Correcting and Enriching Vessel's Noon Report Data Using Statistical and Data Mining Methods

Ali Akbar Safaei, Hassan Ghassemi, Mahmoud Ghiasi

Department of Maritime Engineering, Amirkabir University of Technology, Tehran, Iran

Corresponding Author Contact: gasemi@aut.ac.ir

Abstract

Appraising the ship noon report data is known as one of the approaches and methods to control a ship fuel consumption influential parameters. This ship noon report contains various data such as daily fuel consumption, ship speed, sailing direction, different external forces including wave, wind and current forces recorded on daily basis by ship's captain or chief officer. Because of possible human errors and big interval of data collecting, significant inaccuracy might happen in statistical studies using this raw data. Therefore, in order to correct and enrich the data quality, various mathematical models such as statistical, numerical and analytical models are proposed in different literature reviews. In this study by applying well known models i.e. K-Mean, Self-Organizing Map, Outlier Score Base and Histogram Outlier Score Base into twelve months collected noon reports data of four Very Large Crude Carriers sister ships, the data is analyzed. In the next step, new generated data is compared with the original data. In addition, the mathematical methods are investigated to find out their effectiveness in respect of the owner objective to control the ship fuel consumption. In this line, by calculating expected value and root mean square method, the four above mentioned models are compared and evaluated resulting that the HOSB with 2.11 percent accuracy concerning the original data is known as the most reliable method.

Keywords: Noon Report, Ship Fuel Consumption, Data Mining, Data Enriching.

1. Introduction

So far, much research has been done on reducing fuel consumption. For example, examination of the effects of internal components of the ship, environmental factors, economical & optimal path, operational and managerial factors and other criteria which can be mentioned. Among the researches, the existence of a practical model for predicting fuel consumption along with increasing the productivity using Noon Report (NR) data and Automatic Identifying System (AIS) are rarely observed. Also collecting or generating proper and valid data is an essential part of stablishing a reliable equation to forecast fuel consumption. Required data for this study is found in the NR collected by chief officer or captain of the vessel once a day. Always human error is a part of all reports made by human

therefore in NR data sometimes we face with some odd data that is not in a right harmony with the others. There are many statistical methods to find the roots of the data harmonies or odd data. In the first step, the out ranged or junkie's data are determined. Then they are treated in two different ways i.e. eliminating the existing data or generating new data in the right harmony with the others.

Similar investigations are being carried out in this area, and in this regard, the effects of wind, wave, and loading on the overall resistance and speed of the ship using NR data investigated (Zelazni, 2014). The components of the ship speed for a container ship and for a very large tanker ship can be considered as 3.5 and 4, respectively using existing NR data (Diesel et al, 2011). In addition, another study demonstrated empirically that these components are between 2.7 and 3.3 for a speed of 20 knots (Wang & Meng, 2013). Before them, other research had considered the number 4 for the speed components for the ships that travel at the speeds greater than 14kn (Kontovas & Psaraftis, 2011). The greater the speed of the ship, the shorter the time of ship journey. Therefore, optimizing the speed of the ship is fully related to the planning grid of the ship (Bell et al, 2013). In other researches, many works have been done to plan and design a journey grid of container ship (Agarwal & Ergun, 2008). And subsequently, a model for fuel efficiency in container ships proposed (Meng & Wang, 2016). The displacement path of an empty container to provide multidimensional ship routes by using noon report designed and optimized (Song & Dong, 2012). At the same time, a prediction model for the effects of the increased costs of analyzing relationship between fuel price, speed, and the number of ships on the shipping route offered (Notteboon & Vernimmen, 2012). Other study worked on routing and employing it for a group of a company's ships and conducted a number of surveys on several types of ships with different speeds (Alvarez & Troya, 2016). Also, the speed of different ships of a fleet optimized (Meng, S. Wang, 2015) and in this area Fagerholt chose the shortest route and the most optimal for ships by optimizing the route and the speed of the ship (Fagerholt et al, 2010). In line with the route selection, a revenue management of linear container ships presented (Meng & Wang, 2015).

As far as the ship data is concerned, a new method for enriching, modelling and optimizing data using genetic algorithms was proposed (McCall, 2005). Also in other article, it is tried to state a smart model for planning and modelling ship voyage by statistical method according to available ocean and atmospheric data (Scott, 2012). In a new research, Automatic Identification System deployed to monitor the vessels (Bole, 2014) and later, atmospheric emissions from UK fishing fleet by AIS based approach calculated (Coello, 2015). By increasing, the use of AIS data in the studies, another research teased out the detail to improve understanding of marine AIS data for defining better applications in the industry (Shelmerdine, 2015). Moreover, the importance of using data collection on ship direct other study to create an economical shipping route considering the effects of sea state to lower fuel consumption based on sea state's long term data gathered (Roh, 2017). Finally, a new method optimized different routes for minimizing fuel consumption for a Very Large Crude Carrier (Safaei et al, 2015).

Needless to say that the accuracy of these studies are dependent on the quality of the data particularly for those used NR data therefore, the method deployed to correct and enrich these data specially when the number of available data is not very big is considered crucial. In this study, four mathematical methods examined to correct and enrich NR data, two methods for eliminating and two methods for generating data.

2. Governing Equations

These methods are K-Mean (KM), Self-Organizing Map (SOM), Outlier Score Base (OSB) and Histogram Outlier Score Base (HOSB). K-means and Self-Organized Method deployed to generate new data while OSB and HSOM to eliminating false data. Hereinafter, the basic equations of the four methods used to improve the existing data by eliminating or generating new data are presented in the following sections i.e. equations 1 to 7.

2.1. K-Means Method

Different forms of K-Mean method equation are available for different purposes which have been expressed for this algorithm. But, all of them have a recurrent process that attempts to estimate the followings for a certain number of clusters:

- Obtaining some points as centers of clusters, which are actually the same average points belonging to each cluster.
- Assigning each sample data to a cluster, so that this sample data gives the minimum distance to the center of that cluster.

Thus, by averaging the data for each recurrence, new center is calculated for them, and again, the data are attributed to new clusters. This process continues until there will be no change in the data. The relationship (1) is an objective function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| X_i - C_j \right\|^2 \tag{1}$$

And Xi is j^{th} cluster center. The following algorithm is considered as a basis for this method. As shown in Fig 1 at the start, k points are selected for as centers of the cluster. According to Fig 2 each data sample is attributed to the cluster whose center has the smallest distance to that data sample. So, all data belongs to one of the cluster and a new point is calculated as the center for each cluster; i.e. the mean of points belonging to each cluster is computed. Consequently, steps 3 has similar deviation from step 2 and this cycle can be continued until there will be no change in the centers of the clusters.

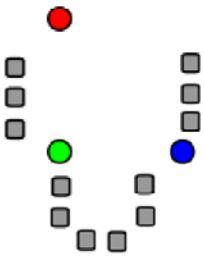

Figure 1. Selection of centers randomly

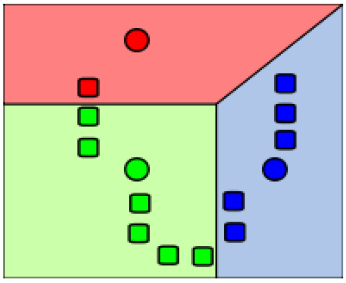

Figure 2. Clustering 3 clusters using initial centers

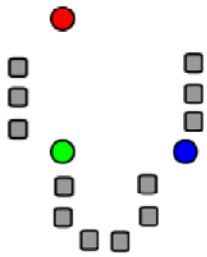

Figure 3. Calculation of initial centers

2.2. Self-Organized Map Method

In the self-organized grid, a competitive learning method is developed and used for training. It is based on some specific characteristics of the human brain. The cells in the

human brain are organized in different regions in such a way that they are presented in different sensory areas with regular and meaningful computational maps. For example, tinnitus and auditory sensory inputs are associated with a different geometry order. The role of self-organization of a neural network has been formed in a regular low-dimensional grid structure. Each neuron has an N-dimensional vector in which N is the dimensions of the input vectors. Weight vectors (synapses) connect the input layers to the output layer (which is called a map or a competitive layer). Neurons are connected to each other by a neighborhood function. Based on the highest similarity, each vector activates the neuron, which is called the winner cell, in the output layer. Similarity is usually measured based on Euclidean distance between two vectors. Close-up observations in the input space activate two close-up units in the map. The training stage continuous until the weight vectors reach the state of stability and do not change anymore.

$$W_{t-j}^{new} = W_{t-j}^{old} + h_{t-j}(X_t - W_{t-j}^{old}) \quad (2)$$

Where X_i is the input sample, W_{t-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_{t-j} is the neighborhood function and $W_{t-j}^{nh_{t-j}ew}$ is the weight vector updated between the input cell I and the 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.

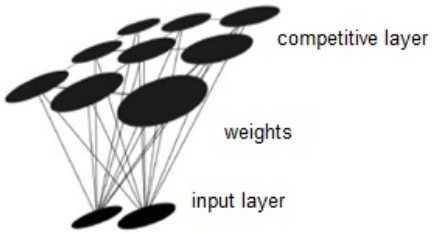

Figure 4. Network structure

2.3. Outlier Score Base Method

Elimination of fussy data will be done using one of the numerical methods, such as neural network, genetic algorithm or numerical nonlinear calculations. One of the most widely used methods in this field is the OSB model. The structure of this model is based on the comparison between two following data and continuation of this comparison to the last data. For example, the ratios of speed and fuel consumption are calculated in different time steps as per formula 3 and 4.

$$r_F^{(i)}/r_F^{(j)} > \max((V^{(i)}/V^{(j)})^2, (V^{(i)}/V^{(j)})^4) \quad (3)$$

$$r_F^{(i)}/r_F^{(j)} < \min((V^{(i)}/V^{(j)})^2, (V^{(i)}/V^{(j)})^4) \quad (4)$$

Then, if the ratio of fuel consumption is at any time step as compared to the second and fourth power of the speed of ship is more than maximum value or less than minimum value, in this stage, the fuel consumption in the given time step will receive a negative score e.g. formula 5 and 6.

$$\text{OutlierScore}^{(i)} = \text{OutlierScore}^{(i)} + 1; \quad (5)$$

$$\text{OutlierScore}^{(j)} = \text{OutlierScore}^{(j)} + 1; \quad (6)$$

Similarly, all scores are calculated for different time steps, and eventually, a percentage of the highest earned scores is considered as out-of-range data. Also, in this regard, the time steps in which the speed of the ship is not in the desired time range can be given negative score using the formula 7.

$$V^{(i)} < 10 \text{ knots OR } V^{(i)} > 30 \text{ knots} \quad \text{OutlierScore}^{(i)} = 10N; \tag{7}$$

After this step, all data will be collected based on the above scoring system from the highest earned score to the lowest earned score. Finally, based on the experience, a percentage of the highest scores are eliminated.

2.4. Histogram Outlier Score Base Method

HOSB is a new neutral network base method which is so similar to OSB but modified by defining a histogram for focusing on reason of fuzziness.

3. Specifications of the Selected Vessels

In this study, four Very Large Crude Carrie (VLCC) vessels' NRs have been selected to be corrected and enriched. Tables 1 and 2 presents the characteristics, name and dimension of the VLCCs.

Table 1. Name and characteristics of ships

Name	Type	Tonnage	Year of construction
Sea Cliff	VLCC	320,000.00	2013
Dune	VLCC	320,000.00	2013
Sana	VLCC	300,000.00	2000
Sonia	VLCC	300,000.00	1996

Other data from the given ships can be used to investigate their speed, power of engine, and fuel consumption. Table 3 illustrates the fuel consumptions in regard to the ship speed and other characteristics. Table 4 and 5 show the reviewed parameters and NR data sample respectively.

Table 2. Dimensions of ships

Name	Dimensions				
	LENGTH O.A.	**LENGTH B.P.**	**WIDTH MLD**	**DEPTH MLD**	**DRAFT MLD**
DUNE	332.95	320.00	60.00	30.50	22.60
SEA CLIFF	333.00	320.00	60.00	30.50	22.64
SANA	330.00	316.00	60.00	28.90	21.58
SONIA	332.00	320.00	58.00	31.00	22.00

Table 3. Consumption-related specifications of ships

VESSELS	MAX POWER	MAX RPM	PITCH	CALCULATED M/E CONSUMPTION PER DAY	AUX. ENGINE CONSUMPTION PER DAY	TOTAL CONSUMPTION PER DAY
DUNE	27160	74	7518	83.300	5.000	88.300
SEA CLIFF	31640	80	7207	88.174	5.000	93.174
SANA	31640	80	7207	88.174	5.000	93.174
SONIA	31640	80	7349.8	86.774	5.000	91.774

Table 4. Reviewed parameters

Distance	**Distance Since Yesterday**
	AVG Speed
Capacity	**Capacity**
	Cargo Quantities
	Ballast
	Mid Draft
Sea	**Sea Force**
	Sea Direction
Wind	**Wind Force**
	Wind Direction
Current	**Current Force**
	Current Direction
Latitude	**Degree**
	Minute
	Direction
Longitude	**Degree**
	Minute
	Direction

Table 5. Example of tables of daily report collection

Name of ship																
Date	**Cargo**	**Ballast**	**Status**	**M.E. Dist.**	**Obs. Dist.**	**Hours Steam**	**Avg. Speed**	**Loading Hour**	**Discharging Hour**	**Ballast Ex. Hour**	**Gas Freeing Hour**	**Tank Washing Hour**	**Port**			
													Load	**Discharge**		
Maneuvering			**Steaming**			**Drifting**			**Fuel**				**Oil**	**EEOI**		
									Main Engine	**Boiler**	**Generator**	**ROB**				
Hour	RPM	Dist.	Hour	RPM	Dist.	Hour	RPM	Dist.								

4. Enriching Existing Data

In this section, Matlab codes of the given methods in the section of governing equations are provided and these codes are used to resolve the problem of raw data. For comparison purposes, the following changes for each model are observed for a sample vessel. In the figure 5 to 12, the original data of fuel consumption and vessel speed vs new generated data mined by writing a program in Matlab are presented. For example, figure 5 shows the reported speeds of the Sonia oil tanker during 12 months in black. The new generated high quality data replaced to original odd data are in red by using K-Mean method. Similarly fuel consumption data are treated as previous mentioned procedure shown in Figure 6. Also figure 7 and 8 presented modified data of vessel speed and fuel consumption by SOM method as following.

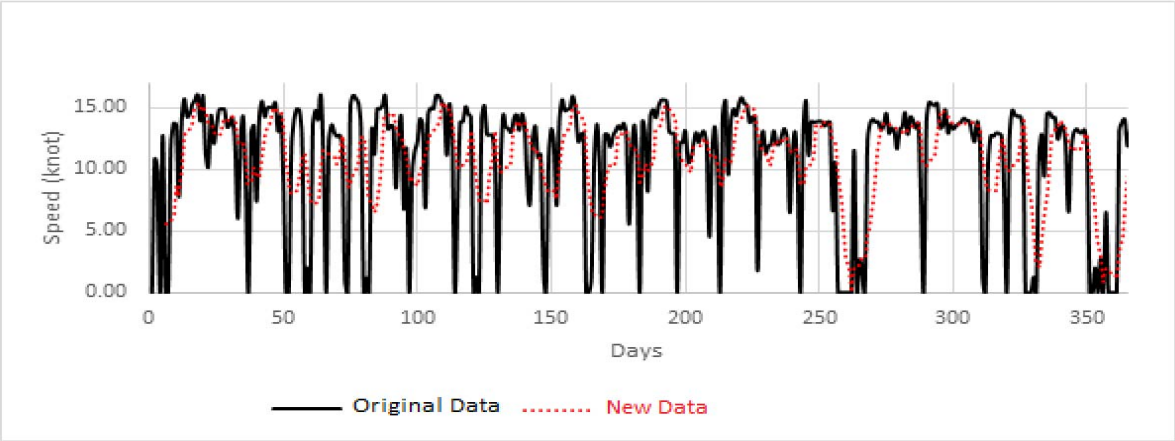

Fig 5. MT Sonia **Speed** variation during one years - (**K Mean Method**)

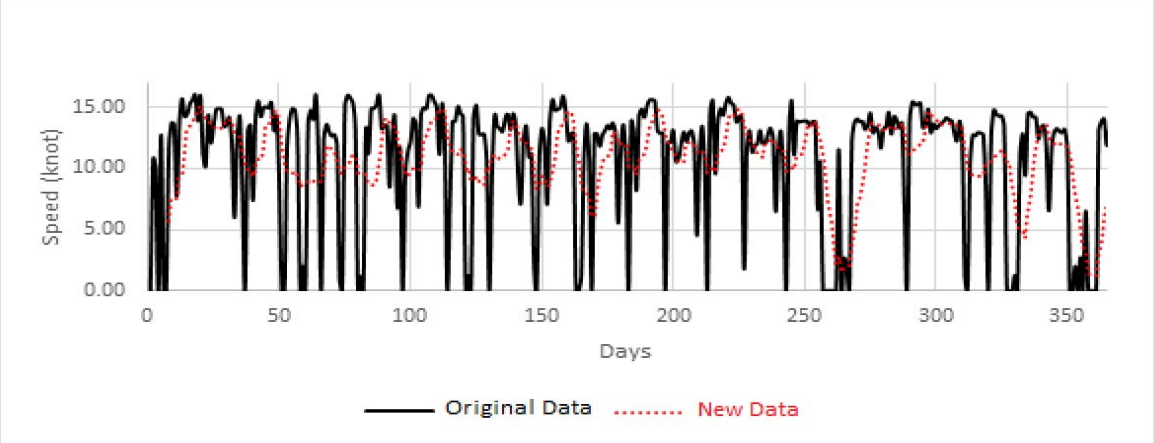

Fig 6. MT Sonia **Fuel Consumption** Variation for One Year (**K_Mean Method**)

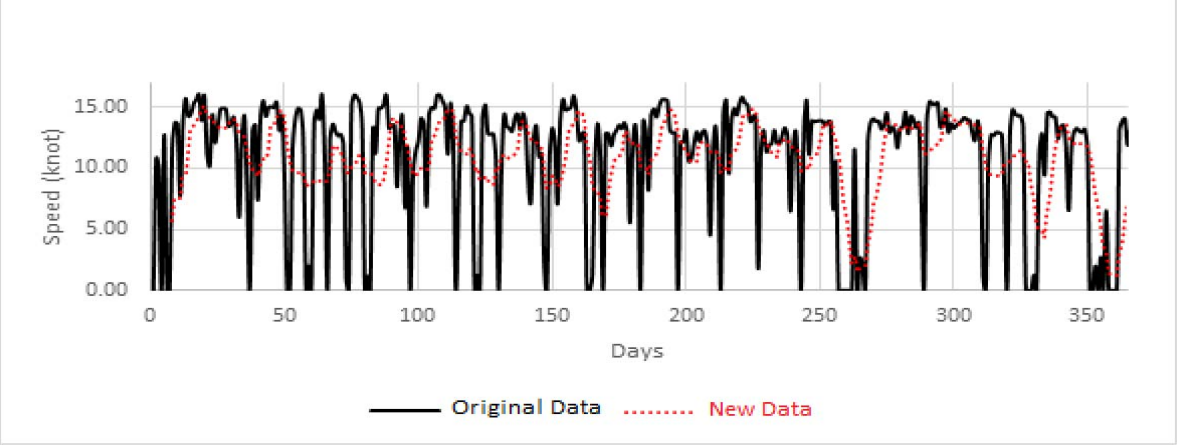

Fig 7. MT Sonia **Speed** variation during one years - (**SOM Method**)

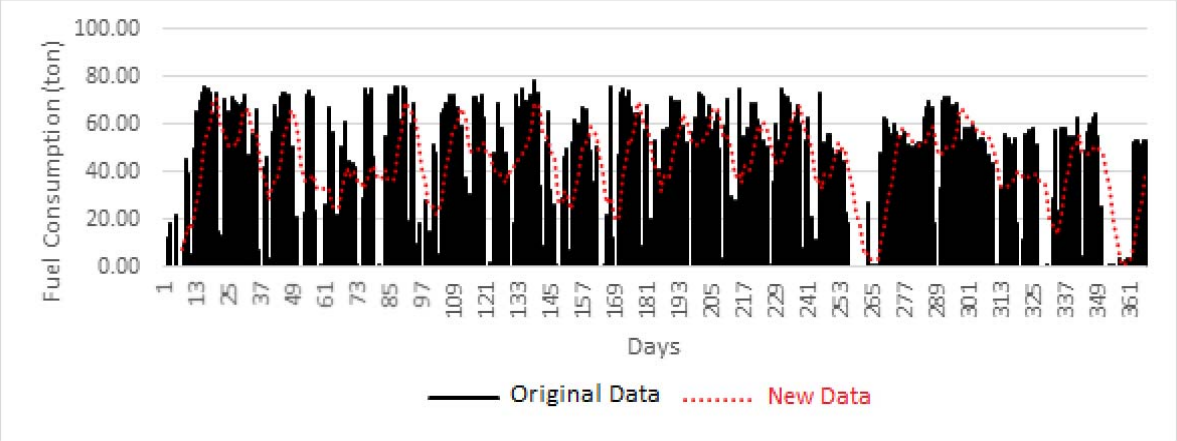

Fig 8. MT Sonia **Fuel Consumption** Variation for One Year (**SOM Method**)

As mention above we used two methods of K-Mean and SOM to generate new high quality data. In the following figures using OSB and HOSB to eliminating fuzzy data are presented. Figure 9 illustrates the treated data using OSB on Sonia for twelve months for speed reported by the NR. Additionally, Figure 10 presents similar method results for eliminating odd date for fuel consumption of Sonia in twelve months. Similarly, Figure 11 ad 12 are depict the new data by using HOSB.

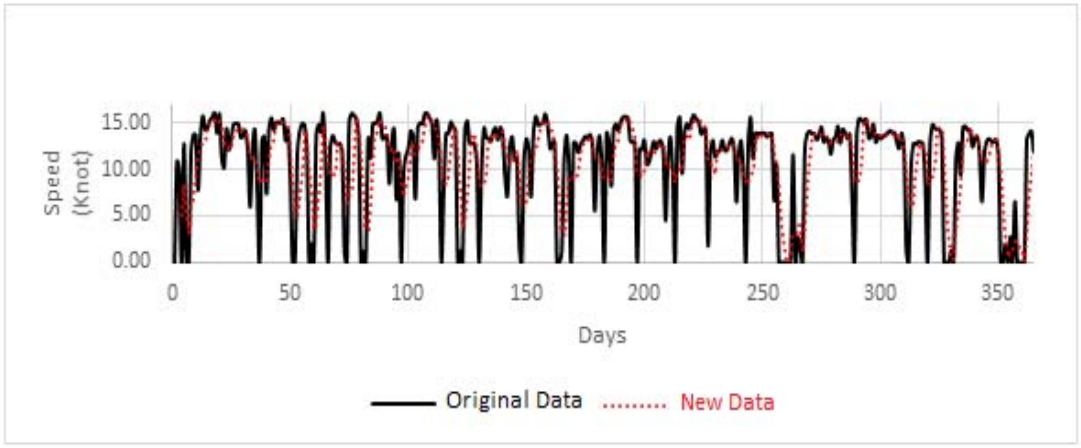

Fig 9. MT Sonia **Speed** variation during one years - (**OSB Method**)

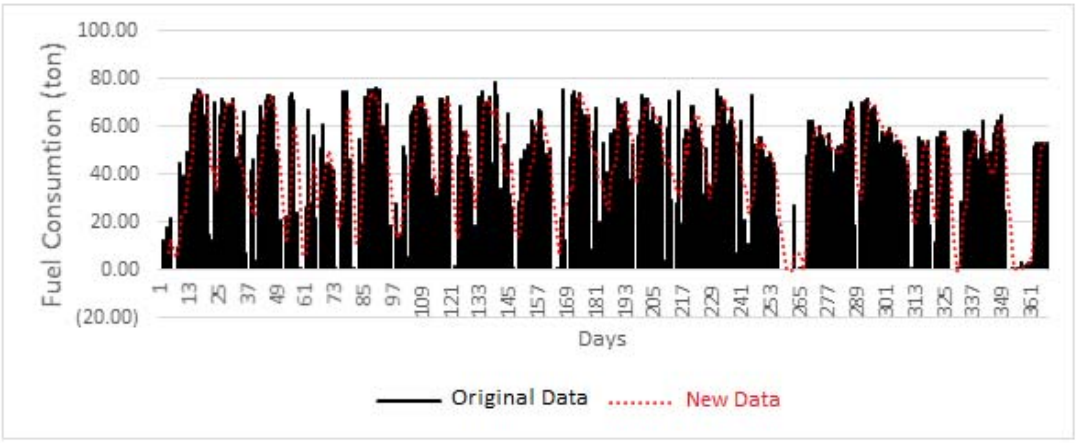

Fig 10. MT Sonia **Fuel Consumption** Variation for One Year (**OSB Method**)

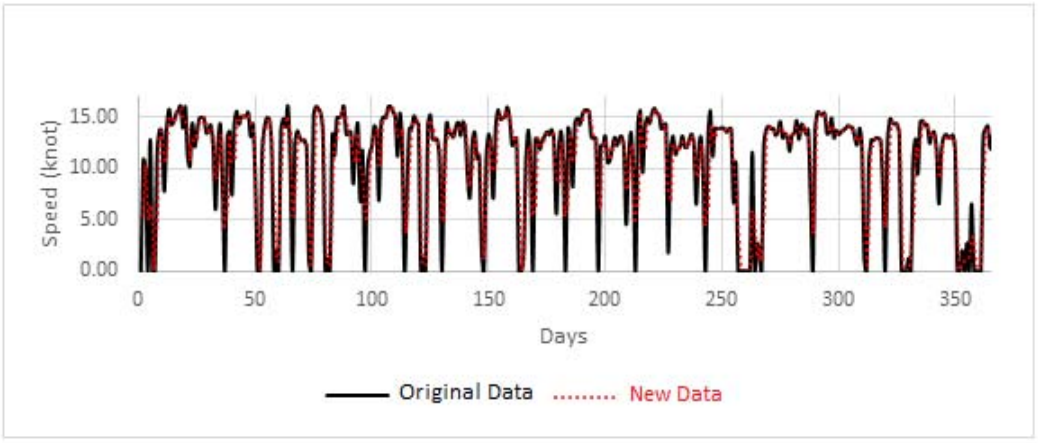

Fig 11. MT Sonia **Speed** variation during one years - (**HOSB Method**)

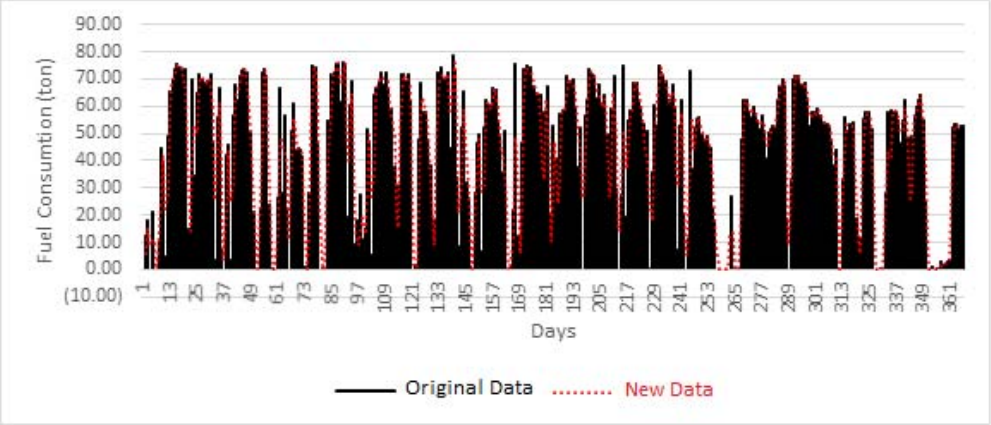

Fig 12. MT Sonia **Fuel Consumption** Variation for One Year (HOSB Method)

As shown in the above figures to some extend all methods are successful in improving the quality of raw data but in different manners. Two methods eliminate the outlier data directly but the others generate and increase the quality of data indirectly. Different usage of various methods is possible but finding the most fitted method to special problem is more important than just deploying a method based on its popularity or being well known. In this line, the following parameters are introduced to be taken into consideration in order to distinguish the most fitted method to solve our problems.

1- Distance to the real original data
2- Harmony of data.

The first parameter can be evaluated by calculating average error of each days for each methods. Average method or expected values formula is:

$$Average\ Erorre = Mean\ of \sum_{i=1}^{365} \frac{\left|\left(Data_{i_{revised}} - Data_{i_{original}}\right)\right|}{Data_{i_{original}}} \quad (8)$$

i=day

In the following figure average error for all previous mentioned methods has been calculated by measuring day-to-day distance of new generated data versus original data. According to the finding of the calculation, the HOSB among others is with the average error percentage less than 15.

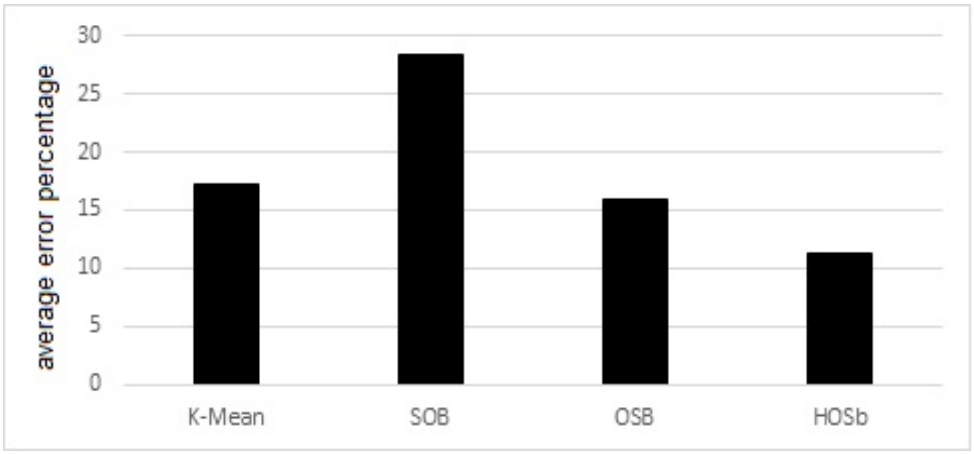

Fig 13. Comparison of the Average error comparing to original data

The second parameter will be satisfied by calculating root mean square of the whole 12 months' data. Root Mean Square formula is:

$$RMS=\sqrt{\sum_{i=1}^{365}\frac{\left(data_{revised}-data_{original}\right)^2}{365}} \quad (9)$$

The following figure shows the error rate of each methods using the RMS index. As it can be seen the HOSB method with the least deviation and error at about 2.11% is better than the other methods. Therefore, we find out that the HSOB with high degree of confidence can be introduced to be deployed for all similar case study in maritime transport.

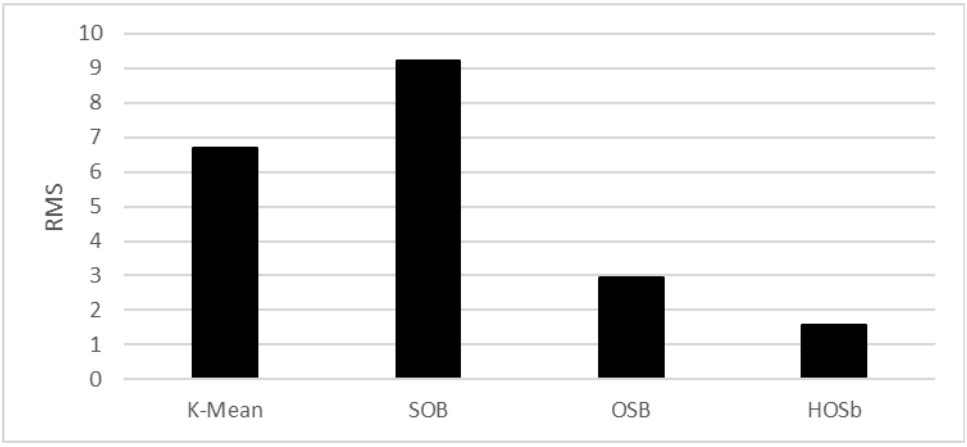

Fig 14. Comparison of RMS average of the used methods vs original data

5. Conclusion

In this study, we focused on the ways of improving vessels NR data. Nowadays NR data is widely used in shipping companies to control and study different parts of their business in order to optimize or minimize the expenses such as fuel consumption. However, fuzziness of the data often misleads the owner of the vessel to appraise the reasons for high consumption vessels due to lack of confidence on accuracy of NR data. In this study, four familiar methods (K-Mean, SOM, OSB and HOSB) were reviewed and deployed. Then four VLCCs have been chosen to gather their NR data. 12-months NR have been collected. On the other hand, four mathematical codes based on the governing equations of reported methods developed using MATLAB. Finally, all methods used to refine the data for all vessels and results are reported. Consequently, it has been observed that best method to enrich the NR data for further purpose is HOSB. Among the others, HOSB not only increases the quality and eliminate the fuzziness of the data but also it is observed that the new generated data is in harmony with the original data collected from the NR. In the future, it is proposed to investigate a new method to improve the quality of data if encountering with the data with long interval between two consecutive gathering points e.g. suing Automatic Identification System (AIS) technology to decrease the errors derive from lack of short or proper interval between two data recorded.

References

1) Zelazny, K. (2014) A Simplified Method for Calculating Propeller Thrust Decrease for a Ship Sailing on a Given Shipping Lane. *Polish Maritime Research*, 2(82) Vol 21; pp. 27-33
2) Disel, M. (2011) Costs and Benefits of LNG as Ship Fuel for Container Vessels. *Engineering the Future*.
3) Wang, S. & Meng, Q. (2013) Bunker Consumption Optimization Methods in Shipping: A Critical Review and Extensions. *Transportation Research Part E: Logistics and Transportation Review* Vol 53.
4) Kontovas, C. & Psaraftis, H.N. (2011) Reduction of Emissions Along the Maritime Intermodal Container Chain: Operational Models and Policies. *Maritime Policy & Management* 38 (4), 451–469.
5) Bell, M.G. Liu, X. Rioult, J. Angeloudis, P. (2013) A Cost-Based Maritime Container Assignment Model. *Transportation Research Part B* 58, 58–70.
6) Agarwal, R. & Ergun, Q. (2008) Ship Scheduling and Network Design for Cargo Routing in Liner Shipping. *Transportation Science* 42 (2), 175–196.
7) Meng, Q. & Wang, Y. (2016) Shipping Log Data Based Container Ship Fuel Efficiency Modeling. *Transportation Research Part B: Methodological*, Vol 83.
8) Song, D.P. & Dong, J. (2012) Cargo Routing and Empty Container Repositioning in Multiple Shipping Service Routes. *Transportation Research Part B* 46 (10), 1556–1575.
9) Notteboom, T. & Vernimmen, B. (2012) The Effect of High Fuel Costs on Liner Service Configuration in Container Shipping. *Journal of Transport Geography* 17 (5), 325–337.
10) Alvarez, C. & Troya, J. (2016) Analyzing the Possibilities of Using Fuel Cells in Ships. *International Journal of Hydrogen Energy*, Vol 41, Issue 4.
11) Meng, Q. & Wang, S. (2015) Optimal Operating Strategy for a Long-Haul Liner Service Route. *European Journal of Operational Research*, 105–114.
12) Fagerholt, K. Laporte, G. Norstad, I. (2010) Reducing Fuel Emissions by Optimizing Speed on Shipping Routes. *Journal of the Operational Research Society* 61 (3), 523–529.

13) Meng, Q. & Wang, Y. (2015) Liner Container Seasonal Shipping Revenue Management. *Transportation Research Part B: Methodological* Vol. 82.
14) McCall, J. (2005) Genetic Algorithms for Modelling and Optimization. *Computational and Applied Mathematics.*
15) Scott, M.E. (2012) Smart Voyage Planning Model Sensitivity Analysis Using Ocean and Atmospheric Models Including Ensemble Methods. *Monterey, California. Naval Postgraduate School*.
16) Bole, A. (2014) Chapter 5 – Automatic Identification System (AIS). *Radar and ARPA Manual*.
17) Coello, J. (2015) An AIS-Based Approach to Calculate Atmospheric Emissions from the UK Fishing Fleet. *Atmospheric Environment*, Vol 114.
18) Shelmerdine, R. L. (2015) Teasing Out the Detail: How our Understanding of Marine AIS Cata Can Better Inform Industries Developments and Planning. *Marine Policy*, Vol 54.
19) Roh, M. (2017) Determination of an Economical Shipping Route Considering the Effects of Sea State for Lower Fuel Consumption. *Naval Architect*, p. 17.
20) Safaei, A. Ghassemi, H. Ghiasi, M. (2015) Voyage Optimization for a Very Large Crude Carrier oil Tanker: A Regional Voyage Case Study. *Scientific Journal of maritime University of Szczecin*, Vol.44, pp83-89.