**AIS: Defining a ship’s journey and sea traffic analysis**

Christina Pierrakou, ELSTAT Greece, [c.pierrakou@statistics.gr](mailto:c.pierrakou@statistics.gr)

Tessa C.J. de Wit, Statistics Netherlands, [tcj.dewit@cbs.nl](mailto:tcj.dewit@cbs.nl)

Marco J.H. Puts, Statistics Netherlands, [m.puts@cbs.nl](mailto:m.puts@cbs.nl)

Anke J.M. Consten, Statistics Netherlands, [a.consten@cbs.nl](mailto:a.consten@cbs.nl)

**Abstract**

*Ships broadcast information on their location and status on a frequent basis by means of a radio signal. This so-called Automatic Identification Signal (AIS) provides a big data source for maritime and emission statistics. Research on the use of AIS for official statistics is part of the ESSnet Big Data project, performed by The Netherlands (Work package leader), Denmark, Greece, Norway and Poland. Here, we present methods to define a ship’s journey and to perform sea traffic analyses.*

*Defining the journey of a ship by using AIS data is needed to obtain insight into all the ports the ship visited. It also makes it possible to calculate the distance a ship travels. Furthermore, a ship’s journey is necessary to improve the calculation of emissions. To determine the journey of a ship, we further build on the algorithm we already developed to determine a port visit. The resulting journey algorithm is output-driven and enables us to define the start of a journey and to deal with noise in the signal.*

*For traffic and economic analyses, we also wanted to explore the possibility of calculating the number of ships during a certain time interval at certain coordinates by using AIS data. We calculated the traffic intensities of ships around Europe. An important prerequisite is that the grid elements all should have the same size, thus the grid was defined as areas of 10,000 square kilometres. Different coordination systems were compared, with the WGS1985 system being selected. To draw the data to a map, the Lambert Azimuthal projection was used. This method preserves surface area under the transformation. The final visualization is done in Shiny in combination with Leaflet.*

*.*

**Keywords:** AIS, Big data, Maritime statistics

**1. Introduction**

Maritime traffic has increased exponentially over the last decades. This led to the development of the Automatic Identification System (AIS), through which nearby ships exchange information on their location and status on a frequent basis by means of a radio signal. Originally, AIS is intended to facilitate safety and traffic management. Therefore, land based stations can also pick up the signals. Later on, it turned out that the signals can also be picked up by satellites. Also, it was realised that the collection of these signals could provide a big data source for maritime and emission statistics. All this information can be used to determine not only the location of ships, but also their routes. Therefore, this data can be proved useful in improving current maritime statistics as described under the Directive 2009/42/EC of the European Parliament and the Council and in composing new statistics. As for current statistics, we know for example that statistics on port visits (the number of ships visiting a port to load/unload goods) are not always complete. Furthermore, interest has been expressed by statistical offices and port authorities to gain insight in intra-port traffic, detecting disruptions, fast economic indicator and emissions. Therefore, we focus here on aspects that can be used to both improve current statistics and make new statistics: defining a ship’s journey and calculating and visualizing sea traffic.

Processing of this large data source, offering historical and streaming maritime data, is still a computational issue. A typical volume of radio and satellite-based worldwide maritime data represents an estimate of 18 million positions per day (Claramunt et al. 2017).Research on the use of AIS for official statistics is part of the ESSnet Big Data project, performed by the Netherlands (Work package leader), Greece, Norway and Poland.

**2. Quality aspects of AIS data**

Ships transmit internationally standardized encoded data fields via AIS.AISreporting requirements are described in Regulation 19 of Chapter V of the International Convention for the Safety of Life at Sea (SOLAS) [IMO, (1974/1980)].AIS data consist of different messages, containing different types of information, each with their own unique ID.The information in each AIS message can be divided into two categories. Static and voyage related information of the vessel, including information on the ship’s identity, such as the Maritime Mobile Service Identity (MMSI), the International Maritime Organization (IMO) Number and the name, transmitted every 6 minutes. Dynamic information such as the positional aspects (current latitude and longitude) are automatically transmitted, depending on the vessels’ speed and course. While the vessel is on the move this information is transmitted every 2 to 10 seconds and while a vessel is anchored every 3 minutes. AIS quality depends on correct installation of AIS device and frequent manual updates of information.

When investigating the quality of AIS data it is important to keep in mind that AIS is a radio signal, which means signals are sensitive to meteorological or magnetic factors. The messages are transmitted encoded. As a result, an error in one transmitted ‘byte’ of the encrypted message could result in an error in one or multiple fields in the decrypted message. Moreover, as their primary use is for the safety at sea and not for producing statistics, the manually entered variables are not always accurate. AIS receivers on land can pick up signals within the range of about 40 sea miles. They have timeslots in which data is received. In busy areas with many ships, not all data from all ships may fit into this time slot. This may result in the loss of data on some ships in that time slot. Finally, there are cases that ships turn off their AIS transponder, resulting in the “disappearance” of a ship.

**3. Defining a ship’s journey**

The first aim was to define the journey of a ship by using AIS data. Data on the routes ships take, provides information needed to determine all the ports a ship visits and it is possible to calculate the distance travelled. Furthermore, a ship’s journey is necessary to improve the calculation of emissions. It can also be used to detect disruptions or (changed) regulations in ports and fairways.

*3.1 The port-visit algorithm*

To determine the journey of a ship, we further built on an algorithm we developed to determine a port visit and calculate the number of ships that have arrived and departed a port using AIS data (De Wit et al., 2017). By adapting some variables, the resulting journey algorithm is output-driven and enables us to define the start of a journey and to deal with noise in the signal (Consten et al., 2017).

This initial port visit algorithm processes data from ships in a reference frame of ships derived from AIS data. Since a number of ships (which do not actually exist) appear due to glitches in the data and non-maritime ships which also transmit AIS data, a reference frame of ships has to be used to filter out existing maritime ships only. Furthermore, not all statistical offices (can) collect MMSI or IMO numbers, some collect call signs as a ship identifier. The reference frame of maritime ships therefore also includes all three ship identifiers: MMSI, IMO and call sign. Note that these three identifiers are not included in the dynamic messages, but only in the static messages. Thus, the static messages are used to build a reference frame. To build a reference frame[[1]](#footnote-1), we initially linked all MMSI-IMO couples and filtered out couples with invalid IMO or MMSI numbers, purely based on length for MMSI and on length and check digit for IMO. By filtering out and coupling ships’ identification codes, maritime ships were selected that were active in the European maritime area (De Wit et al, 2017).

AIS data from these ships is then filtered with a median filter over 10 minutes for latitude, longitude and speed. This 10 minute window reduces the amount of data, enabling faster processing. The algorithm also deals with the noise in the signal, rendering plausible positions and speeds even when a ship is anchored and transmits a signal only every 3 minutes. In that case, a filter with a shorter time interval would result in a single measurement that would not be corrected by other measurements. That is, with a 3 minute filter, only one signal would be picked up. If this particular signal is distorted it will not be corrected by neighbouring signals. In other cases, a different temporal filter could be preferred, so our intention is to optimize the algorithm by making this time interval flexible.

Then, for each data point in this filtered selection, the location is categorized as being at sea or in a port. Coordinates of ports have been defined in advance (we also intend to build an automatic port detection algorithm). Using the variables MMSI, time and location category, the port arrivals and departures are determined by selecting consecutive locations where the category location changes. A port arrival here is defined as a SEA-PORT couple and departure by a PORT-SEA couple. A successive port arrival and departure are then coupled to form a visit interval, resulting in the start and end of a visit, which can be used to calculate total visit time.

If two consecutive visits are separated by one time interval (10 minutes in this case), the intervals are integrated to form a longer visit interval. For the resulting visit interval, latitude, longitude and speed are combined again. As some ships only travel through a port, in the port of Amsterdam for example, speed is used to define an actual visit/stop. Only ships in a port with speed below 0.2 knots are categorized as visiting the port.

Now, using the data from the visiting interval, the distance travelled in the port can be calculated. This is calculated using the haversine function. This function determines the great-circle distance between two points on a sphere given their longitudes and latitudes.

*3.2. The journey algorithm*

The result of the previous algorithm provides an output-driven method to define a journey. The start of the journey is defined as the ship’s departure point from a port. The end point of the journey is determined in three ways:

- The ship arrives in another port (note that this requires the construction of a reference frame of ports which will be used to categorize the ship as being at SEA or in a PORT). Again the successive couple SEA-PORT will signal the arrival in the new port, constituting the end of the journey.

- The ship anchors: if this is at sea, this journey will be continuing after this. Again, the port visit method can be used in this case. Meaning that the algorithm continues until another PORT signal is detected, which will be defined as the end of the journey.

- The ship leaves the area of AIS coverage. For this research, only data covering European maritime waters was available. In that case, the boundary of the data, in other words if a ship travels outside of Europe, is defined as the end of a ship’s journey.

However, there are also areas of sea where satellite reception is minimal or where ships can turn off their AIS signal. In those cases, the next available information is used to combine linearly the data.

The journey algorithm can also be used to determine next destination, to calculate the distance travelled and the duration of the journey. The current algorithm can be optimized on some points. For example, now all the circumferences of ports have to be defined ‘manually’, which, especially in the case of a European or a worldwide level analysis, means a huge amount of work. To deal with this, an automatic port-defining algorithm has to be developed. Also, there are some optimizing steps that can be added, such as filtering out AIS data in which the speed and heading of the ship have not changed since the last signal.

**4. Sea traffic analyses**

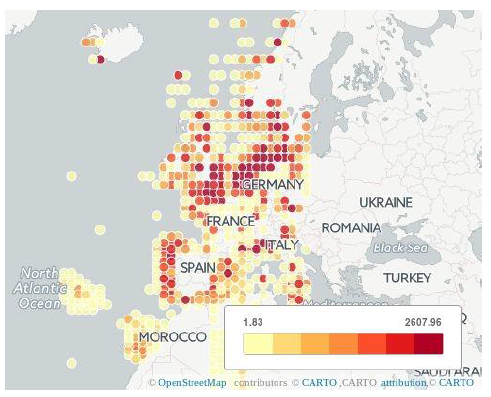
Calculating the intensity of sea traffic can be helpful for traffic management, to detect obstructions and provide a fast indicator of economic development or emissions from maritime ships. Therefore we developed a method to calculate the number of ships during a specific time interval at specific coordinates using AIS-data.

*4.1 Defining the grid*

Traffic intensities for maritime ships in European seas were calculated. The grid was defined as areas of 10,000 square kilometres.

An important prerequisite is that all grid elements have the same size. Otherwise, the calculated number of vessels could depend partly on the size of the grid elements. First it was implemented by defining an area of 100 by 100 kilometres in the middle of Europe and deforming this area for the rest of the map. After creating this grid, the algorithm counted the number of ships in each cell of the grid during one hour on a randomly chosen day. Cartogram 1 shows the result for this mapping. However, this mapping is not a standard cartographic projection and another projection had to be chosen.

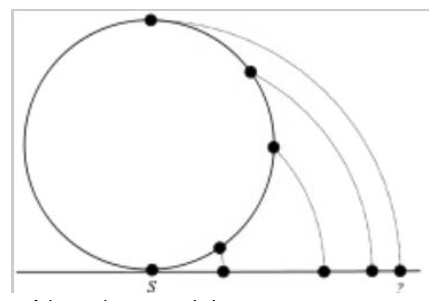
**Cartogram 1. Traffic analyses: the amount of ships in each cell of the grid during one hour on a randomly chosen day based on 100x100 grids**



Source: Statistics Netherlands

For drawing maps, several methods are available to project the earth on a map. One of them is the Lambert Azimuthal projection. In this projection, a surface touching the sphere at point S (see Figure 1) is defined and a point on the sphere is projected on the surface in such a way that the distance between S and the original point on the sphere is equal to the distance between S and the projected point. Thus, this projection preserves the surface area under the transformation (http://mathworld.wolfram.com/LambertAzimuthalEqual-AreaProjection.html).

***Figure 1. Lambert Azimuthal projection***



Within geography, locations are first defined in a coordinate system. The most famous coordinate system at the moment is WGS1985, which describes the earth as an ellipsoid defined at sea level. Hence, this is a suitable system for describing locations at sea. WGS1985 is described in GPS coordinates, since GPS uses WGS1985 for their basis. Other coordinate systems are defined for other uses. For instance, the ETRS89 coordinate system, the one used for the Lambert Azimuthal projection,an ellipsoid is used that models Europe well. ETRS89 and WGS1985 are quite similar. They use the same semi-major axis (“diameter of the earth”). The flattening factors of both systems slightly differ: 298.25722 for WGS1985 and 298.257222101 for ETRS89. Both systems differ slightly due to continental drift, which resulted in 2015 to a difference of 65 centimetres between WGS1985 and ETRS89. For statistical purposes, this difference is negligible but since we are analysing at sea levels and AIS is based on GPS, we keep the WGS1985as a basis for our projections.

Source:https://en.wikipedia.org/wiki/Lambert\_azimuthal\_equal-area\_projection

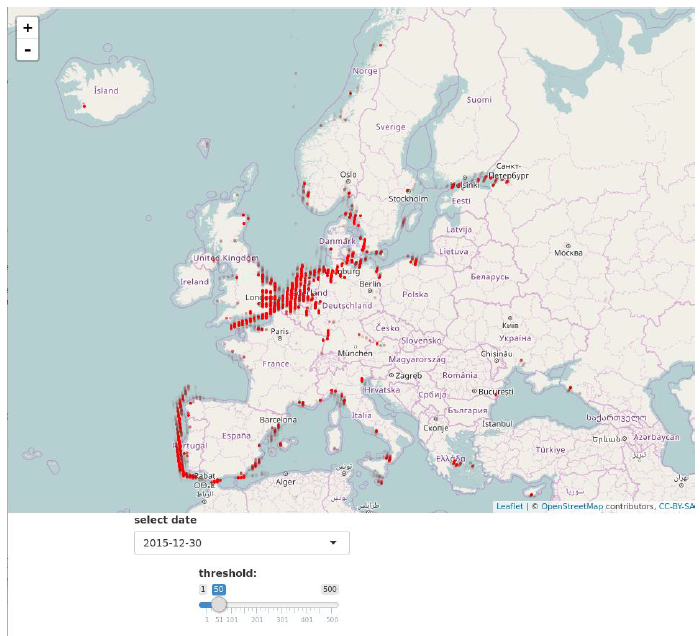
This projection was implemented in Spark and Scala, together with the definition of the grid (file: LAEA. Scala). The bounding box for the grid was defined and subdivided in 200 by 200 grid points. Then, all latitudes and longitudes are projected and mapped to the index of the associated grid point (file: countuniq.scala). Finally, the latitude and longitude of that grid point is calculated. The implementation of the projection can be found on Github: https://github.com/mputs/WP4/locations. The spark code can be found in the subdirectory src/main/scala.Projection.

*4.2 Visualising traffic intensity*

The final visualization is made in Shiny combined with Leaflet. Shiny is a web application framework for R, with which one can create interactive web applications. Leaflet is a JavaScript library for interactive maps, which can also be used from R. The code can be found in https://github.com/mputs/WP4/locations.

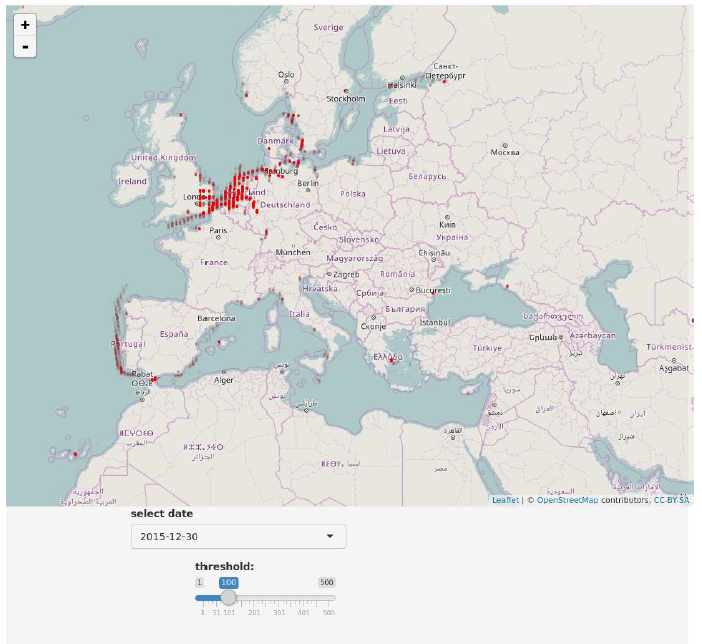
Figure 2 and 3 show two examples of the visualization for different thresholds. In the visualization, one can choose the date out of an available date list in the bottom, where a slider can be used to select a saturation threshold for the traffic intensity. Cells with a higher traffic intensity are displayed as dark red and less visited areas as lighter red. The use of slider gives insight in more and less busy regions in Europe. For the regions that are not displayed in a shade of red, there is no data available in the dataset on that specific day. Also very low intensities are made invisible.

**Figure 2. Traffic analyses for 1 day: number of ships in each cell of the grid (Lambert Azimuthal equal area projection), threshold = 50**



Source: Statistics Netherlands

**Figure 3. Traffic analyses for 1 day: number of ships in each cell of the grid (Lambert Azimuthal equal area projection), threshold = 100*­***



Source: Statistics Netherlands

**5. Conclusions**

Although there is a lot of work already done, further investigation on this promising data source is needed to develop functional production prototypes and to promote and support their implementation at national as well as at European level. The algorithms developed here, to calculate the routes ships take on their journey and to calculate and visualise the intensity of sea traffic, could be further optimized. The results of these analyses can be applied to improve current statistics (e.g. port visits, emissions) and to compose new statistics (e.g., intra-port traffic, detecting disruptions, fast economic indicator and emissions).

**6. Acknowledgment**

This work was supported by the ESSnet Big Data project, which is part of the Big Data Action Plan and Roadmap 1.06 (BDAR) agreed to be integrated into the ESS Vision 2020 portfolio.

**7. References**

Biemer P. and Lyberg L. (2003), Introduction to Survey Quality, J.Wiley, N.Y.

Claramunt et al. (2017), Maritime Data Integration and Analysis: Recent Progress and Research Challenges, Proc. 20th International Conference on Extending Database Technology (EDBT), March 21-24, 2017 – Venice, Italy: ISBN 978-3-89318-073-8,Available at: <https://openproceedings.org/2017/conf/edbt/paper-224.pdf> (accessed: 19 May 2018)

Consten et al., (2017). Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.Available at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4\_Deliverable\_4.3\_2017\_07\_21\_v1.0.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_21_v1.0.pdf%20)(accessed: 1 May 2018)

De Wit et al., (2017). Work package 4 AIS data deliverable 4.2: deriving port visits and linking data from Maritime statistics with AIS-data. Available at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/8/8d/WP4\_Deliverable\_4.2\_2017\_02\_10.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/8/8d/WP4_Deliverable_4.2_2017_02_10.pdf%20)(accessed: 1 May 2018)

<http://mathworld.wolfram.com/LambertAzimuthalEqual-AreaProjection.html>

IMO (1974/1980). International convention for the Safety of Life at Sea (SOLAS), Chapter V: Safety of Navigation, Regulation 19

Little, R. J. A. (1988), Missing-Data Adjustments in Large Surveys, Journal of Business and Economic Statistics, 6, pp. 287-296.

1. The code are available at: https://github.com/mputs/WP4/tree/master/aisframe2 [↑](#footnote-ref-1)