ESSnet Big Data

Specific Grant Agreement No 1 (SGA-1)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata
https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en

Framework Partnership Agreement Number **11104.2015.006-2015.720**

Specific Grant Agreement Number **11104.2015.007-2016.085**

Deliverable 4.3

Report about sea traffic analyses using AIS-data

Version 2017-07-21

Prepared by:

Anke Consten (CBS, Netherlands)
Marco Puts (CBS, Netherlands)
Tessa de Wit (CBS, Netherlands)
Eleni Bisioti (ELSTAT, Greece)
Christina Pierrakou (ELSTAT, Greece)
Anna Bilska (GUS, Poland)
Michal Bis (GUS, Poland)
Olav Grøndal (SD, Denmark)
Øyvind Langsrud (SSB, Norway)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@cbs.nl
Telephone : +31 45 570 7441
Mobile phone : +31 6 5248 7775

List of used abbreviations

AIS	Automatic Identification System
DMA	Danish Maritime Authority
DZ	Dirkzwager
HCG	Hellenic Coast Guard
IMO	International Maritime Organisation
IWW	Inland waterways
KDE	Kernel Density Estimator
MMSI	Maritime Mobile Service Identity
PMA	Polish Maritime Authority
PoC	Proof of Concept
WP	Work Package

Executive summary

In this deliverable we further investigate the quality of AIS (Automatic Identification System) data. In **Chapter 2**, we start with analysing AIS data from data provider Dirkzwager (DZ) specifically, by means of the quality and metadata framework. From this, we concluded that the quality is good. Almost all factors of the quality framework are judged as mostly positive. Only “spatial coverage” and “transparency and soundness of methods and processes for the metadata and the data” are insufficient. Not all European coastal areas are covered and DZ provides partly preprocessed data, but documentation on this is not available to us on how. Privacy is also an issue that needs to be researched further. Finally, in Chapter 2, AIS data from Dirkzwager is compared to national data for Denmark, Greece and Poland. From this analyses it became clear that coverage is insufficient for the European coast area, as data on some ports is completely missing. Furthermore, the frequency of samples per ship is lower for some areas, rendering a less precise journey for ships in these areas.

In **chapter 3**, we describe a methodology to calculate such a journey of a ship, being able to handle noise in the AIS data and defining the start and end of a journey. This output-driven algorithm defines a journey using the departure from a port as the start of a journey. The end of the journey is defined by either a ship entering another port, anchoring or leaving the area of AIS coverage. Processing could be optimized by filtering out AIS data in which the speed and heading of the ship have not changed since the last message. This optimization might be performed in the future, but is not in the scope of the current project.The algorithm for this is still being improved in Scala, but will be available and useable for others to use once finished.

Chapter 4 describes four successful Proof of Concepts (PoC). In the first PoC an algorithm was developed to calculate the intra-port journey by using AIS data. Although coverage/sampling frequency of data would influence the preciseness of the distance calculated, it is still a good indicator. In the future, it would be interesting to develop an algorithm that can automatically detect intra-port movements or detect anomalies in the movements of ships, signalling problems in the ports. In the second PoC, a literature search showed that automatic port detection from AIS is possible by building a data driven algorithm for defining ports. In the near future, Statistics Netherlands and Marine Traffic will collaborate on this to build a reference frame of ports and possibly terminals. From the third PoC we conclude that next destination as reported by captains is not usable, instead the observed next destination from AIS has to be used. More work is needed to handle areas where coverage is not perfect and to compare actual port to port distances to Eurostat’s average distance matrix (perhaps rendering the latter obsolete). Finally, the last PoC shows that AIS data is useful to investigate fluvio-maritime transport. For all these topics good filtering of the data is needed, which we already tackled quite a bit because of handling the noise in the data we used.

Chapter 5 describes how a method for measuring sea traffic analyses was developed. The main issue here was defining the right grid. Also this algorithm will be available and useable for others.

Finally, **chapter 6** describes the lessons learned from this deliverable. Although we are not completely satisfied with the quality of the Dirkzwager data, we conclude AIS data itself can help improve current statistics. The coverage problem does not affect the calculated number of journeys of a ship, but it will affect the calculated distances and traffic estimates. Which would result in an underestimation of the distance (and thus emissions).

Index

1. Introduction

The aim of work package (WP) 4 is to investigate whether real-time measurement data of ship positions measured by the so-called AIS system can be used 1) to improve the quality and internal comparability of existing statistics and 2) to produce new statistical products relevant to the ESS.

Five National Statistical Institutes participate in WP4: the national statistical institutes of The Netherlands (Work package leader), Denmark, Greece, Norway and Poland.

WP4 is subdivided into two phases. SGA-1 focuses on creating a common database, linking AIS-data to maritime statistics and constructing sea traffic analyses (February 2016-July 2017, See Annex 1 for a more detailed description of SGA-1 of WP4.) Methodological, qualitative and technical results, including intermediate findings, will be used as inputs for SGA-2.

SGA-2 then focuses on the calculation of emissions and future perspectives for AIS data as source data for new statistical output (August 2017 until May 2018). See Annex 2 for a more detailed description of SGA-2.

The current SGA has results in three deliverables, of which this deliverable is the final one:

Deliverable 1:"Creating a database with AIS-data for official statistics: Possibilities and Pitfalls" (delivered 21-7-2016)

Deliverable 2: "Deriving port visits and linking data from maritime statistics with AIS data" (delivered 31-1-2017)

Deliverable 3: Sea traffic analyses using AIS-data (the current deliverable)

In this deliverable we further investigate the quality of AIS data. Chapter 2 describes the results on the further investigation of the quality of the AIS data: the quality and metadata framework and comparing it to national data. Chapter 3 describes algorithms developed for handling the noise in the AIS data. Chapter 4 describes the results of the four PoC's we proposed in previous deliverable [1]: calculation of the intraport distance, investigation of the feasibility of using AIS data to define ports, improving next destinations and the distance matrix and completing fluvio-maritime transport. Chapter 5 describes the results of the sea traffic analyses. Finally, chapter 6 describes the lessons learned from this deliverable.

2. Quality

Within this SGA-1 of WP 4 we use European AIS data from Royal Dirkzwager (DZ) of the period October 2015 until April 2016 for our investigations. DZ is a leading Maritime and Nautical service provider situated in the Netherlands. DZ itself has AIS receivers all over the coastline and main ports of the Netherlands (Ouddorp, Rotterdam, Noordwijk, Amsterdam and Petten) and a couple outside the Netherlands: Cherbourg, Gibraltar, Zee Bruges, Antwerp and Hamburg. Besides its own receivers DZ also uses data from its six partners, amongst which AIShub (covering all of Europe), MarineTraffic (covering mostly of Mediterranean: Greece and Italy in the DZ data). The data we bought for this work package constitutes all data of these six partners, without any satellite data.

As described in previous deliverable [1] the coverage of ships in the DZ AIS data, is good but there is also data missing and quite a lot of noise, for example some vessels seemed to be located in the Sahara. We also concluded in the previous deliverable that following a ship during a couple of days gives us a reasonable view of the journey of a ship, but we have also missing data here.
These results made us decide to further investigate different AIS data sources. This was done by subjecting the DZ data to a quality and metadata framework and then comparing DZ to other data sources. We were interested to see how quality of DZ data matched national AIS data. Therefore, national AIS data from Denmark, Greece and Poland was compared to AIS data from DZ.

When investigating the quality of AIS data it is important to keep in mind that:

- AIS is a radio signal, parts of the messages can get lost or scrambled due to factors such as meteorology or magnetics.
- Messages are transmitted encoded. As a result, an error in one transmitted ‘byte’ can result in an error in one or multiple fields in the decrypted message. Most of the times, these errors are detectable as the result yields an invalid variable, but sometimes they result in valid variables. For instance, coincidentally the resulting MMSI (Maritime Mobile Service Identity) can be a technically valid, but incorrect MMSI, resulting from an erroneous detection. These errors can arise for every variable, so this can for example result in erroneous latitude and longitude, yielding faulty locations that are quite far away from the actual location of the ship. In turn, this can result in a very high journey distance of ship.
- Receivers 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. Resulting in the loss of data on some ships in that time slot.
- Ships can turn off their AIS transponder resulting in the disappearance of a ship.
- AIS was intended originally for safety at sea, to warn nearby ships. As it was not meant for producing statistics, the variables entered manually by the shippers are not always reliable.
- AIS receivers on land can only pick up signals within the range of about 40 sea miles. Therefore, land receivers have a very limited coverage of signals transmitted from sea which results in loss of information of ships on open sea.

This chapter describes the results on the further investigation of the quality of the AIS data. First, we describe the quality and metadata framework we have chosen. After that we describe the validation checks performed by the different AIS data providers in more detail. This chapter ends with the comparison of the densities and frequencies of the European AIS data and the national AIS data of Denmark, Greece and Poland.

2.1 Quality framework

To provide an overview of different aspects of the quality of the DZ AIS data, we used a preliminary framework for national statistical offices to conceptualise the quality of big data. This Big Data Quality framework is a deliverable of the Big Data Quality Task Team from the UNECE/HLG project*, The Role of Big Data in the Modernisation of Statistical Production [2]*.

This preliminary framework was developed building on dimensions and concepts from existing statistical data quality frameworks. It provides a structured view of quality at three phases of the business process:

- Input – acquisition, or pre-acquisition analysis of the data;
- Throughput –transformation, manipulation and analysis of the data;
- Output – the reporting of quality with statistical outputs derived from big data sources.

In table 2.1.1 the big data quality framework for the input phase of the business process is reported. This is done only for the European AIS data source of DZ. In table 2.1.1 we filled out the factors to consider for each quality dimension. See annex 3 for a more detailed rating of this quality framework. In this annex we rated different indicators for each quality dimension.

Hyper dimension	Quality Dimension	Factors to Consider	Remarks considering source	Conclusion
Source	*Institutional/ Business Environment*	Sustainability of the entity-data provider	Data is transmitted through the ether. Many data providers available, who deliver the same data.	OK
		Reliability status	Highly reliable	OK
		Transparency and interpretability	Used filters by the data provider are not always clear. However, the data is standardized, which means there is not much room for interpretations	OK
	Privacy and Security	Legislation	Especially on the inland waterways, there are some cases where owners of the vessels have privacy issues	Action needed: solve privacy issues
		Data Keeper vs. Data provider	Data Keeper = Data provider in this case. However, due to legislation, the owner of the data should be the owner of the vessel	OK
		Restrictions	Some ships are owned by a private person, which puts some restrictions on the aggregation level of the data	OK
		Perception	From the owners of the ships, the use of the data might be perceived negative. Using a cautious publication level could minimize negativity	OK
Metadata	*Complexity*	Technical constraints	Data is stored in the NMEA data format. To decode the messages, libraries can be used that are available for different programming languages	OK
		Whether structured or	The data is highly structured	OK

		unstructured		
		Readability	Data is not human readable, although well readable by machines.	OK
		Presence of hierarchies and nesting	No nesting available, but different record types	OK
	Completeness	Whether the metadata is available, interpretable and complete	Available, but one has to search the internet for it. Elements that are not processed automatically have to be put in by owner of the ship, who might not always be aware of all the details	OK
	Usability	Resources required to import and analyse	When dealing with Europe wide data for a longer time span, big data infrastructure and skills are needed	OK
		Risk analysis	Not applicable. These skills and infrastructure are also necessary for other sources	OK
	Time-related factors	Timeliness	The data is high velocity. High quality data can be available in an instance	OK
		Periodicity	Collection time and reference period is about the same.	OK
		Changes through time	Not applicable. The data is collected and stored as is.	OK
	Linkability	Presence and quality of linking variables	Due to transmission errors, the IMO number necessary to link to maritime statistics are sometimes wrongly received. We developed a method to deal with this.	Action needed: develop method do deal with this wrong IMO-numbers
		Linking level	Linking can be at vessel level based on the IMO number and precise geographical locations	OK
	Coherence - consistency	Standardisation	standardized	OK
		Metadata available for key variables (classification variables, construct being measured)	yes	OK
	Validity	Transparency of methods and processes	DZ pre-processes the data, but there is no documentation available on how they do this.	NOT OK
		Soundness of methods and processes	unknown	NOT OK
Data	*Accuracy and selectivity*	Total survey error approach	See deliverable 4.3/4.2	OK
		Reference datasets	See deliverable 4.2	OK
		Selectivity	All ships are present in the data. However, spatial coverage is not sufficient: not all coastal areas of Europe are covered.	NOT OK
	Linkability	Quality of linking variables	Linking is done based on IMO numbers for vessels and Lat/Lon for	OK

			ports	
	Coherence - consistency	Standardized concepts	All key variables are standardized (IMO's, Lat/Lon)	OK
		Coherence with metadata	yes	OK
	Validity	Transparency of methods and processes	Not transparent (see also metadata validity)	NOT OK
		Soundness of methods	unknown	NOT OK

Table 2.1.1: quality framework factors to consider

In line with the beginning of chapter 2, this quality framework also confirms that the quality of the DZ AIS data is good. Table 2.1.1 shows that almost all factors of the quality framework are judged as mostly positive. However, action is needed considering the privacy of the AIS data source. Especially captains of inland waterway vessels have their vessel as their home address. In these cases AIS results in information on private individuals, rendering privacy issues. Before using this source for making statistics these privacy issues should be solved. Another point of action is needed for the quality of the linking variables. Due to transmission errors, the IMO-numbers necessary to link AIS data to maritime statistics are sometimes wrongly received. We already developed a method to deal with this (see previous deliverable [1]). We only judged the spatial coverage, transparency and soundness of methods and processes for the metadata and the data as not ok because not all European coastal areas are covered and DZ pre-processes the data, but the documentation on how they do this is not available to us.

In the next paragraph we describe the validation checks different AIS provides perform on their data.

2.2 Validation checks different AIS providers

Dirkzwager

DZ does multiple validation steps. For example, DZ removes impossible locations from the data and puts a time stamp on the AIS messages. This is not the time of reception of the receiver, but the time of reception at DZ. The time between reception at the receiver and at DZ differs. This timestamp therefore is not always very reliable. When a specific area is covered by multiple partners, DZ orders the trustability of these partners. DZ values the quality of the data of their own receivers as the highest and their partners (see chapter 1) as second best. DZ values the quality of satellite data as the lowest as it contains blind spots (mainly in Finland and around Libya and Egypt) due to weather, magnetism, buildings and for unknown reasons. Moreover, as satellites are not stationary they receive less data on specific locations.

Other AIS providers

The national AIS data of Denmark, Greece and Poland are completely unfiltered and untreated. The raw (unfiltered) data is split into files. Each file includes only one day. For Greece, we received encoded NMEA sentences in CSV format, enriched with the timestamp from the terrestrial station. Each CSV file included approximately monthly data. For Poland there is no timestamp available in the raw data, but they are able to reconstruct the route of the ship based on latitude, longitude and date (day) + sorted by parameter "counter". This last parameter is added during the decoding process for getting the correct sequence of records according to the ship's measurement.

Statistics Norway gets AIS data from Kystverket (The Norwegian Coastal Administration). DNV-GL[1] performs computations for Kystverket. They told us validation checks are an important topic and a known challenge. They constantly keep improving the checks and algorithms. DNV-GL makes validations of the data when they import and organise the AIS data for analyses. They also implemented routines for data cleaning. For example they look for obvious errors in the location data, by validating the speed between two subsequent location reports. The computed speed between successive locations should not exceed the maximum speed of a ship. They also look at ships having too many travelling hours per year and ships with unavailable speeds. They also validate the data by linking the data with other data sources.

Although we did not receive data from Marine Traffic yet we have some information on the validation checks they perform. Marine Traffic performs a number of checks regarding the positions of a vessel in a specific time-window. For example, they check how far a vessel could have travelled in that time frame and they recalculate destination and departure, because AIS is not very valid on that. They also check MMSI/IMO of vessels against other data they purchase from partner association (to verify the details). This is currently done on data entry so they clean as much data before visualization. Following that, a number of processes are executed at specific time intervals on the databases to clean any data that does not make sense. This is done to improve analytics later on. There are also a number of prototype projects currently been built or deployed that run only in specific areas. For example, the machine learning method Support for anomaly detection on specific receivers etc.

In paragraph 2.3 we compare the quality of the different sources of AIS data we have available in this WP: DZ data and national AIS data of Greece, Denmark and Poland. In SGA-2 we will also compare the quality of the data of Marine Traffic and the satellite data of Luxspace on these dimensions. These data sets were not available on time to analyse it before this deliverable. We also hope to have EMSA data available in SGA-2 for a similar analysis.

2.3 Comparing Dirkzwager and National AIS data

To further investigate the quality of the DZ AIS data, we compared this data to the AIS data of the national authorities. Overlapping data was available for Denmark, Greece and Poland. As mentioned, DZ has receivers all over the coastline and main ports of the Netherlands. For the rest of the countries, AIS data comes from partners of DZ, e.g. Marine Traffic, (Greece), AIShub (Poland) and Portvision. Coverage of DZ data might therefore differ between countries.

For Denmark, Greece and Poland, the number of ships and number of messages was compared for different ports for national and DZ data. To this end, the reference frame of ships was used that was already created for the previous deliverable [1] (*/user/tessadew/defframe6all.csv*). This common reference frame of ships was used because we only want to select maritime ships and this reference

[1] This international company is a result of merging an old Norwegian company “Det Norske Veritas” (DNV) and an old German company "Germanischer Lloyd" (GL).

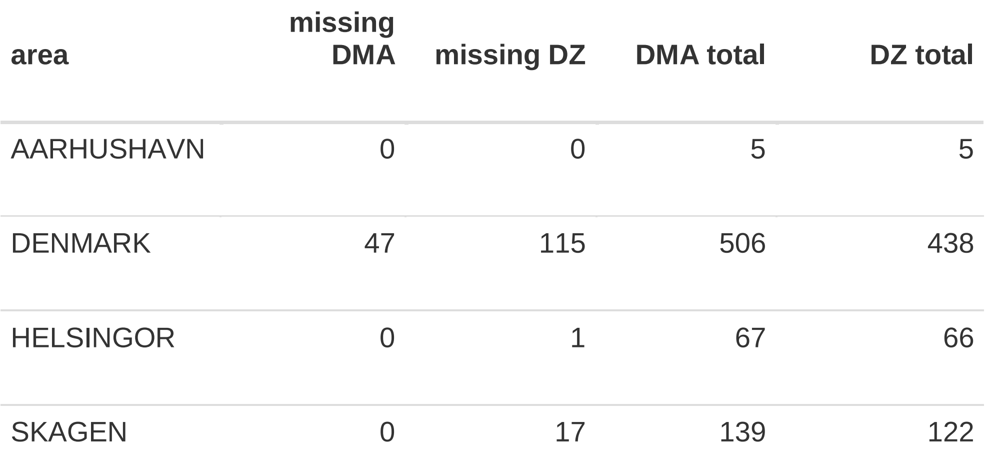
frame is used as a backbone for all our next steps. See Annex 4 for the file with the reference frame of ships and instructions for this analysis.

Denmark

For Denmark the data comes from the Danish Maritime Authority (DMA) and were in a different .csv-format. See Annex 5 for the transformation code. Danish and DZ data for three main ports (Helsingor, Aarhushavn and Skagen) and whole of Denmark were compared for a typical day, December 1st 2015. These ports represent three different area types, namely a regional port without much passing traffic, the busy line between Denmark and Sweden and the passage around Denmark's' most northern point. See Annex 5 for the coordinates of these areas.

The results show that most ships are present in both datasets, table 2.3.1 shows the overlap for all three areas. It can be seen that Skagen shows the biggest loss of data, where 17 ships are only present in the DMA data, and not in the DZ data.

There is also a count for the entire county, but this is not comparable in the same way, since this area contains parts which are known to be out of cover from Denmark, but in cover from Sweden. DZ data does contain data from the Swedish base stations, and should therefore contain some extra ships.



area	missing DMA	missing DZ	DMA total	DZ total
AARHUSHAVN	0	0	5	5
DENMARK	47	115	506	438
HELSINGOR	0	1	67	66
SKAGEN	0	17	139	122

Table 2.3.1: Number of ships in present and absent in 3 ports based on Danish and DZ data

Some ships missing from the DZ data seem to have extremely few messages in the Danish data. It is therefore possible that DZ removes these based on the assumption that they are measurement errors. However, there are also ships with many observations in the national set that are completely missing from the DZ data.

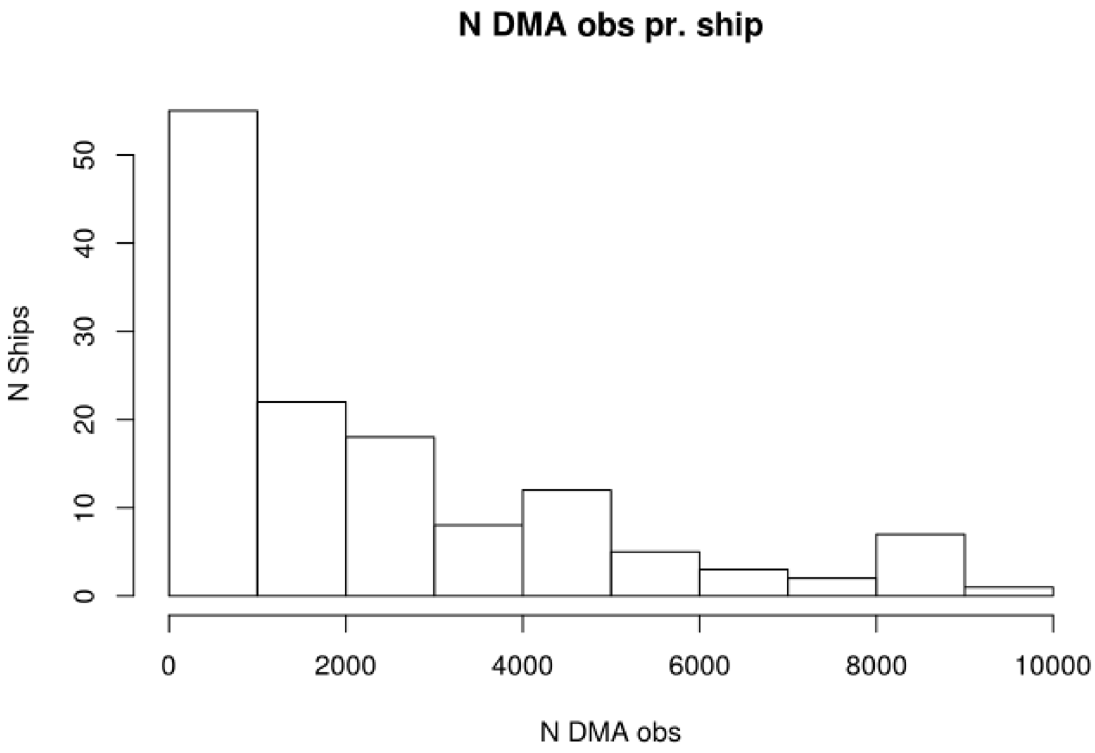


Figure 1: Number of messages by number of missing ships

In general, it seems like DZ data is down sampled compared to the Danish data. Figure 2 shows this down sampling.

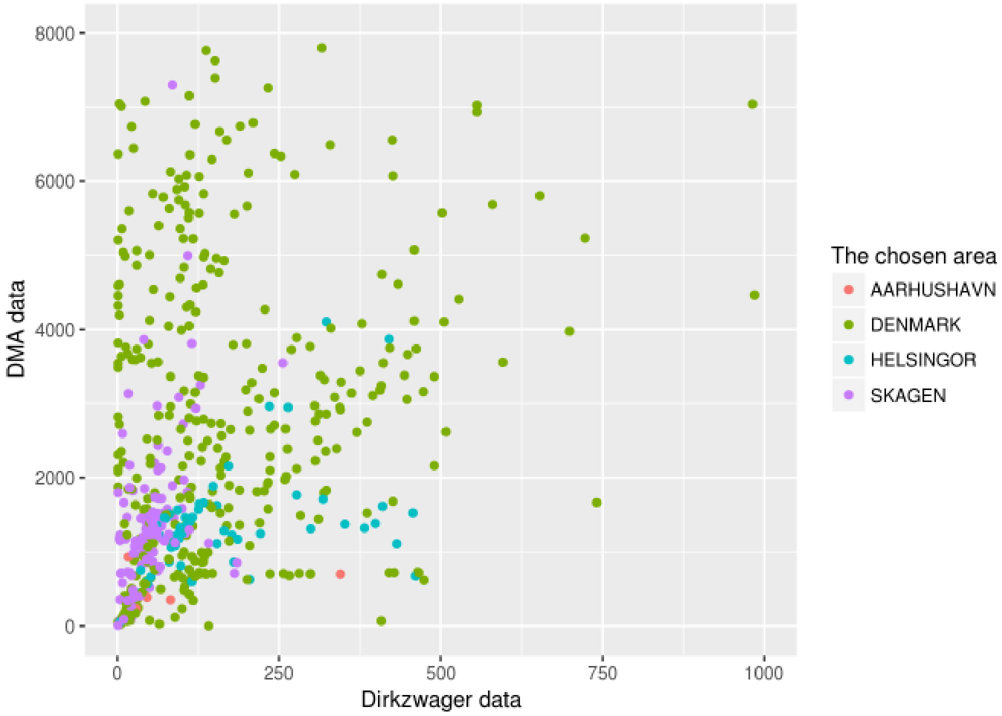


Figure 2: Number of messages per ship for DZ and Danish data

Greece

AIS data from the Hellenic Coast Guard (HCG) was compared to DZ data for a typical day, December 15th2015. The areas of interest were the whole country of Greece and the ports of Piraeus, Thessaloniki, Patras, Volos, and Heraklion (Figure 3) (see Annex 6, Table 1).

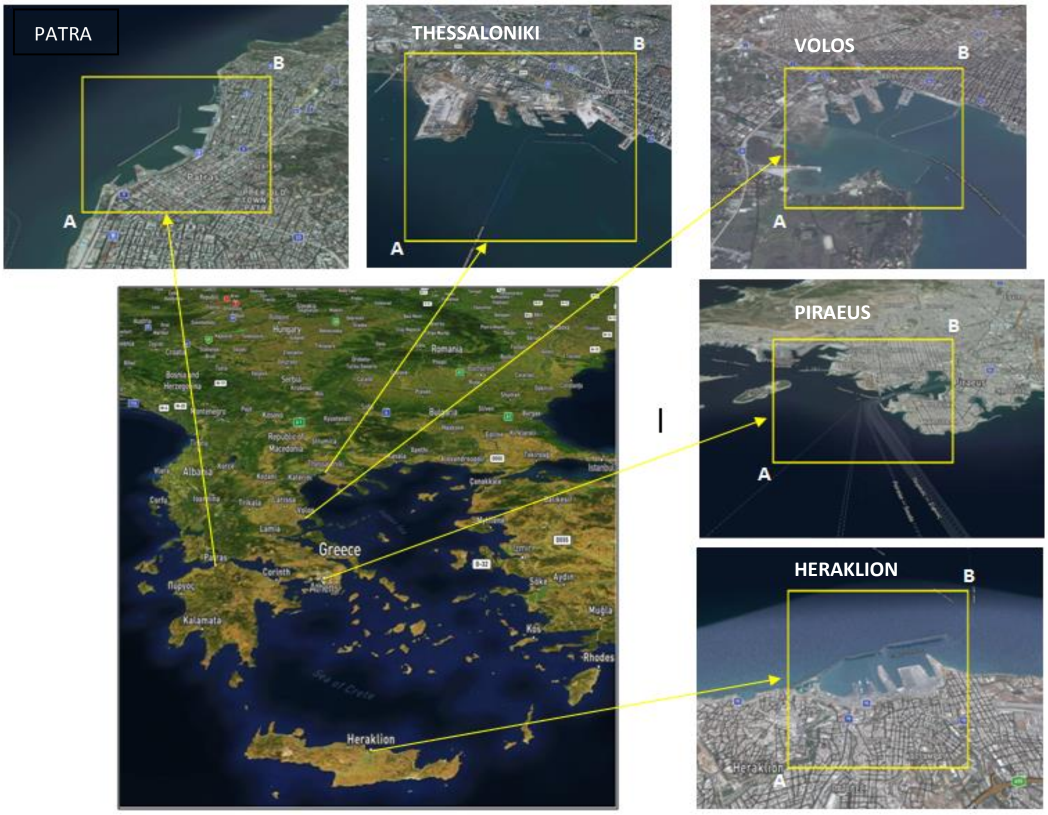
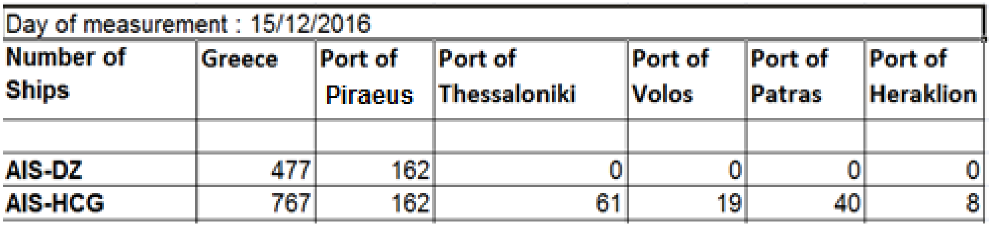


Figure 3: Ports of Patras, Thessaloniki, Volos, Piraeus and Heraklion

The AIS performance was analysed in terms of the number of ships tracked by each dataset across the whole country of Greece and for the five main ports (Piraeus, Thessaloniki, Volos, Patras and Heraklion). Table 2.3.3 shows the results. There is no difference for the Port of Piraeus between the two datasets. However, for the other ports, ships were present only in the HCG dataset.



Day of measurement : 15/12/2016						
Number of Ships	Greece	Port of Piraeus	Port of Thessaloniki	Port of Volos	Port of Patras	Port of Heraklion
AIS-DZ	477	162	0	0	0	0
AIS-HCG	767	162	61	19	40	8

Table 2.3.3: Number of ships tracked by HCG and DZ AIS datasets

Eriksen, Greidanus, Alvarez, Nappo, and Gammieri (2014) examined the sensitivity of tracking ships versus the number of data providers [3]. They found that adding data sources increases the number of ships tracked. However, this increase in the number of ships is smaller as we add providers, because the number of ships detected is getting close to the total number of ships using AIS.

The number of messages was much higher in the HCG compared to the DZ data, as can be seen in Figure 4 which presents the number of messages per ship in the port of Piraeus. There might be reduction method for reducing the number of messages is used to the DZ dataset, but it is not a linear method (see Annex 6, Table 2).

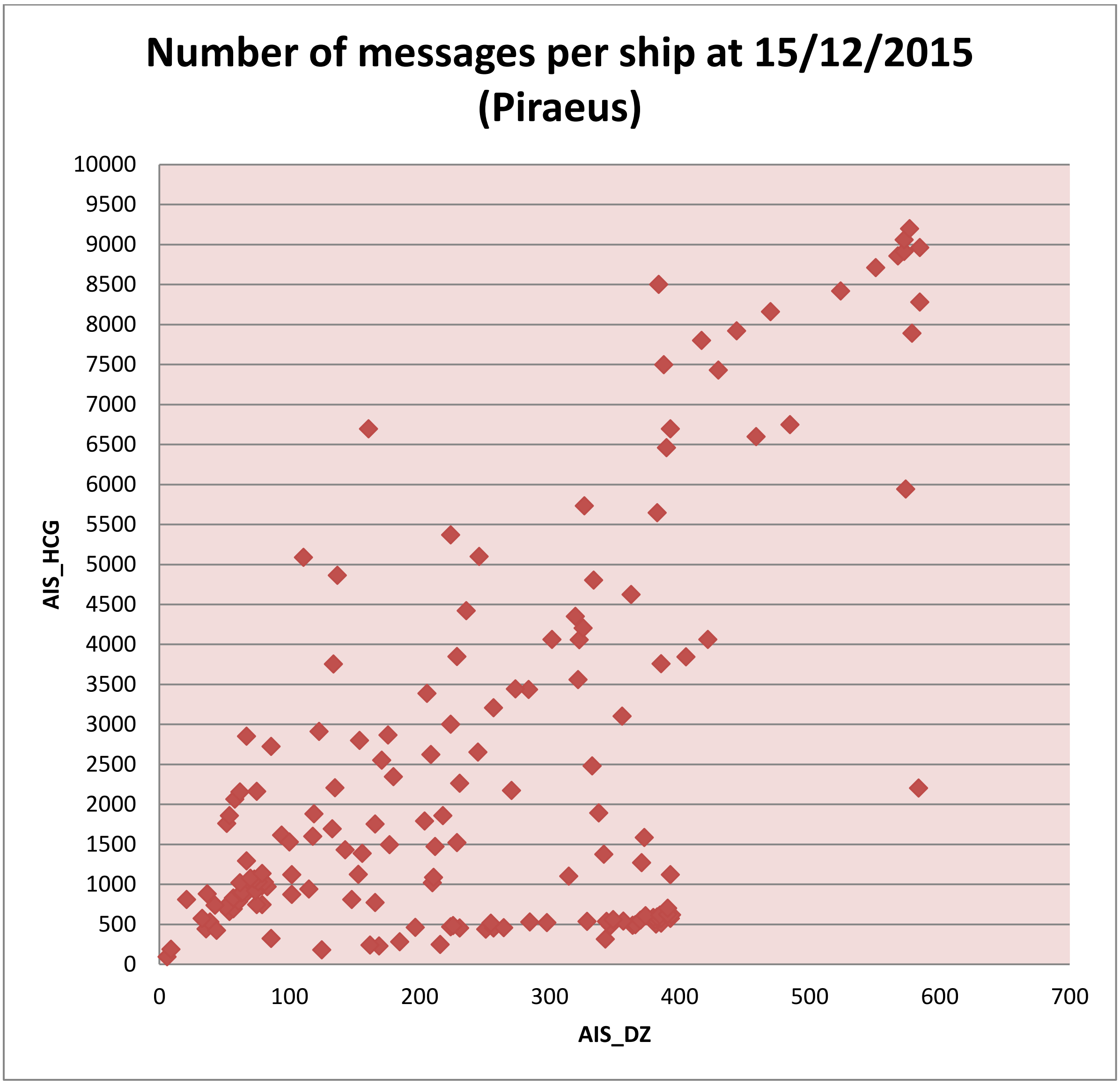


Figure 4: Number of messages per ship for the port of Piraeus

In order to investigate the ships missing from the DZ dataset at **15/12/2015**, we classified the number of messages of these ships using the HCG dataset. As can be seen in figure 5, although for most of these ships the number of messages was less than 1000, there were others with many more messages.

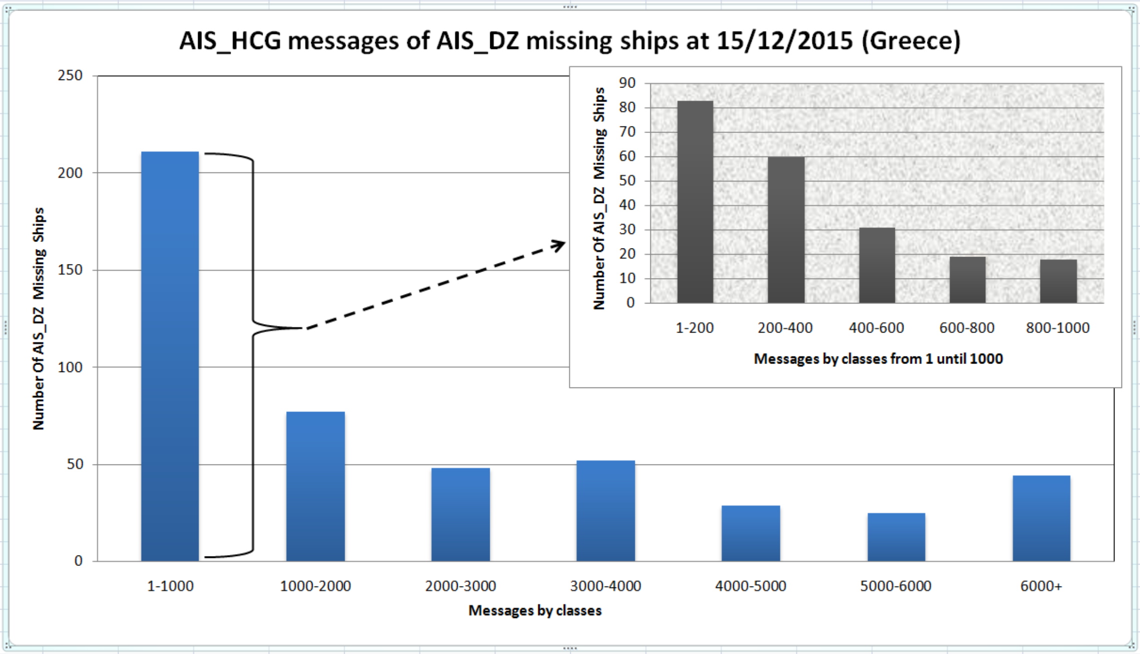


Figure 5: HCG messages of DZ missing ships in DZ data at 15/12/2015

Furthermore, we investigated whether there was a specific type of ship missing in the DZ data. This turned out not to be the case, as all types of ships were absent in the DZ data, see Table 2.3.4. Of course this would affect statistics e.g. when counting the number of port visits.

Type	**Ships**	**%**
Cargo	273	56,17
Tanker	113	23,25
Passenger	57	11,73
Tug	20	4,12
Others	23	4,73

Table 2.3.4: type of ships missing DZ data at 15/12/2015

Given the geography of Greece, which has more than 2000 large and smaller islands most of them located in Aegean sea, we wanted to investigate whether the reduction in the number of messages per ship results in a problem for following the journey of a ship. Figure 6 shows that indeed the reduction of messages in the DZ data may pose difficulties in following the journey of a ship.

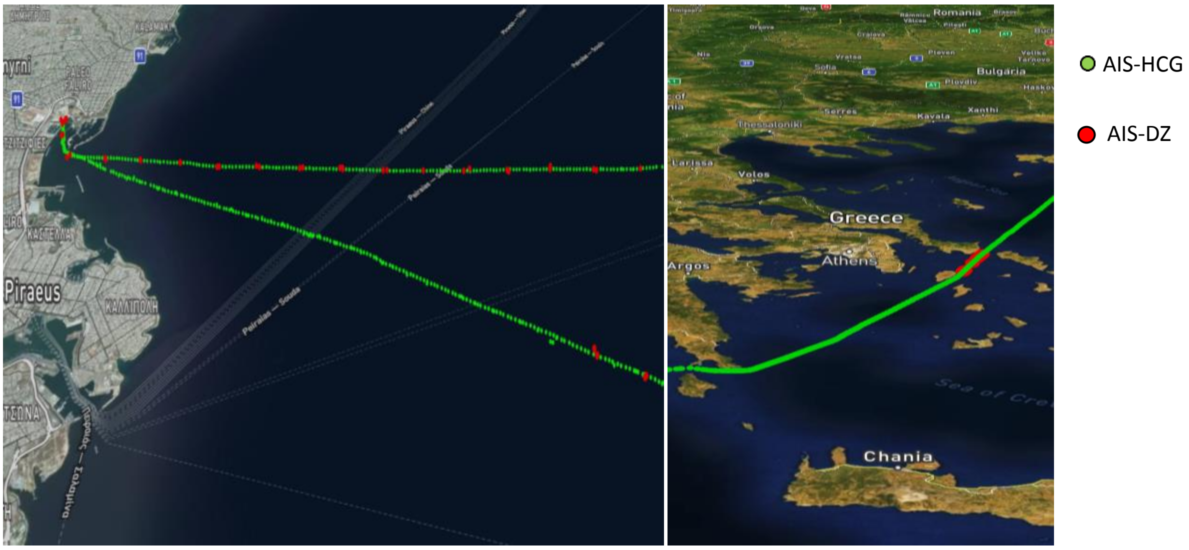


Figure 6: Journey of a ship from HCG and DZ data in the port of Piraeus and in the Aegean Sea

Poland

AIS data from the Polish Maritime Authority (PMA) were compared to DZ AIS data for some typical days, January 1st – 10th 2016. See Annex 7 for checks on the national data. The areas interest were the port of Świnoujście, Szczecin, Gdynia and Gdańsk (Figure 7, see Annex 7 table 1 for the coordinates).

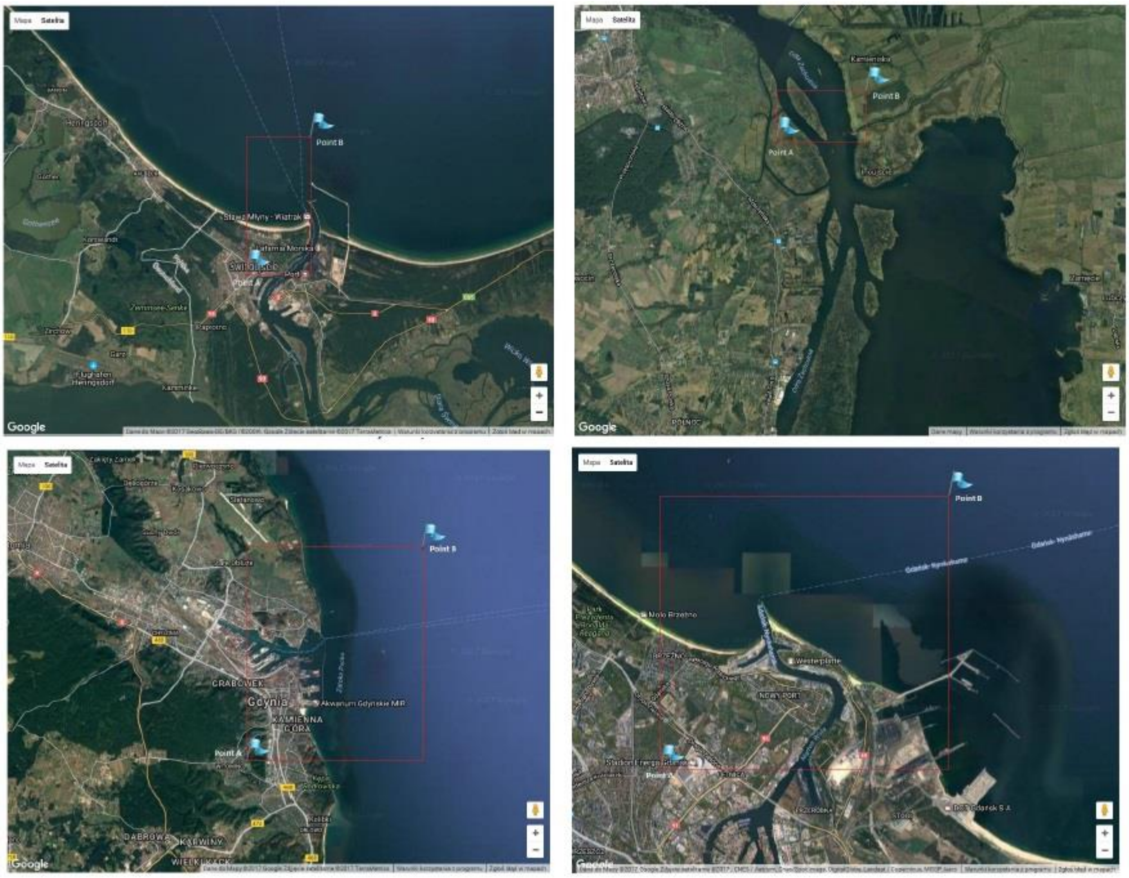


Figure 7: Ports of Świnoujście, Szczecin, Gdynia and Gdańsk

The results showed that for most cases, the number of unique ships for DZ and PMA data is almost the same. Per port and per day, PMA data sometimes contained one more ship than DZ data. However, the number of messages from ships present in both data sets differs greatly, the national data almost always having a higher number of messages. See table 2 till 5 in Annex 7 with detailed results. Here it can be seen that for some ships in some ports, there can be less than 5 per cent of messages of the AIS-PMA data present in DZ data.

Two ports (Świnoujście and Gdańsk) were investigated in more detail for January 1st 2016. For the port of Świnoujście the number of messages of a ship present in both datasets (marked in red in table 2 in Annex 7) was 188 in the PMA data and 11 in the DZ data. Figure 8 shows that that the data does not overlap in latitude and longitude values for PMA and DZ data. However, the route of ship is preserved.

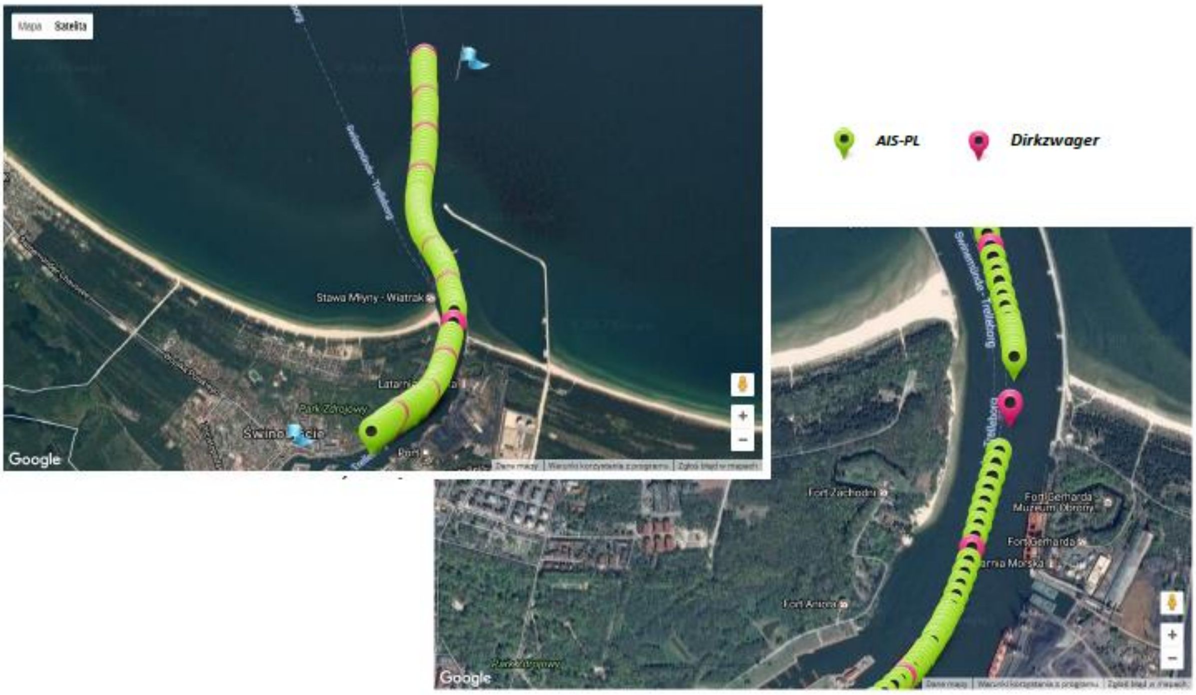


Figure 8: Route of a ship from PMA and DZ data in the port of Świnoujście: two zooms

A similar pattern can be seen for the port of Gdańsk (Figure 9, marked in red in table 5 in Annex 7). Here the number of messages for a ship present in both data sets is 179 in the PMA data, and 14 in the DZ data.

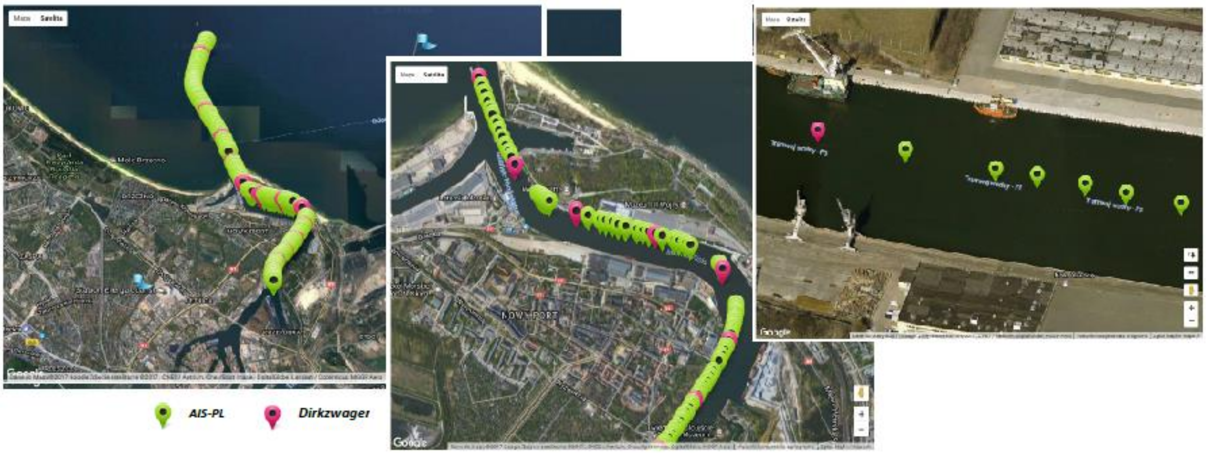


Figure 9: Route of ship from PMA and DZ data, port of Gdańsk: three zooms

As in the previous case, the data from AIS-PMA and DZ does not overlap, but the ship's route is preserved.

All in all, the number of ships is almost the same in the PMA and DZ data. However, mostly PMA contains higher frequent data. National data is much more precise (received data are of a higher frequency) then DZ data and overlaps (over 90%) with statistical survey (TRANSMOR). This 10% difference probably results from errors and methodological differences (e.g., minimal ship size is 100GT for TRANSMOR).

Conclusions

In almost all cases, national data contains much more data than the DZ data. DZ misses data on some ships, which does not seem to be selective for a certain type of ship. This seems to be due to a problem in coverage. For example, for Greece as a whole it has a significant effect, as data on some ports is completely absent in the DZ data. What is more, the number of messages in areas covered by DZ is usually lower in the DZ data compared to the national data. It is clear that DZ data is filtered, but the exact nature of this filtering is not clear, as the reduction of messages per ships differs. This probably differs per provider of the data DZ uses. See also Chapter 2.2 of this deliverable.

In general, we are not satisfied with the present filtering (or information on this filtering), and coverage of the DZ data. Coverage differs per country, which is probably due to DZ's data sources, but if we want to analyze the whole of Europe it does not suffice. If DZ data does cover a port, the data is sufficient to determine the port visits. However, it is not sufficient to determine ships' journeys, especially in areas with a capricious geography. Our algorithm might be able to deal with this, in terms of calculating the right number of journeys, but it will affect the calculated distance. This would result in an underestimation of the distance. This may also have an impact on calculated traffic estimates and emissions. In the near future, we therefore also will investigate other sources like Marine Traffic, Luxspace and hopefully EMSA.

As there was some noise in the data that can result in erroneous measurements (e.g. in terms of distance travelled), we developed robust algorithms to handle this noise. These algorithms are explained in the next chapter. If AIS data in the future would have a better quality, these algorithms would still be applicable.

3. Methodology: defining a journey

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.

The port visit algorithm we have developed calculates the number of ships that have entered and departed a port. This algorithm calculates the time spend in the port and distance travelled within the port. It does this by:

- Processing the data from ships in the reference frame of ships with a median filter over 10 minutes for latitude, longitude and speed. Using this 10 minute window reduces the amount of data, enabling faster processing. It also takes care of the noise in the signal, rendering plausible positions and speeds even in cases where a ship is at anchor and transmits a signal only every 3 minutes. In that case, a shorter time interval would result in a single measurement that would not be corrected by other measurements (see also previous deliverable [1]). In other cases, a different temporal filter could be preferred, so we want to make this time interval variable in the algorithm.
- For each data point in this filtered selection the location is categorized as being at sea or in the port.
- Using (selected) MMSI, time and location category only, the port entries and departs are determined by selecting consecutive locations where category location changes. An entry being defined as a SEA-PORT couple and departure by a PORT-SEA couple. A successive entry and departure are coupled to form a visit interval, already resulting in the total visit time.
- If two visits are separated by one time interval (10 minutes in this case), the interval is coupled to form a longer visit interval.
- For the resulting visit interval, latitude, longitude and speed are combined again. Then, speed is used to define an actual visit/stop as some ships only travel through a port. For example in the port of Amsterdam the port area is also used for ships passing. Thus, only ships are selected that at some point have a speed of <0.2 knots.
- Using the haversine function (which determines the great-circle distance between two points on a sphere given their longitudes and latitude), distance travelled in port is calculated, see also PoC 4.1.

The result of the previous analyses provides us with an output-driven method to define a journey. Using the departure of ships gives us the start of a journey. The end of the journey can be determined in three ways:

- The ship enters another port: a reference frame of ports has to be constructed to be able to categorize the ship as being at SEA or in a PORT. Again the successive couple SEA-PORT will signal the entry of the new port, constituting the end of the journey. The port visit method can be used to determine next destination, time of journey and distance covered during 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.
- The ship leaves the area of AIS coverage. At this time, we have European coverage only. Therefore, we will have to use the boundaries of our data as end of a journey if a ship travels

outside Europe. However, there are also areas of sea where satellite reception is minimal or ships can even turn off their AIS signal. If this is the case, we will use the next available information to linearly combine the data.

Processing can be optimized by filtering out AIS data in which the speed and heading of the ship have not changed since the last message. This optimization might be performed in the future, but is not in the scope of the current project.

The calculation of the intraport distance was part of our first PoC. Additionally, we performed PoC's to investigate the feasibility of using AIS data to define ports, to improve next destinations and to complete fluvio-maritime transport. These four PoC's are described in the next chapter.

4. Proof of concepts

All PoC's are based on DZ data, only the PoC on "Next destination and average distance matrix" is based on Danish national data.

4.1 PoC Calculating intra-port journey distances using AIS

Port authorities do not always have a complete insight in the activities within their port. For example, they do not always know whether ships visit one or multiple terminals during one visit, rendering a higher intra-port journey distance. Using AIS, these journeys in the port, and thus the intra-port distances can be derived. Intra-port travel distances is a new statistical product that could be generated by using AIS data. In this PoC we developed an algorithm to calculate these intra-port journeys.

The algorithm

Using the methodology developed for defining journeys (see Chapter 3), the point and time for a ship entering the port can be defined. Between the time of entering and departing the port, data points are selected. Then the data is median filtered over a 10-minute interval. This 10 minute interval renders sufficient valid measuring points in case a ship is at anchor and only transmits a signal every 3 minutes (see also deliverable 2, chapter 4). For each point and each successive point, the latitude and longitude are fed into the haversine function. The haversine function determines the great-circle distance between two points *on a sphere* given their longitudes and latitude:

$$\operatorname{hav}\left(\frac{d}{r}\right) = \operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)$$

- Where hav is the haversine function:

$$\operatorname{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$$

- *d* is the distance between the two points along a great circle of the sphere
- *r* is the radius of the sphere,
- φ_1, φ_2: latitude of point 1 and latitude of point 2, in radians
- λ_1, λ_2: longitude of point 1 and longitude of point 2, in radians

The algorithm we developed in Scala can be found here:

https://github.com/mputs/WP4/blob/master/Portvisit2/src/main/scala/aisframe.scala

In the future, it would be interesting to develop an algorithm that can detect intra-port movements, where a ship that moves from one terminal to the other within the same port can be automatically detected. Another interesting aspect that could be investigated is the detection of anomalies in the movements of ships that could signal problems in the ports.

4.2 PoC Data Driven Port Definition

At the moment, defining ports for journeys or port visits in AIS is done manually. In other words, to build a reference frame of ports or terminals a bounding box or a combination of bounding boxes has to be defined as the region of interest. We have done this by visually inspecting Google Maps, selecting minimum and maximum latitude and a minimum and maximum longitude as the region of interest. As many countries have a lot of different ports, this constitutes a time costly job. Of course, when the analysis is even more detailed and zooms in on terminals, the number of regions of interest to be defined explodes. Besides the work on the initial reference frame of ports, this frame has to be kept up to date as the locations and functions of ports are dynamic. It is therefore desirable to use actual event data to define ports. Using AIS for this enables us to automate a lot of manual work, but it also facilitates keeping the register of ports up to date. In this PoC we investigated the possibilities of specifying ports data driven using AIS.

Before setting out to develop a port definition algorithm, we investigated what work in this field was already done. Interestingly, Marine Traffic already developed a data driven algorithm for defining the port of Rotterdam by means of AIS [4]. For this, they used a Kernel Density Estimator (KDE). KDE is a non-parametric way to estimate the probability density function of a random variable. It is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample (see [5]). In our case, KDE can be used to infer the probability of a ship being at anchor at a specific location.

We conclude that this PoC has succeeded: it is possible to build a data driven algorithm for defining ports. In the near future, Statistics Netherlands and Marine traffic will collaborate on this and other issues. First, we would like to investigate building a reference frame of ports. Also, taking into account locations where ships anchor, but are not actual ports. Then, it would also be interesting to zoom in on defining the type of terminals. By assuming that type of ship reflects the type of terminal, type of ship will be taken into the model. Resulting in a reference frame of ports and terminals.

4.3 PoC Next destination and average distance matrix

This proof of concept has two goals:

1. Comparing reported and observed destination.
2. Check if distances between port pairs can be estimated.

First analysis is comparing the next destination port, as reported in the AIS messages by the captain, to the actual port visited.

For the purposes of this PoC, a port is defined as a polygon, i.e. a linear ring consisting of four or more points. Only a few Danish ports have been defined manually, and software has been written for defining ports by drawing polygons on maps. For each ship, all observed coordinates are arranged in chronological order and they are each tested for overlap with the polygon. The port visit is then defined as the first time a coordinate set is detected inside the polygon until the next coordinate is observed that is outside the polygon.

For each port visit, the destination of the ship is reported immediately before it entered the port. The next destination of the ship was reported immediately after exiting the polygon. Since the variable destination is part of the voyage related AIS messages, and the coordinates come from the position messages, the time may not be the actual time. Usually the difference only was a few minutes. However, the difference could be up to 7 minutes, in rare cases even more.

The number of coordinate sets contained within the time that the port visit lasted is also recorded, only ships with a close to expected number of recorded observations will be included in the results.

One issue with this algorithm is that corrupted readings can sometimes break port visits up into two, because a single coordinate pair will look like it is outside the polygon. A filter is applied to reject such measurements.

Comparing reported and observed destinations

For the analysis only three ports (Greenaa, Aarhus, Anholt) are used. They have known traffic patterns and represent different sizes. On a single day, 421 port visits were detected, and 209 of these were made by ships with an IMO number from 98 unique ships. Special ships, like coast guard patrol vessels making up the majority of the visits.

The results where discouraging. Many ships had no registered destination information on either entry or exit. Even for ships with IMO numbers a majority of visits had no meaningful destination information. Also, many ships did not have the port they arrived as their reported destination. While the majority of those that entered AARHUS had some entry destination information, it was often misspelled or more text was given such as “TO AARHUS”.

Most of the route ferries had marked their destinations as a constant. So the AARHUS-ODDEN Ferry was always had AARHUS-ODDEN as the destination. Making it harder to determine where it is currently headed. For some reason, the entry destination information was very often missing, relative to the exit information, further analysis will have to be done as to why that is.

This part of the PoC shows that in a lot of cases information on next destination is missing or filled out wrong in the reported data. By AIS data we can observe the next destination. So AIS data is a great source to improve information on next destinations in current statistics.

Estimating travel times and distances between ports

For each port pair (from the three ports), we identified all ships that have visited both ports, with no intermediate visits at other ports. For ships that are always in sight, one can check the route and ensure that the ship travelled only from A to B. In this PoC, only ships like this were considered. A few known Danish routes where studied for this PoC.

The results were good. All the passenger ferries were detected as expected, and times where aligned with expected measurements. The next figure shows the example of the major freight and passenger route Roedby Puttgarden.

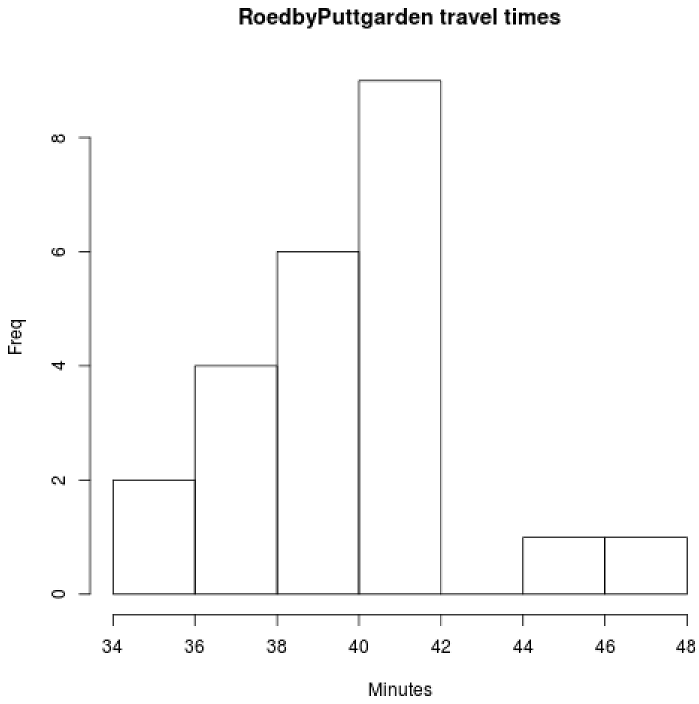


Figure 10: travel times between Roedby and Puttgarden

In conclusion, distance estimation in time and space is viable for a small number of port pairs. For some port pairs with more than a few ships doing regular trips, some distance measures have very noisy distribution. Further work with subject matter experts is needed to decide exactly what kind of distance measure would be relevant for such pairs. In addition, future work is necessary for ships that are not in AIS covered area all the time.

When using AIS covering the whole world (in case of satellite data), AIS detected distances may even replace the distance matrix, as distances can be tracked for every journey of every ship. Perfect world coverage will probably not be accomplished in the near future, but we believe that strong algorithms can be developed to deal with these blind spots.

4.4 Completing Fluvio-maritime transport by using AIS data

Fluvio-maritime transport, transport by ships that travel across both sea and inland waters, has caused discussion on the implementation of the concept for Eurostat. This resulted in an inconsistent measurement of the concept. Eurostat proposes that the concept should be based on the nature of the vessel. For the Netherlands, this might result in incomplete data delivered to Eurostat as maritime vessels performing fluvio-maritime transport will not be part of the data. In this PoC we investigate the extend of the fluvio-maritime transport in terms of number of seagoing vessels. Maritime ships on inland waters for the Netherlands. For the group members of WP4, this idea is only relevant for Netherlands. However, France, Germany, Belgium and the United Kingdom could use these results to improve their current statistics on fluvio-maritime transport.

In so-called fluvio-maritime transport, transport is partly performed on inland waterways (IWW) and partly on sea. For Eurostat, this has created discussion about whether fluvio-maritime transport should be reported as maritime transport, IWW transport, or both. [ESTAT/E-3 Doc.MAR-2016-05.2 Minor corrections] Article 2 of the maritime Directive 2009/42/EC states that:

"(a) 'Carriage of goods and passengers by sea' means the movement of goods and passengers using *seagoing* vessels, on voyages which are undertaken *wholly or partly at sea*. (b) 'Seagoing vessels' means vessels other than those which navigate exclusively in inland waters or in waters within, or closely adjacent to, sheltered waters or areas where port regulations apply."

Similarly, Article 3 of the inland waterways Regulation 425/2007 states that:

(a) "Navigable inland waterways" means a watercourse, *not part of the sea*, which by natural or man-made features is suitable for navigation, primarily by *inland waterways* vessels; (b) "Inland waterways vessel" means a floating craft designed for the carriage of goods or public transport of passengers which navigates predominantly in navigable inland waterways or in waters within, or closely adjacent to sheltered waters or areas where port regulations apply; (d) "Inland waterways transport" means any movement of goods and/or passengers using *inland waterways* vessels which is undertaken *wholly or partly in navigable inland waterways*;

(h) "Inland waterways traffic" means any movement of a vessel on a given navigable inland waterway."

Countries have non-harmonised approaches to report fluvio-maritime data in the maritime and IWW statistics. For example, Germany does not report transit transport through the Kiel Canal. Two countries report fluvio-maritime in both IWW and maritime statistics, while two other countries report fluvio-maritime only in maritime statistics. Some countries use the type of vessel while one country uses the port of loading/unloading to determine between IWW and maritime statistics [6].

Eurostat proposed that fluvio-maritime transport should be reported on the basis of the nature of the vessel (Working Group on Maritime Transport Statistics 23-24 May 2016). In other words, fluvio-maritime transport performed by an IWW vessel should be reported in the IWW statistics and fluvio-maritime transport performed by a seagoing vessel should be reported in the maritime transport statistics. The only exception from this, transport using seagoing vessels undertaken wholly on inland waterways, should be considered as IWW transport. If type of vessel information is unavailable in the source data, related information (such as port of loading/unloading) could be used to determine whether the fluvio-maritime transport is likely to be carried out by IWW or seagoing vessels.

In this PoC, we investigated extent of missing data on fluvio-maritime transport by maritime ships for the Netherlands. Fluvio-maritime transport by IWW ships is already part of the Dutch IWW statistics. However, the Netherlands does not report fluvio-maritime transport if it is performed by maritime ships. The definition we use here is that maritime are ships that have an IMO-number, IWW ships are ships with an EU-number. The extent of this missing data is not clear. We therefore investigated the number of maritime ships travelling to Dutch IWW ports, as these are not port of both maritime and IWW statistics. We also investigated the number of maritime ships travelling over Dutch IWW to Belgian and German ports.

We investigated the number of maritime ships at four locations for a period of 9 days, (March 1-9th):

1. Port of Nijmegen (River Waal)
2. Port of Born/Stein (River Maas)
3. Location on the route to Germany: Hardinxveld-Giessendam (Boven-Merwede)
4. Location on the route to Belgium: Sas van Gent (Ghent-Terneuzen Canal)

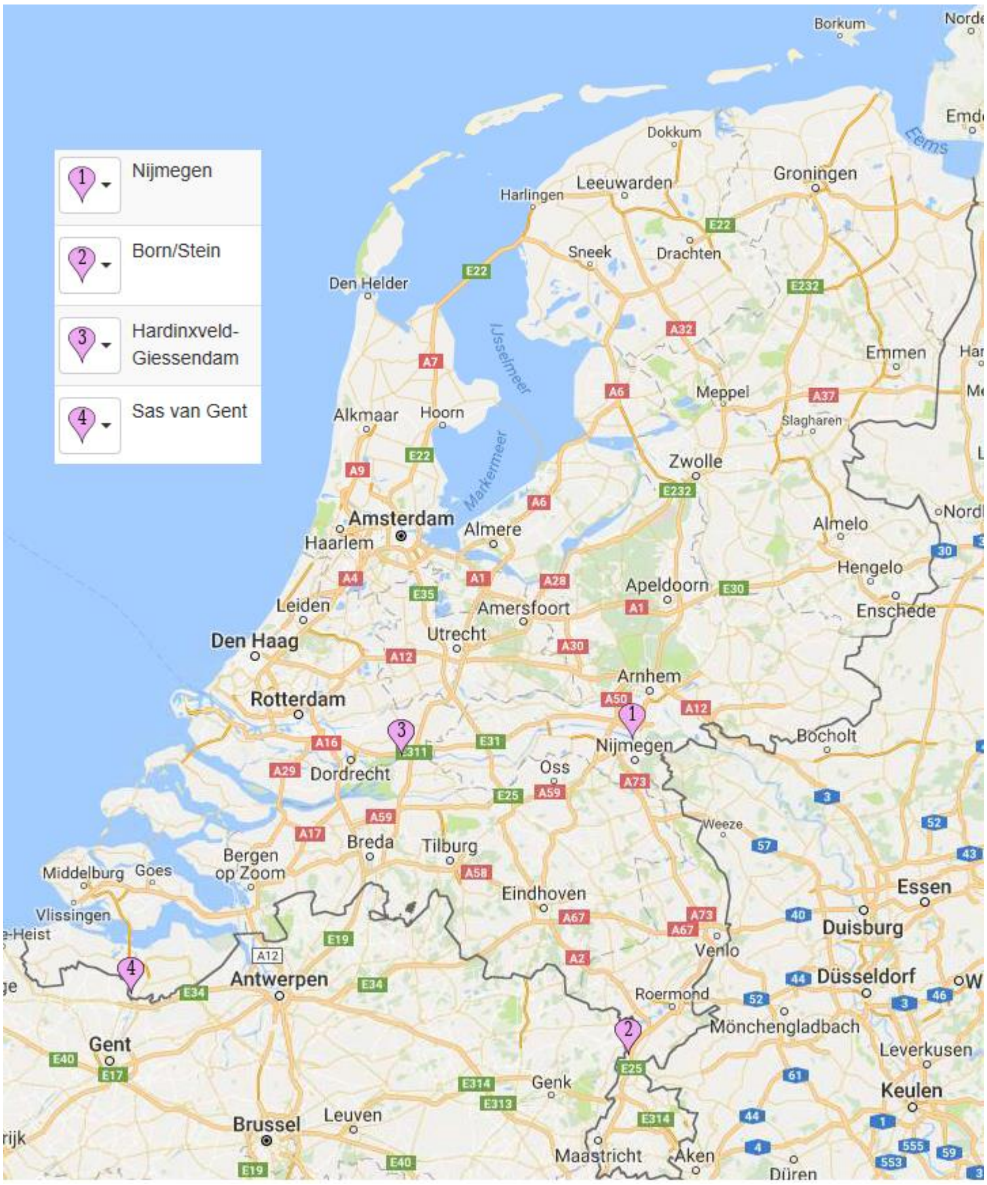


Figure 11: used locations for analysis

Results showed that there were no maritime ships in the ports of Nijmegen and Born/Stein. Thus, there was no fluvio-maritime transport by maritime ships to Dutch IWW ports. However, analysing maritime ships en route to Germany and Belgium by IWW, we found 33 ships going to Germany and 103 ships going to Belgium. If these ships also visited Dutch maritime ports first, they would be part of our maritime statistics. We found that of the maritime ships going to Germany and Belgium, 78 ships loaded and/or unloaded goods in Dutch maritime ports around that period (February and March). The other 58 ships did not visit Dutch seaports in this time period, so the Netherlands has no further data on these ships. They probably came from the United Kingdom or other countries not accessible solely by IWW, constituting fluvio-maritime travel.

We conclude from this that fluvio-maritime transport (with loading/unloading in the Netherlands) probably constitutes a minimal part of transport statistics in the Netherlands. However, from the perspective of traffic intensity, emissions and transit trade, it is interesting to further investigate these fluvio-maritime journeys. What we have not investigated here is the relationship between maritime and IWW transport per se. As AIS can also be used to gain insight in this relationship because direct links might be seen. That is, if goods are loaded and unloaded for these ships around the same terminal and time, it can be seen which goods are shipped from maritime ships to inland waterway ships and the other way around.

Overall, AIS can be useful to investigate fluvio-maritime transport.

5. Results of sea traffic analyses using AIS

We explored the possibility of calculating the number of ships during a certain time interval at certain coordinates by using AIS-data. Because this information could be interesting for traffic and economic analyses.

We calculated the traffic intensities of ships around Europe. First, the grid was defined as areas of 10000 square kilometres.

An important perquisite is that the grid elements all should have the same size. If this would not be the case, the count of vessels would depend 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 program counted the amount of ships in each cell of the grid during one hour on a randomly chosen day. Figure 12 shows the result. From this figure it is clear that we didn't use the right coordinates for defining the whole of Europe. We also were not sure it the defined grid was the most optimal choice.

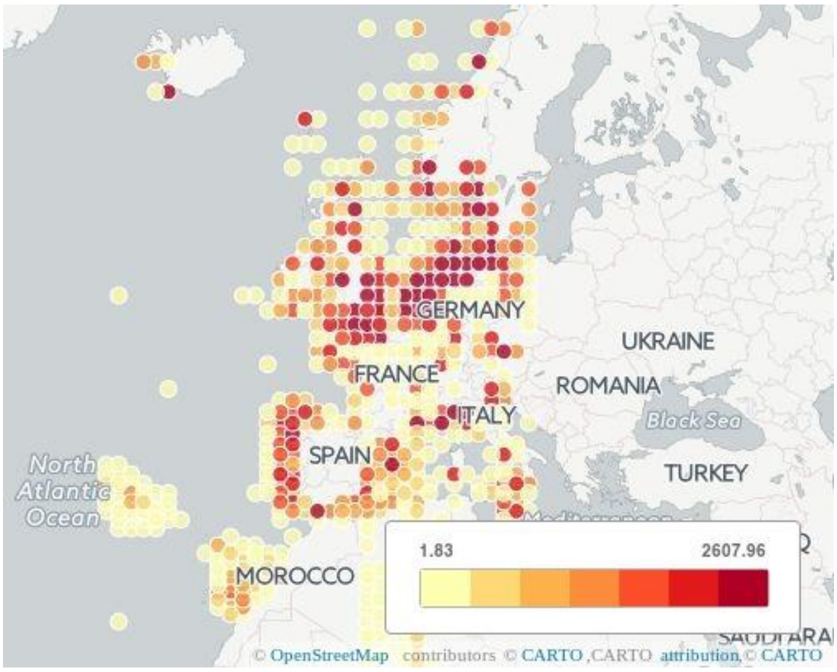


Figure 12: result on traffic analyses: the amount of ships in each cell of the grid during one hour on a randomly chosen day based on 100x100 grids

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 an excellent one for describing locations at sea. WGS1985 is also described as GPS coordinates, since GPS uses WGS1985 for their basis. For other uses, other coordinate systems are defined. For instance, ETRS89 is a coordinate system where the ellipsoid is chosen in such a way that it models Europe well. ETRS89 and WGS1985 do not differ that much. They use the same semi-major axis ("diameter of the earth", which is defined as 6378137.0 meters). 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 and since we are analysing at sea levels and AIS is based on GPS, WGS1985 is preferred.

For drawing maps, several methods are available to project the earth on a map. One of them is the Lambert Azimuthal projection, which preserves the area. In this projection, a surface touching the sphere at point S (see figure 13) 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. This projection preserves surface area under the transformation. The transformation itself is described at [7] and [8].

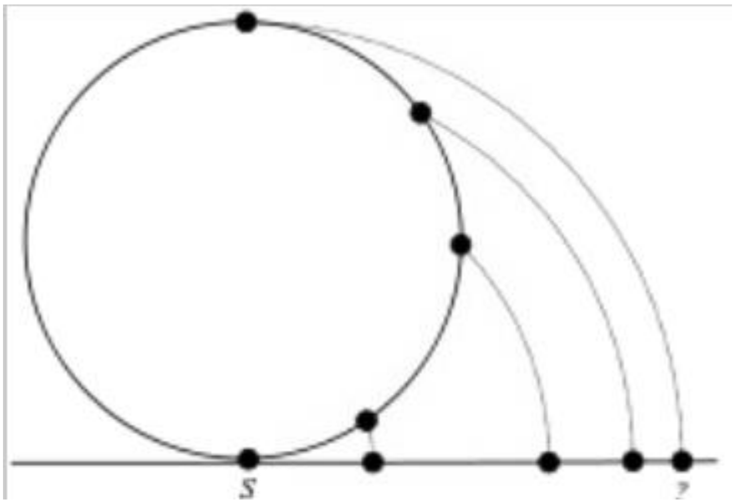


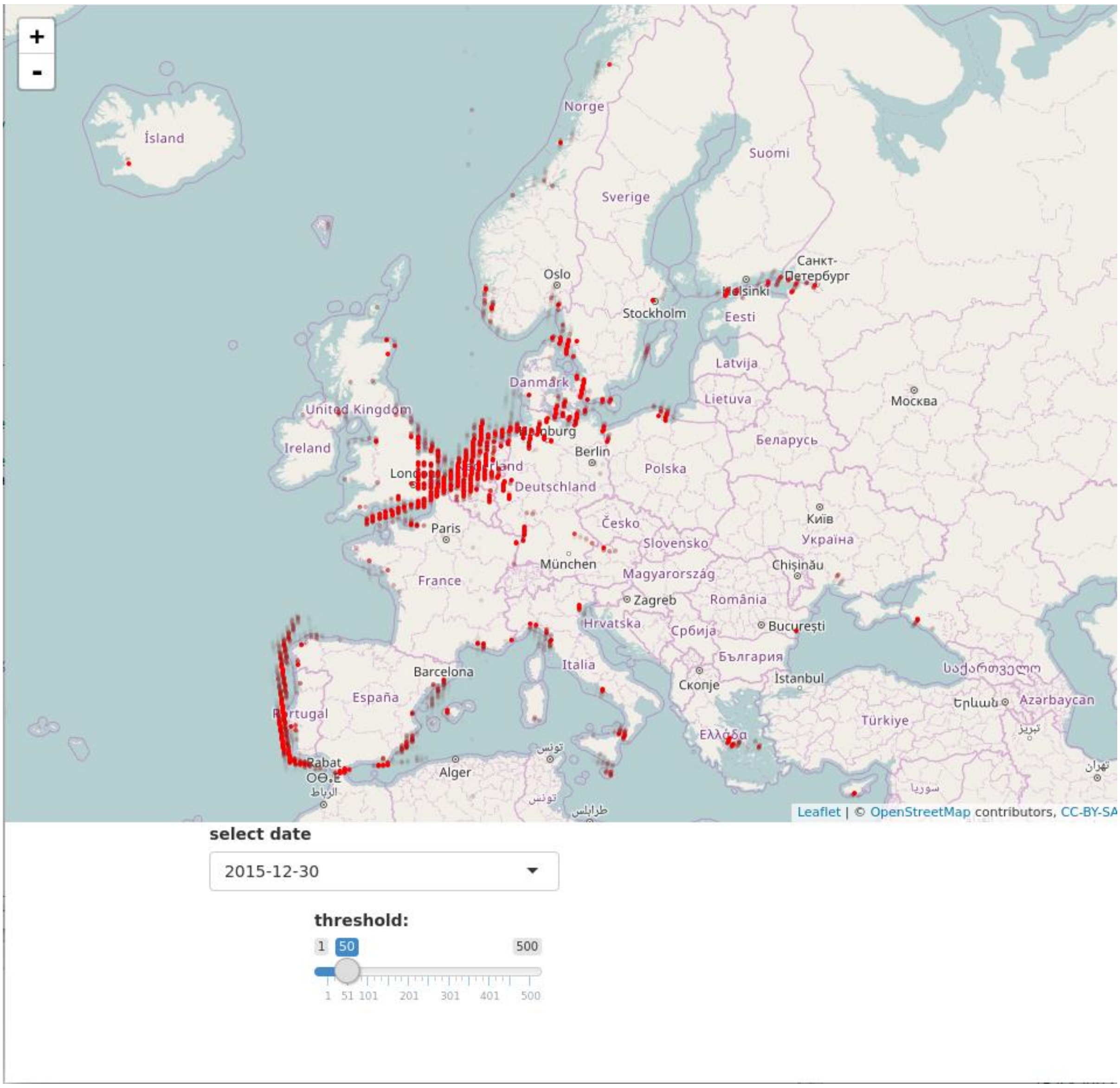
Figure 13: explanation of the Lambert Azimuthal projection

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.

Using this projection in Spark, and henceforth Scala, we implemented the projection (in the file named LAEA. Scala), together with the definition of the grid. For the grid, a bounding box was defined and subdivided in 200 by 200 grid points. Then, (in the file named countuniq.scala), the methods in the file LAEA.scala are used to project all latitudes and longitudes and find the index of the associated grid point. The latitude and longitude of that grid point is given back.

The final visualization is done in Shiny in combination 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 14 and 15 show two examples of the visualization with both a different threshold. In the visualization, one can choose the date out of an available date list in the lower end, where a slider is available for selecting a saturation threshold for the visualization. This means that all cells being more occupied than the threshold are displayed as dark red and all less visited locations are less red. Playing with the slider gives inside in more and less occupied regions in Europe. For the regions that are not displayed in dark or less red, there is no data available in the DZ dataset on that specific day. Also very low intensities are made invisible.

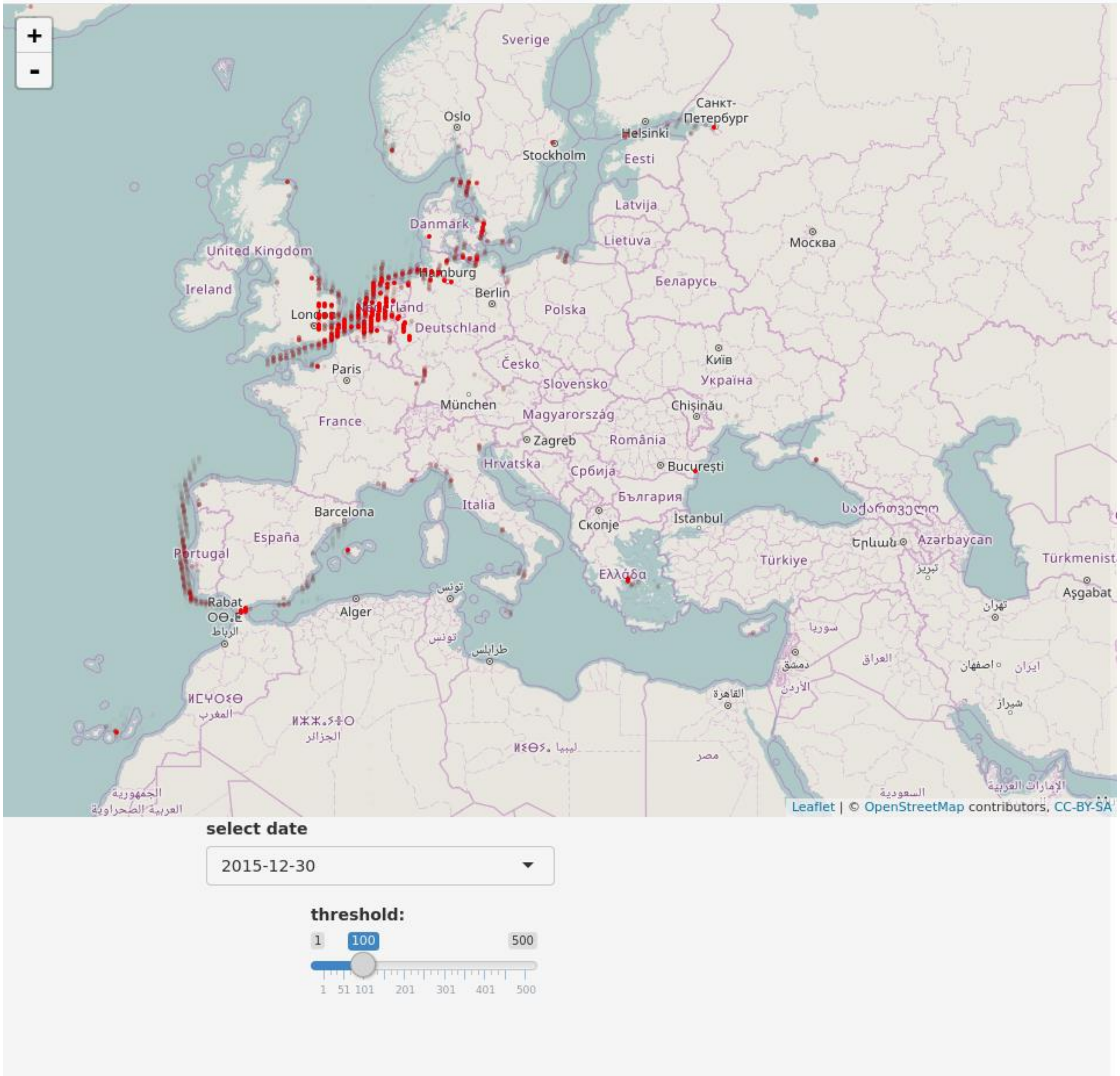


select date

2015-12-30

threshold: 50

Figure 14: result on traffic analyses: the amount of ships in each cell of the grid during one day based on the Lambert Azimuthal equal area projection for a threshold of 50



select date

2015-12-30

threshold:

1 100 500

1 51 101 201 301 401 500

Figure 15: result on traffic analyses: the amount of ships in each cell of the grid during one day based on the Lambert Azimuthal equal area projection for a threshold of 100

6. Lessons learned

In this chapter we will discuss overall conclusions of this deliverable, pitfalls and look into the future by shortly outlining plans on SGA-2.

6.1 Overall conclusions

In this deliverable we further investigated the quality of AIS data. We started with analysing AIS data from data provider DZ specifically, by means of the quality and metadata framework. From this, we concluded that the quality is good. Almost all factors of the quality framework are judged as mostly positive. Only “spatial coverage” and “transparency and soundness of methods and processes for the metadata and the data” are insufficient. Privacy is also an issue that needs to be researched further.

As for validation checks, DZ adds timestamps and performs validation checks on aspects such as position of ships and ordering overlapping data sources. National AIS data of Denmark, Poland and Greece are completely unfiltered and untreated.

In almost all cases, national data contained (much) more data than DZ data. DZ misses data on complete areas in coastal Europe. Thus, ports visits and journeys cannot be analyzed for all European ports and ship routes. The number of messages in areas covered by DZ is usually lower in the DZ data compared to the national data. It is clear that some of DZ data is filtered depending on the data sources, but the exact nature of this filtering is not clear, as the reduction of messages per ships differs. In general, we are not satisfied with this filtering (or information on this filtering), and coverage of the DZ data. Coverage differs per country, but if we want to analyze the whole of Europe it does not suffice. If DZ data does cover a port, the data is sufficient to determine the port visits. However, it is not sufficient to determine ships’ journeys, especially in areas with a capricious geography. Our algorithm can deal with this, in terms of calculating the right number of journeys, but it will result in an underestimation of the calculated distances. The lower frequency of messages can also impact calculated traffic estimates and underestimate emissions.

Because of the noise in the data, we developed robust algorithms to handle this noise. We developed an output-driven method to define a journey. Using the departure of ships gives us the start of a journey. The end of the journey can be determined in three ways:

- The ship enters another port
- The ship anchors
- The ship leaves the area of AIS coverage

Processing could be optimized by filtering out AIS data in which the speed and heading of the ship have not changed since the last message. This optimization might be performed in the future, but is not in scope of the current project.

We also performed four PoC’s. The outcomes are promising. The first PoC, on developing an algorithm to calculate the intra-port journey by using AIS data, succeeded. Intra-port travel distances can become a new statistical product. In the future, it would be interesting to develop an algorithm that can detect intra-port movements, i.e. where a ship that moves from one terminal to the other within the same port can be automatically detected. Another interesting aspect is the detection of anomalies in the movements of ships signalling problems in the ports.

The second PoC, on using AIS data to define ports has succeeded: it is possible to build a data driven algorithm for defining ports. In the near future, Statistics Netherlands and Marine Traffic will collaborate on the possibilities of building a reference frame of ports. Then, it would also be interesting to zoom in on defining the type of terminals.

From the third PoC we conclude that next destination as reported by captains is not a usable variable compared to the observed next destination. We also conclude from this PoC that distance measures in time and space can be done. More work is needed to handle areas where coverage is not perfect. It is also interesting to compare the distances in the port to port distance from Eurostat to the port to port distances based on AIS data maybe resulting in using actual AIS journey data instead of the average distance matrix in the future.

Finally, the last PoC shows that AIS data is useful to investigate fluvio-maritime transport. From the perspective of traffic intensity, emissions and transit trade, it is interesting to further investigate these fluvio-maritime journeys. This could also be used to gain insight in the relationship between maritime and inland waterway transport.

Although we are not satisfied with the quality of the DZ data yet, we conclude AIS data itself can help improve current statistics. AIS data is also useful to analyse sea traffic and to analyse variations in time. By having new data sources like Marine Traffic, EMSA and Luxspace (satellite data) available in the future the possibilities of AIS data seems to be even more promising.

6.2 Pitfalls

In many cases our research included a path of trial and error. Communicating good practices is important but communicating and documenting pitfalls is also important. Some pitfalls we encountered:

- As people like to use software they are used to, it can be tough to get people to use the same software, even if this was agreed upon.
- Using big data that can contain different kinds of errors in combination with the sheer amount of data means thorough research in the kind of errors (technical, human) and with that the process of validating (AIS) data is essential. For example, we thought just checking the validity of the ships identifying number (MMSI/IMO) was sufficient. However, due to large amount of data and errors some erroneous MMSI's and IMO's were valid identity numbers, e.g. in terms of length, but were still incorrect. Therefore, we had to build in an extra step in the process in which the number of occurrences of an MMSI-IMO couple was used as a filter.
- Temporal filtering of a big data set makes it easier to handle the amount of data and remove errors. However, when using this filtering it is important to keep in mind that this filter does not suit all conditions in both geographical en temporal terms. For example, when visualising a journey of a ship, a 10 minute median filter may be too rough: zooming in shows ships going over land. Also, in using a 10 minute filter one loses temporal information needed for short journeys for example for ferries making short trips.

A complete list of pitfalls and bad practices will be presented in the final deliverable of this WP4.

6.3 Future work: SGA-2

In our previous deliverable [1] we described ideas for improving current statistics by using AIS data. See table 6.3.1.

Nr.	Problem
1.	Information on the next destination of departing ships is incomplete. This can also be used to construct new tables with to and from traffic matrixes
2.	Not all ports are well-specified, they are sometimes misclassified by port authorities
3.	Distance travelled per ship is now based on an inaccurate average distance matrix for ports
4.	Fluvio-maritime transport is incomplete
5.	Investigate relationship between maritime and inland waterway transport
6.	Intra-port travel distances are unknown
7.	Missing Information on travel routes for goods to estimate unit prices for transit trade statistics
8.	Current statistics on fuel consumption and emissions are not accurate enough.
9.	Small ports experience response burden from the survey
10.	Customers need faster information on maritime statistics
11.	Experimental ideas: now-cast economic time series on the basis of AIS

Table 6.3.1: Current problems in maritime statistics that could be improved/resolved by European AIS data

For the problems 1 until 6 we performed a PoC, described in chapter 4, of which the results are promising.

In SGA-2 we will focus on describing possibilities of using AIS as a source for making new statistical products, (e.g. like intra-port distances covered in PoC 4.1, sea traffic and variations in time). We also wanted to involve other statisticians working on maritime statistics on thinking about the use of AIS for improving maritime statistics and for new statistical products. To this end, we send out a questionnaire on the use of AIS to maritime statisticians and all member countries of the ESSnet. We will describe the results of this questionnaire, including ideas for making new statistical products by using AIS data, in SGA-2.

In SGA-2 we will also investigate other AIS sources like Marine Traffic, Luxspace (satellite) and hopefully EMSA. This will result in an advice on what data source would best fit analyses for Eurostat's purposes. Furthermore, we will develop a methodology for calculating emissions and report on the impact of this methodology on the (European) level of emissions statistics. All project results of WP4 will result in a consolidated report.

7. References

1. de Wit, et al, Work package 4 AIS data deliverable 4.2: deriving port visits and linking data from Maritime statistics with AIS-data, 10 February 2017
2. UNECE Big Data Quality Task team, A Suggested Framework for the Quality of Big Data, December 2014.
3. Eriksen, T., Greidanus, H., Alvarez, M., Nappo, D., and Gammieri, V. European Commission – Joint Research Centre, Ispra, Italy. Quality of AIS Services for Wide-Area Maritime Surveilance. MAST 2014 conference, Istanbul, 21st May 2014.
4. Millefiori , L.M., Zissis, D, Cazzanti, L., & Arcieri, G. Scalable and distributed sea port operational areas estimation from AIS data, 2016:
   DOI: 10.1109/ICDMW.2016.0060
   Conference: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)
5. Rosenblatt, 1956 and Parzen, 1962 "Remarks on Some Nonparametric Estimates of a Density Function" and "On Estimation of a Probability Density Function and Mode". The Annals of Mathematical Statistics.
6. Working Group on Maritime Transport Statistics 23-24 May 2016 https://circabc.europa.eu/webdav/CircaBC/ESTAT/MOTST/Library/Maritime%20Transport%20Statistics/Meeting%20of%2023-24%20May%202016/Final%20minutes.pdf
7. https://en.wikipedia.org/wiki/Lambert_azimuthal_equal-area_projection
8. http://mathworld.wolfram.com/LambertAzimuthalEqual-AreaProjection.html

Annex 1: SGA-1 of WP4 in more detail

Work package number	4	Start date: End date:			1.2.2016 31.7.2017			
Title	**AIS Data**							
Partner/co-beneficiary	**NL** **120**	DK 60	EL 40	NO 67	PL 40			

Aim of this work package is to investigate whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used 1) to improve the quality and internal comparability of existing statistics and 2) for new statistical products relevant for the ESS. Improvement of quality and internal comparability can be obtained e.g. by developing a reference frame of ships and their travels in European waters and then linking this reference frame, by ship number, to register-based data about marine transport from port authorities. These linked data can then be used for emission calculations. New products can be developed for e.g. traffic analyses. The added value of running a pilot with AIS-data at European level is that the source data are generic word wide and data can be obtained at European level. Challenges ahead with this dataset are: obtaining the data at European level, processing and collecting the data in such way that they can be used for multiple purposes, and visualising the results. A part of this work package is also to look into AIS analyses done by others and to investigate the possibility of obtaining already processed data as input for creating comparable official statistics. Especially it is important to make contact with other public authorities. This work package may require data acquisition in collaboration with Eurostat.

Methodological, quality and technical results of the work package, including intermediate findings, will be used as inputs for the envisaged WP 8 of SGA-2, in case SGA-2 will be realised. When carrying out the tasks listed below, care will be taken that these results will be stored for later use, by using the facilities described at WP 9.

Task 1 – Data access.

AIS-data are available for national territories and the entire European territory. For example, AIS-data in the Netherlands are provided by RijksWaterStaat (a government agency which is part of the Ministry of Infrastructure and the Environment). It is expected that similar agencies in other countries have the AIS-data for their national territories. At the European level, a dataset of AIS data is available at the European Maritime Safety Agency (EMSA). The advantage of using one AIS-dataset for the entire European territory is a) a better comparison of international traffic between the countries and b) more synergy as all participating countries work on the same dataset. A disadvantage is that these data are stored by private companies and handling fees have to be paid. Aim of this task is to decide how European data could be used for this project, to investigate the possibilities of acquiring data from EMSA (to be coordinated with Eurostat) and, if European data are too costly or too hard to obtain, how national datasets can be obtained.

This task will involve:

- Exploration of the possibilities to collect the data at a European (or worldwide) level.

Participants: NL, DK, EL, NO, PL

Task 2 – Data handling

Aim of this task is to process and store the data in such a way that they can be used for consistent multiple outputs, like

- linking AIS-data with data from port authorities
- traffic analyses
- Inference of journeys from AIS data.

Key elements of this task are:

- which programming language and environment should be used for transformation
- where will the data be processed (in each NSI, by NSIs at European level, by data holders)
- how can we create an environment which is easily accessible for all partners

Participants: NL, NO, PL, and possibly DK (if national sources are used)

Task 3 – Methodology and Techniques

Develop traffic statistics: Linking with data from port authorities.

AIS-data may be linked to data from port authorities. Added value of linking AIS-data to data from port authorities is that the same reference population (= ship number) is used in all harbours. As the journeys and harbour visits of ships can be derived from AIS this linking provides the ESSnet information about the origin/destination of the cargo, too.

Aims of this task are:

- to build a reference frame of ships in European water (based on AIS-data)
- to find out how data from port authorities can be linked to AIS-data
- to check whether information improves the quality of current statistical outputs and provides more information about the origin/destination of the cargo.

Participants: NL, NO, PL

Traffic analyses

The number of ships during a certain time interval at certain coordinates (like inland waterways or at certain points at sea) can be calculated by AIS-data. This possibility will be explored because this information could be interesting for traffic analyses and economic analyses.

Aims of this task are:

- calculate the number of ships at certain coordinates
- visualise the results to analyse variations in time

Participants: NL, NO, PL, EL

Estimate emissions (envisaged under SGA-2)

This task involves following individual vessels through time. Consequently we can infer the journeys from the data. Combined with a model to estimate the emission of vessels (which depends on travel distance, speed, draught, weather conditions and characteristics of the vessel itself), emissions of e.g. CO_2 and NO_x can be estimated per ship and per national territory.

An advantage of doing these analyses on a European scale– instead of the national level – is that more precise estimates for emissions at national territories can be made.

Aim of this task is 1) to infer journeys from AIS-data, 2) visualise the results, 3) combine these journeys with a model to calculate emissions and 4) estimate the impact of carrying out these calculations at the European level on the quality of emissions calculations.

Task 4 – Future perspectives (envisaged under SGA-2)

Aim of this task is to summarise the project results and perform a qualitative cost-benefit analysis of using AIS-data for official statistics. These analyses should include aspects like sustainability of the data source, possibilities of improving international comparability, possibilities of data sharing (at micro- or aggregated level), quality improvement of current statistics and a sketch of a possible statistical process and needed infrastructure.

Deliverables (SGA-1 only):

4.1	Report on creating a database with AIS-data for official statistics: possibilities and pitfalls	month 6
4.2	Report about deriving harbour visits and linking data from port authorities with AIS-data	month 12
4.3	Report about sea traffic analyses using AIS-data	month 18

Milestones (SGA-1 only):

4.4	Progress and technical report of first internal WP-meeting	month 4
4.5	Progress and technical report of second internal WP-meeting	month 9

Annex 2: SGA-2 of WP4 in more detail

Work package number	4	Start date: End date:		1.8.2017 31.5.2018	
Title	**AIS Data**				
Partner/co-beneficiary **(person days)**	**NL** **88**	DK 53	EL 66	NO 21	PL 35

Description of the work package

Aim of this work package is to investigate whether real-time measurement data of ship positions (measured by the so-called AIS-system) can be used 1) to improve the quality and internal comparability of existing statistics and 2) for new statistical products relevant for the ESS. Improvement of quality and internal comparability can be obtained e.g. by developing a reference frame of ships and their travels in European waters and then linking this reference frame, by ship number, to register-based data about marine transport from port authorities. These linked data can then be used for emission calculations. New products can be developed for e.g. traffic analyses. The added value of running a pilot with AIS-data at European level is that the source data are generic worldwide and data can be obtained at European level.

Challenges ahead with this dataset are: obtaining the data at European level, processing and collecting the data in such way that they can be used for multiple purposes, and visualising the results. A part of this work package is also to look into AIS analyses done by others and to investigate the possibility of obtaining already processed data as input for creating comparable official statistics. Especially it is important to make contact with other public authorities. This work package may require data acquisition in collaboration with Eurostat.

Methodological, quality and technical results of the work package, including intermediate findings, will be used as inputs for WP 8 of SGA-2. When carrying out the tasks listed below, care will be taken that these results will be stored for later use, by using the facilities described at WP 9.

SGA-2 of WP 4 will deliver the products below:

Task 3 – Methodology and Techniques

Estimate emissions

This task involves following individual vessels through time. Consequently we can infer the journeys from the data. Combined with a model to estimate the emission of vessels (which depends on travel distance, speed, draught, weather conditions and characteristics of the vessel itself), emissions of e.g. CO_2 and NO_x can be estimated per ship and per national territory.

Estimation of emissions based on only AIS data is impossible. At least, we need to know what the emission should be given, a.o., draught, speed and weather conditions. There are several sources that can be used to get information about the emission of vessels. For example Lloyds Register of Shipping.

Also, all ships have to register their emissions, so there is data available at the freight ship companies. A dataset for the same period as the AIS dataset could be used to model the emissions based on the variables mentioned above. We assume getting these needed data for free for this pilot.

Furthermore, satellite data is available for emissions (see, for instance http://www.globemission.eu), which shows very clearly the maritime routes. This satellite data could be used as a more direct source for measuring emissions and maybe for testing our model. Another possibility to test the model is to compare the estimated emissions by the model with the real fuel purchase or information coming from vessel's oil record book.

An advantage of doing these analyses on a European scale– instead of the national level – is that more precise estimates for emissions at national territories can be made.

Aim of this task is to develop and test a methodology for estimating vessel emissions based on AIS data by:

1) inferring journeys from AIS-data
2) visualising the results
3) investigating methodology for calculating emissions
4) combining these journeys with a model to calculate emissions
5) Testing the model and
6) estimating the impact of carrying out these calculations at the European level on the quality of emissions calculations.

Participants: Norway, Netherlands , Denmark, Greece, Poland

Task 4 – Access to and analysing AIS data from EMSA

AIS-data are available for national territories and the entire European territory. In this work package we use European AIS data from Dirkzwager. But for the future we would like to get free European AIS data. That is why we try to get AIS data at the European level from the European Maritime Safety Agency (EMSA).

In SGA-2 we focus on getting the European AIS data from EMSA and compare these two sources (AIS data from EMSA and Dirkzwager) on their coverage and quality. We will assess the quality of both sources by applying the quality framework. We also describe the strengths and weaknesses of both sources compared with existing maritime data.

This work package may require data acquisition in collaboration with Eurostat, because we would like Eurostat to apply for the data at EMSA's. Eurostat already received AIS data from EMSA for other projects and Eurostat can apply EMSA data for all the partner countries once. This will be better than applying for each country separated. If we do not get access to the EMSA data we cannot process this task and we cannot deliver deliverable 4.7, but it has not any consequences for the other deliverables in this WP. Not getting AIS data from EMSA means that we cannot get AIS data on the European level for free.

Participants: Netherlands , Denmark, Greece, Poland

Task 5 – New statistical output

Aim of this task is to explore possibilities of new statistical products (for example intraport statistics) by using AIS data. This task also includes analysing and elaborating scenarios for production of European and national statistics based on one single European data source of AIS data.

Participants: Norway, Netherlands, Denmark, Greece, Poland

Task 6 – Future perspectives

Aim of this task is to produce a consolidated report summarising the contents and the outcomes of WP 4. This report also includes a cost-benefit analysis of using AIS-data for official statistics. The report should also include aspects like sustainability of the data source, possibilities of improving international comparability, possibilities of data sharing (at micro- or aggregated level), possibilities of meeting the needs of both European and national statistics by one European AIS database, quality improvement of current statistics and a sketch of a possible statistical process and needed infrastructure**,** including technical skills required to generate statistical outputs from the source data.

Participants: Norway, Netherlands, Denmark, Greece, Poland

Deliverables (SGA-2 only):

4.6	Report on estimating emissions. This report will describe the investigated methodology for calculating emissions and the reason why we choose for a certain methodology. The report also describes the created model itself (inclusive other needed data sources) and the results of testing the model. Finally this report also describes the impact of carrying out these calculations at the European level on the quality of emissions calculations.	month 13
4.7	Report about the results of comparing the quality and coverage of the European AIS data from Dirkzwager and EMSA (by applying the quality framework). Also the strengths and weaknesses of both sources will be compared with existing maritime data.	Month 15
4.8	Report about possible new statistical output based on European AIS data. The report also describes analysed and elaborated scenarios for production of European and national statistics based on one single European data source of AIS data.	month 16
4.9	Consolidated report on project results including a cost-benefit analysis of using AIS-data for official statistics.	month 17

Milestones (SGA-2 only):

4.10	Progress and technical report of first internal WP-meeting	Month 9
4.11	Progress and technical report of the second internal WP-meeting	Month 14

Annex 3: quality framework in more detail

Hyper dimension	Quality Dimension	Possible indicators	Results	Conclusion
Source	*Institutional/ Business Environment*	1. What is your estimate of the overall risk that the data provider will not meet the quality requirements of the NSO?	Probability: Low, Impact: Medium	OK
		2. What is the risk that the BDS will not be available from the data provider in the future? If it will not, will there be comparable data sources in the future?	Probability: High, Impact: Low (other provider)	OK
		3. How relevant are the data, if they would be available for only a short period of time?	Not very relevant	OK
		4. How long do these data need to be available to be relevant?	Minimally two years, to be able to filter out seasonal effects and see developments	OK
		5. Is it likely that we can replace these data with similar (or next generation) data, once the data source or technology becomes obsolete?	Yes, new technology will still contain information needed.	OK
	Privacy and Security	1. Does the NSO have clear legal authority to obtain the data?	yes	OK
		2. Are there legal limitations or restrictions on the use to which the data can be put?	Yes, privacy issues	Action needed: solve privacy issues
		3. Are the data provider and the NSO willing to enter negotiations to solve any legal issues, if necessary?	yes	OK
		4. Was the data collected in accordance with relevant privacy laws?	yes	OK
		5. Do the NSO’s own confidentiality policies limit the utility of data?	no	OK
		6. Are stakeholders (private sector, public, others) likely to react negatively given the intended use of the data by the NSO?	no	OK
		7. Will there be a need to carry out privacy assessment exercises and public consultations in relation to using this data and its potential impact on the NSO reputation and credibility?	yes	Action needed: solve privacy issues
Metadata	*Complexity*	1. Structure: how easy would it be to render the data source into a useable structure (i.e... one record per unit of observation,)?	easy	OK

		2. Format: Is the data source in standard format (e.g., XLS, XML)? How many different formats were used in the data source? How easy would it be to render the data source variables into a useable format (i.e... parsing, grooming, coding, treatment of outliers or missing values)?	no, only in NMEA format, easy	OK
		3. Data: how many different standards were used in the data source (e.g., ISO-3166 to describe countries)? Is there any non-standard code lists used in the data source that are not unified? How many different code lists were used in the data source?	NMEA itself is an ISO standard	OK
		4. Hierarchies: is there a hierarchical relationship between records or variables?	no	OK
		5. Structure: How many different files or tables are in the data source?	one	OK
	Completeness	1. Qualitative assessment (e.g. score for completeness of metadata for input phase: 0 description missing, 1 description insufficient, 2 description complete)	2	OK
		2. In case of missing/incomplete descriptions what are the consequences/drawbacks for data usability?	Not applicable	OK
		3. Are the population units defined clearly?	yes	OK
		4. Are the variables defined clearly?	Yes	OK
		5. Qualitative assessment of completeness and clarity of metadata	Metadata is not really clear on coverage and filtering	Not OK
		6. In case of unclear/ambiguous descriptions what are the consequences/drawbacks for data usability?	Can be derived from the data	OK
	Usability	1. Will the NSO need to acquire new skills to use and analyse the data?	Yes, big data skills	OK
		2. How much resourcing will cleaning and processing the dataset require?	Processing takes about 20 minutes for 6 months of data	OK
		3. How big is the data set?	400 GB (encoded)	OK
		4. Data transmission: are special arrangements for data transmission required, and if so, can the NSO meet those requirements?	No, when receiving the data in small, daily portions, the transmission should not be an issue	OK

		5. IT requirements: what would be the hardware and software requirements to process and store this data? Will there be a need to develop a specific IT infrastructure?	Yes, Big Data infrastructure (i.e., Hadoop, Spark)	OK
	Time-related factors	1. Time between receipt of data and when the data was collected;	Not applicable	OK
		2. When was the data collected? What is the reference period of the data?	October 2015 –April 2016	OK
		3. Whether data is collected and available periodically. Recurring data provides the opportunity for time series.	Data is available continuously	OK
		4. Could changes in concepts or methods limit the potential use of historical data?	No, methods are developed to process long historical time series.	OK
	Linkability	1. Are potential linking variables present on the file that could be used for data integration with other data files?	Yes, the IMO number is available. This number is a unique number for sea vessels.	OK
		2. Calculate the percentage of units linked and not linked in both the Big Data (BD) and other data sources. The indicator is the percentage of units linked unambiguously (strong link) / percentage of units linked with a soft link (linking requirements were relaxed in order to link more units)	This is done by checking the AIS source with other sources and AIS sources with each other	OK
	Coherence - consistency	1. How do you rate the variables capturing the constructs that are of interest?	High	OK
		2. Are the definitions used aligned with NSO standards?	yes, data is at such a fine grain level that it is easy to create aggregates that are aligned	OK
		3. Do the anomalies in the data indicate important errors that would limit the potential use?	no	OK
	Validity	1. Is the metadata available sufficient to assess the soundness of the methods	It is insufficient. We do not know the	Not OK

		used?	filtering used.	
		2. Are there critical flaws in the processes that would limit potential use of the data?	No, despite some variables are not filled out consistently by the owner of the vessel	OK
Data	*Accuracy and selectivity*	1. If a reference data set is available, assess coverage error. For example, measures of distance between Big Data population and the target population (e.g. Kolmogorov-Smirnov Index, Index of dissimilarity)	see deliverable 4.3	OK
		2. Does the file contain duplicates?	No	OK
		3. Are the data values within the acceptable range?	Yes	OK
		4. Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by Big Data source.	Little vessels do not have AIS on board. Because we were only interested in maritime ships this was not a problem.	OK
		5. Assessment of spatial distribution of measurement instrument and of periodicity of observations	Not applicable	OK
	Linkability	1. Are potential linking variables present on the file that could be used for data integration with other data files?	yes, IMO for linking vessels and Lat/Lon for linking positions to ports	OK
		2. Calculate the percentage of units linked and not linked in both the Big Data (BD) and other data sources. The indicator is the percentage of units linked unambiguously (strong link) / percentage of units linked with a soft link (linking requirements were relaxed in order to link more units)	see deliverable 4.2 and 4.3	OK
	Coherence - consistency	1. How do you rate the variables capturing the constructs that are of interest?	high	OK
		2. Are the definitions used aligned with NSO standards?	yes	OK

		3. Do the anomalies in the data indicate important errors that would limit the potential use?	no	OK
	Validity	1. Is the metadata available sufficient to assess the soundness of the methods used?	It is insufficient. We do not know the filtering used.	Not OK
		2. Are there critical flaws in the processes that would limit potential use of the data?	No, despite some variables are not filled out consistently by the owner of the vessel	OK

Annex 4: code used for analysis

The code used can be found here:
(https://github.com/mputs/WP4/blob/master/ErrorMMSI_loc/src/main/scala/aisframe.scala)
This code identifies the ships (based on MMSI) from the reference frame of ships in a defined area or areas for your defined time period. It then counts the number of appearances of each MMSI in that/those area(s) during your time period. It has to be run for both the national and the DZ data.

We use the reference frame of ships we already created (*/user/tessadew/defframe6all.csv*), because we only want to identify maritime ships and this reference frame is used as a backbone for all our next steps (see previous Deliverables 2). The DZ data run from 10-October-2015 until 10 April 2016, so select a time period within this period. There might be a problem with November 1st 2015, so please do not select that day.

The steps:
1. Define the rectangular area (latitude, longitude) or areas to be compared, see *user/tessadew/areasOI.csv* for the format to be used. Upload this file to your personal folder: *user/<you>/national/ areasOI.csv*

2. Upload your location files to your personal folder e.g. *user/<you>/national/201601*.csv.gz*
The required format is here: *datasets/AIS/Locations/201510102*.csv.gz*

3. Using the bash, login to the Sandbox. Go in to WP4/ErrorMMSI_loc.
- Type *git clone* https://github.com/mputs/WP4.git
- Type *git pull*
- Type *sbt package*

4. Run the code by typing the following code, replacing some code:

```
spark-submit --class "AISframe" target/scala-2.10/ais-frame_2.10-0.1.jar
/user/tessadew/defframe6all.csv datasets/AIS/Locations/201510102*.csv.gz
user/tessadew/areasOI.csv user/tessadew/nrShips.csv
```

Replace the subcommands:
- *datasets/AIS/Locations/201510102*.csv.gz*:
  → for the DZ data select your defined period e.g.:*datasets/AIS/Locations/201601*.csv.gz*
  → with your own data from your personal folder e.g.: *user/<you>/national/201601*.csv.gz*
- *user/tessadew/areasOI.csv*:
  → *user/<you>/areasOI.csv*
- user/tessadew/nrShips.csv:
  → *user/<you>/NatnrShips.csv*
  → *user/<you>/DZnrShips.csv*

This results in a file like *user/tessadew/nrShips.csv*, showing all MMSI's that were in the area(s) and time period defined.

Annex 5: details Denmark

The data from the Danish Maritime Authority are a different cvs-format from the DZ data. To transform the DMA data, the following transformation was used:

```
awk -F ";" '{gsub(/,/, ".") ; print $3 "," $5 "," $4 "," 1 "," $8 "," $9
"," $7 "," 1 "," $1}' aisdk_20151201.csv > test_select.csv
sed -i '1immsi,lon,lat,accuracy,speed,course,rotation,status,timestamp'
test_select.csv
```

Coordinates for the Danish ports:

V1	V2	V3	V4	V5
HELSINGOR	12.3523	13.1076	55.6359	56.1610
AARHUSHAVN	10.1897	10.2529	56.1367	56.1726
SKAGEN	9.9633	10.9677	57.4878	57.9197

Annex 6: details Greece

Two points defining rectangular area

	Point A (latitude, longitude)	Point B (latitude, longitude)
Port of Thessaloniki	40.6062 , 22.8900	40.6505 , 22.9522
Port of Volos	39.3434 , 22.9290	39.3627 , 22.9528
Port of Piraeus	37.9220 , 23.5856	37.9600 , 23.6500
Port of Patras	38.2169 , 21.7075	38.2664 , 21.7407
Port of Heraklion	35.3439 , 25.1348	35.3522 , 25.1577

Table 1: Coordinates of the Ports

Port of Piraeus - 15/12/2015										
					Number of appearances					
a/a of Distinct MMSI's	**AIS- DZ**	**AIS-HCG**		**a/a of Distinct MMSI's**	**AIS- DZ**	**AIS-HCG**		**a/a of Distinct MMSI's**	**AIS- DZ**	**AIS-HCG**
1	6	92		54	148	808		107	171	2549
2	125	179		55	57	827		108	209	2622
3	9	188		56	62	840		109	245	2651
4	169	227		57	102	871		110	86	2723
5	162	240		58	37	879		111	154	2797
6	216	247		59	63	889		112	67	2851
7	185	281		60	67	894		113	176	2866
8	343	312		61	64	899		114	123	2910
9	86	322		62	74	916		115	224	2998
10	44	422		63	115	938		116	356	3100
11	251	439		64	83	970		117	257	3205
12	36	442		65	79	979		118	206	3385
13	257	451		66	210	1018		119	284	3434
14	231	454		67	62	1019		120	274	3441
15	265	456		68	81	1025		121	322	3557
16	197	461		69	76	1031		122	134	3753
17	224	466		70	73	1062		123	386	3755
18	347	482		71	70	1074		124	405	3843
19	226	483		72	211	1086		125	229	3846
20	364	485		73	315	1102		126	323	4055
21	366	495		74	102	1118		127	302	4058
22	382	503		75	393	1121		128	422	4060
23	386	513		76	153	1122		129	326	4201
24	255	514		77	79	1134		130	320	4350
25	298	519		78	371	1270		131	236	4420
26	285	527		79	67	1292		132	363	4620
27	39	528		80	342	1374		133	334	4803
28	329	534		81	156	1384		134	137	4863
29	344	535		82	143	1432		135	111	5088
30	357	537		83	212	1473		136	246	5097
31	370	551		84	177	1493		137	224	5366
32	381	554		85	229	1520		138	383	5645
33	349	559		86	100	1529		139	327	5732
34	384	569		87	373	1583		140	574	5940
35	386	571		88	118	1598		141	390	6459

36	393	573		89	94	1615		142	459	6596	
37	33	574		90	133	1694		143	161	6693	
38	380	583		91	166	1751		144	393	6694	
39	383	590		92	52	1760		145	485	6746	
40	374	605		93	204	1789		146	430	7426	
41	385	613		94	218	1856		147	388	7495	
42	394	619		95	54	1859		148	417	7799	
43	391	624		96	119	1878		149	579	7890	
44	54	658		97	338	1892		150	444	7920	
45	57	692		98	58	2064		151	470	8157	
46	391	700		99	62	2153		152	585	8278	
47	51	716		100	75	2160		153	524	8417	
48	43	736		101	271	2172		154	384	8500	
49	79	744		102	584	2201		155	551	8710	
50	75	746		103	135	2206		156	568	8856	
51	166	772		104	231	2261		157	573	8912	
52	61	787		105	180	2345		158	585	8958	
53	21	807		106	333	2479		159	573	9057	
								160	577	9195	

Table 2: Number of messages per ship in AIS-DZ and AIS-HCG at 15/12/2017 in the Port of Piraeus

Annex 7: details Poland

Checks on Polish national data

National data has been validated on the basis of:

- IMO and MMSI number

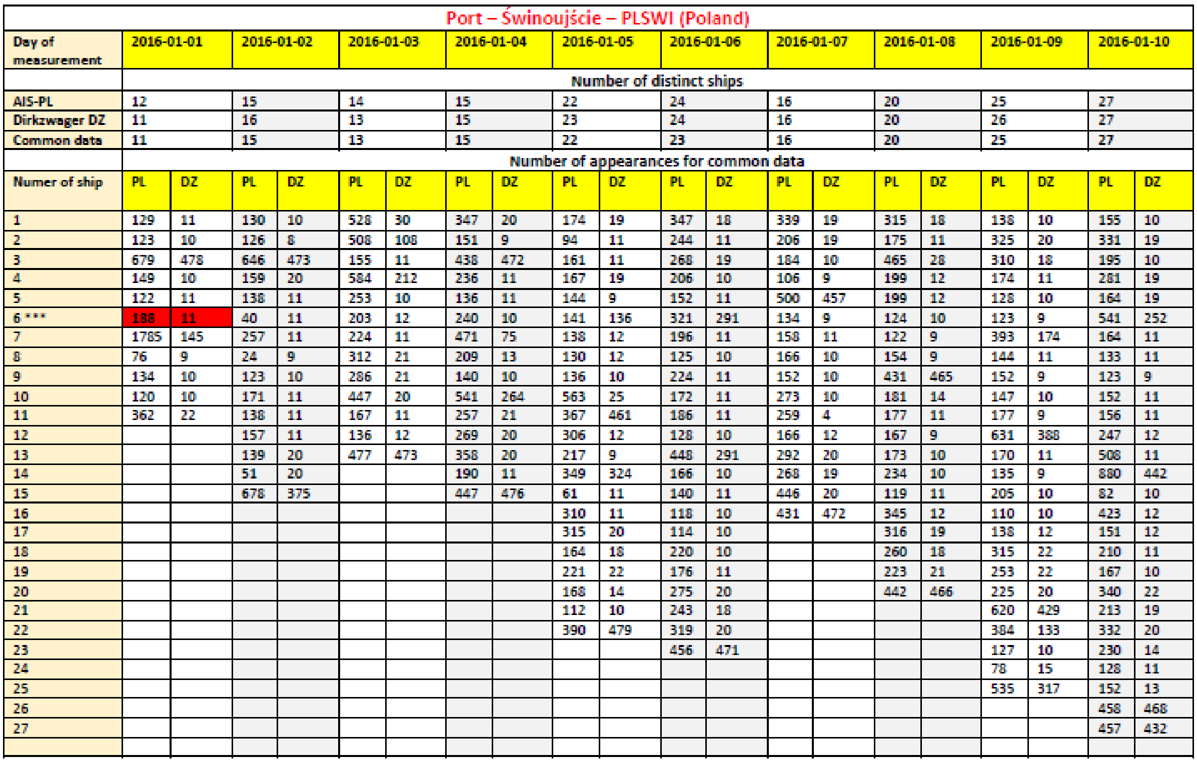
- Range of value for latitude and longitude coordinates

- The length of the ship (sum of dim_a + dim_b < 410)

- The width of the ship (sum of dim_c + dim_d < 70)

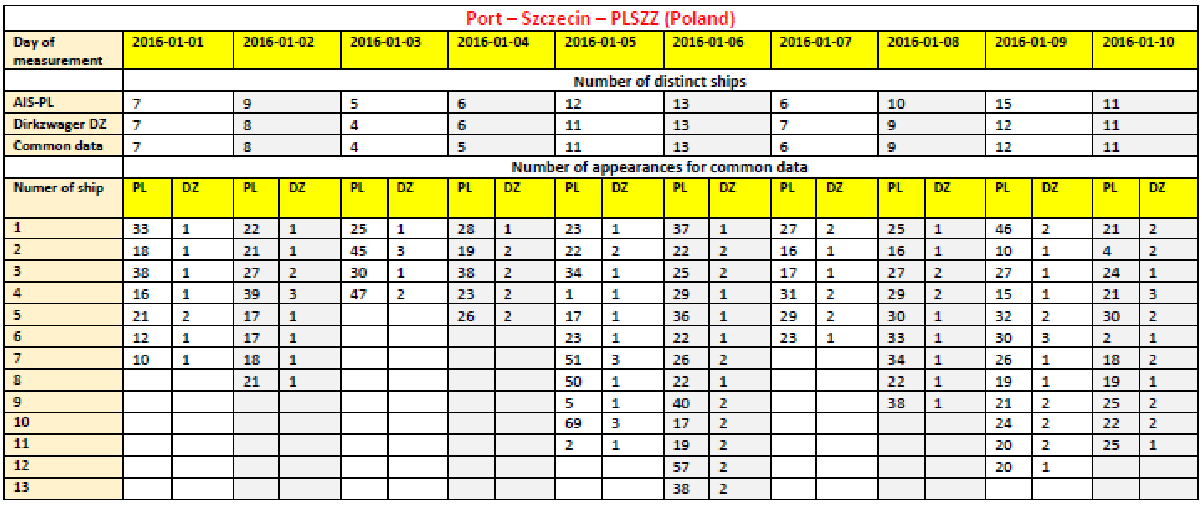
Port of Świnoujście	Point A (latitude, longitude)	Point B (latitude, longitude)
	53.907008, 14.250708	53.951968, 14.286217
Port of Szczecin	Point A (latitude, longitude)	Point B (latitude, longitude)
	53.532288, 14.617817	53.540470, 14.642990
Port of Gdynia	Point A (latitude, longitude)	Point B (latitude, longitude)
	54.498334, 18.525681	54.569087, 18.625133
Port of Gdańsk	Point A (latitude, longitude)	Point B (latitude, longitude)
	54.388773, 18.633629	54.433563, 18.715090

Table 1: Coordinates of the Ports



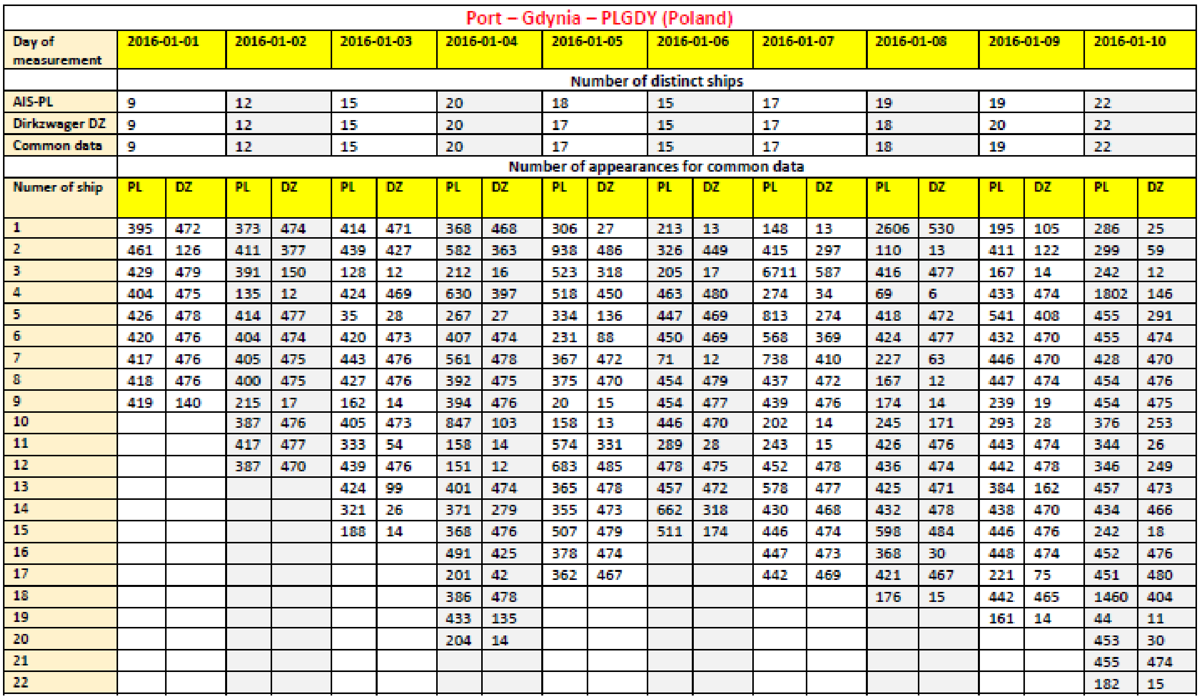
Port – Świnoujście – PLSWI (Poland)																				
Day of measurement	2016-01-01		2016-01-02		2016-01-03		2016-01-04		2016-01-05		2016-01-06		2016-01-07		2016-01-08		2016-01-09		2016-01-10	
	Number of distinct ships																			
AIS-PL	12		15		14		15		22		24		16		20		25		27	
Dirkzwager DZ	11		16		13		15		23		24		16		20		26		27	
Common data	11		15		13		15		22		23		16		20		25		27	
	Number of appearances for common data																			
Numer of ship	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ
1	129	11	130	10	528	30	347	20	174	19	347	18	339	19	315	18	138	10	155	10
2	123	10	126	8	508	108	151	9	94	11	244	11	206	19	175	11	325	20	331	19
3	679	478	646	473	155	11	438	472	161	11	268	19	184	10	465	28	310	18	195	10
4	149	10	159	20	584	212	236	11	167	19	206	10	106	9	199	12	174	11	281	19
5	122	11	138	11	253	10	136	11	144	9	152	11	500	457	199	12	128	10	164	19
6 ***	188	11	40	11	203	12	240	10	141	136	321	291	134	9	124	10	123	9	541	252
7	1785	145	257	11	224	11	471	75	138	12	196	11	158	11	122	9	393	174	164	11
8	76	9	24	9	312	21	209	13	130	12	125	10	166	10	154	9	144	11	133	11
9	134	10	123	10	286	21	140	10	136	10	224	11	152	10	431	465	152	9	123	9
10	120	10	171	11	447	20	541	264	563	25	172	11	273	10	181	14	147	10	152	11
11	362	22	138	11	167	11	257	21	367	461	186	11	259	4	177	11	177	9	156	11
12			157	11	136	12	269	20	306	12	128	10	166	12	167	9	631	388	247	12
13			139	20	477	473	358	20	217	9	448	291	292	20	173	10	170	11	508	11
14			51	20			190	11	349	324	166	10	268	19	234	10	135	9	880	442
15			678	375			447	476	61	11	140	11	446	20	119	11	205	10	82	10
16									310	11	118	10	431	472	345	12	110	10	423	12
17									315	20	114	10			316	19	138	12	151	12
18									164	18	220	10			260	18	315	22	210	11
19									221	22	176	11			223	21	253	22	167	10
20									168	14	275	20			442	466	225	20	340	22
21									112	10	243	18					620	429	213	19
22									390	479	319	20					384	133	332	20
23											456	471					127	10	230	14
24																	78	15	128	11
25																	535	317	152	13
26																			458	468
27																			457	432

Table 2: results comparison DZ and Polish data for Świnoujście



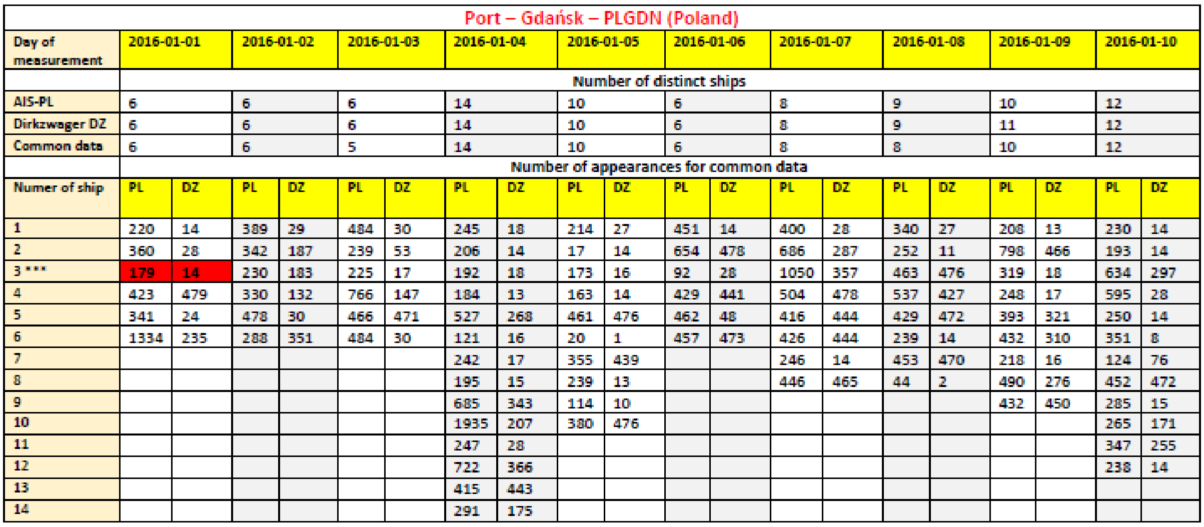
Port – Szczecin – PLSZZ (Poland)																				
Day of measurement	2016-01-01		2016-01-02		2016-01-03		2016-01-04		2016-01-05		2016-01-06		2016-01-07		2016-01-08		2016-01-09		2016-01-10	
	Number of distinct ships																			
AIS-PL	7		9		5		6		12		13		6		10		15		11	
Dirkzwager DZ	7		8		4		6		11		13		7		9		12		11	
Common data	7		8		4		5		11		13		6		9		12		11	
	Number of appearances for common data																			
Numer of ship	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ
1	33	1	22	1	25	1	28	1	23	1	37	1	27	2	25	1	46	2	21	2
2	18	1	21	1	45	3	19	2	22	2	22	2	16	1	16	1	10	1	4	2
3	38	1	27	2	30	1	38	2	34	1	25	2	17	1	27	2	27	1	24	1
4	16	1	39	3	47	2	23	2	1	1	29	1	31	2	29	2	15	1	21	3
5	21	2	17	1			26	2	17	1	36	1	29	2	30	1	32	2	30	2
6	12	1	17	1					23	1	22	1	23	1	33	1	30	3	2	1
7	10	1	18	1					51	3	26	2			34	1	26	1	18	2
8			21	1					50	1	22	1			22	1	19	1	19	1
9									5	1	40	2			38	1	21	2	25	2
10									69	3	17	2					24	2	22	2
11									2	1	19	2					20	2	25	1
12											57	2					20	1		
13											38	2								

Table 3: results comparison DZ and Polish data for Szczecin



Port – Gdynia – PLGDY (Poland)																				
Day of measurement	2016-01-01		2016-01-02		2016-01-03		2016-01-04		2016-01-05		2016-01-06		2016-01-07		2016-01-08		2016-01-09		2016-01-10	
	Number of distinct ships																			
AIS-PL	9		12		15		20		18		15		17		19		19		22	
Dirkzwager DZ	9		12		15		20		17		15		17		18		20		22	
Common data	9		12		15		20		17		15		17		18		19		22	
	Number of appearances for common data																			
Numer of ship	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ
1	395	472	373	474	414	471	368	468	306	27	213	13	148	13	2606	530	195	105	286	25
2	461	126	411	377	439	427	582	363	938	486	326	449	415	297	110	13	411	122	299	59
3	429	479	391	150	128	12	212	16	523	318	205	17	6711	587	416	477	167	14	242	12
4	404	475	135	12	424	469	630	397	518	450	463	480	274	34	69	6	433	474	1802	146
5	426	478	414	477	35	28	267	27	334	136	447	469	813	274	418	472	541	408	455	291
6	420	476	404	474	420	473	407	474	231	88	450	469	568	369	424	477	432	470	455	474
7	417	476	405	475	443	476	561	478	367	472	71	12	738	410	227	63	446	470	428	470
8	418	476	400	475	427	476	392	475	375	470	454	479	437	472	167	12	447	474	454	476
9	419	140	215	17	162	14	394	476	20	15	454	477	439	476	174	14	239	19	454	475
10			387	476	405	473	847	103	158	13	446	470	202	14	245	171	293	28	376	253
11			417	477	333	54	158	14	574	331	289	28	243	15	426	476	443	474	344	26
12			387	470	439	476	151	12	683	485	478	475	452	478	436	474	442	478	346	249
13					424	99	401	474	365	478	457	472	578	477	425	471	384	162	457	473
14					321	26	371	279	355	473	662	318	430	468	432	478	438	470	434	466
15					188	14	368	476	507	479	511	174	446	474	598	484	446	476	242	18
16							491	425	378	474			447	473	368	30	448	474	452	476
17							201	42	362	467			442	469	421	467	221	75	451	480
18							386	478							176	15	442	465	1460	404
19							433	135									161	14	44	11
20							204	14											453	30
21																			455	474
22																			182	15

Table 4: results comparison DZ and Polish data for Gdynia



Port – Gdańsk – PLGDN (Poland)																				
Day of measurement	2016-01-01		2016-01-02		2016-01-03		2016-01-04		2016-01-05		2016-01-06		2016-01-07		2016-01-08		2016-01-09		2016-01-10	
	Number of distinct ships																			
AIS-PL	6		6		6		14		10		6		8		9		10		12	
Dirkzwager DZ	6		6		6		14		10		6		8		9		11		12	
Common data	6		6		5		14		10		6		8		8		10		12	
	Number of appearances for common data																			
Numer of ship	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ	PL	DZ
1	220	14	389	29	484	30	245	18	214	27	451	14	400	28	340	27	208	13	230	14
2	360	28	342	187	239	53	206	14	17	14	654	478	686	287	252	11	798	466	193	14
3 ***	179	14	230	183	225	17	192	18	173	16	92	28	1050	357	463	476	319	18	634	297
4	423	479	330	132	766	147	184	13	163	14	429	441	504	478	537	427	248	17	595	28
5	341	24	478	30	466	471	527	268	461	476	462	48	416	444	429	472	393	321	250	14
6	1334	235	288	351	484	30	121	16	20	1	457	473	426	444	239	14	432	310	351	8
7							242	17	355	439			246	14	453	470	218	16	124	76
8							195	15	239	13			446	465	44	2	490	276	452	472
9							685	343	114	10							432	450	285	15
10							1935	207	380	476									265	171
11							247	28											347	255
12							722	366											238	14
13							415	443												
14							291	175												

Table 5: results comparison DZ and Polish data for and Gdańsk