Chapter 12
Methods and Techniques for Automatic Identification System Data Reduction

Claudia Ifrim, Manolis Wallace, Vassilis Poulopoulos, and Andriana Mourti

Abstract Extensive use of sensors and system automations tend to become a trend in our everyday life. In this manner, maritime traffic is monitored by advanced sensor systems, one of which is the Automatic Identification System, called AIS. The use of AIS systems in shipping is extensive and large numbers of data are produced; thus the task of analyzing this information becomes a laborious procedure. In this paper, we analyze AIS data sets in order to propose methods and techniques for reducing the data without losing important information in order to produce sets of data that can be easily processed in real time. Furthermore, the deriving data are visualized so as to present that data reduction does not affect the visualized data which is a very common application of AIS data.

12.1 Introduction

Nowadays, sensors and intelligent systems are widely used and extensive research and development activities are performed in this domain. The outcomes of the research that is done is available to our everyday lives, even without being able to realize the existence or the complexity of the systems behind daily tasks. An important intelligent sensor system that is used primarily for identifying and locating vessels is Automatic Identification System (AIS). The Automatic Identification

C. Ifrim
University Politehnica of Bucharest, Bucharest, Romania
e-mail: claudia.ifrim@hpc.pub.ro

M. Wallace (✉) · V. Poulopoulos · A. Mourti
Knowledge and Uncertainty Research Laboratory, University of Peloponnese, Tripoli, Greece
e-mail: wallace@uop.gr

V. Poulopoulos
e-mail: vacilos@uop.gr

A. Mourti
e-mail: andrianamourti@uop.gr

F. Pop and G. Neagu (eds.), *Big Data Platforms and Applications*,
Computer Communications and Networks,
https://doi.org/10.1007/978-3-030-38836-2_12

System (AIS) is an automated tracking system used on ships and by vessel traffic services (VTS) that broadcasts in an interval of seconds information, such as unique identification of the ship, position, course, speed, and navigation status, to other nearby ships, AIS base stations and satellites [1].

Acting as an enhanced two-way positioning system, it is able to assist the tracking and monitoring of vessel movements to watch-standing officers, by providing a number of useful metadata. On board, it is integrated with other navigation aids, such as:

- Global Positioning System (GPS);
- Radio Detection And Ranging (RADAR);
- Electronic Chart Display Information System (ECDIS);
- Voyage Data Recorder (VDR);
- Automatic Radar Plotting Aid (ARPA).

As it is obvious from several different options, the number of positioning and navigation aiding systems is large and, unfortunately, do not utilize a unique standard. Such a policy would offer a global solution for actively managing the increasing traffic and provide solutions for efficient resource planning. This means that the analysis of maritime traffic data would lead to the analysis of a very large number of heterogeneous data. In specific, AIS data received by base stations or satellites is recorded to stations for extended use, either real-time or future use. According to the protocol that defines the generation of AIS data, their number is vast and can easily overload a storage system in a short period of time; therefore, leading to an increase of the amount of time needed for data processing and information retrieval.

This paper presents methods and technologies applied on the nature of AIS data sets in order to achieve storage of much lower amounts of data without losing the important information about monitored vessels. The actual size of data produced for a number of vessels can easily become unmanageable, leading to an increase in the needs of resources for the applications that use this data for historic or real-time purposes. We believe that lossless data reduction can decrease the response time of such applications and provide better quality in the combination of results.

The rest of the paper is organized as follows: The next section presents projects and research that rely on analysis of AIS data. Section 12.3 contains an overview of AIS technology, what types of errors can be detected on AIS data, and how they can be corrected, as well as references to the existing applications; Sect. 12.4 presents methods and techniques for reducing the amount of AIS data without losing important information; in Sect. 12.5, we present the results of applying our techniques on AIS data sets and we finalize with Sect. 12.6 discussing the results of our procedure and future enhancements on the proposed methodologies.

12.2 Related Work

A number of research fields is related on maritime traffic each one for different purposes. As the AIS data is useful for providing details about the ship, its cargo and its route information, they are utilized for their quality of information.

As part of the project "Emissions from shipping in the Arctic" in [2] they clearly state that the availability of satellite based AIS data for ship tracking makes it possible to set up detailed fleet specific emission inventories in a high temporal and spatial resolution for the Arctic and carry out dispersion calculations with enhanced precision. Similar to this, a project that analyzed emissions in Asia was conducted and presented in [3].

An unsupervised and incremental learning approach to the extraction of maritime movement patterns is presented in [4] to convert from raw data to information supporting decisions. This is useful, for example, in counter piracy applications to identify risk areas associated with the joint predicted presence of white shipping density (e.g., commercial merchant traffic) and Pirates Action Groups (PAG) [5].

The possibility of using AIS data for fisheries research and the provision of an analysis about the level of uptake of the AIS by the EU fishing fleet is explored in [6]. The specific work is part of a more long term objective of producing a high resolution map of fishing effort for Europe using AIS. A similar procedure for improving the detection of fishing patterns from Satellite AIS data using Data Mining and Machine Learning is presented in [7].

From the aforementioned, it is clear that the applications of AIS data analysis can be multidimensional. It includes from simple fleet management or monitoring, environmental impact of vessels to more complex and important tasks that include real-time collision avoidance systems. As the regulations are such that AIS has to be installed to every vessel, it is a technology that will formulate the future of data analysis considering maritime traffic.

12.3 AIS Technology

The Automatic Identification System (AIS) [8] is an automated, autonomous tracking system which is extensively used in the maritime world for the exchange of navigational informational between AIS-equipped terminals [9]. As its definition implies, it is a two-way navigation and metadata exchange system designed to be applied on vessels. In order to operate as a two-way system, it consists of AIS receivers and transmitters on board, satellite, and ashore; which enable the electronic exchange between AIS stations. In fact, all passengers' vessels and all commercial vessels over 299 Gross Tonnage (GT) that travel internationally are required to have an AIS transponder aboard. Two types of transponders exist, Class A, which should be used generally by all vessels, while smaller vessels can use Class B transponders which provide fewer data and have smaller ranges of transmission (5–10 min) [1].

The AIS format uses Time-division multiple access (TDMA) which is a channel access method for shared-medium networks. It allows several users to share the same frequency channel by dividing the signal into different time slots [1]. Each user is assigned a time slot during which the transmission is possible, making the procedure able to be performed at the same time to a single station from a large number of transmitters. On the other side, the time slots are just 4,500 per minute; assuming each vessel requires a time slot for its transmission, this numbers is equal to the maximum capacity of a station. In case of larger number of transmitters, interference between the signals will occur, and as a result a number of errors will be produced to the recorded data. In order to overcome such a problem, a grid of receivers is used in cases that is expected to have transmissions of large number of data from multiple vessels at the same time and place. Apparently, due to the nature of maritime traffic, this is expected to occur in specific times and places which makes it possible to foresee and anticipate the problem. In more detail, multiple base stations are installed in high traffic areas, ports and channels; all of them aggregate their input into a single stream of data while in parallel they correct multiple instances of the same information (duplicates) or errors that may occur.

Technically, AIS data is defined as ASCII[1] packets as byte stream using the NMEA 0183 [10] or NMEA 2000 [11] data formats. As a broadcasting transmission protocol, it has an indicator for packets concerning other ships ("!AIVDM") and packets concerning current ship ("!AIVDO"). The standard about the so-called AIVDM/AIVDO messages is ITU1371 [12], which was expanded and clarified by ITU-R [13]. The ASCII format for AIVDM/AIVDO representations of AIS radio messages have been set by IEC-PAS [14] and a common AIVDM byte stream could be like the following:

!AIVDM,1,1,B,177KQJ5000G?tO'K > RA1wUbN0TKH,0*5C

The following Table 12.1 presents the definition of each part of the message.

*The *-separated suffix (*5C) is the NMEA 0183 checksum for the sentence, preceded by "*". It is computed on the entire sentence including the "AIVDM" excluding the "!"*

As the AIS protocol is used for defining many different parameters under many different circumstances, there are numerous different types of messages. Twenty-seven different message types exist, most of which are used to report position. Regulations exist in order to define the frequency of transmitting by type of message. As such, *safety* messages must be sent as required, *long range* messages have to be sent every 30 min, *static* messages have to be sent every 6 min or when data is amended upon request, and finally, the most important *dynamic* messages have very detailed rules that must be followed in order to achieve the scope of the AIS protocol usage [15]. The following Tables 12.2 and 12.3 present the dynamic message frequency regulations by type:

From the definition of the regulations, it is obvious that messages can occur as often as every 2 s per vessel. The regulation and definition of the protocol is such that

[1] ASCII character set—reference definition: https://tools.ietf.org/html/rfc2046.

Table 12.1 Definition of AIS data example

Field	In example	Definition
#1	!AIVDM	Identifies this as an AIVDM packet
#2	1	Is the count of fragments in the message. The size of a sentence is limited to an 82-character maximum, so it sometimes has to be split over multiple sentences
#3	1	Is the fragment number of this sentence, one based. So a message with fragment count of 1 and fragment number of 1 is a complete message
#4	Empty	Is a sequential message ID for multi-sentence messages
#5	B	Is a radio channel code. AIS uses the high side of the duplex from two VHF radio channels: AIS Channel A is 161.975 Mhz (87B); AIS Channel B is 162.025 Mhz (88B). Codes 1 and 2 may also be encountered instead of A or B
#6	177KQJ5000G?tO‘K> RA1wUbN0TKH	Is the data payload
#7	0	Is the number of fill bits requires to pad the data payload to a 6 bit boundary, ranging from 0 to 5

Table 12.2 Dynamic Class A

Ships dynamic conditions	Not changing course	Changing course
At anchor or moored and moving less than 3 knots	3 min	3 min
At anchor or moored and moving faster than 3 knots	10 s	10 s
0 to 14 knots	10 s	3 1/3 s
14 to 23 knots	6 s	2 s
Over 23 knots	2 s	2 s

it acts as the medium to achieve a number of important issues that exist in maritime traffic. In short, the transmission of the messages can support:

- Collision avoidance
- Vessel traffic services
- Aids to navigation
- Search and Rescue
- Maritime security
- Cargo tracking
- Fleet tracking
- Fishing fleet monitoring.

Table 12.3 Equipment other than Class A shipborne mobile

Platform's condition	Nominal reporting interval
Class B "SO" shipborne mobile equipment not moving faster than 2 knots	3 min
Class B "SO" shipborne mobile equipment moving 2–14 knots	30 s
Class B "SO" shipborne mobile equipment moving 14–23 knots	15 s
Class B "SO" shipborne mobile equipment moving > 23 knots	5 s
Class B "CS" shipborne mobile equipment not moving faster than 2 knots	3 min
Class B "CS" shipborne mobile equipment moving faster than 2 knots	30 s
Search and rescue aircraft (airborne mobile equipment)	10 s
Aids to navigation	3 min
AIS base station	10 s

Each of the aforementioned usages of the AIS data may have different prerequisites. In our algorithmic analysis, we focus both on the kind of information that is needed for real-time application, such as collision avoidance, as well as data analysis application, which can occur in research fishing fleet analysis.

12.4 Algorithm Analysis

From the analysis and definition of the protocol, we observe that a large number of data is produced in a unit of time, in which we should be able to apply algorithms and procedures either to reduce the data themselves or to predict data. By analyzing the types of messages existing (as presented in the following list), we put the focus on position messages which are the most frequent (messages 1, 2, 3, 18 and derivatives 5, 19, and 24). The following Table 12.4 displays the AIS messages.

Table 12.4 AIS message analysis

Message	Analysis
1, 2, 3	Position reports
4[a]	Base station report
5	Ship static and voyage related data
18	Standard Class B equipment position report
19	Extended Class B equipment position report
24	Static data report

[a]Message 4 will be treated just to display the general geographic distribution of base stations

Important information that helps us recognize and analyze the messages that are transmitted are the number of ships or equipment of the transmitter, a number that is unique; an identification of the message transmitted is helpful for distinguishing messages as well as a repeat indicator that was designed to be used for repeating messages over obstacles by relay devices.

We are mainly focusing on messages 1, 2, and 3 (position reports for class A) as they contain navigational information that include longitude and latitude, time-stamp, heading, speed, ship's navigation status (under power, at anchor, etc.), which are the messages on which we will apply our algorithmic procedures.

A surface observation of the data leads to more than 2 million records in a small but crowded area in the time of a month. In parallel, more and more vessels install AIS devices on board leading to an exponential increase of the data in the period of time. Consequently, it is expected that data will only increase and the effort of the applications analyzing data will be affected. As a matter of fact, it is inevitable that reducing the amount of data without losing the valuable information should be a procedure that is essential.

A first procedure of our algorithm is the identification of the parameters that are included in the majority of messages and could be candidate data to be compressed. As already mentioned in the previous paragraphs, messages 1, 2, and 3 containing navigational information will be processed.

From the data observation, as well as the nature of the protocol, we make some important assumptions in order to examine if there are situations under which we should be able to reduce data. We conclude the following four situations which are candidates to help us reduce the data.

Repeated data. This is a very common situation that can occur when we receive data from "stopped" vessels. A vessel can be in multiple different conditions in order to be stopped (e.g., vessel in port, or standing for refueling). In that case, we can apply a simple algorithm that rejects new data (e.g., when latitude and longitude remain unchanged) and update the time-stamp of the latest stream.

Easily calculated data. This approach has the philosophy of Video compression protocols. What remains unchanged or can be easily calculated (e.g., standard route) can be omitted as information. *Huge number of data*. For specific occasions where the vessels' speed is too low, despite the fact that the protocol itself predicts lower rates of transmission, we can enhance the procedure with additional algorithmic procedures in order to further reduce the data. *Custom data provision*. As a matter of fact, we are able to provide an API through which we perform custom queries on the data in order to return as results portions of the data and not the complete data set. In this manner, we should be able to perform all our techniques on the data that are demanded and only the part of data that is useful is provided as an answer.

In this manner, we utilize a library in order to be able to decode AIS streams from a specific AIS data set. This is useful as we will be able to have the encoded byte stream as "human readable" information on which we are able to perform algorithmic

actions. It is mC++[2] decoder for Automatic Identification System for tracking ships and decoding maritime information.

By analyzing our AIS data set, we can easily observe that the records count can be 2 million or more per month. Considering the increased number of vessels that install AIS devices on board, the number of records that will be stored can only increase and the only solution that we have is to propose a reduction technique that could be applied on AIS data records in order to reduce its size without losing any valuable information and to be able to optimally store and query historic AIS data.

We identified that navigational messages are the ones that will be processed in order to reduce either their number or their information. A vessel broadcasts such messages within specific time intervals that are based on its speed and course. By analyzing several different messages transmitted by a single vessel on voyage, we make the following assumptions:

- Attributes frequently changing over time: *latitude*, *longitude*, *time-stamp*
- Attributes rarely changing over time: *speed*, *heading*.

From the aforementioned, we extract the following information. When we are aware of a starting location, the heading and the speed we may locate the final location if we are aware of the starting and final time-stamp. This assumption helps us remove a large number of the frequent messages and replace them with a start and final location and the period of time. This also means that a vessel is traveling in straight lines, and thus we should be able to have enough information to reproduce the vessel's path with a starting and ending point. Under other conditions, if the speed remains under 0.1 knots, we should be assured that a vessel is not moving or that the change of its location should be imperceptible.

12.4.1 Analyzing the Data Set

From the analysis described, it is more than obvious that each vessel's records must be treated separately. If we obtain a data set from records deriving from a base station, that are the transmitted data from the vessels in proximity, then we should be able to recognize each vessel's transmissions. At first, we extract all the unique MMSI values in order to obtain the number of vessels that transmitted AIS data. For each of the extracted unique identifier, we export the byte streams transmitted in chronological order. We are utilizing the technique of "windows of information" a procedure similar to the way video compression is implemented. We assume that we have *i-records* that are records which include the complete set of data for all the parameters, *d-records* that are records which include only data for parameters that were differentiated and finally *p-records* that are records whose data for the parameters can be fully predicted. As in the video compression procedure some "key records" contain all the data while every other record's data are reduced significantly. Depending on the cruising point

[2] Libais C++ AIS data decoder: https://github.com/schwehr/libais/tree/master/src/libais.

of a vessel, it is expected to have a small number of *i-records* whenever a ship is en-route that include a large number of d or p-records, while in the situation of ship maneuvering more i-records are essentials while predictions or differentiation cannot be expected in the transmitted data.

12.5 Experimental Evaluation

The AIS data set that we used for our experiments contains information retrieved from the area of the Black Sea. It includes a number of 136,008,000 records. At first, we perform data correction, as well as data cleansing, a data preprocessing procedure in which we remove incorrect information. After this procedure, we apply the data reduction techniques and we visualize the results before and after the reduction in order to present the differentiation of the procedure.

12.5.1 Analyzed Data

The initial data analysis of the information leads to the creation of two different database tables. A PostgreSQL database with PostGIS extension is used to store the information within the messages. The information recorded was the decoded information and thus the two tables include:

- *static information*. Includes all the data that are related to the physical information of a vessel (type of vessel, length, year of construction, etc.)
- dynamic information (latitude, longitude, speed, etc.).

Despite the fact that the static information can provide us with a large number of qualitative data, which could be, for example, the maximum speed of the vessel and can be helpful for determining data anomalies, we analyze the dynamic data as they are much larger in number and are the data that can be used for applying data reduction techniques.

12.5.2 Data Reduction Applied on AIS Data Set

As mentioned in the previous paragraphs, some information is eliminated as an initial cleanup procedure. This includes searching for the following malformed data and removing them:

- Coordinates greater that 180, −180 latitude, and 90, −90 longitude
- The 0, 0 location.

Table 12.5 Break down of record types produced by algorithm application

Initial	Information records	Difference records	predicted records
568,934	96,719	223,488	248,727
	17%	39%	44%

This procedure removes almost 20% of the records in the data set, leading to the assumption that a large number of the transmitted data contain errors, that cannot possibly be corrected. The experimental evaluation is focused around a specific open sea area between Constanta and Sevastopol, which is crowded with AIS data transmissions (point: 43.70, 26.60). As a matter of fact from the starting set of data, after removal of the false and by applying a proximity query, we are able to obtain a number of 568,934 records.

The experimental evaluation consists of algorithm application on these records to obtain the final number of records that occur, indicating the compression that can be done on data. The application of the algorithm implies that it is possible to receive a number of almost 17% (96,719) of the data as information positions, and thus we need to keep the complete set of data. In order to be assured that we will have a number of detailed information records, it is essential to record one i-record at least every 20 min. These data cannot be compressed as they are essential in order to calculate the differentiated data (d-records) as well as the predicted data (p-records).

According to the algorithm, whenever we have slight changes to the speed then we are able to remove any other information and store the speed difference only. This is a d-record. In case of unchanged speed or heading, then a p-record is produced which is actually a virtual records as we assume that data for the specific time-stamp can be predicted. As the vessels tend to have standard speed and straight paths when being in open sea, then we are able to have large number of prediction and an amount of differentiated data. The following Table 12.5 presents the number of records per record type.

By analyzing the size of data, we shall be able to compute the level of compression. The information data include a number of data that need to be stored in the database in order to recognize and analyze AIS data. On the other hand, the data stream that is sent when a transmission exists could be a small number of bytes compared to the data stored in a database record. We will make a comparison with both the database info and the byte data stream. The data stream as defined by NMEA 0183 protocol is limited to a maximum of 82 characters which can be equivalent to the number of bytes needed to store the "sentence". On the other hand, a database human readable record of the data is more than 240 byte long, as a number of metadata is stored. The total number of 568,934 records would require 46MB of AIS data sentences or 136MB of data in the database. The number is approximate based on the maximum data. Our algorithm can possibly remove at all almost 45% of the messages, which are messages that are able to be predicted by combination of the information records and the differentiations records. Thus, the space required is only 55% of the initial. In

Table 12.6 Initial records versus reduced records

Initial number of records	Unique MMSI	Speed/heading difference	Final number of records
752,552	458	Less than 0.1/less than 5 degrees	248,743
752,552	458	Less than 0.15/less than 3.5	204,338
752,552	458	Less than 0.2/less than 1	202,248

parallel, when we need to store only the differentiation, then we put the focus on the speed parameter. In almost 40% of the cases, we can keep only the speed difference related to the information record of reference. Storing this data can be as large as 3 bytes in extreme occasions. This means that in 40% of the cases we can have a compression of 96% compared to the data stream and 98% compression compared to the data stored in the database records. Applying the compressions in the initial data, we conclude that we have a compression of 37% in the case of the byte stream and 38% in the case of records in the database. The total compression achieved by application of the algorithm is considered to be almost 80%. This compression can be easily achieved in areas that are open sea points of sea traffic where vessels are expected to have standard route and speed.

Our algorithm treats differently areas where large amount of diverse traffic occurs, which can be the are of the port. In this case, we examine the Constanta port. The initial set for the area has 752,552 records occurred from 458 different vessels. For this set, we followed the algorithm using different parameters for speed and heading of the vessels. The algorithm implies that when the speed or heading values change slightly then we omit the records considering them as ones that can be predicted or just be thrown away. As an example, we can consider that a speed of 0.1 knots means that a vessel will move around 150 m within an hour. At this amount of time, the system will certainly have a number of updates, thus making it possible to have a working data set without the need of this data (Table 12.6).

In this case, we observe that by completely omitting the records that have low speed or slight difference in speed heading, we can achieve a compression of almost 75%. The following section presents the visualization of the real and differentiated data set on a map. By viewing the initial versus reduced amount of data, it is possible to realize that our data reduction algorithm does not seem to cause any real change in the data, and especially of what is needed by AIS and is the definition of data, speed, heading, and location of vessels. From the visualization, it is furthermore possible to apply higher margins in speed and heading, and have a much large compression (more than 85%).

12.5.3 Data Visualization

In order to understand the differences that occur in the initial data set compared to the data set produced after application of our algorithm, we present a set of visualizations. By viewing the representation of the data on the map, we actually depict each record of the database as a dot on the map. The first set of visualizations present how the initial data set is presented on the map. In Fig. 12.1, we can see the row data presented on the map, while Fig. 12.2 presents the same data organized by MMSI, noting that MMSI is unique for each different vessel. Using a visual comparation we can easily outline the impact of our solution on data size and data quality. In this section, we present the visualization of the initial data set, the visualization of the reduced data set using different parameters for the speed difference, and in the last image we will have a representation of the initial data set compared with the reduced data set resulted after we applied our reduction algorithm.

Finally, Fig. 12.3 presents the density of the transmitted messages related to their position on the map. It is obvious that there is a huge number of transmitted data and the density—deriving from the port of Constanta—is large, as it is expected to be in huge sea traffic areas.

After the initial figures, we proceed with visualization of the data that are generated after the application of the algorithm on the initial data set. We will focus on the port of Constanta as it was the area that was analyzed in the experimental evaluation producing data reduced by 75% compared to the initial.

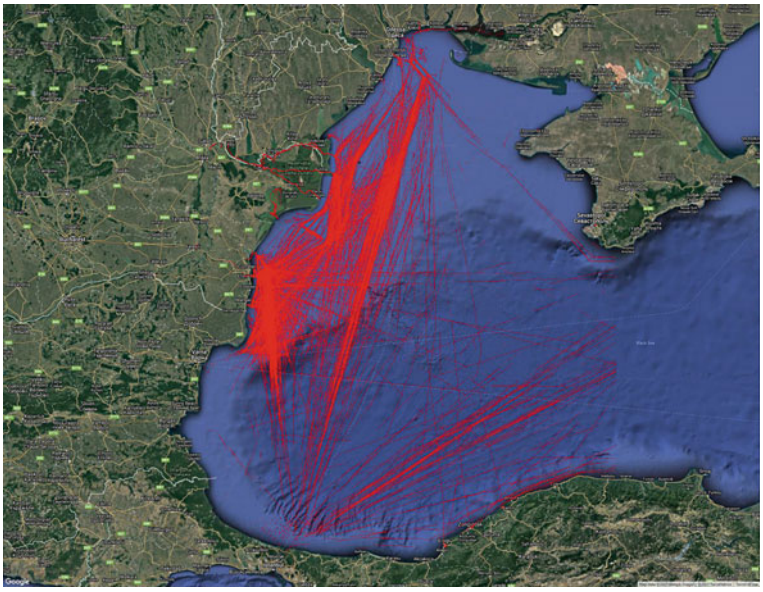

Fig. 12.1 Initial data set—routes visualization

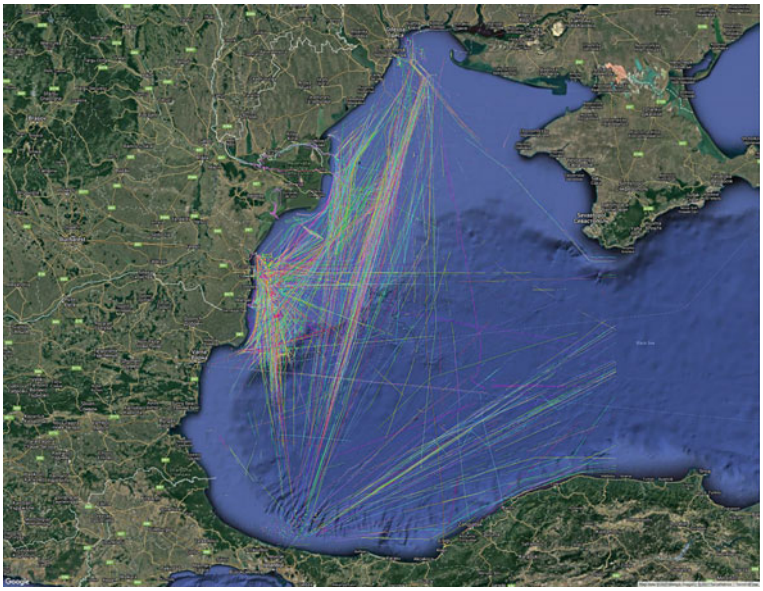

Fig. 12.2 Initial data set—routes visualization by MMSI

Fig. 12.3 Initial data set—density map

Fig. 12.4 Compressed data set—speed difference less than 0.1 knots

It is obvious from Fig. 12.4 that for the simplest application of our algorithm, which is expected to have the worst compression rate we observe that it seems that the visualization remains almost the same. As the density map depicts in Fig. 12.3, the data are so dense in the specific area—expected to be the case of any busy port—that omitting a large number of data does not seem to affect the data set.

We furthermore performed the visualization in extreme values of the parameters of the algorithm, which is altering the speed parameter limit to 0.5 knots and 1 knot. *We should note that with 1 knot speed a vessel can move 1.8 km within an hour*. It is obvious from Figs. 12.5 and 12.6 that the data loss is such that the result can be acceptable as long as we are able to predict and reproduce the omitted data, a case that is possible according to our algorithm definition.

Finally, we calculate the aggregated results of comparison between the initial data set and the compressed data. In this occasion, we decided to compare the differences that occur in situations of parameter limit for speed being below 0.2 knots.

It is obvious from Fig. 12.7 that within the area of the port, where speeds tend to be lower the data set is reduced significantly. In the area outside the port, the reduction is less with the application of the speed limit parameter, though in this area, the compression of data occurs from the algorithm part that creates differentiation records and prediction records.

Fig. 12.5 Visualization of the reduced data set for extreme cases|speed difference less than 0.5 knots

Fig. 12.6 Visualization of the reduced data set for extreme cases|speed difference less than 0.5 knots

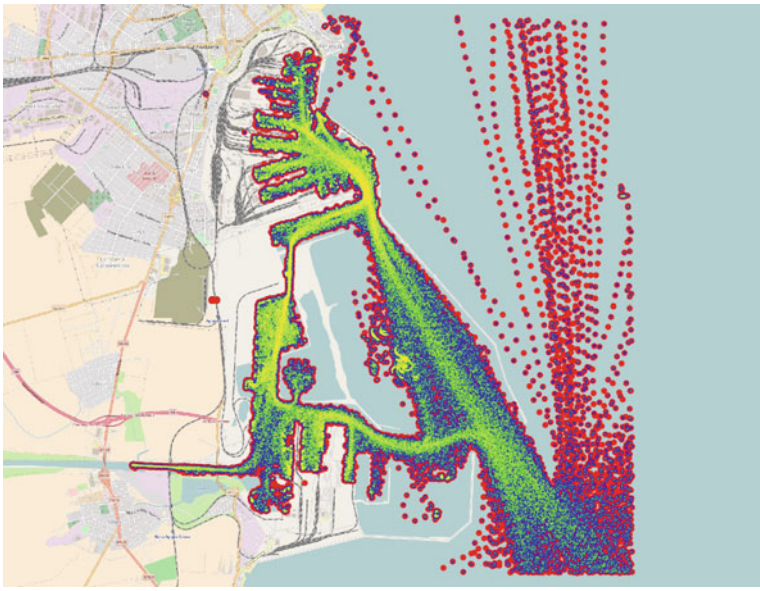

Fig. 12.7 Aggregated results—initial data set versus reduced data set (speed difference less than 0.2 knots)

12.6 Conclusion and Future Work

We described the Automatic Identification System (AIS) a system that is utilized in maritime traffic in order to provide a set of functionality including, among others, procedures that can help even special occasions including, collision avoidance and fleet monitoring. In this manner, we discussed the necessity of having real-time or near real-time application in order to be able to perform the set of procedures that the AIS protocol was designed for. In order to serve its scope, the protocol has strict regulations on the data and amount of data produced in the unit of time, which leads to generation of huge databases within a short period; thus making the analysis of the information a big data analysis problem.

In this scope, we presented a novel approach for significantly reducing the amount of data produced by AIS without losing the information that could be needed in order to perform real-time data analysis and actions required by it. Our algorithm is able to analyze data and create different kinds of records similarly to the video compression algorithms; that is creation of information records, differentiation records, and prediction records.

The experimental evaluation was performed on a large data set produced in a dense traffic area and more precisely in the port of Constanta located Romania (Black Sea). We divided our algorithm into two main occasions: places with very high density of data due to low speeds (usually ports) and areas with high density of data

due to standard routs. We presented how we can reduce the produced data in both occasions without losing the information. Our experiments prove that we can perform compression of at least 75% in both types of areas. We presented visualization of the initial and derived data in order to prove the information persistence after the compression.

As a future work, we plan to create a real-time service for analyzing in real time the data produced by AIS and provide a near real-time API that will be able to provide compressed AIS data. We will furthermore elaborate on the parameters of the algorithm in order to achieve more efficient levels of compression.

Acknowledgements This work has been partially supported by COST Action IC1302: Semantic keyword-based search on structured data sources (KEYSTONE); we particularly acknowledge the support of the grant COST-STSM-IC1302-36978: "Curating Data Analysis Workflows for Better Workflow Discovery".

References

1. What is the automatic identification system (AIS)? https://help.marinetraffic.com/hc/en-us/articles/204581828-What-is-the-Automatic-Identification-System-AIS
2. Winther M, Christensen JH, Plejdrup MS, Ravn ES, Eriksson ÓF, Kristensen HO (2014) Emission inventories for ships in the arctic based on satellite sampled AIS data. Atmos Environ 91:1–14, ISSN 1352-2310
3. Chen D, Wang X, Li Y, Lang J, Zhou Y, Guo X, Zhao Y (2017) High-spatiotemporal-resolution ship emission inventory of China based on AIS data in 2014. Sci Total Environ 609
4. Pallotta G, Vespe M, Bryan K (2013) Vessel pattern knowledge discovery from AIS data: a framework for anomaly detection and route prediction. Entropy 15(6):2218–2245
5. Hansen J, Jacobs G, Hsu L, Dykes J, Dastugue J, Allard R, Barron C, Lalejini D, Abramson M, Russell S et al (2011) Information domination: dynamically coupling METOC and INTEL for improved guidance for piracy interdiction. NRL Rev 2011:110–119
6. Natale F, Gibin M, Alessandrini A, Vespe M, Paulrud A (2015) Mapping fishing effort through AIS data. PLoS ONE 10(6):e0130746
7. de Souza EN, Boerder K, Matwin S, Worm B (2016) Correction: improving fishing pattern detection from satellite AIS using data mining and machine learning. PLOS ONE 11(9)
8. Greene M (1993) Radio frequency automatic identification system. U.S. Patent No. 5204681
9. Automatic identification system. https://en.wikipedia.org/wiki/Automatic_identification_system
10. Definition of the NMEA 0183 Standard. http://www.nmea.org/content/nmea_standards/nmea_0183_v_410.asp
11. Definition of the NMEA 2000 Standard. http://www.nmea.org/content/nmea_standards/nmea_2000_ed3_10.asp
12. ITU Recommendation M.1371, Technical characteristics for a universal shipborne automatic identification system using time division multiple access [ITU1371]
13. IALA Technical Clarifications on Recommendation ITU-R M.1371-1
14. IEC-PAS 61162-100, Maritime navigation and radio communication equipment and systems [IEC-PAS]
15. Harati-Mokhtari A et al (2007) Automatic identification system (AIS): data reliability and human error implications. J Navig 60(03):373–389