Maritime Insights: Improving Italian Port Traffic Analysis

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

[1]Italian Institute of Official Statistics (Istat), Italy

Abstract

The Istat TRAMAR survey on Maritime Transport (TRAMAR) generates statistics about several features of maritime traffic. TRAMAR gathers data related to goods and passengers movements, from a questionnaire, and data related to vessels' travels from Port Authorities registers, via "machine to machine". These data feeds into an integrated database for high quality maritime statistics, including the Eurostat-requested F2 table, reporting vessels' arrivals in major Italian ports. In this regard, Eurostat proposed transmitting the F2 table quarterly instead of annually. Italy, while not responding to this request yet, has initiated a study to combine Automatic Identification System (AIS) data with traditional sources, to enhance data dissemination timeliness. AIS is an automatic tracking system on ships extensively used in the maritime world, for safety and management purposes, that produces a huge amount of real time data, containing information regarding navigation status of vessels. Other studies used AIS to calculate covariates that allow estimating arrivals using a statistical model. Our goal, while ambitious, is to reconstruct the complete routes of all vessels visiting Italian ports. Vessel's identifier and two consecutive port visits, the first one at the departure port, the second one at the arrival port, define a route. Finding a port visit, given by a ship being in a port area at speed zero, is the first step to obtain a route from AIS data. Unfortunately, this simple idea faces several obstacles in practice, mostly due to errors in the data values. Thus, we propose a methodology to overcome these issues in order to avoid losing port visits, while reconstructing the complete route of the vessel. To this end, we developed an algorithm to detect missing AIS data and attempt deterministic or probabilistic imputation. We assessed the quality of our methodology by comparing the results with statistics produced by traditional sources. We achieved good results for most ports, but we still observe a partial coverage of the phenomenon for smaller ports, characterized by frequent and short vessel travels.

Keywords: AIS, Big Data, Maritime Transport Statistics, Imputation

1. Introduction

The TRAMAR (MARitime TRAnsport) survey, conducted by Istat, provides statistics on the transport of goods and passengers by ship for commercial purposes in Italian ports [1]. The survey in question is a census, which refers to ships with a gross tonnage of at least 100 tonnes that are engaged in commercial activities. The statistical production of TRAMAR is achieved through the integration of survey data with an administrative source, namely PMIS (Port Management Information System). PMIS is the digital registry of the Italian General Command of the Port Authorities. The integration of sources is essential to guarantee the quality of the indicators, given that neither source provides comprehensive coverage of the phenomenon under analysis. Note that the PMIS only covers the most important ports for commercial and passenger transport (i.e. a list of 38 ports), whereas the TRAMAR survey covers all relevant

ports. However, it does not encompass every aspect of ship traffic, due to a non-response rate in certain ports that is not negligible.

The objective of this study is to present a nonconventional source of data, namely Big Data, as a means of enhancing the quality of maritime traffic statistics. AIS (Automatic Identification System) is an automatic tracking system on ships extensively used in the maritime world for safety and management purposes. AIS provides the geo-localization of all vessels navigating the world's seas, with regular and frequent intervals. The continuous storage of this data in a Big Data archive theoretically allows for the reconstruction of the history of ships' voyages. Since sensors (GPS) are the source of this data, we may categorise the AIS database as a source within the Internet of Things (IoT), although AIS also contains data generated by humans. The integration of the AIS data source into the TRAMAR production process could enhance the timeliness of crucial outputs, including the F2 table required by Eurostat. This table must report the number of vessels, with a focus on inward movements, for each reporting port by type of vessel and size of vessel expressed in GT (Gross Tonnage). The PMIS source, which is released on a monthly basis via machine-to-machine communication, is available in advance of the TRAMAR survey. However, the F2 table cannot be derived from the aforementioned PMIS data alone since it does not encompass all ports. Therefore, the AIS source, providing near-real-time data, may be integrated with the PMIS data to create an F2 table that is both timely and comprehensive.

The use of AIS data to support the official statistics has been already investigated in several works and projects. In [2] the Port Visits Geo-Solution prototype monitors the movement of ships within the Piraeus Central Port, defined by a polygon, in order to compute the number of arrivals and departures. In the United Nations Global Platform (UNGP) Handbook page [3], several case studies on the use of AIS are described, such as the experimental statistics of the daily number of vessels visiting Danish ports using AIS data published from Statistics Denmark [4]. In [5] two methodologies for generating port visits are compared. The first one, in particular, uses polygons to identify ships inside a port area, as in [2]. The results of these studies were used in this work, although none of them is aimed at our objective, i.e. to reconstruct the entire voyages of ships from departure to arrival, with the aim of reducing the occurrence of manual errors in assigning reporting ports.

2. Methodology

This section outlines the methodology employed to process AIS data, which is designed to yield a comprehensive list of all maritime voyages involving vessels arriving in or departing from Italian ports. The vessels of interest are all passenger or commercial vessels.

2.1 Data Source

The AIS data used in this study were provided by the Task Team on AIS Data of the UN Committee of Experts on Big Data and Data Science for Official Statistics (Task Team on AIS Data - UN-CEBD [6]). These data are accessible through the UN Global Platform (UNGP) [7], which holds a global (UN-AIS) repository of live and historical AIS data. The UN-AIS dataset contains regular observations of all types of vessels, with a gross tonnage of more than 300 tonnes, since 1st December 2018. The temporal interval between two observations of the same vessel is approximately 10 minutes. The attributes available in each AIS observation (Figure 1) include three categories of information: static (MMSI code, IMO code, ship's name and type), dynamic (ship's position coordinates, navigation status, speed and course) and voyage related (destination and draught).

Figure 1: AIS observation attribute.

Static data					Dynamic data					Travel related data	
IMO	MMSI	CALLSIGN	VESSEL NAME	VESSEL TYPE	TIME	COORDINATES	NAVIG. STATUS	SPEED	COURSE	DESTINATION	DRAFT
8401561	20110115	ZAD4L	FROJDI II	Cargo	04/06/2023 19:45	41.1323 16.8530	MOORED	0	258	Ravenna	null

The UN-AIS dataset also provides vessel position at multiple resolutions through the H3 index. H3 is a geospatial indexing system, developed by Uber Technologies [8], that approximates the GPS coordinates using a hexagonal tessellation of the earth's surface. The H3 index identifies the hexagon containing the ship's coordinates, with the hexagon size depending on the resolution chosen. In addition to the AIS dataset, we used a second dataset of world ports [9], which contains geographical coordinates (latitude, longitude) for each port.

2.2 Pipeline

In order to fulfil the objectives of this work, it is necessary to identify and quantify the ships visiting Italian ports. The desired output is a dataset of ships' voyages, as in Figure 2.

Figure 2: Dataset of ships' voyages.

IMO	VESSEL TYPE	DEPARTURE PORT	ARRIVAL PORT	DEPARTURE DATE	ARRIVAL DATE
8401561	Cargo	ITBRI (Bari)	ITRAN (Ravenna)	04/09/2023	05/09/2023
9483712	Passenger	ITGOA (Genova)	ILOLB (Olbia)	05/09/2023	06/09/2023

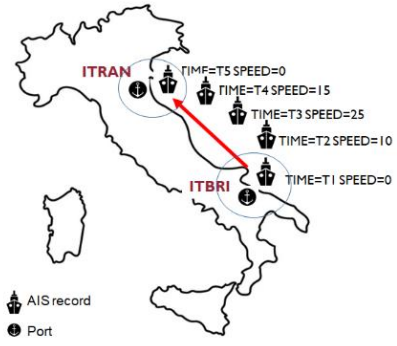
A voyage is defined by three key elements: the ship (IMO and vessel type), the departure date and port, the arrival date and port. In the following sections, we will refer to all generic visits of a vessel in a port as "port calls". The pipeline employed to process AIS data comprises three principal logical stages. The first stage of the process is the selection of vessels to be analysed.

This entails accessing the AIS database, the Italian ports list, and the ship register. Information from the ship register (available on the UNGP), such as the ship's type and gross tonnage, is used to choose the vessels. The analysis focuses on voyages that depart from or arrive at Italian ports, but the initial analysis is not limited to AIS records in Italian waters. Both the port of departure and the port of arrival are considered, even if they are not both Italian. However, if a ship does not visit an Italian port during the specified period, it may be excluded from the subsequent analysis. Therefore, the list of ships in the output will include all ships, except fishing and yacht, which have visited an Italian port, based on AIS data, and are listed in the ship register with a gross tonnage of more than 100 tonnes. The second stage is the calculation of voyages. Further details regarding the generation of this data will be presented in Section 2.3. The same section also describes a computer algorithm that calculates all the journeys that ships undertake over the specified time period. This process requires the AIS database and the lists of Italian and non-Italian ports. In the third and final stage, the ports imputation, missing data in the voyages list (the unknown ports) are imputed through the algorithm described in Section 2.5. The process pipeline described herein generates two final outputs: the list of voyages, which adheres to the format presented in Figure 2, and the statistics for the F2 table, which displays the number of voyages grouped by port of arrival.

2.3 Voyages calculation algorithm

A departure and an arrival are two consecutive port calls by the same vessel. As an example, Figure 3 shows a voyage from port ITBRI at time T1 to ITRAN at time T5. The intermediate AIS observations at T2, T3, and T4 are irrelevant to the definition of the voyage.

Figure 3: Voyage between the ports ITBRI (Bari) and ITRAN (Ravenna) in AIS



A port call in AIS data occurs when a ship's speed is 0 and its location is within a port area. The present study focuses on ships with a gross tonnage of at least 100 tonnes that have visited an Italian port at least once. By arranging port calls in sequence, we are able to map all the ship's voyages. The initial issue to be addressed is the identification of the areas associated with ports. To this end, each area was constructed from hexagons of resolution 8, marked by

H3 indexes. The initial step involved the selection of a hexagon at each port's coordinates and the surrounding ring of hexagons, which were then combined to form the area. Subsequently, we collected AIS data over a six-month period in the reference year (e.g., 2022), filtering for stationary ships, including cargo, tankers, and passenger vessels, within these hexagons. Finally, the port areas were augmented with the hexagons containing the positions of the filtered stationary ships (Figure 4), thereby simplifying the port visit visualization.

Figure 4: Hexagons identifying a port area.

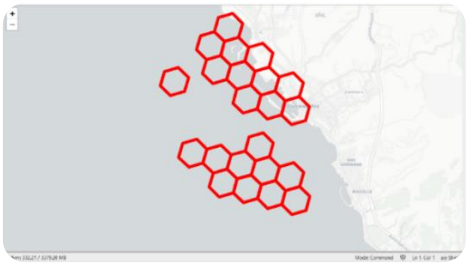
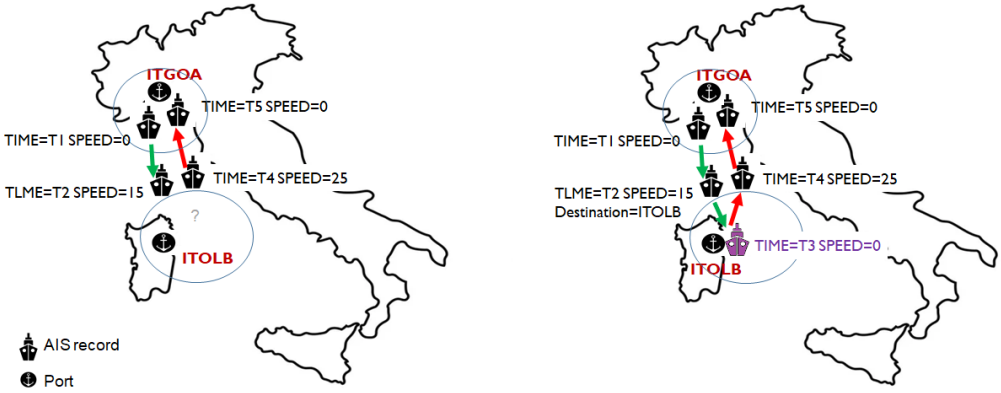


Figure 5: Example of missing data (on the left) and the solution (on the right).



A second issue that affects AIS database is the lack of data. This refers to a period during which no AIS observations were taken for a vessel. For example, Figure 5 shows a regular series of observations until time T2. Following a period of no data for several hours, the vessel reappears at time T4. If between times T2 and T4, the vessel enters the ITOLB area, we lose both the two voyages ITGOA-ITOLB and ITOLB-ITGOA. To resolve this issue, it is necessary to reexamine the AIS observations that were previously considered less significant (for example, records T2, T3, and T4 in Figure 3). The records are evaluated against the previous one, taking into account the differences in time (Dtime) and coordinates between two records. Assessing whether the vessel made a port call during this period is a challenging task, which is discussed in detail in Section 2.4. As Figure 5 on the right shows, if data is missing, a dummy entry is added to represent the missing arrival. This entry should include a time value (between

T2 and T4, in the example) and a location represented by an H3 hexagon of the arrival port area. The algorithm for identifying the probable arrival port is described in Section 2.5.

2.4 Dealing with missing AIS data

This study aims to reconstruct the complete routes of all vessels visiting Italian ports from AIS data. Addressing the absence of signals due to outages is a necessary part of the process, requiring targeted data preparation. To this end, we organize the data according to the timestamps, t_i of recorded vessel positions. This let us to introduce three key variables for our analysis:

- Δt_i which is the time interval between two successive timestamps at the *i-th* recorded location of the vessel, which helps in identifying periods of signal absence.
- Δs_i which is the distance between the consecutive locations of the vessels at timestamps t_i and t_{i+1} (computed as the geodesic distance between the two points on the earth's surface, by the Haversine formula to approximate the earth's curvature).
- $\bar{v}_i = \frac{\Delta s_i}{\Delta t_i}$ which is the computed average velocity of a vessel during Δt_i.

We emphasize the distinction between the average speed $\bar{v}_i$ from the 'speed over ground' (SOG) variable directly available in AIS data. The 'SOG' shows the vessel's speed at a particular moment, while our calculated average speed provides a view of the vessel's overall speed across the time interval Δt_i. To detect docking events when AIS signals are missing, our algorithm uses a specific heuristic based on calculating average speed. First, we identify segments where the time intervals are unusually long (say $\Delta t_i > 60'$), suggesting a lack of signal and thus a gap in tracking ship movements. Within these segments, we focus on 'suspect' intervals that might contain undetected docking events, and we exclude intervals where docking seems very unlikely. Normally, a vessel's movement during the outage interval Δt_i involves traversing the geodesic distance between two points, resulting in a standard average speed for that segment. However, if a vessel docks during Δt_i, its actual journey will cover a significantly greater distance to dock and then return to the next recorded location. This means the straight-line distance Δs_i would be notably less than the actual navigated distance, resulting in a ratio $\frac{\Delta s_i}{\Delta t_i}$ (average computed speed) that is lower than expected for normal travel. Our heuristic „flags“ (i.e. identifies) a travel segment as a potential docking event if it satisfies two criteria:

1. The time interval between signals Δt_i exceeds a maximum threshold: $\Delta t_i > \Delta t_{Max}$
2. The average speed $\bar{v}_i$ is below a minimum threshold: $\bar{v}_i < v_{min}$

The aforementioned parameters can be adjusted according to the desired level of strictness in searching for potential docking events during signal outages. We set the maximum time interval Δt_{Max} at approximately 60 minutes. Meanwhile, the minimum velocity v_{min} is set at a low percentile - specifically, around the 25th percentile of the average velocity distribution, which generally equates to about 10 knots. It should be noted that a flagged travel segment does not automatically indicate a docking event. Conversely, it is extremely unlikely that a non-flagged travel segment, in which the ship is moving at its usual average speed, would conceal a docking event, because there simply would not be enough time for it to occur.

2.5 Ports Imputation algorithm

If the attribute "destination" in the last AIS record before the missing one (T2 in the example in Figure 5) clearly indicates a port, this is designated as the port of arrival. If the "destination" is unclear or missing, a special code denotes an unknown port. The voyage dataset generated by the voyages calculation algorithm may contain several records with unknown departure or arrival port. In this case, the value must be imputed later. To do this, we use the historical data of the trips from the previous three months. The port of arrival will be the most frequent port in the vessel's history, given the port of departure. Similarly, the port of departure will be the most probable given the port of arrival. To ensure the reliability of the imputations, this rule applies only to 'scheduled' routes.

3. Results and conclusion

We present the initial results of AIS data processing. In particular, we compare the number of arrivals obtained by the proposed pipeline with the number of arrivals in the F2 Table for Eurostat. The results for 2022, the latest version of the F2 table to date, are shown in Table 1. Due to space limitations, only a sample of the total number of ports is shown. First, we observe that 14% (on average) of voyages calculated from AIS data have an unknown port of arrival, due to a missing data problem. In 3% of cases, the arrival is imputed using the destination field, while in about 1% of cases it is imputed using the vessel's historical data. In the remaining 10% of cases, the port of arrival remains unknown and is not included in the statistics. The total number of arrivals obtained from AIS for the whole year 2022 is 295,819 for 56 ports, 41% less than the F2 total (504,411). However, we note that most lost voyages are concentrated in ports with short and frequent routes, such as "La Maddalena" (ITMDA). In these cases, missing signals for a few hours means that several voyages are lost and calls are more easily missed. If we don't consider these routes, AIS arrivals even exceed F2: 117,685 instead of 114,262. In some ports, such as 'Porto Foxi' (ITPFX), an important commercial port not included in the PMIS source, we observe a higher value of arrivals than the official statistics In other cases

such as Gaeta (ITGEA) it is doubtful and probably due to ships excluded as out of scope. Another special case is the port of Gela (ITGAE). Only 3 voyages were found by AIS for a value in F2 of 103. The problem is probably due to the incomplete port area. Indeed, Gela's traffic is mainly oil tanker traffic and the dedicated dock is far from the rest of the port.

Table 1: Comparison between F2 and arrivals obtained from AIS for a list of Italian ports in 2022

UNLOcode	Port Name	F2 Arrivals	AIS Arrivals	Difference
ITCAG	Cagliari	1,690	1,691	+0,1%
ITCVV	Civitavecchia	2,815	2,558	-9,1%
ITGAE	Gaeta	226	365	+61,5%
ITGEA	Gela	103	3	-97,1%
ITLIV	Livorno	6,035	6,122	+1,4%
ITMDA	La Maddalena	16,442	944	-94,3%
ITNAP	Napoli	34,367	32,225	-6,2%
ITPFX	Porto Foxi (Sarroch)	769	1,095	+42,4%
ITPRJ	Capri	14,730	13,725	-6,8%

Future works include solving special cases, as mentioned above, and implementing a machine learning model to impute remaining missing ports. Eventually, we expect to improve the quality on long tracks, and we may need a more sensitive algorithm on short tracks.

References

[1] Istat TRAMAR survey - https://www.istat.it/it/archivio/14330

[2] ESSnet Big Data II WPE – Tracking ships – Deliverable E4: Consolidated report on project results (2020-11-19)

[3] United Nations Global Platform (2020). AIS Handbook Online. Geneva: United Nations - https://unstats.un.org/wiki/display/AIS/Case+studies

[4] AISDAG: Daily number of vessel (2019) - https://www.statistikbanken.dk/aisdag

[5] Port Visits Using Real-Time Shipping Data – CSO (2022) https://www.cso.ie/en/releasesandpublications/fp/fp-pvrts/portvisitsusingreal-timeshippingdata/datasourcemethodsandquality/

[6] UN AIS Data Task Team - https://unstats.un.org/bigdata/task-teams/ais/index.cshtml

[7] UN Global Platform - https://unstats.un.org/bigdata/un-global-platform.cshtml

[8] H3 (Hexagonal hierarchical geospatial indexing system) - https://h3geo.org/

[9] C. Merrien (2021), Worldwide list of seaports. http://dx.doi.org/10.12770/59ab5f6f-79ea-425d-830e-be5ecdb7bdbe