**Coverage of AIS data: comparison of privately held to national datasets - Poland and Hellenic experiences.**

Michał Bis, Statistics Poland, Centre for data engineering, M.Bis@stat.gov.pl

Anna Bilska, Statistics Poland, Marine Statistics Centre, A.Bilska@stat.gov.pl

Eleni Bisioti, Statistics Greece, Organization, Methodology and International Relation Division, e.bisioti@statistics.gr

Christina Pierrakou, Statistics Greece, Organization, Methodology and International Relation Division, c.pierrakou@statistics.gr

**Abstract**

*AIS data, real-time measurement data of ship positions, is one of the potential Big Data sources investigated by the ESSnet Big Data project. The main aim of the specific work package (WP4) is to explore the potential use of AIS data in the production of official statistics, due to their advantage of being generic worldwide and obtainable at European level. Five National Statistical Institutes participate in WP4: the national statistical institutes of the Netherlands (Work package leader), Denmark, Greece, Norway and Poland.*

*This paper outlines the results of the comparison of national public AIS data to privately held AIS data, in terms of quality and metadata. Various Big Data technologies were used, to store, manage and process the huge volumes of AIS data such as the Distributed File System of Hadoop, the large-scale data processing engine Apache Spark, Scala language, the noSQL database Elasticsearch using GeoData. Moreover, the exploration and visualization tool Kibana was used.*

*The main conclusion drawn from this paper is that AIS is a big data source with great potential to improve official statistics, however more work is needed in the area of ensuring transparency and soundness of methods and processes for the privately held data to be incorporated in the statistical production.*

**Keywords:** AIS, position data, quality of big data source, big data technologies

**1. Introduction**

This paper was written based on the results of the work within the ESSnet Big Data project of work package (WP) 4 related to AIS data. The aim of this work package was to investigate whether real-time measurement data of ship positions measured by the so-called AIS data system can be used: to research on the use of AIS data as a potential big data source for official statistics; to improve the quality and internal comparability of existing statistics; to produce new statistical products relevant to the ESS. Five National Statistical Institutes participated in WP4: the Netherlands (Work package leader), Denmark, Greece, Norway and Poland.

Using one AIS-dataset for the entire European territory (EU terrestrial AIS) for the production of statistics would result to improve the comparability and enhance the synergy among Member States. Therefore, we focus here on quality and technological aspects of this large AIS data source as a potential Big Data source for the production of official statistics at ESS level.

At the start of the WP4, after exploring potential access to several data sources (public and private), we decided to obtain commercial European terrestrial raw AIS data, for a 6-months period (8 October 2015 - 12 April 2016) [1].

We started by investigating the coverage and quality of commercial and national AIS data. While working with AIS data, we noticed a lot of noise that prevented proper analysis. This was the basis for us to investigate quality of these data.

Moreover, AIS data may contain:

* Technical errors - related to dynamic data such as position of ship, speed, course, rotation which comes from AIS device (sensors, cables and antenna)
* Human errors – related to static data (MMSI, IMO number, ship’s name, call sign, type, length) or voyage data (draught, destination) which are manually entered in the AIS device [2].

**2. Comparing of privately held to national datasets**

*2.1 Approach*

While working with AIS data from Εuropean data source, apart from the ‘noise’ of data, there was, also, lack of data for specific coastal European areas. To investigate, further, the issue of coverage, we compared the European dataset to national AIS data from Denmark, Greece and Poland and satellite data from Luxspace (LS). For the process of data comparison a common assumption and approach was adapted. A common reference frame of ships was built to filter out existing maritime ships only [2]. Then, we defined the areas to be compared by selecting ports for the best coverage of AIS base stations. For each port a rectangular area (latitude, longitude) was defined. A code developed in Scala and processed by Apache Spark was used. The results were saved in a Distributed File System of Hadoop at Sandbox. This code identifies the ships (based on MMSI) from the reference frame of ships in a defined area(s) for a defined time period. Then, it counts the number of appearances of each MMSI in that/those area(s) during that time period. It was run for both data sources.

*2.2. Poland*

AIS data from the Polish Maritime Authority (AIS-PL) were compared to European terrestrial AIS data for some typical days, January 1st – 10th 2016. The areas interest were the port of Świnoujście, Szczecin, Gdynia and Gdańsk.

**Table 1. Coordinates of the Ports**

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

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

The results showed that for most cases, the number of unique ships for both data sources is almost the same. Per port and per day, AIS-PL data sometimes contained one more ship than European data source. However, the number of messages from ships presented in both data sets differ greatly, the national data almost always having a higher number of messages.

**Table 2. Port – Świnoujście (Poland) – PLSWI, Term of measurement – January, 2016**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day of measurement | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|  | Number of distinct ships |
| AIS-PL | 12 | 15 | 14 | 15 | 22 | 24 | 16 | 20 | 25 | 27 |
| EU terrestrial AIS | 11 | 16 | 13 | 15 | 23 | 24 | 16 | 20 | 26 | 27 |
| Common data | 11 | 15 | 13 | 15 | 22 | 23 | 16 | 20 | 25 | 27 |

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

**Figure 2. Number of apperances for common data (11), Port Świnoujście for 2016-01-01**

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

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-PL data present in European data source.

**Table 3. Port –Szczecin – PLSZC (Poland), Term of measurement – January, 2016**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day of measurement | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|  | Number of distinct ships |
| AIS-PL | 7 | 9 | 5 | 6 | 12 | 13 | 6 | 10 | 15 | 11 |
| EU terrestrial AIS | 7 | 8 | 4 | 6 | 11 | 13 | 7 | 9 | 12 | 11 |
| Common data | 7 | 8 | 4 | 5 | 11 | 13 | 6 | 9 | 12 | 11 |

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

**Table 4. Port –Gdynia – PLGDY (Poland), Term of measurement – January, 2016**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day of measurement | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|  | Number of distinct ships |
| AIS-PL | 9 | 12 | 15 | 20 | 18 | 15 | 17 | 19 | 19 | 22 |
| EU terrestrial AIS | 9 | 12 | 15 | 20 | 17 | 15 | 17 | 18 | 20 | 22 |
| Common data | 9 | 12 | 15 | 20 | 17 | 15 | 17 | 18 | 19 | 22 |

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

**Table 5. Port –Gdańsk – PLGDN (Poland), Term of measurement – January, 2016**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Day of measurement | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 | 09 | 10 |
|  | Number of distinct ships |
| AIS-PL | 6 | 6 | 6 | 14 | 10 | 6 | 8 | 9 | 10 | 12 |
| EU terrestrial AIS | 6 | 6 | 6 | 14 | 10 | 6 | 8 | 9 | 11 | 12 |
| Common data | 6 | 6 | 5 | 14 | 10 | 6 | 8 | 8 | 10 | 12 |

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

**Figure 3. Number of apperances for common data (6), Port Gdańsk for 2016-01-01**



Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-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 figure 2) was 188 in the AIS-PL data and 11 in the European data source. Data does not overlap in latitude and longitude values from the two sources. However, the route of ship is preserved.

A similar pattern can be seen for the port of Gdańsk (marked in red in figure 3). Here the number of messages for a ship present in both data sets is 179 in the AIS-PL data, and 14 in the European data source.

**Figure 5. Route of ship from AIS-PL and EU terrestrial AIS data, port of Gdańsk: three zooms**

Source: Own study on the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

As in the previous case, the data from AIS-PL and EU terrestrial AIS data does not overlap, but the ship's route is preserved. All in all, the number of ships is almost the same in the AIS-PL and European data source. However, mostly AIS-PL contains higher frequent data. National data is much more precise (received data are of a higher frequency) than European 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).

*2.3. Greece*

For Greece we had three sources: national terrestrial AIS data from Hellenic Coast Guard (HCG), European terrestrial AIS data and satellite data. Therefore, we also take into account in this analysis.

*2.3.1 Comparing national terrestrial AIS data to EU terrestrial AIS data*

National terrestrial AIS data from the Hellenic Coast Guard (HCG) was compared to EU terrestrial AIS data for a one day, December 15th 2015. The data was compared for the whole country of Greece and the ports of Piraeus, Thessaloniki, Patras, Volos and Heraklion.

**Table 6. Coordinates of the Ports**

|  |  |  |
| --- | --- | --- |
| **Name of port** | **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 |

Source: On the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

The AIS data was analysed in terms of the number of ships across the whole country of Greece and for the five main ports (Piraeus, Thessaloniki, Volos, Patras and Heraklion). The results showed that for most cases, the number of unique ships for EU terrestrial AIS data and national terestrial AIS data is diferent. There was no difference only for the Port of Piraeus.Table 7 shows the results.

**Table 7. Ports of Greece, Term of measurement – 2016-12-15**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number of ships** | **Greece** | **Port of Piraeus** | **Port of Thessaloniki** | **Port of Volos** | **Port of Patras** | **Port of Heraklion** |
| **EU terrestrial AIS** | 477 | 162 | 0 | 0 | 0 | 0 |
| **AIS HCG** | 767 |  162 | 61 | 19 | 40 | 8 |

Source: On the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

Only Port of Pireaus could be investigated in more detail. It was analyse for December 15th 2016. The number of messages was much higher in the HCG data. Effect for the Port of Piraeus shows Figure 7.

**Figure 7. Number of apperances for common data (160), Port of Piraeus for 2016-12-15**

Source: On the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

Greece has more than 2000 large and smaller islands most of them located in Aegean sea and the reduction in the number of messages per ship results in a problem for following the journey of a ship. Figure 8 shows that the reduction of messages in the European terrestrial AIS data may pose difficulties in following the journey of a ship.

**Figure 8. Journey of a ship from HCG and EU terrestrial AIS data in the port of Piraeus and in the Aegean Sea**

 AIS-HCG

 AIS-EU

Source: On the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

*2.3.2 Comparing satellite to land based AIS data*

To analyse the satellite data the number of ships and number of messages was compared for different ports for European terrestrial source and satellite data from Luxspace (LS). It was the same date - for December 15th, 2016 and the same reference frame of ships.

In Table 8 it can be seen that coverage of the data based national terrestrial AIS data from HCG and satellite data is better than EU terrestrial AIS data. Only for Port of Piraeus the number of ships is identical. For the most other ports there were more ships on HCG data. The satellite data from LS completely covers the Mediterranean Sea. However, satellite data structurally detects a lower number of ships per port compared to the high quality national data of Greece.

**Table 8** **The number of ships per port – 2016-12-15**

|  |  |  |
| --- | --- | --- |
| **Name of port** | **Satellite AIS** | **Land based AIS** |
| **LS** | **EU terrestrial AIS** | **HCG** |
| ***Port of Heraklion*** | 5 | - | 8 |
| ***Port of Patras*** | 24 | - | 40 |
| ***Port of Piraeus*** | 162 | 162 | 162 |
| ***Port of Thessaloniki Thessaloniki Thessaloniki Thessaloniki*** | 33 | - | 61 |
| ***Port of Volos*** | 20 | - | 19 |

Source: On the basis of Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data.

 **3. Technology**

One of the aims of the projects was to obtain practical experience in the use of methods and big data technologies for official statistics. For our work, we chose a common working environment based on Hadoop Stack, in order to ensure high scalability and performance. Data comes from European terrestrial AIS was decoded by Statistics Netherlands and uploaded to the UNECE sandbox and stored in the Hadoop File System (HDFS) and made available for use from all participating countries in our project [1]. Generally, we used the code developed in Scala and processed by Apache Spark which was designed as a computing platform. In Spark, the speed comes from in-memory computing that allows a much faster processing and response on data stored in the HDFS. The process of comparing datasets (PL, GR and EU terrestrial AIS) was made on the same code prepared in Scala.

In addition, we investigated the possibility of using and implementing noSQL databases for AIS data. Main advantages of NoSQL database are high scalable and performance, they run well on clusters and are mostly open-source. We focused on Elasticsearch which is a tool for indexing, analyzing, searching and storing JSON data and Kibana which searches and visualizes information in a table, graph or cartogram. A great benefit of Elasticsearch is that it can store Geo Data [4] in the database and can be use to filter data for distances between points and to search within bounding boxes and aggregations. Geo Data can be represented as a geo-shape or a geo-point which allows to store data as latitude and longitude coordinate pairs. By using geo-point or geo-shape, Elasticsearch automatically finds the coordinates, validates them according to the needed format, and index them. In Elasticsearch many useful methods based on Geo Data are available, eg. geo-bounding box, geo-distance or geo-shape. For the comparison of privately held to national datasets geo-bounding box with geo-point was used where two points (top-left, bottom-right) were set to defining rectangular area for ports and the number of appearances of each ships (based on MMSI number) was counted [5].

For the presentation of AIS-data we have used analysis tools like R (directly reading data from HDFS) and Kibana (reading data from Elasticsearch).

**4. Conclusions**

* In almost all cases, national AIS data and satellite data contained (much) more data than European terrestrial AIS data: both in terms of coverage and number of messages per region. If European terrestrial AIS data does cover a port, data is sufficient to determine the port visits. However, it is not always sufficient to determine ships’ journeys, especially in areas with a capricious geography. European terrestrial AIS data source provided filtered data from multiple partners – with no documentation available (metadata);
* Satellite data structurally detects a lower number of ships per port however they are indispensable to following the route of ships across oceans;

**5. Acknowledgment**

This work was supported by the ESSnet Big Data project, [WP4: AIS data], which is part of the Big Data Action Plan and Roadmap 1.06 (BDAR) agreed to be integrated into the ESS Vision 2020 portfolio.

**6. References**

 [1] Puts et al., (2016). Work package 4 AIS data deliverable 4.1: Creating a database with AIS for official statistics: possibilities and pitfalls. Available at: <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/1/18/WP4_Deliverable_4.1_2016_07_28.pdf>

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

 [3] Consten et al., (2017). Work package 4 AIS data deliverable 4.3: Report about sea traffic analysis using AIS-data v1.0. Available at: [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4\_Deliverable\_4.3\_2017\_07\_21\_v1.0.pdf](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/5/5c/WP4_Deliverable_4.3_2017_07_21_v1.0.pdf%20)

[4] Elasticsearch - geodata, <https://www.elastic.co/blog/enriching-searches-open-geo-data>

[5] Bis et al., (2017). Work package 4 AIS, Milestone 4.10: Progress and technical report of 1st internal WP-meeting SGA-2, Annex 1: How to use geo\_point in the queries available for Elasticsearch, Available at: <https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/76/WP4_Milestone_4.10_2017_11_13.pdf>