SCIENTIFIC JOURNAL OF POLISH NAVAL ACADEMY
ZESZYTY NAUKOWE AKADEMII MARYNARKI WOJENNEJ

2018 (LIX) 4 (215)

sciendo

DOI: 10.2478/sjpna-2018-0024

BIG DATA FOR ANOMALY DETECTION IN MARITIME SURVEILLANCE: SPATIAL AIS DATA ANALYSIS FOR TANKERS

Dominik Filipiak, Milena Stróżyna, Krzysztof Węcel, Witold Abramowicz

Poznan University of Economics and Business, Department of Information Systems, Niepodległości 10 Ave., 61-875 Poznań, Poland; e-mail: {dominik.filipiak; milena.strozyna; krzysztof.wecel; witold.abramowicz}@ue.poznan.pl

ABSTRACT

The paper presents results of spatial analysis of huge volume of AIS data with the goal to detect predefined maritime anomalies. The maritime anomalies analysed have been grouped into: traffic analysis, static anomalies, and loitering detection. The analysis was carried out on data describing movement of tankers worldwide in 2015, using sophisticated algorithms and technology capable of handling big data in a fast and efficient manner. The research was conducted as a follow-up of the EDA-funded SIMMO project, which resulted in a maritime surveillance system based on AIS messages enriched with data acquired from open Internet sources.

Key words:
maritime surveillance, AIS data, anomaly detection, big data.

Research article

INTRODUCTION

In the past decade, along with an increasing number of commercial ships, maritime protection and surveillance have become more challenging. Governments have been facing tasks such as protection of sovereignty and infrastructure or counteracting terrorism and piracy. Therefore, detection or anticipation of maritime illegal activities at sea — anomalies — became one of the main issues of maritime surveillance. The challenge to be faced in the maritime domain is collection and analysis of maritime-related data, required to detect these anomalies. An important source of information is AIS, which generates huge volume of data every day. Before leveraging them, they have to be pre-processed (decoded), stored, and analysed. Within a timespan of a few years, they stack up to terabytes of data. To date, scant attention has been paid to this issue. Advanced and efficient analytical methods have the potential to enable real-time detection of potential threats and support users in decision-making [26], hence the motivation for this article.

AIS messages provide only basic information about a given ship. In order to complement the AIS data other relevant information about ships must be retrieved from external sources, which are often scattered across different systems and applications. Open internet sources can provide general ship data (flag, detailed type, length, gross tonnage, capacity, technical specifications, and construction details), ownership data (current and past owners), classification status, classification history, former ship name, flag history, or security related information (bans, detentions, port state controls). All this information is relevant to characterise a ship and its movement — it gives some context that might be important in anomaly detection. However, currently such additional information is rarely exploited in analyses aimed at anomaly detection.

Taking into account these challenges, a set of methods for detection of selected maritime anomalies based on analysis of big number of AIS messages that are enhanced with data retrieved from open internet sources has been developed. The conducted analyses concerned anomalies related to the movement of ships and their characteristics. The aim of this paper is to present the results of performed calculations using the presented methods on real data. They were implemented using big data technologies and thus the potential resulting from efficient analysis of a vast amount of maritime-related data has been demonstrated.

The paper is structured as follows. In next part we review related research, sequent presents the research method, as well as describes used dataset. The main

part of the article is included in part 'Anomaly detection', in which we describe maritime anomalies selected for this study and results of the conducted analyses. The paper is concluded with a short summary.

RELATED WORK

There is no strict definition for an anomaly — in general, it represents a deviation from a standard behaviour [6, 20] or which appears to be inconsistent with the remainder of a dataset [10]. In the maritime domain, an anomaly can be defined as a deviation from the expected [14] and its detection refers to the problem of finding vessels that behave differently from the majority of other vessels. The review of the literature showed that there is a large variety when it comes to anomalous behaviour of ships. Andler et al. gathered different types of maritime anomalies in a taxonomy that classifies anomalies in a few groups, such as rendezvous, movement-, cargo-, history-, owner-, and crew-related [2]. Similar analysis was made by Roy and Davenport, who additionally paid attention to the quality of transmitted AIS data (e.g. missing or impossible data), criminal activities of ships, ship's motion (e.g. drifting, loitering, too high speed), and its position (e.g. proximity to other objects, presence in restricted zones, travelling outside normal or historical route) [31]. Riveiro studied anomalies that are particularly important for Vessel Traffic Service (VTS) centres: grounding situations, collisions, vessel entering restricted zones, identification of unknown vessels, vessels not following a standard route, vessels carrying dangerous goods, vessels with a history of being involved in illegal activities, suspicious flag or port, discovering oil spills and floating object [29]. Another issue is a trend of registering merchant vessels in open registries (Flag of Convenience) [27].

Previous line of research has established a number of various maritime anomalies whose detection is crucial for different entities from the maritime domain. In the maritime domain, the anomaly detection is a complex task that requires acquisition of information from various sources and its analysis (often with operator's expertise) in order to detect an abnormal behaviour [24]. This process must often be tailored to an application domain and properties of data [5]. Possible patterns in data should also be considered. An anomalous behaviour at sea depends on the context, therefore its detection requires application of different approaches, techniques, and data. The anomaly detection techniques can be divided into two groups: data-driven and knowledge-driven approaches [12, 15]. Knowledge-driven

techniques use the external source of knowledge (e.g. expert knowledge) and include different reasoning paradigm such as rule-based, description logic, and case-based reasoning. Data-driven approaches can be divided into several classes: classification-based, nearest-neighbour-based, clustering-based, statistical (parametric, non-parametric), spectral, and information theoretic [6]. The vast majority of anomaly detection techniques used in maritime domain is data-driven. Considering the machine learning methods, the Bayesian network seems to be particularly useful in detection of single-point vessel anomalies (such as location, course, speed) [9, 11, 22, 23], as well as to detect anomalies in the whole scenario of cooperation between a ship and pilot-boat [8] or piracy on oil platforms [4]. Neural networks are another popular classification-based technique, used to detect anomalies in ship's speed, position, and course [3, 28]. Among clustering-based techniques, DBSCAN algorithm was applied to detect anomalies in ship's speed [13, 26] or to identify entry and exit points to a particular area [26]. K-means method was used to detect various activities of ships [37]. Laxhammar examined unsupervised methods for analysis of normal sea traffic patterns [16, 17, 18]. Statistical methods include both parametric (such as regression and Gaussian Mixture Models [13, 15, 16, 30]) and non-parametric (such as kernels Gaussian Process) approaches [5, 18, 34].

The process of anomaly detection also depends on factors other than the applied methods, such as data preparation. It consists of two major stages: representation and transformation. For instance, a chosen representation restricts what can be learnt and what kind of anomalies can be discovered. Mascaro et. al. studied possible advantages of integrating additional data to AIS reports [23]. They added variables related to a ship itself, the weather, time-related factors, and information on vessel interactions. Additional maritime data can also be obtained from open data [12]. Transformation of maritime data is required in identification of ships' trajectories — an ordered vector of positions readings for a single vessel. Mascaro proposed to separate AIS data into tracks: each record was assigned to a separate track, based on MMSI [23]. Another approach is grid-based, which tessellates the area into equal sized tiles (referred as segments hereafter), and calculates a number of properties for them (such as speed) [25, 33, 38]. There are a lot of approaches focusing on maritime anomaly detection. Still, there is a lack of methods that would integrate sensor and positional data (e.g. AIS) with ancillary sources or detect maritime anomalies based on retrospective analysis of AIS from a relatively long period of time (e.g. one- or two-year) using big data techniques. A single, comprehensive method for the detection of loitering has not been proposed as well.

RESEARCH METHOD

The defined research problem positions this study in the area of data science with the focus on data mining and data analysis. Less formally, data mining starts with raw data to come up with a solution for a given problem. Formal steps have been encapsulated within the Cross-Industry Standard Process for Data Mining (CRISPDM) [32]. Regardless of the considered domain, CRISP-DM consists of 6 phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. In this paper, we deliberately focus on two phases — data preparation and modelling. Other phases are described in other papers stemming from our SIMMO project (e.g. [19, 35, 36]).

We used two types of data sources in our research: AIS messages and selected open internet sources. Our AIS dataset consisted of class A position reports (AIS message types 1, 2, and 3) and static and voyage-related data reports (AIS message type 5) from satellite and terrestrial receivers. The data scope of this study was limited to messages sent by tankers[1] in 2015. In total, we collected 569,079,486 class A position reports matching these criteria (only 19GB with Parquet's compression[2]). Due to the focus on anomalies with a dynamic nature, we later selected a subset consisting of messages in which a navigational status was set to 0 or 4 (under way using engine and constrained by her draught respectively), which resulted in 313,747,021 position reports (message type 1–3). These messages formed trajectories of vessels. Following other studies [25, 33, 38], we adopted the grid-based approach. Therefore, the whole world was tessellated into 64,800 segments of 1°×1° each (including the virtual one for null coordinates), which were further characterised by parameters derived from the ship position reports. Segments with no received messages have been excluded from further calculations, leaving us with 33,123 active segments. Each point of each trajectory of each vessel was evaluated — we checked whether it was a valid point or whether it was an occurrence of an anomaly (along with a type of an anomaly). Regarding static and voyage-related data reports (message type 5), we collected 255,349,946 messages from tankers in 2015 (7.2GB parquet file). In this dataset, there were 34,662 tankers with unique MMSI numbers, of which 32,262 had meaningful trajectories, i.e. trajectories for which it was possible to calculate the maximum speed over ground (max(SOG) > 0).

[1] A tanker is a vessel whose reported two-digit *ship type* field starts with 8.

[2] https://github.com/apache/parquet-format.

The second dataset consisted of data collected from selected open internet sources. The purpose of this step was to enrich information about a given vessel with data not available in AIS messages. However, internet sources are known for their issues with data quality. Data might be incomplete, inconsistent, or may contain contradictory facts. Therefore, appropriate methods were developed to alleviate these quality issues. The process of internet sources selection and data fusion was described in detail in separate papers ([35] and [19] respectively). As a result a vast amount of ancillary data for all types of vessels was acquired, such as tonnage, dimensions, detailed type, built year, builder, home port, data about detentions and inspections of ships as well as data about classification statuses of ships and their affiliation to a classification society. Notably, we collected 57,193 maritime companies, 85,652 classification surveys from 134 classification societies, and 29,011 events of bans and detentions of vessels. Regarding MID country codes, we collected 23 black--listed flags, 30 grey-listed flags, and 50 flags marked as the Flag of Convenience.

The analysis of a big amount of maritime data is a complex task that requires application of appropriate technologies and tools for processing and storing them. Previous research showed that the processing time of large sets of AIS data is a real challenge. Since it is a continuous data stream, traditional analysis methods constrained to one physical machine lead to a computational inefficiency. Some studies reported that processing of one month's AIS data takes one day [25, 33]. Wu et al. covered few years of AIS data on a global scale in their research [38], though they did not leverage big data techniques. Namely, the solution proposed by Wu et.al., based on the MySQL technology, does not scale well. This problem was addressed in our research. We developed a scalable solution that allows to efficiently analyse a huge amount of AIS data. As a rule of thumb, a reliable big data cluster should provide a near-linear scalability, in-memory computing, stream processing, and efficient data compression. The Hadoop-compliant processing framework — Apache Spark [39] — has been chosen. It enables fast in-memory computing and facilitates a number of analytical tasks. In the case of our research, 257.5GB RAM were used to conduct the analysis in an in-memory manner, whereas a set of 40 cores was responsible for parallel task execution. AIS messages were stored using space-efficient column--oriented data format — Parquet. Technically speaking, maritime anomaly detection requires appropriate processing of vast amounts of immutable data (AIS) in order to infer correct findings. Our system is inspired by the Lambda architecture [21], since static anomalies can be processed in a stream, while the traffic analysis and loitering detection requires a batch processing. A final output depends on a combination of these three approaches. Similarly to other big data solutions, adding more worker

nodes to a cluster is expected to increase a computational power and storage capacity in a linear manner. The former is the main bottleneck in legacy maritime surveillance systems.

ANOMALY DETECTION

This section describes the selected maritime anomalies that were considered in this study, followed by a presentation of the results of the developed anomalies detection methods for tankers in 2015. Maritime anomalies were selected in the course of the SIMMO project [1]. A user requirements analysis, conducted in the project, allowed us to identify a list of anomalies that are particularly interesting for potential users of the SIMMO system. This paper contains a subset of these anomalies, which were further grouped into three types of analysis: traffic analysis, static anomalies, and loitering detection. Traffic analysis identifies the busiest routes, determines an average and maximum speed within a segment, as well as an average relative speed and its standard deviation. This part does not detect anomalies, though it is necessary for further analysis. Static anomalies concern detection of ships that fly a black or grey flag or Flag of Convenience, have the IMO number in a banned or detention list, have a certificate issued by a low-performing Recognised Organisation, belong to a low-performing company, and have a withdrawn or suspended classification status. These anomalies rely on data from open internet sources. Loitering-related anomalies form the last type and they are divided into 7 subcategories: an invalid coordinates, location or speed, a sharp change of course, an unpredicted location, and an unusually low or high speed.

Traffic analysis

One of the categories of anomalies considered in this research is 'loitering', which mainly concerns anomalous speed of a vessel. When a vessel moves at the high sea with a speed that is too slow for its class, it is an indicator of anomalous behaviour. Altogether, when the speed deviates, then a warning should be issued. In order to be able to carry out such reasoning and detect 'loitering', we have to define the notion of a normal speed, which then can be used as a reference point to indicate the anomalous speed in a given area. Such speed should be location-specific, i.e. it should be defined for a certain geographical area. In general, the normal speed can be inferred from historical AIS data. To this end, in our research we divided the globe in smaller

segments of size 1°×1°. Then, we calculated the average speed of vessels in a given area. In the first place we took an absolute speed, but the results were not satisfactory. Therefore, we decided to introduce a relative speed assuming that vessels travel at full steam at high sea (if otherwise, it can be considered as an anomalous behaviour). A variation of speed was calculated with relation to the standard deviation — variability is higher closer to ports, straits, etc., so the anomaly warning should not be raised. The results of our approach for determination of a normal speed are presented in the series of figures. The proposed method is easily parametrised and precision can be increased.

We started with an overview of a number of messages sent within each segment (fig. 1). In total 313,747,021 position reports received from tankers with navigational status set to 0 or 4 in 2015 were analysed. It is important to notice that colours are based on logarithmic scale, so the supremacy is even higher than visually interpreted from the figure. For example, the English Channel is a region that stands out with respect to the number of reported messages. The vast amount of messages in this region may result from the popularity of this waterway. Moreover, it is important that in areas with a dense ship traffic we observed problems with synchronisation of time between various AIS devices. Therefore, some anomalies that were detected in such regions might be false positives.

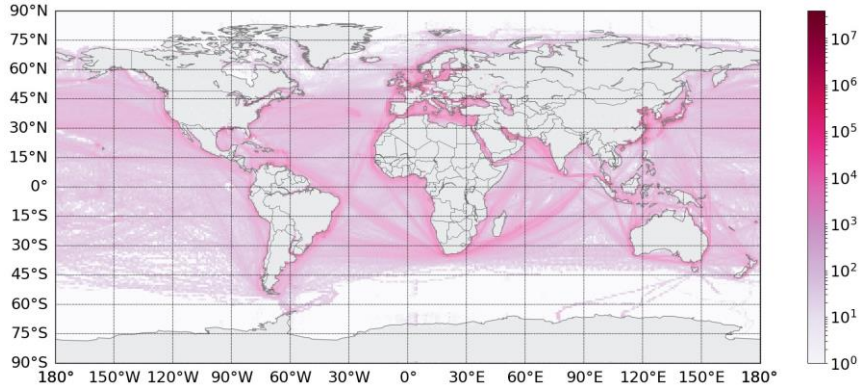

Fig. 1. Number of received AIS position reports per segment (log scale)

To perform further analyses, we calculated an average speed of all vessels based on AIS messages sent in a given region. Results are presented in fig. 2. The analysis of the chart does not reveal anything specific about the results. There are some areas that are characterized with a relatively low average speed, though in general an average speed at high sea is relatively stable (e.g. Atlantic Ocean).

Nevertheless, the average speed is not a good measure since a maximum speed significantly varies among different types of ships. Therefore, a second, more sophisticated method for an evaluation of speed was proposed. Instead of looking at the absolute value of speed, a relative speed was taken into account, i.e. a current speed with relation to a maximum observed speed (not the one declared by a ship-yard). The results of the relative speed calculation are presented in fig. 3. For example, if a relative speed in a given segment equals 0.13, it means that vessels travelled on average at 13% of their maximum historical speed in a given segment. It is worth to mention that at the high sea this speed was not much higher than in coastal regions. It means that in general vessels rarely travel at full steam — usually, it is 61% of their maximum speed.

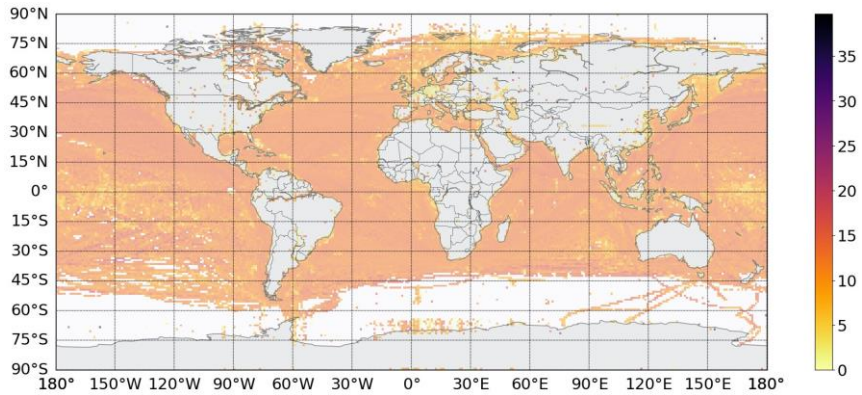

Fig. 2. Average speed over ground in knots per segment

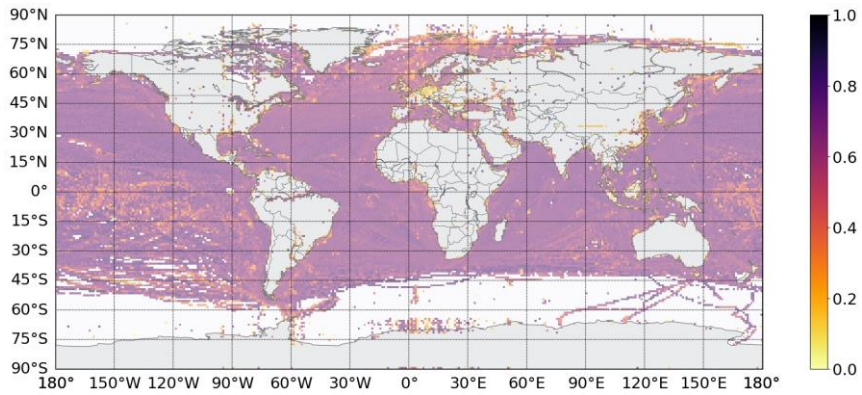

Fig. 3. Average relative speed per segment

If we want to detect speed anomalies in a given segment, in fact we need to consider the standard deviation of a relative speed. On the one hand, a speed of vessels can be relatively stable and, thus, even a minor deviation may indicate an occurrence of loitering. On the other hand, in some regions, like near ports, there were ships that travelled both with a high and a close-to-zero speed. Looking at the world map (fig. 4), a general conclusion can be drawn that a variability of speed is higher at coastline and in regions near to ports, as expected. Tab. 1 gathers basic statistics about ships' speed in all the segments of the world.

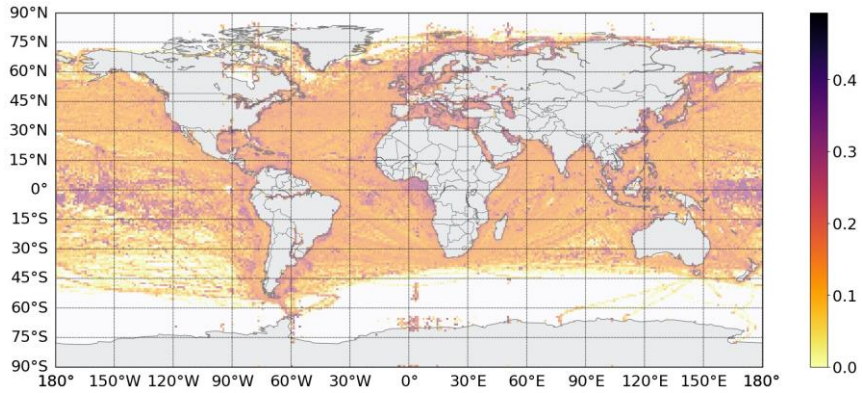

Fig. 4. The standard deviation of a relative speed per segment

Tab. 1. Statistics of tankers' traffic in 33,123 segments of 1°×1°

	Mean	**σ**	**Min**	**25%**	**50%**	**75%**	**Max**
Messages	9449.12	240,904.28	0.0	149.00	554.00	1670.00	39,260,666.00
Mean SOG	11.63	3.10	0.0	10.97	12.53	13.31	39.70
Max SOG	17.70	4.91	0.0	14.70	18.20	20.50	40.00
Rel. SOG	0.61	0.16	0.0	0.59	0.66	0.70	1.00
Rel. SOG (std)	0.14	0.08	0.0	0.10	0.14	0.17	0.49

Static anomalies

As mentioned in part 'Research method', in order to detect static anomalies, data from selected internet sources were used.

In calculations presented in this section, the number of detected anomalies in a given segment was divided by the number of AIS position reports received in this segment. Such an approach enabled to spot anomalies also in the areas with a dense ships' traffic and resulted from the fact that if we would use a nominal number

of ships with a given anomalous characteristic, the map would be biased due to the high standard deviation of messages received in segments with a dense traffic. We refer this approach as a relative.

The analysis of static anomalies started with classification of flag states into the three categories: black (high risk), grey, and white (low risk). The colours of flags were assigned based on data published by well-known maritime organisations, such as: the Paris MoU, the Tokyo MoU, and the US Coast Guard. These organisations determine colours of flags based on a risk assessment that reflects the safety performance of ships registered to each flag state, measured as the number of port state inspections and detentions recorded over a three-year period. If a ship is flying a black or grey flag, it is considered as an anomaly.

Spatial distributions of black flags (Anomaly S1) and grey flags (Anomaly S2) are presented in fig. 5 and 6 respectively. Blacklisted tankers are particularly active in the area between the Indian Ocean and the Pacific Ocean. It results probably from a localization of their home ports – for example, we registered 555 vessels from Indonesia, which is a black-listed country. The highest activity of the grey flags was observed on the coast of the Indian Ocean, particularly around Thailand, Philippines, and India — probably for the same reason as above. Interestingly, Madagascar, which is not a grey-listed country, observed a very high activity of such vessels.

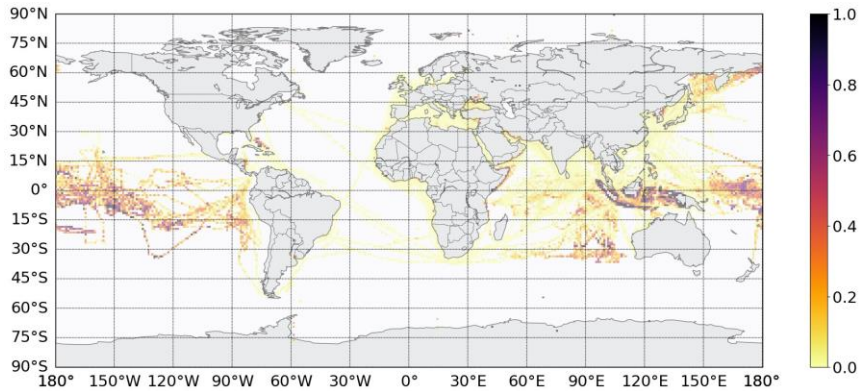

Fig. 5. Anomaly S1 — traffic of tankers of the black-listed flags (relative)

Flag of Convenience (Anomaly S3) is a business practice of registering ship in a sovereign state, different from that of the ship's owners. FoC allows ship-owners to be legally anonymous, what hinders prosecution in civil and criminal actions. There are examples of FoC ships that have been found engaged in crime, offering

substandard working conditions, and negatively impacting the environment. Therefore, the fact that a ship is of a FoC country indicate that it might be dangerous. Besides, such ships are targeted for special enforcement by other countries they visit. Apparently, the spatial distribution of FoC tankers are very high across the whole world, since they constitute nearly 20% of all the analysed vessels (fig. 7). Marshall Islands, Liberia, and Panama were the most popular FoCs.

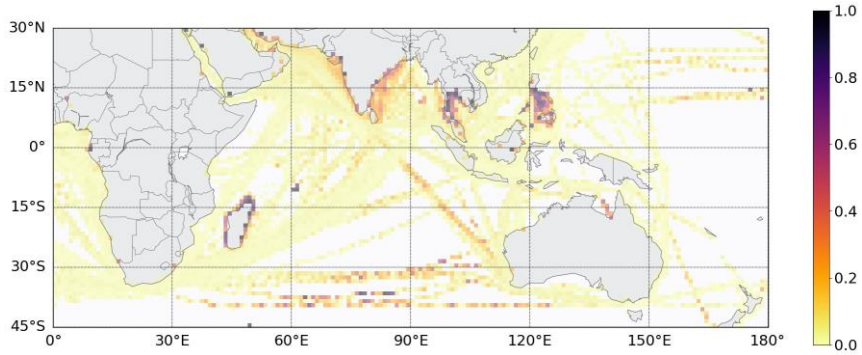

Fig. 6. Anomaly S2 — traffic of tankers of the grey-listed flags (relative)

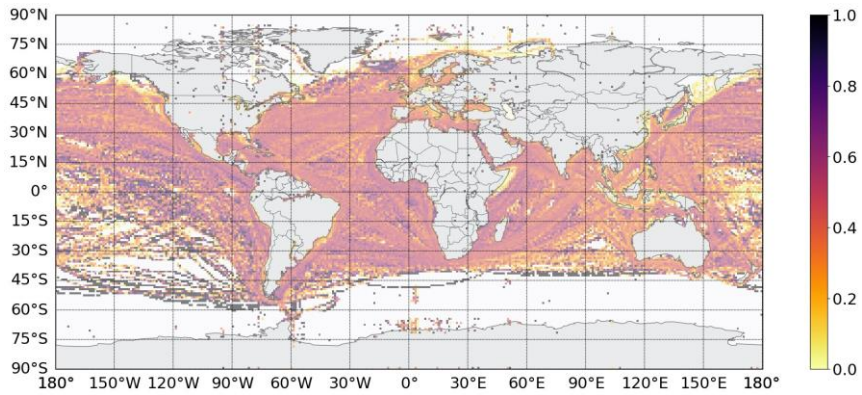

Fig. 7. Anomaly S3 — AIS position reports sent by FOC tankers (relative)

A ship can be subject to Port State Control (PSC), after which, in the case of an occurrence of any deficiencies that are clearly hazardous to the safety of a state or to the environment, a ship can be detained (Anomaly S5). If a ship was detained for three or more times by a maritime authority during the last 12 or 24 months, it is classified as banned or added to a list of an under-performing ships by a given MoU (Anomaly S4). These ships are subjected to more frequent inspections at ports

within a MoU region. In the course of the analysis a single banned tanker was found in the area of the Gulf of Oman (fig. 8). On the contrary, detained tankers were found across the whole globe. However, it seems that they were active mostly nearby the Micronesia and the Marshall Islands (fig. 9).

Classification societies are non-governmental organisations that establish and maintain technical standards for construction and operation of marine vessels. The primary role of a classification society is to validate if a design and technical equipment of a ship are in accordance with the published standards. If a ship meets all the requirements, a classification society issues a classification certificate. Nevertheless, each ship can be further inspected by a flag states with regard to the security and safety standards. Paris MoU maintains a list of classification societies whose total number of inspections over a 3-years period does not meet the minimum of 60. Such classification societies are called Recognised Organizations (RO). ROs that do not meet the criteria for their ships to qualify as Low Risk Ships, are listed as low--performing RO (Anomaly S6). Thus, ships having classification certificate issued by a low-performing RO are potentially dangerous. Our analysis showed that in 2015 such tankers concentrated mostly at the Chinese coast and particularly near Taiwan (fig. 10).

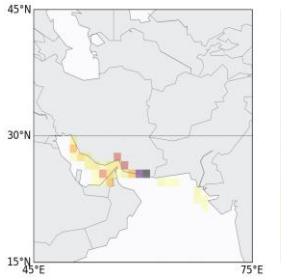

Fig. 8. Anomalies S4/S8 — AIS position reports sent by a banned or suspended tankers (relative)

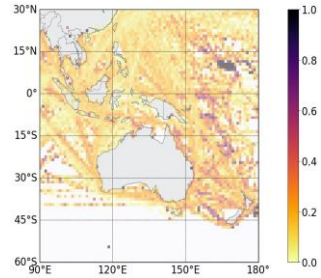

Fig. 9. Anomaly S5 — AIS position reports sent by tankers marked as detained (relative) [

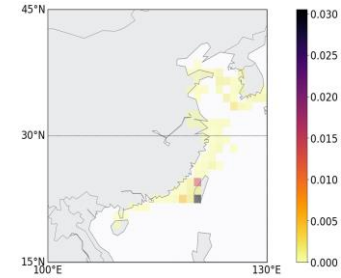

Fig. 10. Anomaly S6 — AIS position reports sent by tankers belonging to the low-performing RO (relative)

The European Maritime Safety Agency (EMSA) publishes a list of poor--performing companies (Anomaly S7). The list includes entities that 'having had the vessels for which they are responsible subject to Port State control inspections and on the basis of those inspection results demonstrate an unwillingness or inability to comply with the international conventions on maritime safety and on the protection

of the marine environment' [7]. In the course of the analysis, 24 tankers matching that criterion were identified. They were particularly active in some parts of the Pacific Ocean, south of Hawaii (fig. 11).

Classification societies are also responsible for granting a ship a classification status. This status is designated based on a periodical survey of a ship, which aims at ensuring that it continues to meet the classification standards. There are five classification statuses that may be granted: delivered, suspended, reinstated, withdrawn, or reassigned. The ships with a withdrawn and suspended status should can be regarded as an anomaly due to the fact that they do not comply with the standards and thus pose a potential threat (Anomaly S8). We detected only one tanker that matched this criterion — it was the same vessel as presented in fig. 8.

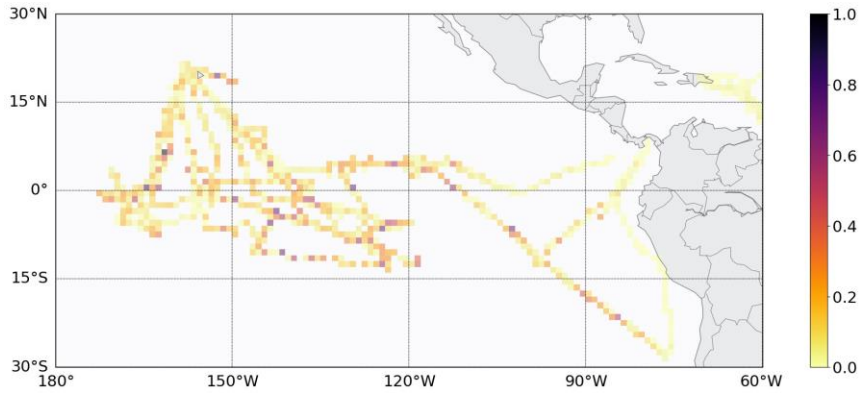

Fig. 11. Anomaly S7 — AIS position reports sent by tankers belonging to low performing companies (relative to the number of all considered position reports in a segment)

Tab. 2 summarizes the results presented in this section, providing the statistics on detected static anomalies that relate to selected characteristics of tankers.

Tab. 2. Static anomalies related to tankers in 2015

ID	Anomaly	Tankers	%
S4	IMO in banned list	1	0.003%
S8	Withdrawn or suspended	1	0.003%
S6	Low performing RO	5	0.014%
S7	Low performing company	24	0.069%
S1	Blacklisted flag	1512	4.362%
S2	Graylisted flag	1521	4.388%

ID	Anomaly	Tankers	%
S5	IMO in detention list	1983	5.721%
S3	Flag of Convenience	7097	20.475%
	Tankers without static data anomalies	24345	70.235%
	Tankers total	34662	100.000%

Loitering-related anomalies

Having calculated the statistics concerning all the regions of the world (relative speeds in particular segments presented in Section 'Traffic analysis'), the next step was detection of anomalies related to a movement of ships. We call this type of anomalies as loitering. First, we performed a simple check if correct coordinates were provided in AIS messages (Anomaly L1). Since we worked on a pre-processed subset of data, this particular anomaly should not be referred as a good indicator of the percentage of incomplete AIS data — it was used just to filter out incorrect messages from a dataset used in next stages. After that, we performed a simple check whether a speed over ground reported in a message was within expected limits (Anomaly L2). We set a threshold to 25 knots, meaning that a speed above this value was perceived as an anomaly. Although this approach may seem to be trivial and the threshold is rather arbitrary, it was helpful for further data cleansing process, since segments with the highest relative amount of reports with invalid speed, presented in fig. 12, indeed seem to be anomalous.

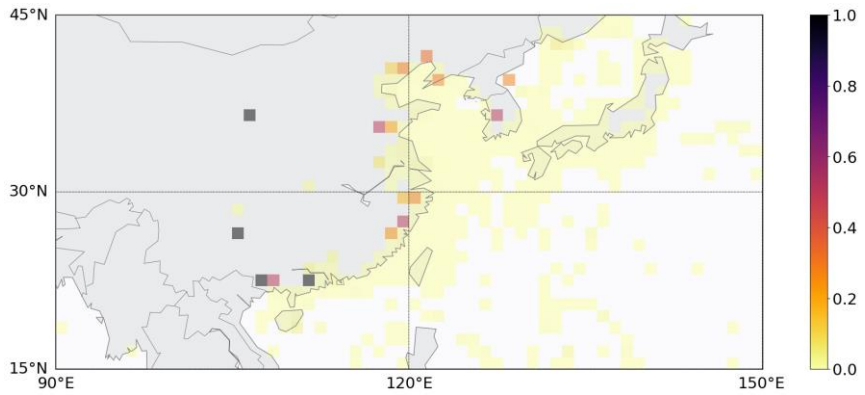

Fig. 12. Anomaly L2 — AIS position reports from tankers with an invalid speed (relative)

In the next step we checked if the actual position of a ship is reliable taking into account its potential speed over ground (Anomaly L3). This method was also

helpful for further data cleansing, since it eliminated problems with incorrect AIS reading — it filtered out situation of a sudden 'teleportation' of a ship. Fig. 13 presents some interesting patterns, as the marked segments tend to cluster along parallels and meridians.

The next analysis concerned an angle anomaly (Anomaly L4), which detects a sharp change of course (over 90 degrees). We assume that a ship should not change its course so rapidly and if it happens, it might be interpreted as loitering. Although such behaviour seems to be very evenly distributed across the seas (fig. 14), it was possible to discover some interesting examples, as in fig. 15. The vessel, while waiting for an entry to a port, was travelling in circles. Perhaps a more rational behaviour would be to stop at high sea, therefore this example may be treated as an example of detected loitering behaviour.

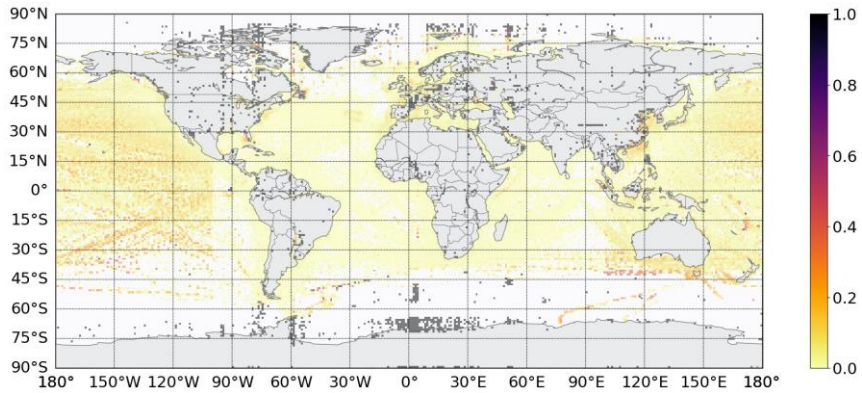

Fig. 13. Anomaly L3 — AIS position reports from tankers with an invalid location (relative)

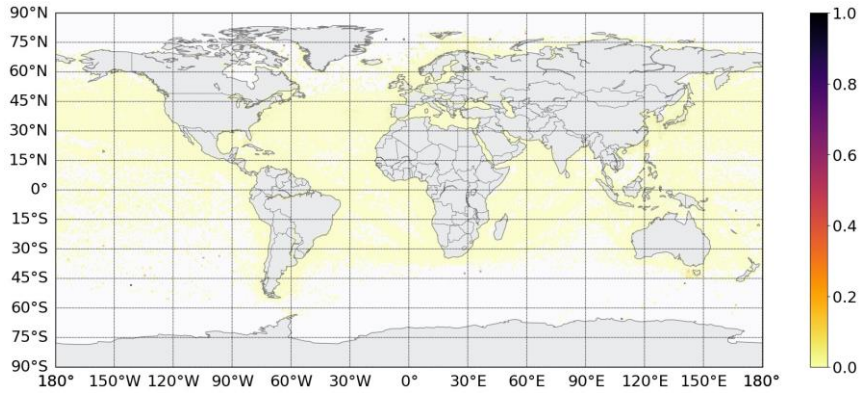

Fig. 14. Anomaly L4. AIS position reports with an anomalous angle (relative)

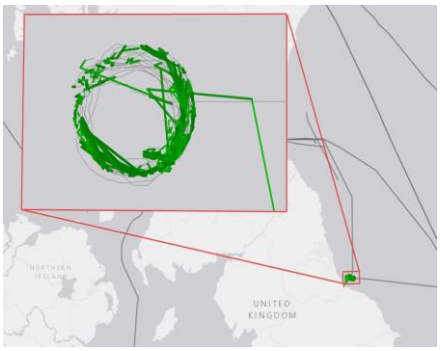

Fig. 15. Example of an angle anomaly (L4) — trajectory of a vessel travelling in small circles

Another detected anomaly concerned a situation when a ship was found in another location than inferred from its previous course — we marked this as an unpredicted location (Anomaly L5). Compared to L3, more sophisticated rules were checked here. The expected location was predicted based on two previous points and times, so it was assumed that a vessel continues its trajectory. A location other than the predicted one with a margin of 3 miles was marked as anomalous, however ships that did not move for 1 hour were excluded. The analysis of the results provided based on this rule shows that it might have produced a number of false positives (mostly in Europe — fig. 16).

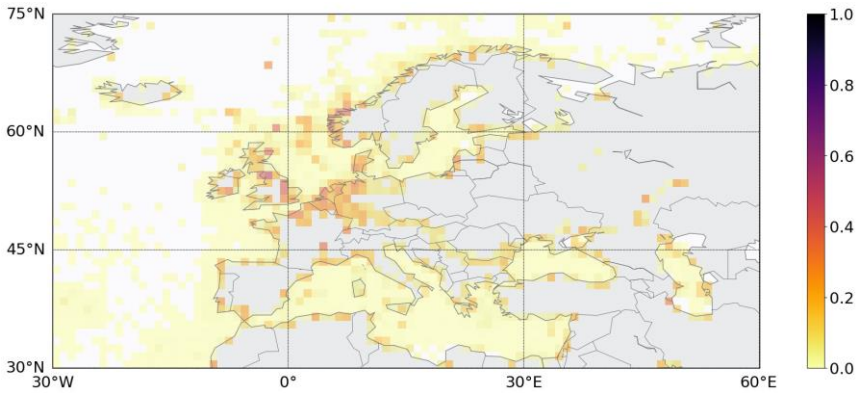

Fig. 16. Anomaly L5 — AIS position reports with an unpredicted location (relative)

The last method tested whether a ship was sailing with an unusually low or high speed (Anomalies L6 and L7). Loitering occurs when a ship being at the high sea starts sailing with a low speed. This method takes an average relative speed *rs* and its standard deviation $\sigma(rs_s)$ in a given segment *s*. For a position report *i* sent

in segment s, a ship's relative speed v is compared with an average relative speed in a given segment. If the difference exceeds a defined threshold, this position report was marked as anomalous, i.e. $\frac{s_{v,i}}{\max(s_v)} > rs_s + 2\,\sigma(rs_s)$ in the case of L7. This algorithm required definition of an appropriate threshold (it had to be decided what is a reasonable deviation). After a set of experiment, the 2σ approach was used since using based on this value less false positives were returned that for σ. The results of these analyses are presented in fig. 17 and 18. The former is slightly prone to report false positives in a dense areas (such as the English Channel). Though, the latter seems to yield positions worth further investigation. A brief summary of all the loitering--related anomalies is provided in tab. 3.

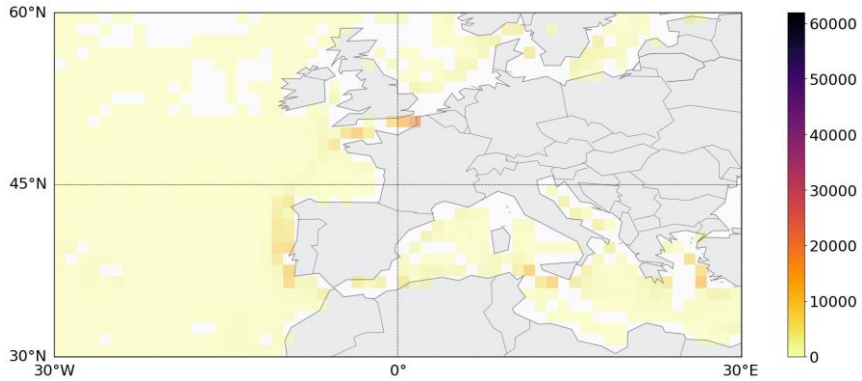

Fig. 17. Anomaly L6 – AIS position reports from tankers with an unusually low speed (nominal)

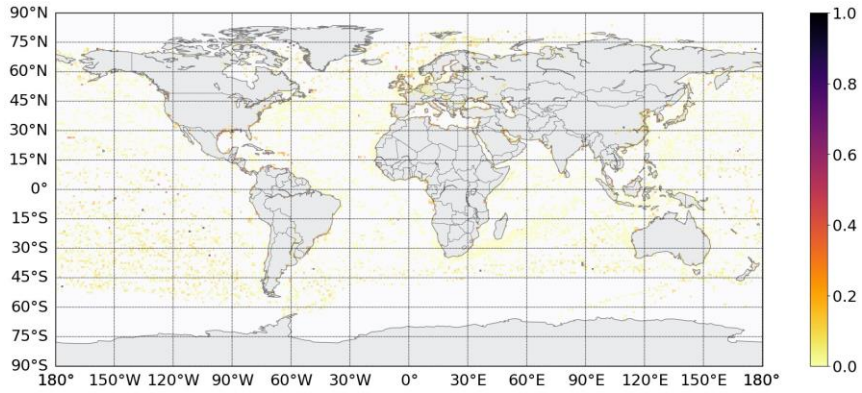

Fig. 18. Anomaly L7 — AIS position reports from tankers with an unusually high speed (relative)

Tab. 3. Results of loitering detection for tankers in 2015

ID	Anomaly	Reports	%
L1	Invalid coordinates	126,483	0.040%
L3	Invalid speed	808,339	0.258%
L5	Unpredicted location	2,746,783	0.875%
L7	Speed unusually low	3,434,848	1.095%
L2	Invalid location	11,849,397	3.777%
L6	Speed unusually high	24,495,390	7.807%
L4	Sharp course change	37,105,095	11.826%
	Messages without anomalies	235,543,722	75.074%
	Messages total	313,747,021	100.000%

CONCLUSIONS

In this article we proposed the big data approach to the problem of maritime anomaly detection based on AIS data fused with data retrieved from open internet sources. We conducted a large-scale spatial analysis of behaviour of tankers in 2015. The aim was to perform a traffic analysis and to discover static and loitering related anomalies.

The summary of the obtained results is presented in tab. 4, which shows the most frequent anomalies detected in segments. Since in our study we examined only tankers, our findings are rather not generalisable to a larger population of ships (other types of ships). Future research will have to clarify whether this statement is true for the presented anomalies. Among the static anomalies, flying a Flag of Convenience (S5) dominated the ranking by a huge margin. With regard to loitering, an unpredicted location was the most frequently reported anomaly. Though, this method (L5) needs more study, since it seems that many of the messages were false positives. The most interesting results were yielded by the invalid location (L3) and angle anomaly (L4) detection methods. The latter provides a good starting point for a process of searching for loitering, since some detected trajectories indeed resembled such behaviour. The invalid location (L3) anomaly detection method revealed some interesting patterns. Further research should identify mechanisms why these anomalies are generated. A possible interpretation of these findings might stem from quality--related issues, such as data integrity. Unfortunately, AIS data collection is prone to noise — not all data are transmitted correctly. Normally, the noise is coming from terrestrial AIS sources, not from satellite ones. Moreover, technical limitations may

cause other problems. Some AIS messages are lost, particularly when ships are not in the scope of a satellite or in high-density zones like the North Sea, the Mediterranean Sea, the Caribbean, and the Yellow Sea. Also, intentional tampering of data can occur. Therefore, future studies on maritime anomaly detection method should isolate the effects of incorrect AIS messages and incorporate more sophisticated methods to discover them.

Tab. 4. Statistics on detected anomalies for 33,123 segments of 1°×1°

ID	Anomaly	Mean	σ	Min	25%	50%	5%	Max
S1	Blacklisted flag	74.29	4408.66	0	0	0	3	557416
S2	Graylisted flag	331.08	7591.80	0	0	0	20	1121255
S3	Flag of Convenience	1881.67	13355.65	0	24	207	710	1170285
S4	IMO in banned list	0.20	13.44	0	0	0	0	1733
S5	IMO in detention list	435.08	4331.25	0	0	32	136	497589
S6	Suspicious RO	1.38	166.22	0	0	0	0	29932
S7	Suspicious company	7.45	204.06	0	0	0	0	26253
S8	Withdrawn/suspended	0.20	13.44	0	0	0	0	1733
L1	Invalid coordinates	3.82	694.97	0	0	0	0	126483
L2	Invalid location	24.40	678.81	0	0	0	0	68999
L3	Invalid speed	357.74	7530.97	0	1	7	44	1098903
L4	Sharp course change	82.93	1306.01	0	0	1	5	89904
L5	Unpredicted location	1120.22	72458.99	0	0	0	2	12327223
L6	SOG unusually high	103.70	642.88	0	0	8	63	61911
L7	SOG unusually low	739.53	23069.14	0	0	0	0	3342353

The performed experiments provided a preliminary evidence that incorporation of big data techniques and the Lambda architecture in AIS data processing increases the speed and efficiency of analyses, and as such should be the preferred solution used in the maritime domain. The whole process of data analysis (described in part 'Anomaly detection'), starting from reading of data in Parquet format from disk up to the plotting of figures took less than two hours, which is several orders of magnitude less than using standard database queries. The conducted analysis were very data-intensive and these operations required a high throughput. We believe that the presented results might still be improved with a proper hardware, such as fast (at least 10 Gbps) network switches and solid-state drives. It is worth to mention that plotting the maps was a small bottleneck — we used the standard Matplotlib library. A dense map of the whole world (such as in fig. 1) needed more than 30 seconds to be generated. In the future studies we plan to further evaluate

incorporation of big data technologies in the process of analysis of AIS and other maritime-related data, especially in terms of meticulously measuring of analysis time using different frameworks and storage formats.

REFERENCES

[1] Abramowicz W., Filipiak D., Małyszko J., Stróżyna M., Węcel K., *Maritime Domain Awareness System Supplied with External Information-Use-Case of the SIMMO System*, Publication materials of 7th International Science and Technology Conference NATCON on 'Naval Technologies for Defence and Security", ed. T. Szybrycht, Polish Naval Academy, Gdynia 2016, pp. 1–20.

[2] Andler S. F., Fredin M., Gustavsson P. M., van Laere J., Nilsson M., Svenson P., *SMARTracIn: A Concept for Spoof Resistant Tracking of Vessels and Detection of Adverse Intentions*, 'Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VIII', 2009, Vol. 7305, 73050g-73050g-9, DOI: 10.1117/12. 818567.

[3] Bomberger N. A., Rhodes B. J., Seibert M., Waxman A. M., *Associative Learning of Vessel Motion Patterns for Maritime Situation Awareness*, Proc. of 9th International Conference on Information Fusion, FUSION 2006, pp. 1–8, DOI: 10.1109/Icif.2006.301661.

[4] Bouejala A., Chaze X., Guarnieri F., Napoli A., *A Bayesian Network to Manage Risks of Maritime Piracy Against Offshore Oil Fields*, Safety Science, Elsevier, 2014, Vol. 68, pp. 222–230, DOI: 10.1016/j.ssci.2014.04.010.

[5] Brax C., *Anomaly Detection in the Surveillance Domain*, PhD Thesis, Örebro Universitet, 2011.

[6] Chandola V., Banerjee A., Kumar V., *Anomaly Detection: A Survey*, 'ACM Comput. Surv.', 2009, Vol. 41, No. 3, pp. 15:1–15:58, DOI: 10.1145/1541880. 1541882.

[7] European Maritime Safety Agency, *Important Information Regarding The Publication of Low and Very Low Performance Companies According to Article 27 of Directive 2009/16/EC on Port State Control and Commission Regulation (EU) 802/2010 as Amended Implementing Article 10(3) and Article 27 of Directive 2009/16/EC of the European Parliament and of the Council as Regards Company Performance*, [online], https://portal.emsa.europa.eu/web/thetis/company-performance-legal-information [access 30.04.2015].

[8] Fooladvandi F., Brax C., Gustavsson P., Fredin M., *Signature-Based Activity Detection Based on Bayesian Networks Acquired from Expert Knowledge*, Proc. of 12th International Conference on Information Fusion, FUSION 2009, pp. 436–443.

[9] Helldin T., Riveiro M., *Explanation Methods for Bayesian Networks: Review and Application to a Maritime Scenario*, Proc. of The 3rd Annual Skövde Workshop on Information Fusion Topics, SWIFT 2009, pp. 11–16.

[10] Hodge V. J., Austin J., *A Survey of Outlier Detection Methodologies*, 'Artificial Intelligence Review', 2004, Vol. 22, No. 2, pp. 85–126.

[11] Johansson F., Falkman G., *Detection of Vessel Anomalies — a Bayesian Network Approach*, 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, 2007, pp. 395–400.

[12] Kazemi S., Abghari S., Lavesson N., Johnson H., Ryman P., *Open Data for Anomaly Detection in Maritime Surveillance*, 'Expert Syst. Appl.', 2013, Vol. 40, No. 14, pp. 5719–5729, DOI: 10.1016/J.Eswa.2013.04.029.

[13] Kraiman J. B., Arouh S. L., Webb M. L., *Automated Anomaly Detection Processor*, Proc. of SPIE: Enabling Technologies for Simulation Science VI, 2002, eds. A. F. Sisti, D. A. Trevisani, pp. 128–137.

[14] Laere J., Nilsson M., *Evaluation of a Workshop to Capture Knowledge from Subject Matter Experts in Maritime Surveillance*, Proc. of 12th International Conference on Information Fusion, FUSION 2009, pp. 171–178.

[15] Lane R. O., Nevell D. A., Hayward S. D., Beaney T. W., *Maritime Anomaly Detection and Threat Assessment*, Proc. of 13th International Conference on Information Fusion, FUSION 2010, pp. 1–8.

[16] Laxhammar R., *Anomaly Detection for Sea Surveillance*, Proc. of 11th International Conference on Information Fusion, FUSION 2008, pp. 1–8.

[17] Laxhammar R., Falkman G., *Conformal Prediction for Distribution-Independent Anomaly Detection in Streaming Vessel Data*, Proc. of 1st International Workshop on Novel Data Stream Pattern Mining Techniques, 2010, pp. 47–55.

[18] Laxhammar R., Falkman G., Sviestins E., *Anomaly Detection in Sea Traffic-A Comparison of the Gaussian Mixture Model and The Kernel Density Estimator*, Proc. of 12th International Conference on Information Fusion, FUSION 2009, pp. 756–763.

[19] Małyszko J., Abramowicz W., Stróżyna M., *Named Entity Disambiguation for Maritime-Related Data Retrieved from Heterogeneous Sources*, 'TRANSNAV: International Journal on Marine Navigation and Safety of Sea Transportation', 2016, Vol. 10, No. 3, pp. 465–477.

[20] Martineau E., Roy J., *Maritime Anomaly Detection: Domain Introduction and Review of Selected Literature*, Technical Report, DTIC Document, October 2011, [online], https://apps.dtic.mil/dtic/tr/fulltext/u2/a554310.pdf [access 18.03.2018].

[21] Marz N., Warren J., *Big Data: Principles and Best Practices of Scalable Real-time Data Systems*, Manning Publications Co., 2015.

[22] Mascaro S., Korb K. B., Nicholson A. E., *Learning Abnormal Vessel Behaviour from AIS Data with Bayesian Networks at Two Time Scales*, Clayton School of Information Technology, Monash University, August 2010, [online], https://bayesian-intelligence.com/publications/TR2010_4_AbnormalVesselBehaviour.pdf [access 20.03.2018].

[23] Mascaro S., Nicholson A. E., Korb K. B., *Anomaly Detection in Vessel Tracks Using Bayesian Networks*, 'Int. J. Approx. Reasoning', 2014, Vol. 55, No. 1, pp. 84–98, DOI: 10.1016/J.Ijar.2013.03.012.

[24] Matthews M., Martin L. B., Tario C. D., Brown A. L., *A Non-Intrusive Alert System for Maritime Anomalies: Literature Review and the Development and Assessment of Interface Design Concepts [Systeme D'alerte Non Intrusive en Cas D'anomalies Maritimes: Examen de la Documentation et Elaboration/Evaluation de Concepts D'interface]*, Technical Report, DTIC Document, March 2009, [online], http://cradpdf.drdc-rddc.gc.ca/PDFS/unc88/p531847.pdf [access 18.03.2018].

[25] MMO, *Mapping UK Shipping Density and Routes from AIS*, Technical Report, Marine Managment Organisation, MMO Project No. 1066, Newcastle 2014.

[26] Pallotta G., Vespe M., Bryan K., *Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction*, Entropy', 2013, Vol. 15, No. 6, pp. 2218–2245.

[27] Pozo F., Dymock A., Feldt L., Hebrard P., Monteforte F., Sanfelice D., *Maritime Surveillance in Support of CSDP*, Technical Report, European Defence Agency, April 2010, [online], http://www.offnews.info/downloads/wisePenReportApril2010.pdf [access 20.03.2018].

[28] Rhodes B. J., Bomberger N. A., Seibert M., Waxman A. M., *Maritime Situation Monitoring and Awareness Using Learning Mechanisms*, Proc. of Military Communications Conference, IEEE, Milcom 2005, pp. 646–652.

[29] Riveiro M., *Visual Analytics for Maritime Anomaly Detection*, PhD Thesis, Örebro Universitet, 2011.

[30] Riveiro M., Falkman G., Ziemke T., *Improving Maritime Anomaly Detection and Situation Awareness Through Interactive Visualization*, Proc. of 11th International Conference on Information Fusion, FUSION 2008, pp. 1–8.

[31] Roy J., Davenport M., *Categorization of Maritime Anomalies for Notification and Alerting Purpose*, Technical Report, Defence R & D Canada — Valcartier, October 2009.

[32] Shearer C., *The CRISP-DM Model: The New Blueprint for Data Mining*, 'Journal of Data Warehousing', 2000, Vol. 5, No. 4, pp. 13–22.

[33] Shelmerdine R. L., *Teasing out the Detail: How Our Understanding of Marine AIS Data Can Better Inform Industries, Developments, And Planning*, 'Marine Policy', 2015, Vol. 54, pp. 17–25.

[34] Smith M., Reece S., Roberts S. J., Rezek I., *Online Maritime Abnormality Detection Using Gaussian Processes and Extreme Value Theory*, ICDM, 2012, pp. 645–654.

[35] Stróżyna M., Eiden G., Abramowicz W., Filipiak D., Małyszko J., Węcel K., *A Framework for the Quality-Based Selection and Retrieval of Open Data — A Use Case from the Maritime Domain*, 'Electronic Markets', 2018, Vol. 28, Issue 2, pp. 219–233.

[36] Stróżyna M., Małyszko J., Węcel K., Filipiak D., Abramowicz W., *Architecture of Maritime Awareness System Supplied with External Information*, 'Annual of Navigation', 2016, Vol. 23, No. 1, pp. 135–149.

[37] Tun M. H., Chambers G. S., Tan T., Ly T., *Maritime Port Intelligence Using AIS Data*, 'Recent Advances in Security Technology', 2007, No. 33, pp. 33–43.

[38] Wu L., Xu Y., Wang Q., Wang F., Xu Z., *Mapping Global Shipping Density From AIS Data*, The Journal of Navigation', 2017, Vol. 70, No. 1, pp. 67–81.

[39] Zaharia M., Chowdhury M., Das T., Dave A., Ma J., McCauley M., Franklin M. J., Shenker S., Stoica I., *Resilient Distributed Datasets: A Fault-tolerant Abstraction for in-Memory Cluster Computing*, Proc. of 9th Usenix Conference on Networked Systems Design and Implementation, April 2012, pp. 2–2.

BIG DATA I WYKRYWANIE ANOMALII W RUCHU MORSKIM: PRZESTRZENNA ANALIZA DANYCH AIS DLA TANKOWCÓW

STRESZCZENIE

W artykule zaprezentowano wyniki przestrzennej analizy dużej ilości danych AIS z jednego roku w celu wykrycia wybranych anomalii morskich. Anomalie podzielono na trzy grupy: związane z ruchem, statyczne i wykrywanie tzw. loiteringu — każda z nich została przetestowana na

podstawie raportów wysyłanych przez tankowce w 2015 roku. Analizę przeprowadzono przy użyciu zaawansowanych algorytmów i technologii big data pozwalających na szybką ocenę dużych wolumenów danych morskich. Badanie zostało przeprowadzone jako kontynuacja projektu SIMMO, w ramach którego opracowano system nadzoru morskiego oparty na wiadomościach AIS wzbogaconych o dane pozyskiwane z otwartych źródeł internetowych.

Słowa kluczowe:
nadzór morski, dane AIS, wykrywanie anomalii, big data.

Article history
Received: 08.05.2018
Reviewed: 27.09.2018
Revised: 09.11.2018
Accepted: 12.11.2018