ALI AKBAR SAFAEI, Ph.D. Candidate[1]
E-mail: alisafaei123@aut.ac.ir
HASSAN GHASSEMI, Ph.D.[1]
(Corresponding Author)
E-mail: gasemi@aut.ac.ir
MAHMOUD GHIASI, Ph.D.[1]
E-mail: mghiasi@aut.ac.ir
[1] Maritime Engineering Department
Amirkabir University of Technology
Tehran, Iran

Traffic Engineering
Preliminary Communication
Submitted: 4 Mar. 2018
Accepted: 4 Apr. 2019

METHODOLOGY OF ACQUIRING VALID DATA BY COMBINING OIL TANKERS' NOON REPORT AND AUTOMATIC IDENTIFICATION SYSTEM SATELLITE DATA

ABSTRACT

Fuel consumption of marine vessels plays an important role in both generating air pollution and ship operational expenses where the global environmental concerns toward air pollution and economics of shipping operation are being increased. In order to optimize ship fuel consumption, the fuel consumption prediction for her envisaged voyage is to be known. To predict fuel consumption of a ship, noon report (NR) data are available source to be analysed by different techniques. Because of the possible human error attributed to the method of NR data collection, it involves risk of possible inaccuracy. Therefore, in this study, to acquire pure valid data, the NR raw data of two very large crude carriers (VLCCs) composed with their respective Automatic Identification System (AIS) satellite data. Then, well-known models i.e. K-Mean, Self-Organizing Map (SOM), Outlier Score Base (OSB) and Histogram of Outlier Score Base (HSOB) methods are applied to the collected tankers NR during a year. The new enriched data derived are compared to the raw NR to distinguish the most fitted methodology of accruing pure valid data. Expected value and root mean square methods are applied to evaluate the accuracy of the methodologies. It is concluded that measured expected value and root mean square for HOSB are indicating high coherence with the harmony of the primary NR data.

KEY WORDS

marine transport; voyage; noon report;
Automatic Identification System; fuel consumption;

1. INTRODUCTION

Nowadays, world widespread concerns on air pollution and energy efficiency persuade scientists to put a lot of efforts into reducing the maritime engine fuel consumption of ships. In this regard, to establish an energy prediction model for the forthcoming voyage, the determination of relationship between fuel consumption in different sea states and weather conditions, as external factors, along with internal factors such as ship speed and displacement, have remained a challenging topic. The most popular approaches to achieve this challenging goal are deployment of suitable mathematical models. In order to use any mathematical methods, the input statistical NR data need to qualify for further fuel consumption analysis and prediction. In this study combining the NR with AIS data is proposed. This is because the NR data collected by ship staff on a daily basis [1] involves risk of human error such as fatigue, reduced vision, written errors, etc. Consequently, gathering and generating high quality data remain crucial. The objective of this paper is to develop a new methodology in which NR data become pure and valid relying on predicting the ship forthcoming fuel consumption rate.

This study consists of mathematical basis steps, in order to acquire suitable and valid NR data, to establish an accurate relation between the tanker fuel consumption and independent variables such as vessel speed and displacement. In the first step, the NR data of two VLCCs are composed with their respective AIS data. In the following step the out-ranged data are determined by different methods i.e. K-Mean, SOM, OSB and HOSB. Furthermore, the data are treated by eliminating the existing out-ranged data or by being replaced by new generated in-ranged data. For the purpose of validation or error estimation, the expected value and root mean square methods are implemented.

In literature, the sailing speed of a ship as an important independent variable constitutes the main factor in fuel consumption. Fuel consumption and emissions on a shipping route are typically a cubic function of speed [2]. Also, in experimental studies, the relation of fuel consumption and sailing speed shows an exponent of 3.5 in equation for small container vessels [3]. Additionally, for the ship with a sailing speed of less than 20 knots, the fuel consumption relation is in order of 2.7 to 3.3 in exponent function [4]. Meanwhile, by increasing the sailing speed in excess of 20 knots, fuel consumption relation arises in the order of exponent function to four and more [5]. Moreover, the sailing speed optimization problem for a ship operating on a route having a specified sequence of calling ports with time windows for calling time is then addressed. [6]. Furthermore, a regional voyage case study aiming at optimization of a VLCC shows the route of vessel due to different weather conditions can change consumed fuel due to change in wave height and wind direction [7].

The voyage time has direct linear correlation to the ship speed and by increasing speed the voyage time will be decreased. Consequently, the vessel can acquire a better efficiency score in the net amount of the carried volume cargo due to the voyage time. It means that the ship owner can transport more volume cargo annually. But the efficiency of vessel operational expenses (OPEX) such as fuel consumption is a big dilemma due to its weight factor among ship operational costs items. Accordingly, optimization of the speed of a ship has direct correlation with the elements of the shipping line network including: ship routing and scheduling, service frequency, number of vessels and capacity of the fleet, selection of the appropriate ships for each trip and cargo planning [8]. In this regard, a model for the running costs of the ship with a view to analyse the relation between fuel price, ship sailing speed, voyage frequency and the number of ships employed has been proposed [9]. The following issue to guarantee timely arrival in a destination, is finding the most suited, safe and optimized path which might not be the shortest one [10].

Having in mind the paramount importance of the ship speed and its optimization in respect of fuel consumption, the effect of the weather and maritime environment condition should not be ignored [11]. Another study demonstrated the parameters that have influences on ship route decision, e.g. environmental forces, ship configurations and operational conditions emphasizing the effect of weather condition on fuel consumption [12]. In reality, most of the extant studies in connection with the effect of the weather condition on the fuel consumption focus mainly on ship routing. The importance of routing selection and its problems led to studying the drawbacks of the existing methods of selecting routes, i.e. plotting courses in maritime navigation, and giving recommendations of how to improve them [13].

Data mining in a general sense means discovering the underlying relation between various data. Advancement in information technology provided new sources of information to humankind and the elapse of the time added to their complications. Accordingly, there emerged the need for analysing the data [14]. Therefore, detection of errors and outliers has become one of the important issues facing data mining. Outlier is data with great divergence with the extant data giving rise to doubt that it might be recorded or generated in a different method or unusual mechanism [15]. The detection of outlier was studied in the early 19th century in a statistical population and the related techniques were gradually developed. Some of these techniques have particular application and some of them are general techniques. Outlier data mining points to a problem of finding the unusual models in a large set of data which do not match the existing models [16]. In other words, the NR database which is collected manually by ship staff, might involve a set of wrong data as a result of handmade method of recording. In mathematics the out-ranged data named as noise are identifiable [17]. A simple method to recognize the out-ranged data is to calculate the mean and variance of all data. Afterwards, the maximum allowable distance to the mean would be evaluated. Then, the false outliner records would appear and they would be detected using different statistical models. Then, a new qualified NR database is available for further study [18, 19].

2. CHARACTERISTICS OF SELECTED VESSELS

Fuel consumption monitoring results are occasionally misleading and can lead to questionable judgments being made by industrial specialists on the real fuel efficiency of ships [11]. In *Figure 1* the fuel consumption rate (Mt/Month) recorded in NR for two VLCCs (labelled as "Ship D1" and "Ship D2", two sister vessels) are plotted from 1 January 2016 to 31 December 2016. Without further consideration, it seems that Ship D2 is more efficient compared to Ship D1 in fuel consumption rate at first glance. Nevertheless, taking into consideration the result of carefully analysing ship and sea conditions during the past voyages done by the abovementioned ships indicate that when the sea waves encountered by the ships were higher than 4 metres, the fuel consumption of Ship D1 was often less than of Ship D2 providing that both ships

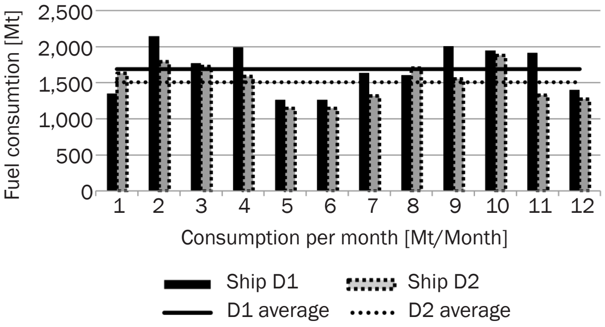


Figure 1 – Fuel consumption rate comparison of two sister ships, 320,000 Dead Weight Tanker

were sailing at the same speed. As a result, although according to fuel efficiency index Ship D1 consumes more, she encountered severe weather conditions during the past voyages. In other words, Ship D1 is more fuel-efficient than Ship D2.

In this study, two VLCCs of D Class order which is called D1 and D2 herein, have been selected. *Table 1* shows the characteristics, name and dimension of the ships. Other data from the given tankers that have been used in this investigation i.e. speed, power of engine, and fuel consumption were collected from NR and AIS data of a reliable oil tanker company. In addition, AIS database is collected by one of the famous members of the International Association of Classification Societies (IACS) named DNV-GL.

Tables 2 and *3* show the reviewed parameters of NR and AIS data sample, respectively. As indicated in *Table 2*, Fuel Consumption Rate (FCR), speed, distance, wave, current and wind force are reported. Furthermore, *Table 3* shows the real position of vessels, ground speed and draft of the ships.

Table 1 – Basic information of NR data of two D Class oil tankers

Ship	D1	D2
Capacity	317,519	317,519
Log duration	2016	2016
Type	VLCC	VLCC
Year of construction	2013	2013
Length overall	332.95	320.00
Length B.P.	332.95	320.00
Width MLD	60.00	60.00
Depth MLD	30.50	30.50
Draft MLD	22.60	22.60
Max power	16,870	16,869
Max RPM	74	74
Main engine consumption (P/D)	91	91

3. COMBINATION OF AIS WITH NR

In this section, integrating AIS reports with NR datum to increase the quality and accuracy of NR data by replacing AIS satellite data is explained. AIS data are reported by Global Positioning System (GPS) installed on the vessels on a daily basis. Because of the lack of accuracy and vast interval of the NR data gathered from the two tankers, the reported speed data from AIS have been deployed in this study to enrich the average speed in NR. For creating one speed for each day, a simple formula of averaging was used:

$$Average\ speed = \frac{\sum_{i=1}^{n} All\ recorded\ speed\ in\ one\ day}{n} \quad (1)$$

where n is the number of recorded speeds per day.

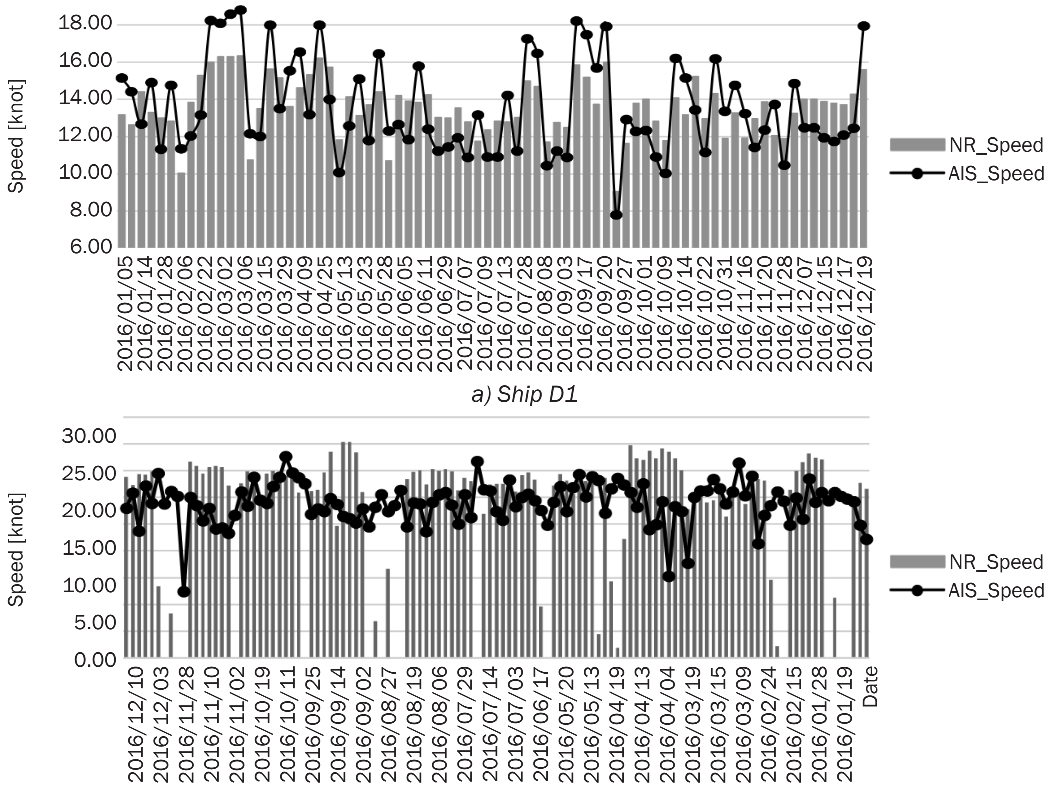
In addition, according to the author's experience as a CEO of an International Tanker Company, the issue of reliability of AIS compared to NR data was investigated from his employed key officers. The result pointed out that the quality and accuracy of the AIS data are higher than NR; therefore, it is declared that in this way NR data are promoted to a more reliable position for being implemented in the research. In the following section NR speed quality is obtained by replacing AIS reported speed for the selected tankers. Hereinafter, in *Figures 2* and *3*, this process is depicted for the two selected ships.

Table 2 – NR sample of ship D1

ID	Date	FCR	Speed	Distance	Wave force	Wave direction	Wind force	Wind direction	Current force	Current direction
1	2016/01/03	66.80	11.67	329.32	5	SE	6.000	NW	1.1	NW
2	2016/01/04	66.90	13.39	358.91	3	SE	4.000	NW	1.1	NW
3	2016/01/05	74.10	13.17	368.87	3	SE	4.000	NW	1.6	NW
...	...	...	...	...	...	...	...	...	...	...

Table 3 – AIS historical record sample of Ship D1

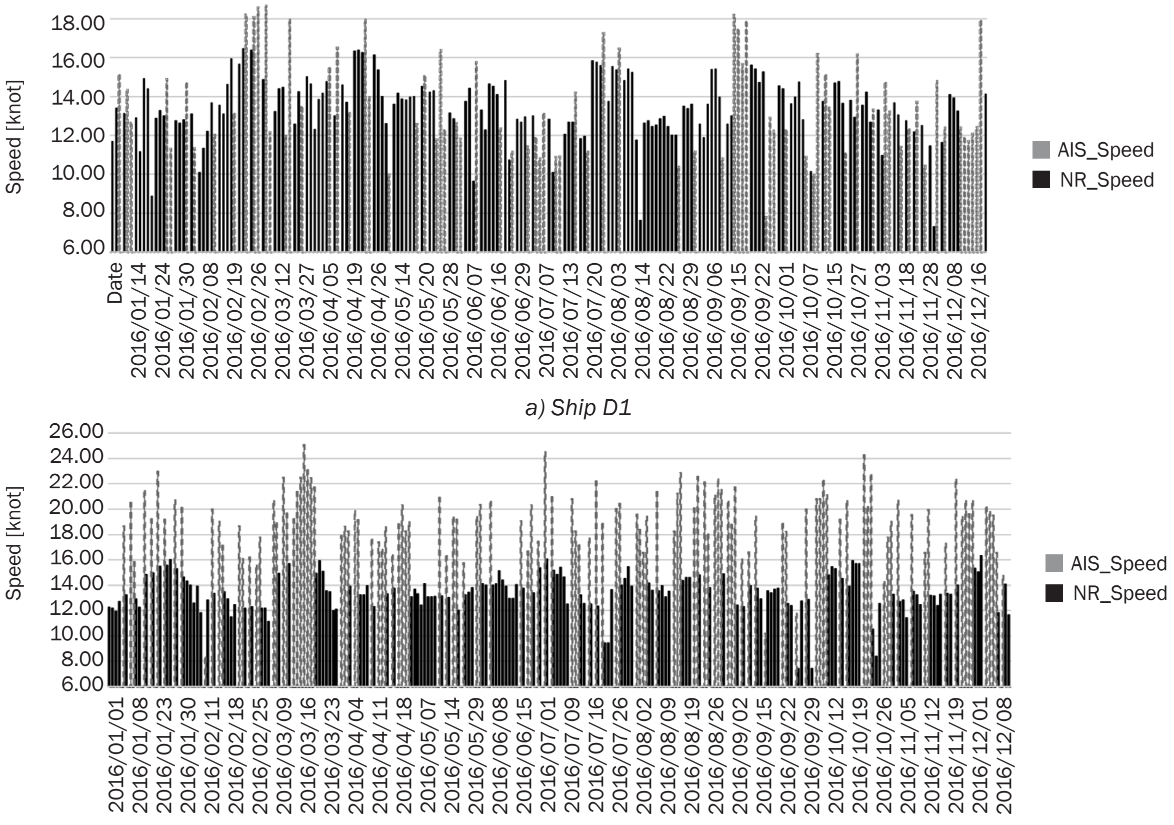
ID	Date	Time	Longitude	Latitude	Draft	Reported speed over ground
1	2016/01/03	13:00	55.193	26.117	11.5	14.5
	2016/01/03	16:00	56.014	26.345	21.6	13.5
	2016/01/03	18:00	55.154	26.121	11.5	14.6
2	2016/01/04	07:00	58.829	25.25835	11.5	13.2
	2016/01/04	10:00	62.041	24.547	11.5	13.6
	2016/01/04	21:00	58.429	25.294	11.5	13
3	2016/01/05	07:00	66.373	22.873	21.6	13.4
	2016/01/05	16:00	66.534	22.81061	21.6	12.9
	2016/01/05	19:00	67.062	22.60417	21.6	12.8
...	...	...	...	...	...	...



a) Ship D1

b) Ship D2

Figure 2 – Out-range NR compared to AIS vessel speed datum



a) Ship D1

b) Ship D2

Figure 3 – Combining NR data with AIS

As the scope of this study is to enrich the quality of independent variables data of ships to estimate fuel consumption in the running mode, the related data to the anchorage or berthing condition have been omitted. In the following sections, combined data of NR with AIS are enriched by different mathematical methods

4. GOVERNING EQUATIONS

In this Section, the process of acquiring valid data and enhancing the NR data quality through the four method equations are explained. Hereinafter, while the K-Means and SOM methods are organized to produce fresh data, the OSB and HOSB are to eradicate wrong data.

4.1 K-Means method

K-Means clustering known as a method of vector quantization originated from signal processing. This method has become one of the most reputable data mining methods to cluster analysing. The objective of K-Means clustering is to separate the observations into clusters. In these separations, each observation pertaining to the cluster with the closest mean serves as a prototype of the cluster. The Voronoi cells will be the outcome of a partitioning of the data space. Because of the nature of the method, it is difficult to use it computationally. Nevertheless, the existence of an efficient heuristic algorithm that is based on the empirical theory can be deployed to converge rapidly to a local optimum. This is equivalent to the expectation maximization algorithm for mixtures of Gaussian distributions through an interactive refinement approach utilized by both heuristic and expectation maximization algorithms. Moreover, in order to model the data, both algorithms employ cluster centres. However, K-Means clustering will normally find clusters of comparable spatial extent [20]. This is done while the expectation maximization mechanism permits clusters to form various shapes. The K-nearest neighbour classifier has a loose relationship with the algorithms known as a machine learning technique used for classification. Because of using the same character K with K-Means this algorithm usually leads to confusion. It is possible to exercise K-nearest neighbour classifier on the cluster centres derived from K-Means. This result is classified as new data into the existing clusters. This process is named as the nearest centroid classifier or Rocchio algorithm. Suppose having a series of observations characterized with d-dimensional and named X_1, to X_n. K-Means clustering objectives is to separate n observations into k ($\leq n$) sets S = {S_1, S_2, ..., S_k in order to minimize the within cluster sum of squares i.e. variance. So the aim is to [20]:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \arg\min_{S} \sum_{i=1}^{k} |S_i| Var S_i \tag{2}$$

where μ_i is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{S} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} \| x - y^2 \| \tag{3}$$

Because the total variance is constant, this is also equivalent to maximizing the squared deviations between points in different clusters Between-Cluster Sum of Squares (BCSS). The K-Means algorithm is also called Lloyds algorithm, especially in the computer science community. This also uses iterative refinement technique and contains ambiguity. Given an initial set of k means $m_j(1)$,...,$m_k(1)$ the algorithm proceeds by alternating between two steps [20]. Each observation should be assigned to a cluster, which contains the least squared Euclidean distance to its mean. This will be instinctively the nearest mean. Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.

$$S_i^{(t)} = \left\{ x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_i^{(t)} \|^2 \forall j, 1 \leq j \leq k \right\} \tag{4}$$

where each x_p is assigned to exactly one $S_i^{(t)}$ even if it could be assigned to two or more of them. As an update step, the new means to be the centroids of the observations in the new clusters can be calculated by:

$$m_i^{(t+1)} = \frac{1}{|s_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{5}$$

As long as the assignments do not change, the algorithm is converged. Therefore, finding the optimum is not guaranteed by deploying this algorithm. The algorithm is shown as the assignment objectives to the closest cluster by distance characteristics. In order to stop the algorithm to converge, it might use various distance functions other than the squared (Euclidean) distance. Spherical K-Means and K-Medoids known as different kinds of modifications of K-Means are normally employed to permit using other distance measures.

As described above, different forms of K-Means method equation are available for different purposes expressed for this algorithm. However, all of them have a recurrent process that attempts to estimate the following for a certain number of clusters:

- Finding several points as the cluster centres means actually identical average points fitting to each cluster.
- Allocating each trial data to a cluster, and then the trial data provide the smallest distance to the middle of that cluster.

Therefore, by obtaining the data average for each recurrence, a novel middle is designed for them, and another time, the data are credited to the novel clusters.

By continuing this process, a point is reached where there are no changes in the data. The objective function is demonstrated by *Equation 6*.

$$J=\sum_{J=1}^{K}\sum_{i=1}^{n}\| X_i - C_j \|^2 \tag{6}$$

In addition, X_i is the j-th cluster centre and the presentation of how this method works is shown by the algorithms below. *Figure 4a* at the start demonstrates the selection of K points as the middle of the cluster. Each data sample is grouped to the cluster bearing in mind that this has the smallest distance to the data sample. Therefore, when all data are categorized to different clusters for each cluster a new point is calculated again, that is the means of points presented in *Figure 4b*. According to this, the process continues until no changes are to be achieved in the centre of the clusters shown in *Figure 4c* [20].

4.2 Self-organized map method

Since the self-organized grid founded on several physiognomies of the human brain, for the training purposes, a competitive learning method has been developed. The compartments in the human brain are systematized in different areas in the way in which they are presented in varied sensory parts with systematic and meaningful computational charts. A neural network character of self-organization is shaped in a systematic low-dimensional network arrangement. N means the dimensions of the input vector and every neuron has an N-dimensional vector. Weight vectors (synapses) link the input sheets to the output sheets called a map or a competitive sheet. Neurons are linked to each other by a neighbourhood function. As per maximum similarity, every vector stimulates a neuron, which is called the winner cell, in the output layer. The Euclidean distance between two vectors is often a base to calculate the similarity. Close-up remarks in the input space stimulate two close-up units in the chart. Until the weight vectors touch the stability level and no changes are to be repeated the training stage continues [21].

Being iterative is the basisc of SOM methodology. For each neuron represented by i, dimension d prototype vector W_i=[W_{i1},...,W_{id}] is assumed, which is also the weight of the i-th neuron. A sample data vector x is selected from the training set occasionally in each training step. The computation of distance between x and all prototype vectors is to be performed resulting in the Best Machine Unit (BMUW). This is also called a winner unit marked by x_{i*} which is the map unit carrying the prototype closest to x.

$$|w_{i*} - x| \leq |w_{i*} - x|, \quad \forall i \neq i^* \tag{7}$$

In the next step, the updating of the prototype vectors is performed and then the BMU and its respective topological neighbours have to be transferred near to the input vector in the input space using

$$\Delta w_{i*}^{r} = \eta(x^r - w_{i*}) \tag{8}$$

where: η is learning rate; Δw_{i*}^{r} is i-th neuron weight modification and lastly the update of *Equation 8a* represents for all vectors of unit i as presented below:

$$\Delta w_i^r = \eta \wedge (|i - i^*|)(x^\tau - w_i) \tag{8a}$$

$\wedge(|i - i^*|)$ is a neighbourhood kernel centred on a winner unit. For example, the Gaussian is the kernel on the equation $\wedge(a) = \exp\left(-\frac{a^2}{\sigma^2}\right)$; where: σ is the neighbourhood radius.

As time goes the learning rate η and neighbourhood radius σ decrease steadily. In the training process, SOM moves as a flexible net being created by the training data. Neighbouring prototypes are dragged to identical course for the reason of neighbourhood relations. Therefore, prototype vectors of neighbouring units look like one another. The number of neurons in output layer means maximum difference of model vectors. At this stage the trained SOM is prepared to classify its inputs. Thus, the class of input vectors is defined by BMU.

$$W_{i\text{-}j}^{new} = W_{i\text{-}j}^{old} + h_{i\text{-}j}(X_i - W_{i\text{-}j}^{old}) \tag{9}$$

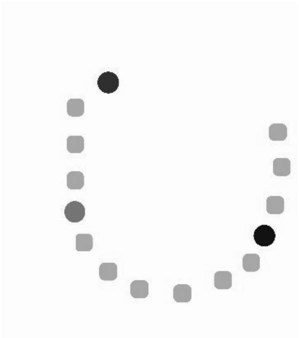


Figure 4a – Selection of centres randomly

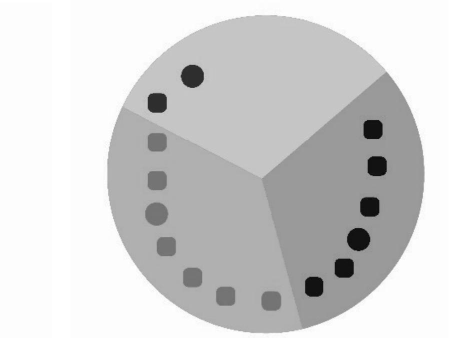


Figure 4b – Clustering 3 clusters using initial centres

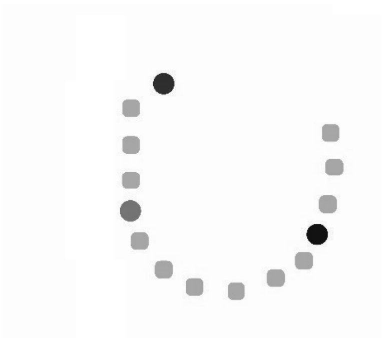


Figure 4c – Calculation of initial centres

where X_i is the input sample, $W_{i\text{-}j}^{old}$ is the previous weight vector between the input vectors X_i and the weight vector connected to the output neural cell $(j.h_{i\text{-}j})$ is the neighbourhood function and $W_{i\text{-}j}^{nhi\text{-}jew}$ is the weight vector updated between input cell i and output cell j. After the training stage, i.e. at the mapping stage, there will be the possibility of automatic ranking of each input data vector [21].

4.3 Outlier score base method

Mathematical methods, e.g. neural network, genetic algorithm, or numerical non-linear calculations, made it possible to level up the quality of raw data. In this section the OSB method is implemented to remove fuzzy data. The basic structure of OSB defines the comparison between two subsequent data. This comparison process is continual until the last data; for instance, having ship NR data the speed ratio and fuel consumption are calculated in different time steps using *Equations 10* and *11*.

$$\frac{rF(i)}{rF(j)} > \max\left(\frac{v(i)}{v(j)}\right)^2, \left(\frac{V(i)}{V(j)}\right)^4 \tag{10}$$

$$\frac{rF(i)}{rF(j)} < \min\left(\frac{v(i)}{v(j)}\right)^2, \left(\frac{V(i)}{V(j)}\right)^4 \tag{11}$$

Applying the two equations, if the fuel consumption ratio at any time step compared to the second and fourth power of the ship speed is more than the maximum value or less than the minimum value, in this stage, the fuel consumption in the given time step will receive a negative score e.g. *Formulas 12* and *13*.

$$OutlierScore(i) = OutlierScore(i) + 1 \tag{12}$$

$$OutlierScore(j) = OutlierScore(j) + 1 \tag{13}$$

Likewise, all scores are calculated for different time steps, and ultimately, a percentage of the uppermost earned scores is measured as out-of-range data. Moreover, in this respect, the time steps in which the speed of the ship is not in the desired time range are given negative score using *Formula 14*.

$$\begin{gathered} V(i) < 10 \text{ knots} \quad \text{OR} \quad V(i) > 30 \text{ knots} \\ OutlierScore(i) = 10N \end{gathered} \tag{14}$$

Then, the entire data are gathered according to the abovementioned scoring structure from the uppermost earned score to the lowermost earned score. At the end, the experimental basis, a percentage of the uppermost scores are removed.

4.4 Histogram outlier score base method

HOSB is a neutral network base method. The difference of HOSB vs OSB is that the accuracy of HOSB is improved by defining a histogram for concentrating on the cause of fuzziness. Because of the fact of having various distributions of the feature values in the real data, both methods are presented in HOSB. This will be more practical when the value ranges contain big gaps. In addition, the fixed bin width approach can calculate the density inaccurately when a few bins may cover most of the data. Since anomaly detection tasks usually involve such gaps in the value ranges, due to the fact that outliers are far away from the normal data, we recommend using the dynamic width mode, especially if the distributions are unknown or long-tailed. In addition, several bins need to be set. An often-used rule of thumb is setting K to the square root of the number of instances N.

Then, for each dimension d, an individual histogram has been computed, regardless of categorical, fixed-width or dynamic-width where the height of each single bin represents the density estimation. The histograms are then normalized in such a way that the maximum height will be 1.0. This ensures an equal weight of each feature to the outlier score. For every data for example p, $hist_i\ (p)$ is calculated by multiplication of the inverse of the estimated densities of neighbourhood data to the independence factor, p. The equation could be written as:

$$HOSB(p) = \sum_{i=0}^{d} \log\left(\frac{1}{hist_i(p)}\right) \tag{15}$$

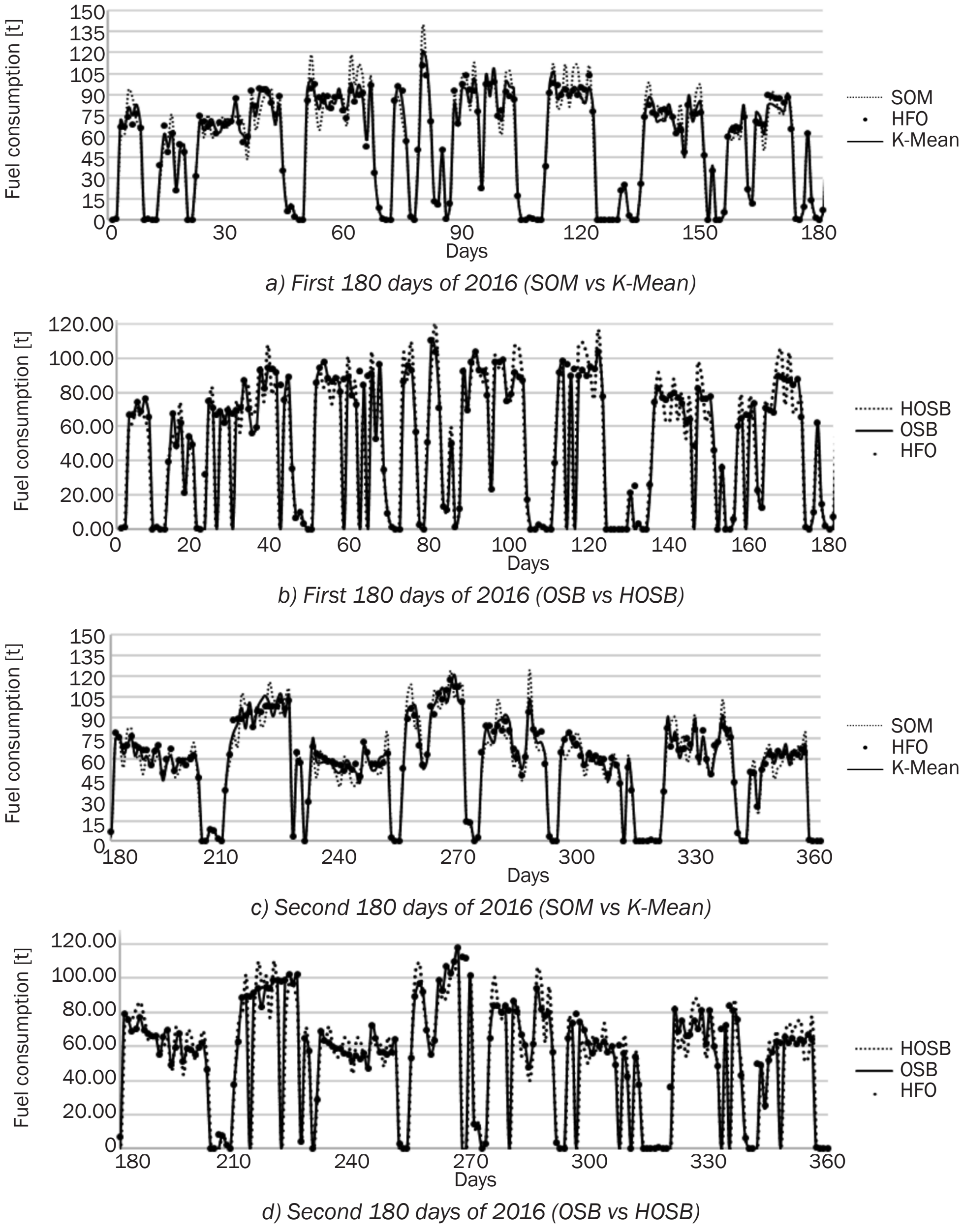
In fact, this method is a discrete method based on the probability theory Naïve Bayes. In other formulation, the sum of the logarithms can be taken as $(\log(a - b) = \log(a) + \log(b))$ and by applying this new formula to simplify *Equation 15* by separating the logarithm part. By this separation, new equations have low sensitivity to errors according to precision of the floating points that cause high scores in unbalanced distributions [21].

5. ENRICHING THE EXISTING DATA

Hereinafter, the aim is to solve the problem of raw and fuzzy NR data of the tankers D1 and D2 using the explained mathematical methods by writing the automatic code in MATLAB due to a high number of available statistical data. For clarification purposes, a comparison model is shown, with the changes for each model. *Figures 5* and *6* illustrate the original data of fuel consumption vs new data enriched for two VLCCs. *Figures 5a-5d* show the reported fuel consumption of the D1 oil tanker during 12 months (each chart divided into two 180-day parts for visualization purpose). The new generated high-quality data replaced to original odd data are in black line by using K-Mean and OSB method while SOM and HOSB are depicted in the dash line. In addition, real original data are in scatter black points. Similarly, *Figures 6a-6d* are fuel consumption treatment for one of the ship D2. As mentioned above, two methods of K-Mean and SOM have been deployed to generate new high-quality data, and OSB and HOSB to eliminate fuzzy data.

As shown above, heavy fuel oil consumption (HFO) of vessel named D1 is depicted in two amplitudes for each 180 days of the year. The beginning step is the first 180 days of 2016 and the rest of 2016 occurred in the last 180 days, presented in *Figures 5a-5d*. SOM and K-Mean are fully dependent on the raw data for generating a rich primary generation. Therefore, poor harmony of data collecting can cause ambiguity and fuzziness in generating new high-quality data generation. Fortunately, as indicated in *Figures 5a* and *5c*, the established generation has an acceptable quality because of high frequency of data collecting of the NR available data. Then, the developed program based on the mentioned mathematic models successfully removed 15 percent of out-ranged data and the last produced generation created pure valid new data in the range of original raw data by a parallel harmony using MATLAB. In addition, as shown in *Figures 5b* and *5d* using HOSB and OSB methods removed the 15 percent of outlier NR data that are far from the mean of original data. In *Figures 6a-6d* a similar concept can be derived for new VLCCs named D2.

According to the fact presented in *Figures 5* and 6, it can be judged that to some extent all methods successfully improve the quality of raw data in different manners. Two methods remove the outlier data directly, while the others by generating new data increase



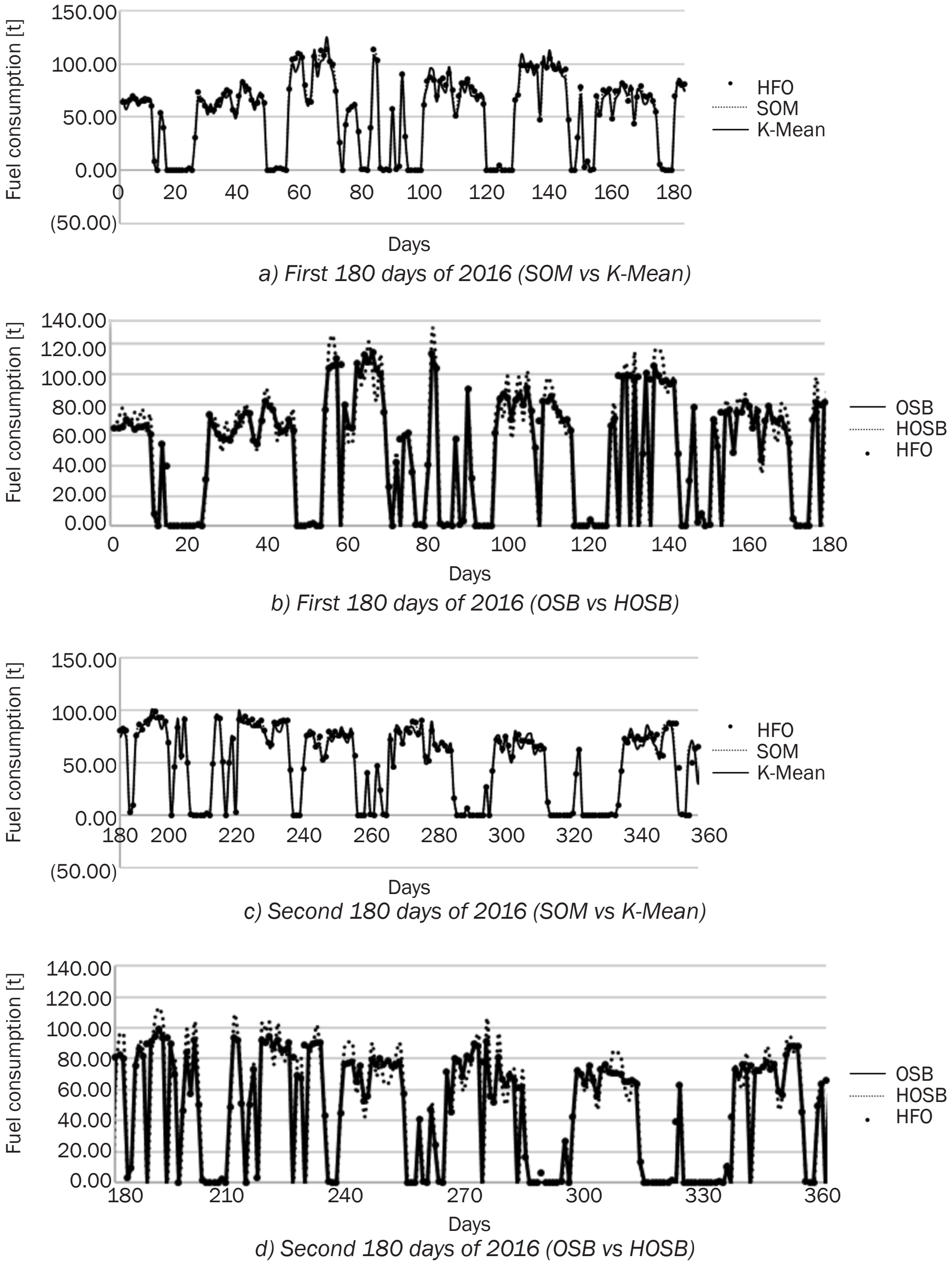
a) First 180 days of 2016 (SOM vs K-Mean)

b) First 180 days of 2016 (OSB vs HOSB)

c) Second 180 days of 2016 (SOM vs K-Mean)

d) Second 180 days of 2016 (OSB vs HOSB)

Figure 5 – D1 fuel consumption variation



a) First 180 days of 2016 (SOM vs K-Mean)

b) First 180 days of 2016 (OSB vs HOSB)

c) Second 180 days of 2016 (SOM vs K-Mean)

d) Second 180 days of 2016 (OSB vs HOSB)

Figure 6 – D2 fuel consumption variation

the quality of data indirectly. In reality, based on the type of the usage, different methods can be deployed. However, the importance is to find the best fitted method to address a special problem rather than just deploying a popular or even well-known method.

In this regard, the criteria below are to be considered when aiming at distinguishing the most suitable method to solve the problems.

- Distance to the real original data;
- Harmony of data.

The distance to the real original data is assessed by calculating the average error of the days for each method. *Equation 16* is represented by an average method or expected values.

$$Average\ Erorr = Mean\ \ of\ \sum_{i=1}^{365} \frac{\left| Data_{irevised} - Data_{ioriginal} \right|}{Data_{ioriginal}} \qquad (16)$$
$$i = day$$

Figure 7 pointed out the calculated average error for all the previous mentioned methods. This calculation is done by measuring day-to-day distance of new generated data versus original data. According to the finding of the calculation, HOSB is, among others, with the average error percentage less than 6.25.

The root mean square calculation result of the entire 12 months data satisfies the second criterion. *Equation 17* demonstrates the root mean square.

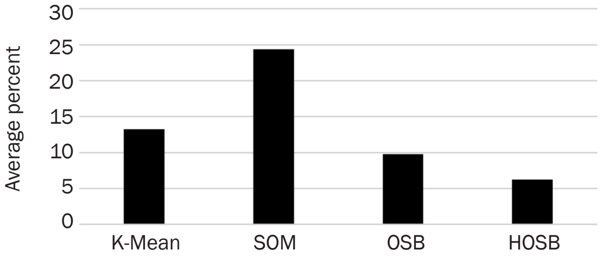


Figure 7 - Comparison of the average error vs original data

$$RMS = \sqrt{\sum_{i=1}^{365} \frac{(data_{i revised} - data_{i original})^2}{365}} \qquad (17)$$

Figure 8 demonstrates the error rate of each method using the root mean square index. As it can be seen, the HOSB method with the least deviation and error at about 0.4 is better than other methods. Therefore, HOSB is a successful method with high degree of confidence to be deployed for all similar ships to optimize the fuel consumption in maritime transport.

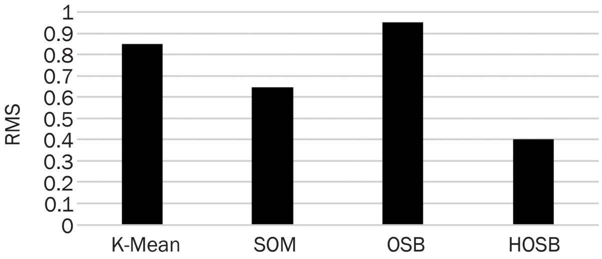


Figure 8 – Comparison of RMS average of the used methods vs original data

6. CONCLUSION

In maritime transportation, ship operational expenses are a considerable factor for charterers and owners in which fuel consumption has the highest share among other operational cost elements. Meanwhile, global concern on air pollution brings experts to predict and decrease fuel consumption of ships. Herein, the lack of worthwhile data along with a high accurate method is sensible. In this study, for two sister ships the NR databases are gathered and enriched by combining with AIS report for a year. The qualified database is treated first by eliminating the fuzzy values of NR data. Furthermore, four well-known methods including K-Mean, SOM, OSB and HOSB were deployed and compared to validate and obtain the best methodology. In addition, based on the stated four mathematical governing equations, a program is generated using MATLAB. The output of the program as a result of this study indicates that still the combination of AIS and NR enriched by HOSB model is known as the most reliable methodology to be applied. The least deviation and error using the root mean square index derived is about 0.4 indicating the high accuracy of the method. In future, it is proposed to investigate a proper relation between reported parameters, i.e. fuel consumption rate, vessel speed, waves, current, route etc. in order to derive fuel consumption prediction formula using the enriched pure valid NR data.

روش جدید برای به دست آوردن داده های خالص معتبر با ترکیب گزارش روزانه تانکر نفت و شناسایی سیستم های ماهواره ای با استفاده از روش های آماری و داده کاوی

علی اکبر صفایی، دانشجوی دکتری مهندسی دریا، دانشگاه صنعتی امیرکبیر
حسن قاسمی، استاد دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر
محمود غیاثی، دانشیار دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر

چکیده

با نوجه به اهمیت قابل توجه سهم هزینه های سوخت در هزینه های عملیاتی کشتی و تاثیر ان بر محیط زیست و آلودگی هوا، موضوع بهره وری در مصرف سوخت همواره بعنوان یکی از موضوعات مهم بوده است. بنابراین پیش بینی میزان مصرف سوخت کشتی در یک سفر دریایی دارای اهمیت است. یکی از ابزارهای پیش بینی مصرف سوخت استفاده از تکنیک تحلیل داده های گزارش روزانه کشتی میباشد. گزارش روزانه کشتی بدلیل جمع آوری روزانه توسط افسر اول کشتی دقیق نمیباشد و احتمال خطای انسانی در ان زیاد است. برای حل این مشکل در این مطالعه و بمنظور تلخیص داده های گزارش روزانه، داده های استخراج شده از گزارش روزانه یازده کشتی تانکر های بزرگ با داده های همان کشتی ها از سیستم شناسایی اتوماتیک کشتی ها ترکیب گردیده است. سپس با استفاده از مدلهای معروف **K Mean**، **Self-Organizing Map** ، **Histogram Outlier Score Base** داده های دوازده ماه تانکرهای موصوف بازنگری و داده های صحیح استخراج گردیده است. داده های پیشنهادی هر مدل با داده های اولیه مقایسه گردیده تا میزان اعتماد هر مدل برای استخراج داده های تلخیص شده و صحیح ارزیابی گردد. بدین منظور روش های **Expected Value** و **Root Mean Square** برای ارزیابی هر مدل بکار گرفته شده است. در نتیجه گیری انجام شده **Expected Value** و **Root Mean Square** برای مدل **HOSB** بترتیب **6.25** و **0.4** بدست آمده است که نشاندهنده تطابق خوب و هارمونی قابل توجه بین داده های استخراج شده و داده های خام اولیه استخراج شده از گزارش های روزانه است.

کلمات کلیدی: گزارش روزانه کشتی، سیستم شناسایی اتوماتیک کشتی، مصرف سوخت کشتی، تحلیل داده ها، تلخیص داده

REFERENCES

[1] Bole A. *Radar and ARPA Manual*. Chapter 5 – Automatic Identification System (AIS); 2014.

[2] Fagerholt K, Laporte G, Norstad I. Reducing fuel emissions by optimizing speed on shipping routes. *Journal of the Operational Research Society*. 2010;61(3): 523-529.

[3] Man Diesel & Turbo. *Costs and Benefits of LNG as ship fuel for container Vessels. Engineering the Future;* 2011.

[4] Wang S, Meng Q. Bunker consumption optimization methods in shipping: A critical review and extensions. *Transportation Research Part E: Logistics and Transportation Review.* 2013;53: 49-62.

[5] Kontovas C, Psaraftis HN. Reduction of emissions along the maritime intermodal containerchain: operational models and policies. *Maritime Policy & Management*. 2011. p. 451-469.

[6] Kim JG, Kim HJ, Lee PTW. Optimizing ship speed to

minimize fuel consumption. *The International Journal of Transportation Research.* 2014. p. 109-117.

[7] Safaei AA, Ghassemi H, Ghiasi M. A Voyage Optimization for a Very Large Crude Carrier Oil Tanker: A Regional Voyage Case Study. *Scentific Journal of Maritime University of Szczecin.* 2015;44(116): 83-89.

[8] Meng Q, Wang S. Optimal operating strategy for a long-haul liner service route. *European Journal of Operational Research*. 2011;215(1): 105-114.

[9] Notteboom T, Vernimmen B. The effect of high fuel costs on liner service configuration in container shipping. *Journal of Transport Geography*. 2009;17(5): 325-337.

[10] Nie Y, Wu X. Shortest path problem considering on-time arrival probability. *Transportation Research Part B: Methological*. 2009;43(6): 597-613.

[11] Meng Q, Du Y, Wang Y. Shipping Log Data Based Container Ship Fuel Efficiency Modeling. *Transportation Research Part B: Methological*. 2016;83: 207-229.

[12] Fang MC, Lin YH. The optimization of ship weather-routing algorithm based on the composite influence of multi-dynamic elements (II): Optimized routings. *Applied Ocean Research.* 2015;50: 130-140.

[13] Lusic Z, Kos S, Galic S. Standardization of Plotting Courses and Selecting Turning Points Maritime Navigation. *Promet – Traffic & Transportation*. 2014;26(4): 313-322.

[14] Vijendra S, Shivani P. Robust Outlier Detection Technique in Data Mining: A Univariate Approach. *Computer Vision and Pattern Recognition Journal*. 2014.

[15] Williams G, Baxter R, He H, Hawkins S, Gu L. A Comparative Study for RNN for Outlier Detection in Data Mining. In: *Proceedings of the 2nd IEEE International Conference on Data mining, 9-12 Dec 2002, Maebashi City, Japan*; 2002.

[16] Kantardzic M. *Data mining concepts, models, methods, and algorithms.* 2nd ed. Johan Wiley & Sons, Inc; 2011.

[17] Han J, Kamber M, Pei J. *Data mining concepts and techniques.* 3rd ed. Morgan Kaufmann Publishers; 2006.

[18] Ghosh-Dastidar B, Schafer JL. *Outlier Detection and Editing Procedures for Continuous Multivariate Data.* ORP Working Papers, Working Paper No. 2003-07, 2003.

[19] Safaei AA, Ghassemi H, Ghiasi M. Correcting and Enriching Vessel's Noon Report Data Using Statistical and Data Mining Methods. *European Transport*. 2018: no 67.

[20] Kriegel HP, Schubert E, Zimek A. The art of runtime evaluation: Are we comparing algorithms or implementations?. *Knowledge and Information System*. 2016;52(2): 341-378.

[21] Kind A, Stoecklin M, Dimitropoulos X. Histogram Based Traffic Anomaly Detection. *IEEE Transaction on Network and Services Management*. 2009;6(2): 110-121.