Automatic Identification System Data Quality: Outliers Detection Case

Sara El Mekkaoui and Abdelaziz Berrado
Equipe AMIPS, Ecole Mohammadia d'Ingénieurs,
Mohammed V University in Rabat, Morocco
saraelmekkaoui@research.emi.ac.ma, berrado@emi.ac.ma

Loubna Benabbou
Département Sciences de la Gestion
Université du Québec à Rimouski, Lévis
Qc, Canada
loubna_benabbou@uqar.ca

Abstract

The Automatic Identification System (AIS) a vessel tracking system. It provides rich information on vessel particulars in addition to dynamic navigational and voyage details. AIS data have significantly contributed in the digitization of the shipping industry, but still are prone to measurement and collection errors. As poor data quality leads to inaccurate analysis and affects decision making, a thorough preprocessing of AIS data is needed before any use. In this paper, we present the main quality issues encountered when dealing with AIS data. This concerns noise, outliers, duplicates, inconsistent data, and out of range values. We also provide some errors examples and how to overcome them. As an application, we address the problem of outliers' detection in an unsupervised way using clustering and anomaly detection techniques, which attribute an anomaly score for each observation. The case study shows promising results for spatial outliers' detection, which can be further explored for other anomalies detection tasks.

Keywords

Maritime intelligence, Automatic Identification System, Data quality, Outliers detection, Unsupervised learning.

1. Introduction:

In this paper we investigate issues related to data provided by the Automatic Identification System (AIS) which is a vessel tracking system. AIS provides information about vessels locations, as well as their static, dynamic, and voyage details (Bole et al. 2014). Although, AIS is meant to be a safety and security tool, the system provides useful information on vessels motions for maritime intelligence. AIS data have been well explored in studying trajectories extraction and estimation, and anomaly detection to deal with safety and security issues especially in busy areas near ports, straights and inland waterways (Svanberg et al. 2019). AIS data have also contributed to improve vessels routing and scheduling and ports logistics analysis and forecasting.

In spite of the important contribution of AIS data in supporting shipping intelligence, they have some serious limitations (Harati-Mokhtari et al. 2007). AIS data can be unreliable because of failures ascribed to human error, equipment malfunction, programming mistakes or wrong installation and configuration. The fails were observed in almost all attributes. Wrong vessels identification numbers, confusing vessel types, inexact navigational status, unsound vessels length and beam, erroneous positions, incorrect draughts, misleading destination and ETA, as well as missing values and spoofing problems. Therefore, particular attention should be given to data cleaning and preprocessing before undertaking any analysis.

In real-world applications, data is often imperfect and their quality may have significant impact on the performance of data mining and learning algorithms. Quality problems are emphasized especially when dealing with extended geographical areas for applications such as trade routes detection or prediction of real-time vessels arrival times. In this paper we highlight AIS data issues falling into popular data quality problems including duplicates, noise, outliers, missing values, and inconsistent entries (Emmens et al. 2021). Understanding and improving AIS data quality typically

improves the decision making as AIS data quality and reliability issues can affect the performance of Machine Learning models.

In this paper, we focus on the case of spatial outliers. Outliers are either rare observations with different characteristics from the other observations, or unusual feature values with regards to the typical values for one feature. Outliers can be anomalous instances due to error entries or legitimate data points that have to be detected and studied (Tan et al. 2018). In AIS data, spatial outliers happen when the vessel position (latitude and longitude) is within the normal range but does not reflect the actual vessel position. Our goal is to find unusual spatial points from among a large number of normal ones. For that we will use model free techniques which do not explicitly characterize the distribution of the normal or anomalous class. More specifically, we use clustering and anomaly detection techniques to produce anomaly scores for each observation. We report the results of different techniques on an AIS dataset of vessels positions calling a particular port.

2. Automatic Identification System:

2.1. Overview of the Automatic Identification System:

AIS is based on the Global Positioning System (GPS) technology and broadcasts vessels location and information using a radio channel. By that, it is possible to detect ships all over the world, as long as radio signals could be exchanged. Information can be exchanged between receiving stations onboard of ships, along coasts and by satellite. Therefore, any internet user with a Very High Frequency (VHF) antenna and an AIS receiver onshore can contribute to a global AIS network. This is how AIS community-based networks work. A vessel onboard transceiver can have a coverage ranging between 15 to 20 nautical miles, onshore stations have a range of 40 nautical miles, and satellites can cover up to 1000 nautical miles (Tu et al. 2017). AIS data is globally used by harbors and port authorities through Vessel Traffic Services (VTS), which is a real-time control system similar to air traffic control, to improve the safety and efficiency of vessel traffic in related waters. The increasing relevance of AIS data in addressing many problems has motivated many commercial companies to build a large global AIS network. The network is usually composed of coastal receivers enhanced by satellites for a better coverage. AIS messages are received and transmitted to a central server where they are processed and provided to end-users in different forms (Bole et al. 2014). The offer includes some free services such as live vessel positions maps, last received position, photos and weather information. Other reports as historical data, port calls, and predictive ETAs are provided with charge. A selection of AIS data providers can be found in the work of Tu et la. (2017).

AIS data require several processing phases. The information is first received in the National Marine Electronics Association (NMEA) format, a standard for communication between wired electronic ships devices, which is a text encoded in binary format as shown in Figure 1. AIS information are then decoded and presented in many simple exploitable formats like CSV, XML, and geodatabase files. Messages could be decoded using for instance, the python libais library (Schwehr 2018).

```
!AIVDM,1,1,,A,13HOI:0P0000VOHLCnHQKwvL05Ip,0*23
!AIVDM,1,1,,A,133sVfPP00PD>hRMDH@jNOvN20S8,0*7F
!AIVDM,1,1,,B,100h00PP0@PHFV`Mg5gTH?vNPUIp,0*3B
!AIVDM,1,1,,B,13eaJF0P00Qd388Eew6aagvH85Ip,0*45
!AIVDM,1,1,,A,14eGrSPP00ncMJTO5C6aBwvP2D0?,0*7A
```

Figure 1. A sample of raw AIS data.

2.2. AIS data attributes:

AIS messages are classified into four categories which are static, dynamic, voyage, and safety messages (Bole et al. 2014) as summarized by Table 1. Static information are about vessel characteristics such as vessel identification number, length, beam, and ship type. Voyage related data include details about current voyage such as draught, Estimated Time of Arrival (ETA) and destination. Static and voyage data are transmitted each 6 minutes or when requested. Dynamic messages report about vessel movements including position (latitude and longitude), Course Over Ground (COG), and Speed Over Ground (SOG). Their transmission intervals depend on factors such as speed and course, the interval varies between 2 and 10 seconds when vessel is moving, and every 3 minutes when vessel at anchor. Finally, the safety text message is transmitted as required. The different types of messages are transmitted in different time slots, their association is performed by the Maritime Mobile Service Identification (MMSI) number.

Table 1. AIS data attributes details.

Category	Item	Description	Units
Static	MMSI	Maritime Mobile Service Identity	-
	Vessel name	Maximum of 20 characters	-
	Call sign	Unique designation used in radio communications	-
	IMO	International Maritime Organization vessel number	-
	Length	Length of vessel	meters
	Beam	Width of vessel	meters
	Type of vessel	Type of vessel (e.g. cargo, tanker, tug)	-
Dynamic	Timestamp	Position timestamp in Coordinated Universal Time (UTC)	-
	Longitude	Position longitude	decimal degrees
	Latitude	Position latitude	decimal degrees
	COG	Course Over Ground	degrees
	SOG	Speed Over Ground	knots
	Heading	True heading angle	degrees
	ROT	The rate the ship is turning	degrees
	Navigation status	Navigational Status (e.g. underway by engines, at anchor)	-
Voyage	Cargo	Hazardous cargo type	-
	Draft	Current draft depth of the vessel	meters
	Destination	Port of destination	-
	ETA	Estimated Time of Arrival at the port of destination	-

2.3. AIS data providers:

AIS data plays an important role in maritime intelligence with a wide range of applications (Svanberg et al. 2019). However, getting to the right source of data can be time consuming. Depending on the application scope, different commercial and non-commercial sources can be used. A survey of AIS data sources with an assessment of their quality was provided by Tu et al. (2017) and Stróżyna et al. (2018). Choosing the right source can be difficult as information completeness differs from a source to another (Tu et al. 2017).

In order to avoid missing important attributes or wasting time during data acquisition, we recommend to follow a methodology for selecting and assessing the quality of data sources. Such methodology can be found in Stróżyna et al. (2018), and also can be adapted or developed depending on the considered application. A first step is to learn about AIS data attributes, understand them and define the variables needed for the study. The second step is to get acquainted with the different AIS sources and the information they can provide. A third and final step would be to assess the quality of the sources and choose the one that satisfies the study need or combine different sources to get the required data.

Also, one important element to consider while searching for AIS data is data cost. AIS data can be provided by non-commercial or commercial sources. Non-commercial sources usually provide free regional coverage of AIS data such as the U.S. Coast Guard and the Australian Maritime Safety Authority. Other private and commercial AIS data sources which relies on terrestrial receivers and satellites provide data with global coverage but in most cases are charged (e.g. MarineTraffic, VesselFinder and FleetMon).

3. AIS Data issues and challenges:

In spite of the important contribution of AIS data in supporting shipping analysis, they have some serious limitations. AIS data can be unreliable (Harati-Mokhtari et al. 2007) because of human errors, equipment malfunctions, programming mistakes, and wrong installation and configuration. AIS data transmission can also be altered by external factors such as severe weather (Emmens et al. 2021). The fails were observed in almost all attributes. Wrong identification numbers, confusing vessel types, inexact navigational status, implausible vessel lengths and beams, erroneous positions, incorrect draughts, misleading destinations and ETAs, along with missing values and spoofing problems. A single attribute can be affected by many issues, for instance, ETAs can be missing, incorrect or outdated. As ETAs are manual information in AIS, vessel crew may omit refreshing it. Therefore, further work on pre-processing AIS data is needed.

AIS data corruption can also be caused by packet collision especially in the case of satellite receivers (Poļevskis et al. 2012). Actually, AIS was initially developed for local situation awareness to exchange data between ships and with onshore stations. In order to allow a global coverage of AIS, satellites are used to improve the transmission range. However, when the satellite receives more than two messages at the same time, it results in incorrect interpretation of received messages leading to issues such as erroneous positions or speeds.

AIS data quality and reliability issues can affect the performance of data mining and learning algorithms. In the following subsections, we highlight the main issues that should be investigated in the AIS data context to improve data quality.

3.1. Noise:

Real world data are prone to measurement and data collection errors. AIS rely on many sensors in addition to manual entries which make noise an unavoidable problem. Noise can be defined as the modification of the original value. As an example we can find random errors of positions in vessels tracks (Zhao et al. 2018), speed from AIS not matching the speed of the vessel, and unstable values of speed and course from AIS sensor data (Emmens et al. 2021). Removing noise is usually difficult, yet it can be addressed using robust Machine Learning algorithms for efficient results in presence of noise.

3.2. Outliers:

Outliers are data points that differs significantly from the other observations, but, unlike noise they can be legitimate observations that should be detected and studied (Tan et al. 2018). The volume of outliers is usually very small compared to the dataset, and failing to remove them can harm the performance of many Machine Learning based models. For instance, spatial outliers can occur in AIS data (Zhao et al. 2021) when the GPS position does not reflect the real location of the vessel. However, assessing position quality can be difficult, as spatial outliers hide within normal positions. To tackle this issue, anomaly detection methods can be used to detect outliers using anomaly scores (Tan et al. 2018).

3.3. Missing values:

In AIS data, missing values is a recurring problem that touches any attribute. For example, missing messages can result in discontinuous trajectories or large geographical coverage gaps when the time interval between consecutive observations is very large. An example of incomplete trajectory is given in Figure 2. In this case, interpolation methods can be used to complete missing trajectories (Mao et al. 2016).

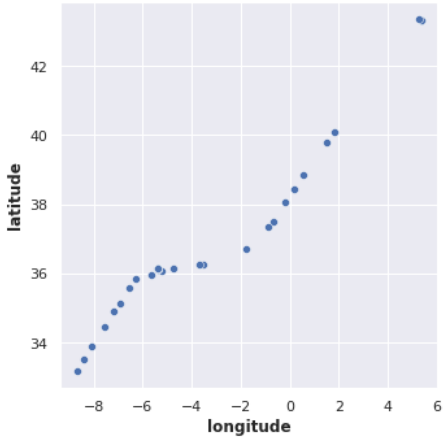

Figure 2. Illustration of missing observations in a vessel trajectory.

Also, static information from AIS suffers from this issue and can be filled using a database of vessel particulars to complete missing information related to vessel beam, length or type. Other dynamical information such as navigational

status, rate of turn and draft may also be missing. They can be filled with default values, last or next available values in a trajectory, or interpolated.

3.4. Duplicate data:

AIS data may include redundant observations especially for position messages obtained from terrestrial receivers. This situation occurs when a vessel is in the range of more than one receiver. Duplicates also may appear in the attributes level. For example different vessels may have the same MMSI. In this case we can observe different tracks of the same vessel in contrasting areas (Zhao et al. 2018). This phenomena can be caused unintentionally by a configuration error or equipment malfunction. However, it happens that the MMSI or other information get intentionally modified to mislead AIS data receivers, which is called spoofing.

3.5. Inconsistent values:

In AIS messages, information that are specific to the vessel should remain consistent. For example, a change in vessel type, size or cargo type throughout the same voyage should not happen (Emmens et al. 2021). Also, the vessel beam and length should go together with draught values. Another issue is the destination text format that can take any structure. A vessel can use as destination the name of the port, the code of the port, the port's country, the port of origin and destination, unknown abbreviation, or leave it blank. In our case, we have found about 200 different formulations of the destination for a single port.

3.6. Out of Range Values:

Out of range values may appear in many AIS data attributes and can be simply handled using filters. For instance, latitude of more than 90° or longitude of more than 180° (Harati-Mokhtari et al. 2007) can be easily detected. Another example is negative or very large values of speed, for instance it generally doesn't exceed 15 knots for a bulk vessel.

In Table 2 we can see an example of speed jumps with out of range values. The table represents three consecutive observations for two different bulk vessels. For both of them, there is a significant change in the speed reaching unrealistic values. Speed jumps can be detected and corrected by evaluating distances between consecutive observations and setting a threshold (Mao et al. 2018). Another example concerns vessels course and heading values that should be comprised between 0° and 359°. A course value of 511 means not available and a heading value of 360 means unknown. However, values between 360 and 511 may show up even though they are not allowed.

Table 2: Example of speed errors.

Vessel ID	Position Timestamp (UTC)	Latitude (°)	Longitude (°)	SOG
01	2020-05-21 00:28:52	33.14	-8.70	5.7
01	2020-05-21 12:14:51	33.13	-8.63	**102.2**
01	2020-05-21 17:18:01	33.13	-8.63	0.0
02	2020-08-07 11:34:40	53.97	4.25	11.4
02	2020-08-07 12:01:19	53.92	4.14	**51.2**
02	2020-08-07 15:01:20	53.48	3.54	11.5

4. Case study of outliers' detection:

4.1. Problem description:

Outliers are rare and distinct observations different from the other observations. Outliers can be hard to define and identify as they depend on each dataset specifications. Outliers' detection is an important step in cleaning real datasets. The objective is to separate regular observations or inliers from different observations or outliers. However, in real world applications we don't have a clean dataset of only regular observations to use for learning a representation of the normal class and predict either new data points are normal or outliers, such as in the one-class classification task.

AIS data tracks usually show deviations from expected vessels trajectories. When plotting vessels positions of AIS datasets we can see abnormal tracks, deviating points, overland positions and incomplete trajectories (Emmens et al. 2021). For position drift in a certain geographical range, Mieczynska and Czarnowski, (2021) proposed to use K-mean clustering for signal reconstruction to assign damaged observations to related cluster, i.e. trajectory. Also, Zhao et al. (2018) provided an algorithm to improve the quality of spatial data and time information. In our case, we deal with a dataset with extended geographical area for applications such as trade routes definition (Spiliopoulos et al. 2017) and

the prediction of vessels arrival times (Kim et al. 2017). In this case, the deviation of vessels positions may be very important.

In our case, the AIS dataset represents over 1400 trajectories of vessels calling a particular port, with more than 78 000 observations (or vessels' positions). The attributes include AIS information such as speed over ground, course over ground and true heading. We have manually inspected the dataset and classified each point as normal or abnormal, following the correctness of the geographical position. For instance, vessels that were supposed to be anchored at the studied port with far positions from the port were classified as abnormal points. This classification will serve to assess the performance of the outliers' detection methods. Figure 3 is a visualization of vessels trajectories representing normal data points by 1 and outliers by -1. A visual inspection reveals that vessels lying far away from the rest of the observations might have erroneous GPS positions.

In the shipping context, visual inspection of vessels positions can help identify scattered outliers but it is hard to detect anomalies surrounded by normal points. In this section, we test some outliers' detection methods on our dataset and compare their performance.

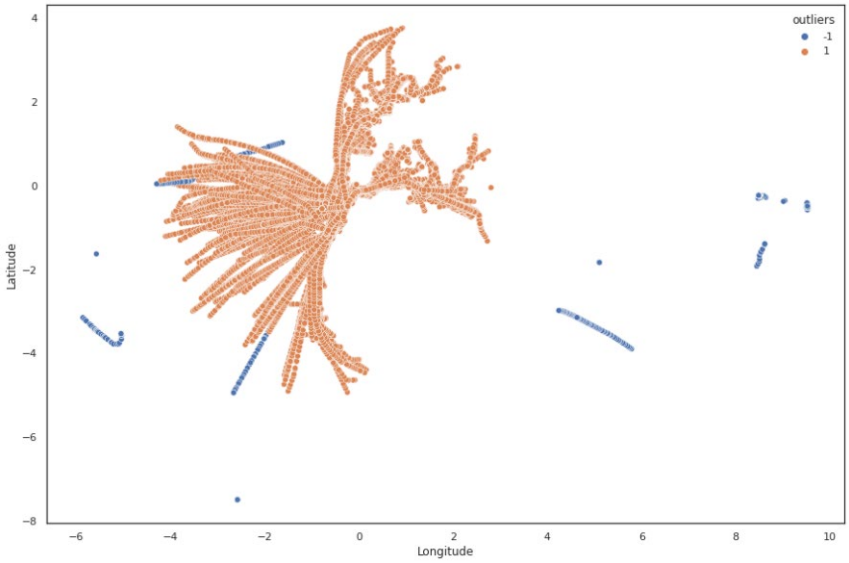

Figure 3. Visualization of vessels positions.

4.2. Outliers' detection methods:

4.2.1. DBSCAN:

Outliers are different from normal data points. If we suppose that outliers lie far away from normal data and don't form dense clusters, we can detect them using a clustering technique. In this paper we propose to use the Density-Based Spatial Clustering with Applications with Noise (DBSCAN) (Tan et al. 2018). DBSCAN locates regions of high density that are separated from one another by regions of low density. It is based on a center approach which estimates density for a particular point based on the number of points comprised within a radius of that point. Using this approach, we can classify a point as being a core, border or noise point. Core points are points onside a dense region, border points are on the extremity of a dense region, and noise points are in sparsely occupied regions. The DBSCNA algorithm puts any two core points that are close enough in the same cluster, and any border point to the cluster of the closest core point.

4.2.2. Isolation Forest:

Using density or distance based methods for outliers' detection suppose that normal points occur in dense regions or are close to their neighbors. Isolation Forest or iForest (Liu et al. 2008, 2012) is a tree-based anomaly detection algorithm based on the isolation of anomalies without using any measure of density or distance. The isolation of anomalous observations from the normal ones uses an anomaly score. It follows two stages of, first building isolation

trees using subsamples of the data and second calculating an anomaly score for each test instance by passing through isolation trees.

4.2.3. One class Support Vector Machine:

The One class Support Vector Machine (SVM) (Schölkopf et al. 1999, 2001) was proposed for novelty detection. It is an extension of Support Vector Machines to the case of unlabeled data. The One Class SVM tries to find a binary function that would be nonzero in regions with the most of the data points. In practice, it defines a frontier using a kernel and a scalar parameter known as the margin which corresponds to the probability of finding new regular data points outside the frontier.

4.2.4. Local Outlier Factor:

The Local Outlier Factor (LOF) (Breunig et al. 2000) determines a score of abnormality for each observation. It is based on the measure of the density deviation of an observation with regard to its neighbors. Based on a local density measure with regards to k-nearest neighbors, the LOF score is obtained for each observation by computing the ratio of the average local densities of the k-neighbors and the observation local density.

4.3. Experimental results:

In our case, we have 761 data points identified as outliers representing about 1% of the dataset. For each observation, we define the positive class as being an inlier and the negative class as outlier. We compare the performance of the pre-mentioned outliers' detection methods using different metrics. In Table 3, we report the number of True Positives (TP) and False Positives (FP), the True Positives Rate (TPR), False Positives Rate. TP are outliers that were correctly detected and FP are inliers that were misidentified. TPE and FPR reflects how many outliers were found and how many misleading points were returned. The objective is to increase the TPE and decrease the FPR so that we avoid wasting time in inspecting misleading false outliers and we have enough detected outliers.

In Figure 4, we provide the visualization of the results. On the left side we plotted the outliers together with the anomaly scores of the different detection methods. On the right side we did the same for the normal data points. We can see that the One Class SVM performs well for detecting about 89% of the outliers. However this performance is interfered by the important number of false positives that can be hard to inspect. LOF also suffers from the same issue. DCBSCAN provides a better performance for detecting outliers than the isolation Forest, the latter however offers a low number of false positives.

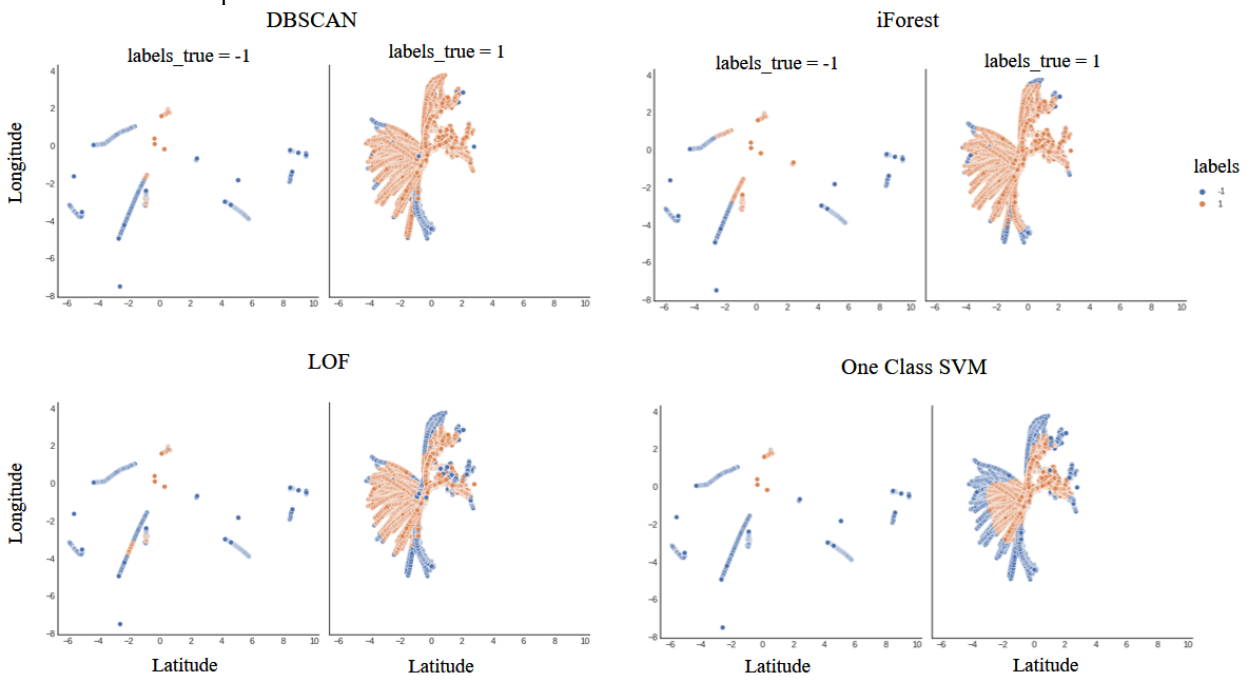

Figure 4. Visualization of the performance of the outliers' detection methods.

Table 3. Performance of the outliers' detection methods.

Method	TP	FP	TPR	FPR
DBSCAN	630	1276	82.79	1.64
Isolation Forest	515	273	67.67	0.35
One-Class SVM	676	7184	88.83	9.21
LOF	645	7685	84.76	9.85

5. Conclusion:

AIS data represent an important element in maritime intelligence. To get full insight from the data and use it properly, AIS data should be first prepressed before performing any analysis. Otherwise, AIS data quality and reliability issues may harm the performance of many data mining or Machine Learning techniques. An appropriate use of AIS data suppose that we have learned about their attributes and understood their characteristics. Depending on the problem we want to solve, we should define the needed variables and target the right data sources to avoid wasting time in data sourcing.

AIS data veracity is an important challenge as vessels are entirely responsible for the quality of transmitted messages. Any fail in AIS equipment, error in the implementation or manual entry can lead to false and incorrect data. Hence, when analyzed, AIS data should be preprocessed to deal with issues such as noisy, missing, and uncertain information. In this paper we covered the main AIS data issues and provided some examples on how to overcome them. AIS data errors can be induced by technical or human factors and can touch all the attributes. For this reason, a rigorous checking and preprocessing of the data is paramount. However, AIS data volume and velocity can make the preprocessing step very challenging, as they are characterized by high transmission frequency and very large amount of information.

Finally, we proposed a case study of spatial outliers' detection in an unsupervised way. This happens when the position transmitted by AIS does not reflect the exact position of the vessel. In our case, we deal with an extended geographic area and we are interested in detecting false positions with a significant gap from the real ones. Using an AIS dataset of vessels calling a particular port, we applied clustering and anomaly detection techniques to find erroneous vessel positions. The study shows promising results for this outliers' detection task and could be further investigated for issues related to detecting false, falsified or spoofed AIS information.

Acknowledgements

The authors would like to thank the Port Authority (Agence Nationale des Ports) of Jorf Lasfar for their support and guidance.

References:

Bole, A., Wall, A., and Norris, A., Chapter 5 - Automatic Identification System (AIS), *Radar and ARPA Manual*, 3rd Edition, Elsevier, pp. 255–275, 2014.

Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J., LOF: identifying density-based local outliers, *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93-104, 2000.

Emmens, T., Amrit, C., Abdi, A., and Ghosh, M., The promises and perils of Automatic Identification System data, *Expert Systems with Applications,* vol. 178, pp. 114975, 2021.

Harati-Mokhtari, A., Wall, A., Brooks, P., and Wang, J., Automatic Identification System (AIS): Data Reliability and Human Error Implications, *Journal of Navigation*, vol. 60, no. 3, pp. 373-389, 2007.

Kim, S., Kim, H., and Park, Y., Early detection of vessel delays using combined historical and real-time information, *Journal of the operational research society,* vol. 68, no. 2, pp. 182-191, 2017.

Liu, F.T., Ting, K.M., and Zhou, Z.H., Isolation forest, *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 413-422, 2008.

Liu, F.T., Ting, K.M., and Zhou, Z.H., Isolation-based anomaly detection, *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1-39, 2012.

Mao S., Tu E., Zhang G., Rachmawati L., Rajabally E., and Huang G.B., An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining, Proceedings of the ELM-2016, vol. 9, Springer, Cham, 2018.

Mieczyńska, M., and Czarnowski, I., K-means clustering for SAT-AIS data analysis, *WMU Journal of Maritime Affairs*, vol. 20, no. 3, pp. 377-400, 2021.

Osekowska, E., Henric J., and Bengt C., Maritime vessel traffic modeling in the context of concept drift, *Transportation research procedia,* vol. 25, pp. 1457-1476, 2017.

Poļevskis, J., Krastiņš, M., Korāts, G., Skorodumovs, A. and Trokšs, J., Methods for Processing and Interpretation of AIS Signals Corrupted by Noise and Packet Collisions, *Latvian Journal of Physics & Technical Sciences*, vol. 49, no. 3, 2012.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C., Estimating the support of a high-dimensional distribution, *Neural computation,* vol. 13, no. 7, pp. 1443-1471, 2001.

Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., and Platt, J.C., Support vector method for novelty detection, In S. Solla and T. Leen and K. Müller (eds.), *Advances in Neural Information Processing Systems*, vol. 12, pp. 582-588, 2000.

Schwehr, K., Library for decoding maritime Automatic Identification System messages, Available: https://github.com/schwehr/libais, October 30, 2021.

Spiliopoulos, G., Chatzikokolakis, K., Zissis, D., Biliri, E., Papaspyros, D., Tsapelas, G., and Mouzakitis, S., Knowledge extraction from maritime spatiotemporal data: An evaluation of clustering algorithms on Big Data, *Proceedings of the 2017 IEEE International Conference on Big Data*, pp. 1682-1687, 2017.

Stróżyna, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., and Węcel. K., A framework for the quality-based selection and retrieval of open data-a use case from the maritime domain, *Electronic Markets*, vol. 28, no. 2, pp. 219-233, 2018.

Svanberg, M., Santén, V., Hörteborn, A., Holm, H., and Finnsgård. C., AIS in maritime research, *Marine Policy*, vol. 106, pp. 103520, 2019.

Tan, P.N., Steinbach, M., Anuj, K., and Kumar, V. *Introduction to data mining*, 2nd edition, Pearson, 2018.

Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., and Huang, G.B., Exploiting AIS data for intelligent maritime navigation: A comprehensive survey from data to methodology, *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. pp. 1559-1582, 2017.

Zhao, L., Shi, G., and Yang, J., Ship trajectories pre-processing based on AIS data, *The Journal of Navigation*, vol. 71, no. 5, pp. 1210-1230, 2018.

Biographies

Sara El Mekkaoui is a Ph.D. student in the Department of Industrial Engineering at Ecole Mohammadia D'ingénieurs (EMI), Rabat, Morocco. She received her engineering degree in Industrial Engineering from EMI School of Engineering in 2011. She has an experience of more than four years working as Purchasing & Procurement Engineer in sugar industry and two years as Scheduling Senior Analyst in shipping & port logistics. Her research focuses on using Machine Learning for Maritime Logistics.

Dr. Loubna BENABBOU is a Professor of Management Sciences at Université du Québec à Rimouski (UQAR) at Lévis campus. Her research work lies in the application of decision/ management sciences and machine learning techniques to transform data for making better decisions and improving operational processes. Dr. Benabbou has been supervising several undergraduate and graduate students in projects for different Industries related to the areas of Decision Sciences, Machine Learning and Operations Management. Her research related to these fields has been published in international scientific journals and conferences' proceedings. Dr. Benabbou is an industrial engineer from EMI School of Engineering; she earned an MBA and Ph.D. in Management and Decision Sciences from Laval University.

Abdelaziz BERRADO, Ph.D. is a Professor of Industrial Engineering in EMI School of Engineering at Mohamed V University in Rabat. He holds degrees in Decision Systems and Industrial Engineering. He is interested in the areas of d Machine Learning, Industrial Statistics, Operations and Supply Chain Modelling, Planning and Control with applications in healthcare and other industries. He published several papers in research journals and conferences with local and international funding. He is a fellow of IEOM society and a member of INFORMS and IEEE. Previously, he was also a senior engineer at Intel.