POLISH MARITIME RESEARCH 1 (97) 2018 Vol. 25; pp. 44-50
10.2478/pomr-2018-0005

OUTLIER DETECTION IN OCEAN WAVE MEASUREMENTS BY USING UNSUPERVISED DATA MINING METHODS

Kumars Mahmoodi, PhD student
Hassan Ghassemi, Prof.
Amirkabir University of Technology, Tehran, Iran

ABSTRACT

Outliers are considerably inconsistent and exceptional objects in the data set that do not adapt to expected normal condition. An outlier in wave measurements may be due to experimental and configuration errors, technical defects in equipment, variability in the measurement conditions, rare or unknown conditions such as tsunami, windstorm and etc. To improve the accuracy and reliability of an built ocean wave model, or to extract important and valuable information from collected wave data, detecting of outlying observations in wave measurements is very important. In this study, three typical outlier detection algorithms:Box-plot (BP), Local Distance-based Outlier Factor (LDOF), and Local Outlier Factor (LOF) methods are used to detect outliers in significant wave height (H_s) records. The historical wave data are taken from National Data Buoy Center (NDBC). Finally, those data points are considered as outlier identified by at least two methods which are presented and discussed. Then, H_s prediction has been modelled with and without the presence of outliers by using Regression trees (RTs).

Keywords: ocean wave data, data mining, outlier detection, data correction

INTRODUCTION

Appropriate data collection with the least uncertain instances is very important in the field and experimental researches. That is because such data are usually used for different objectives e.g. validation and calibration of numerical and mathematical models. Outliers affect the analysis and their results and also cause perturbation in process of concluding. Also they contain useful information underlying the abnormal behaviour. The data can arise from several different mechanisms or causes, such as human errors, errors in data recording or collection, environmental conditions, sampling errors, unusual phenomena in measurement conditions, faulty or non-calibrated equipment, Incorrect assumptions about the distribution of the data [1], natural variations in the population and etc. So detecting of outlying observations in collected data is necessary. After detection of outliers, they can be removed or corrected.

There are different instruments for measuring sea data. These instruments are able to gather the metrological and oceanography data such as speed, direction and duration time of wind, air and water temperature, air pressure and humidity, density and salinity of sea water, and historical sea water level changes. Wave buoys are commonly used instruments to this end [2]. The devices can measure important wave data, such as height and period. Various factors may cause outliers in the buoys' measured values. It is necessary to detect outliers and remove or correct them before extracting any information from collected data.

Outlier detection in wave measurements can be used for automatic identification of hurricanes and typhoons [3, 4], identification of areas with the influence of storms, which should be further studied by using appropriate models with higher accuracy, detect potential height wave energy resources, awareness of ocean dynamics and climatic variability, definition of operable conditions in shipping routes, maintenance and repair strategies for offshore constructions, extreme wave analysis [5], eliminate data related to malfunctions of buoys, so the results will be accurate enough to make them suitable for ocean studies etc. [6].

The main aim of this study is to detect outliers in the historical data taken from buoys by using some unsupervised data mining approaches. Unsupervised outlier detection methods do not require training data, and assume that normal instances are far more frequent than outliers. So these techniques are most useful, when training data is not available. The discussed methods are: Box-plot, LDOF, and LOF. Here, outliers in measured significant wave heights (H_s) are selected as a case study. H_s is an average measurement of the largest one-third of wave heights, which is a useful way to describe the sea state. This parameter is the basis for many computations in the coastal and marine engineering. Therefore its correct measuring is very important.

OUTLIER DETECTION APPROACHES

Outlier detection, also known as anomaly detection or data cleansing [7], is an important research problem in data mining and a pre-processing step in any data analysis application that aims to discover useful abnormal and irregular patterns hidden in data sets [8, 9]. The discussed methods are presented below.

BOX-PLOT (BP)

Box-plot test [10] is the commonly used outlier test for normal distributions. It is a useful type of graph used to show the shape of distribution by using the five major attributes: smallest non-outlier observation (*Min*), lower quartile (*Q*1), median, upper quartile (*Q*3), and largest non-outlier observation (*Max*). The quantity *Q*3 – *Q*1 is called the Inter Quartile Range (*IQR*). A box-plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. In this method a data point will be labelled as an outlier if it is located 1.5 × *IQR* times lower than *Q*1 or 1.5 × *IQR* times higher than *Q*3.

LOCAL DISTANCE-BASED OUTLIER FACTOR (LDOF)

LDOF [11] is a distance – based method which uses the relative location of an object to its neighbours to determine the degree to which the object deviates from its neighbourhood. In this method only the objects with the highest LDOF values are regarded outliers. LDOF implementation includes the following steps:

1) Find -nearest neighbours distance of object *p*: it equals the average distance from *p* to all objects in $N_k(p)$. The *k*-nearest neighbours distance of object *p* is defined as:

$$\bar{d}(p) = \frac{1}{k} \sum_{q \in N_k(p)} dist\ (p, q) \qquad (1)$$

where *p*, and *q* are some data points in the data set *D*, *dist*(*p, q*) denotes the distance between point *p* and *q* (in this research Euclidean distance), $N_k(p)$ is the set of the *k*-nearest neighbours of object *p* (excluding *p*), and *k* is a user-specified parameter which is selected according to the type of problem and nature of data set by trial-and-error method, but it is beneficial to use a large neighbourhood size *k*. kNN algorithm [12] is used to find nearest neighbours of object *p*. In data mining, the kNN algorithm is a very useful non-parametric method used to analyze a data object with respect to its nearest neighbours..

2) Find kNN inner distance of object *p*: given $N_k(p)$, the *k*-nearest neighbours inner distance of *p* is defined as the average distance among objects in $N_k(p)$:

$$\bar{D}(p) = \frac{1}{k(k-1)} \sum_{q,q' \in N_k(p)} dist\ (q, q') \qquad (2)$$

3) Find LDOF of object *p*: the local distance-based outlier factor of *p* is defined as:

$$LDOF_k(p) = \frac{\bar{d}(p)}{\bar{D}(p)} \qquad (3)$$

If $LDOF_k(p)$>1, it means that *p* is outside its neighbourhoods and can be an outlier candidate.

LOCAL OUTLIER FACTOR (LOF)

LOF is the first major density-based outlier detection method proposed by Kriegel and Ng [13]. It is possible to detect local outliers by assigning an outlier score (LOF) to any given data point depending on its distance from its local neighbourhood. Assume that *p*, *q* and are some data points in the data set *D*, LOF is computed in following procedure:

1) Find *k*_distance of object *p*: given a positive integer user-specified parameter *k*, *k_distance(p)*, is defined as the distance between object *p* and object *o*, denoted with *dist*(*p*,*o*), such that:
   i) for at least *k* objects *o'* ∈ *D*\{*p*} it holds that *dist*(*p, o'*) <u>≤</u> *dist*(*p, o*), and
   ii) for at most *k*–1 objects *o'* ∈ *D*\{*p*} it holds that *dist*(*p, o'*) <_*dist*(*p, o*).
2) Find *k*_distance neighbourhood of object *k_distance neighbourhood of p* contains every object whose distance from *p* is not greater than the *k_distance*, i.e:

$$N_k(p) = \{q \in D\{p\} | dist(p, q) \leq k\_distance\} \qquad (4)$$

3) Find reachability distance of object p w.r.t. object o: The reachability distance of object p with respect to object o is defined as:

$$reach_{dist_k}(p,o) = \ max\ \{k\_distance(o), dist(p,o)\} \quad \textbf{(5)}$$

4) Find local reachability density of object p: the local reachability density of p is defined as:

$$lrd_k(p) = 1/\left(\frac{\sum_{o\epsilon N_k(p)} reach_{dist_k}(p,`o)}{|N_k(p)|}\right) \quad \textbf{(6)}$$

5) Find local outlier factor of object p: the $LOF(p)$ is defined as:

$$LOF_k(p) = \frac{\sum_{o\epsilon N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \quad \textbf{(7)}$$

The LOF of an object reflects the density contrast between its density and those of its neighbourhood. So $LOF_k(p)$ captures the degree of a point being an outlier. If the LOF value becomes larger, then the degree of outlierness will be risen. In general, if $LOF_k(p)$ were be close to 1, then object p is normal; and if $LOF_k(p)$ were be larger than 1, object p is outlier candidate.

DATA AND STUDY AREA

The historical wave data were taken from National Data Buoy Center (http://www.ndbc.noaa.gov). Tab. 1 shows the main characteristics of four buoys considered in this study at the Western coast of the USA. All historical data were collected in the year 2015. Missing records are removed from all the studied stations. The statistics and histogram bar plot with 20 bins of H_s data sets are presented in Tab. 2 and Fig. 1.

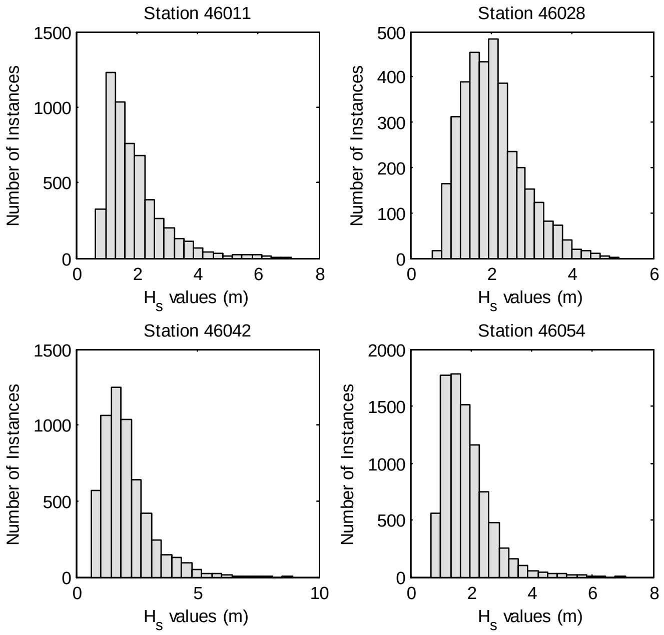

Fig. 1 Histogram bar plot of H_s with 20 bins of all the studied stations

EXPERIMENTAL RESULTS

In this section, the effectiveness of the three above mentioned outlier detection methods when applied on H_s data sets, is investigated. At first, outliers in all the data sets are detected individually by each method, then those data points are considered as outlier identified by at least two methods. All algorithms were implemented in Matlab software. The results of Box-plot method are shown in Tab. 3.

Tab. 1. Buoy's geographic coordinates and description

Characteristics	Station 46011 (34°57'22" N 121°1'7" W)	Station 46028 (35°42'42" N 121°51'30" W)	Station 46042 (36°47'29" N 122°27'6" W)	Station 46054 (34°15'53" N 120°28'37" W)
Site elevation	sea level	sea level	sea level	sea level
Air temperature height	4 m above site elevation	4 m above site elevation	4 m above site elevation	4 m above site elevation
Anemometer height	5 m above site elevation	5 m above site elevation	5 m above site elevation	5 m above site elevation
Barometer elevation	sea level	sea level	sea level	sea level
Sea temperature depth	0.6 m below water line	0.6 m below water line	0.6 m below water line	0.6 m below water line
Water depth	464.8 m	1036 m	2098 m	382.3 m

Tab. 2. Statistics of H_s(m) data sets corresponding to all the studied stations

Station	Number of instances	Max	Min	Mean	Median	Std
46011	5355	7.08	0.66	1.9098	1.66	0.9308
46028	3593	5.14	0.56	2.0372	1.95	0.7468
46042	5714	8.92	0.63	2.0724	1.87	0.9865
46054	8700	7.08	0.69	1.8481	1.69	0.7610

To implement LDOF to data sets it is necessary to determine the neighbourhood size k. Based on the rule for selecting k value, suggested by Zhang and *et al.*, $k = 150$ was assumed in all data sets. The data points with $LDOF > 1$ (Eq. 3) are considered outliers. Fig. 2 demonstrates the LDOF coefficient values in all the data sets along with the threshold limit value (horizontal line). The data points falling above the horizontal line have been considered the outlier candidates. The experimental results are listed in Tab. 3.

Similar LDOF method, to implement LOF for data sets requires to determine the neighbourhood size *k* to compute the density in the neighbourhood of data points. The value of this parameter is application-dependent and selected based on the nature of studied data sets. A heuristic method is proposed to pick the right k values for the LOF computation [13]. Its authors provided several guidelines for picking the range of *k* values. Following such guidelines, *k* = 100 is selected for lower bound and *k* = 500 for upper bound in the experiments. Fig. 3 shows the mean of data point LOF values of all data sets, with increasing *k* at the step of 20. It can be seen from this figure that for all the data sets the mean LOF values change a little and is almost stable when the *k* value is higher than 220. In fact, if *k* value is selected from the range of [150, 500], the mean LOF values and thus the results will not change much. For this reason and to reduce the amount of calculations and computation time, *k* is set to 220 for all data sets. In the LOF method, the data points with *LOF* > 1 (Eq. 7) are considered outliers. Here, the number 1.9 is considered a threshold for Stations 46011, 46028 and 46042, and number 2.15 - for Station 46054 by trial and error method. A data point is labelled an outlier if its LOF coefficient exceeds the thresholds. Tab. 3 provides the results of LOF method applied to the data sets. LOF coefficients of data sets are plotted in Fig. 4. The data points falling above the horizontal line have been considered the outlier candidates.

COMPRESSION OF RESULTS

As mentioned above, outliers are patterns which deviate from an expected normal behaviour. This definition looks simple but is highly challenging because it is difficult to define what is the normal behaviour or a normal region. Some of the difficulties are: the uncertainty in the exact boundary between normal and abnormal behaviour, the absence of a comprehensive definition of outliers, various definitions of outliers in the different science fields and applications, natural behaviours of the studied phenomenon in certain circumstances which tends to be similar to the actual outliers, availability of labelled data, and other additional factors. Due to the challenges, the outlier detection problem is not easy and is usually problematic. For this reason, most of the existing outlier detection techniques lead to different results, based on their formulation, definition of outliers and type of outliers to be detected. This matter is also observed in this study because different approaches produce different results, as seen in Tab. 3. Based on this table, Box-plot method detected more objects as being outlier; In fact, the sensitivity of this method was higher than other methods. However, depending on the parameters of each method, their sensitivity may also change. For example in LOF method, if threshold parameter is selected lower, more objects will be selected as being outlier. However, it is possible that some normal data are considered to be outliers by this method. Hence the right choice of each method parameters is always very important and may be changing the results. To detect potential outliers, it is better to consider data points to be outlier candidates which are detected by most methods. In fact, such data are most likely to be considered outliers. In this research the voting method is used to better detect the outliers . Voting is not a new method and uses the results of other methods to detect outliers. In the voting method, outliers are data points which have been selected as the outlier by most methods. As a result, the voting method leads to more accurate and reliable results. To implement voting into results of the discussed outlier detection methods, those data points are considered as outlier which have been identified by at least two methods. The voting results are presented in Tab. 3. The final outlier candidates detected in all data sets, obtained by voting, are presented in Fig. 5. In this figure detected outliers are distinguished with circle marks.

After detecting outliers the reason of their occurrence should be carefully investigated because outliers in ocean wave measurements result from various reasons such as meteorological events like windstorm, or fail to function properly in sensor data streams, rare phenomena like tsunami, defects occurring in measurement devices and data transmission, etc. This way, valuable information can be achieved from the detection of outliers.

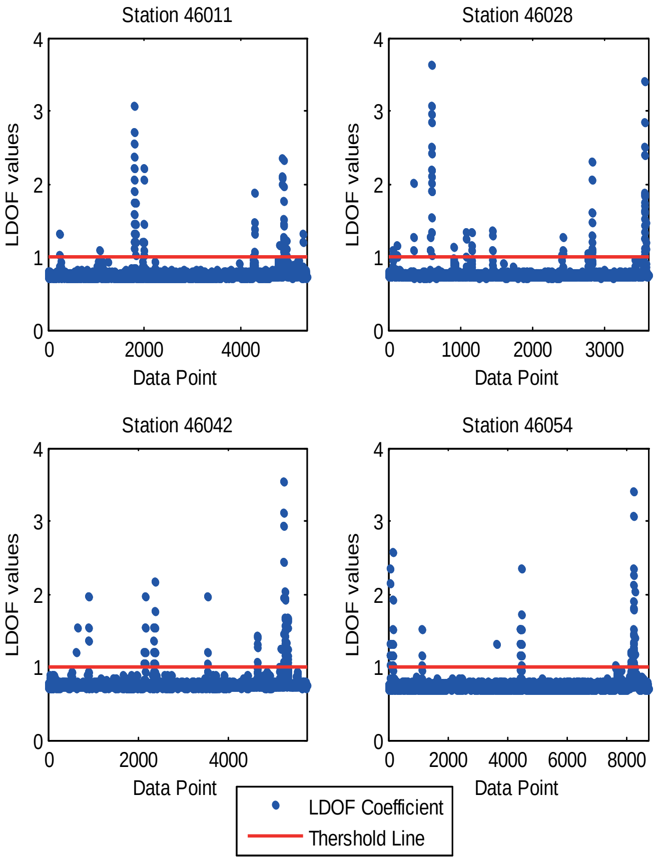

Fig. 2 LDOF coefficients for all the studied stations

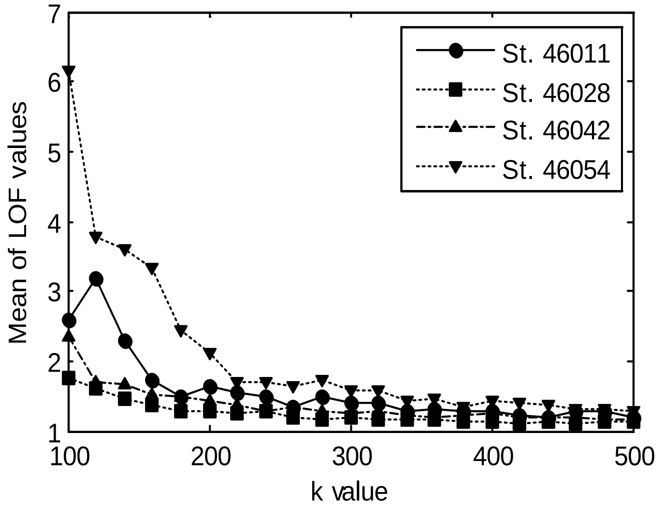

Fig. 3 Mean of data point LOF values with different k, for all the studied stations

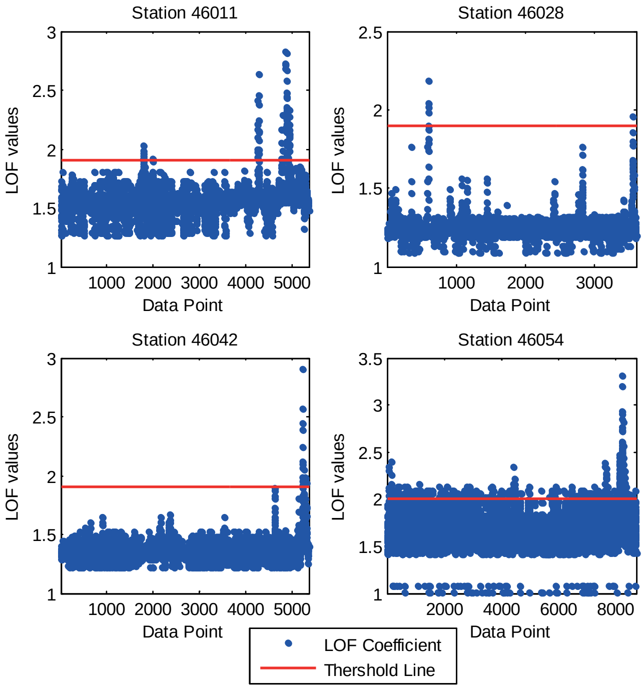

Fig. 4 LOF coefficients of all the studied stations

Tab. 3. Number of outliers in H_s data sets, detected by all methods

Station	BP	LDOF	LOF	Voting
46011	271	69	81	69
46028	69	78	6	1
46042	269	82	18	48
46054	299	67	83	71
Sum	908	296	188	189

SIGNIFICANT WAVE HEIGHT PREDICTION

Outlier detection is one of the major issues in preparing the data for data mining classification and prediction problems. In this study to show the importance of outlier detection in the ocean wave studies, significant wave height prediction using wave parameters is considered with and without the presence of outliers. Regression trees (RTs) [14] are used to model H_s. Tab. 4 shows details of the predictive variables for this problem. In all created models, the output variable is WVHT (H_s), and other variables are considered inputs. In this research two error criterion measurements are considered in order to evaluate the performance of the created models, namely , the Root Mean Square Error (*RMSE*) and the Pearson's correlation coefficient (R^2), according to the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i - O_i)^2} \quad (8)$$

$$R = \frac{\sum_{i=1}^{n}(T_i-\bar{T})(O_i-\bar{O})}{\sqrt{\sum_{i=1}^{n}(T_i-\bar{T})^2\sum_{i=1}^{n}(O_i-\bar{O})^2}} \quad (9)$$

where n represents the total number of instances, while T_i and O_i represent the experimental and predicted values, respectively; $\bar{T}$ and $\bar{O}$ are the average values of these data.

Regression trees are the well-known predictive modelling approaches in data mining, which are built through a process known as binary recursive partitioning. It is necessary to set a series of parameters for the training of regression trees, such as tree depth etc. The optimal architecture of the developed RTs is presented in Tab. 5. The choice of the tuning parameters comes from preliminary tests carried out on the studied data sets. RTs results are presented in Tab. 6. Also in this case the experimental results demonstrated that after removing outliers the accuracy of the created models is increased.

Tab. 4. Variables used in the experiments

Acronym	Predictive variable	Unit
WDIR	Wind direction	[degree]
WSPD	Wind speed	[m/s]
GST	Gust speed	[m/s]
WVHT	Significant wave height	[m]
DPD	Dominant wave period	[sec]
APD	Average period	[sec]
PRES	Atmospheric pressure	[hPa]
ATMP	Air temperature	[°C]
WTMP	Water temperature	[°C]

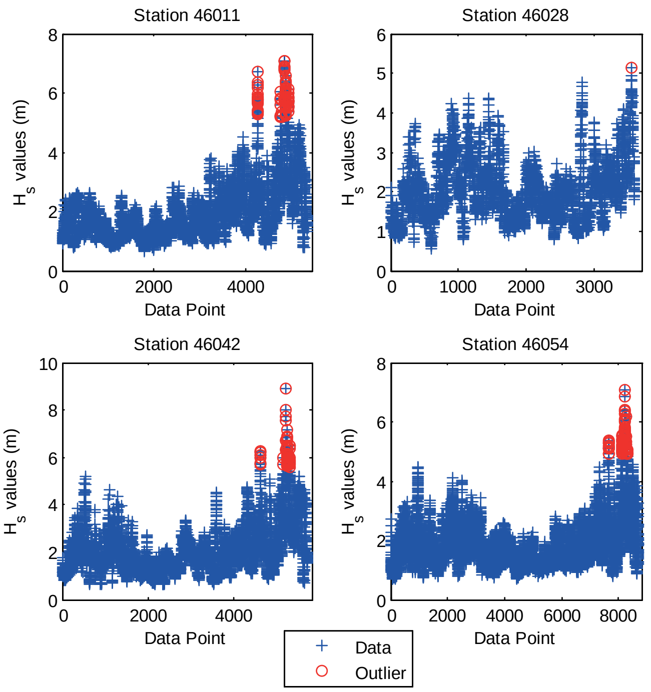

Fig. 5 Final outliers detected in all the studied stations by using voting method

Tab.5. RTs selected structure for all the studied stations

Parameter	Setup
Tree maximum depth	Automatic
Minimum leaf	1
Minimum parent	10
Split Criterion	Mean Squared Error

Tab. 6. Comparative performance of the H_s estimation by the RTs with and without the presence of outliers

Station	With outliers		Without outliers	
	RMSE (m)	*R*^{2}	*RMSE* (m)	*R*^{2}
46011	0.1530	0.9697	0.1410	0.9729
46028	0.1434	0.9629	0.1433	0.9630
46042	0.1709	0.9651	0.1697	0.9699
46054	0.1375	0.9640	0.1310	0.9673
Mean	0.1512	0.9654	0.1462	0.9683

CONCLUSIONS

In this paper, the outlier detection problem in data of the significant ocean wave height H_s was presented and discussed. Three outlier detection approaches, i.e. the Box-plot, LDOF, and LOF, were described and their performance were compared. Box plot can be used for one- dimensional data sets, while LDOF and LOF are methods for identifying outliers in multi-dimensional data sets , which utilize data density estimation concepts. Based on the experiments, the mentioned methods presented different results. Each method has several parameters which specify its performance. Input parameters have great influence on the outlier detection performance and should be carefully selected based on the nature of studied data sets. Based on the obtained results it can be stated that the LOF and Box plot were of low and high sensitivity in outlier detection in studied data sets, respectively. In this research the voting method was used to achieve a better outlier identification. The experiment shows that the voting method can achieve a high performance in detecting outliers, compared to other methods. In general, outliers are caused due to an error or rare behaviour of studied phenomenon containing valuable information about some unexpected events. For example, it is possible to detect tsunami by recognizing the outliers. In H_s measurements the outlying observations are mostly related to the presence of typhoons and/or hurricanes, which must be removed to avoid wrong analysis of incorrect results and create accurate analytical and numerical models. To demonstrate the effect of outliers in wave measurement data collection, the significant wave height was modelled based on metrological and wave parameters. The results showed that the accuracy of the created models increased in case of absence of outliers. So far, many outlier detection methods which can be used to detect outliers in ocean wave measurements have been proposed by researchers. Experimental results demonstrated that the proposed approach is capable to better detect outliers. It is suggested that outliers should be identified before any analysis and deduction of data sets by using the presented procedure. A future direction of research is to apply this procedure to multi-dimensional data sets of sea variables with the use of different outlier detection methods.

REFERENCES

1. Iglewicz, B., Hoaglin, D.C.: *How to detect and handle outliers*. Milwaukee, WI.: ASQC Quality Press, 1993.

2. Sun S. Z., LI, H., Sun, H. : *Measurement and analysis of coastal waves along the north sea area of China*. Polish Maritime Research, 3 (91) 2016, 23, pp. 72-78.

3. Whan Lee, J., Park, S. C., Kee Lee, D., Ho Lee, J. : *Tsunami arrival time detection system applicable to discontinuous time-series data with outliers*. Journal of natural hazards and earth sciences, 2016, 16 (12), pp. 2603-2016.

4. Mínguez, R., Reguero, B.G., Luceño, A., Méndez, F.J. : *Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases*. Journal of Atmospheric and Oceanic Technology, 2012, 29, pp. 267–285.

5. Lucas, C., Muraleedharan, G., Soares, C. G. : *Outliers identification in a wave hindcast dataset used for regional frequency analysis*. Maritime Technology and Engineering, 2015, pp. 1317-1327.

6. Reguero, B.G., Menéndez, M., Méndez, F.J., Mínguez, R., Losada, I. J. : *A Global Ocean Wave (GOW) calibrated reanalysis from 1948 onwards*. Coastal Engineering, 2012, 65, pp. 38–55.

7. Chandola. V., Banerjee, A., Kumar, V. : *Anomaly detection – a survey.* ACM Comput Surv. 2009, 4 (3), pp. 1–58.

8. Barnett, V., Lewis, T. : *Outliers in Statistical Data*. John Wiley, 3rd edition 1994.

9. Zhang, Ji. : *Advancements of Outlier Detection: A Survey*. ICST Transactions on Scalable Information Systems, 2013, 13 (1), pp. 1-26.

10. Muraleedharan, G., Lucas, C., Guedes Soares, C.: *Regression quantile models for estimating trends in extreme significant wave heights*. J. Ocean Engineering. 2016, 118, pp. 204–215.

11. Zhang, K., Hutter, M., Jin, H. : *A new local distance-based outlier detection approach for scattered real-world data*. Proc. 13th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, 2009, pp. 813-822.

12. Chen, Y., Miao, D., Zhang, H. : *Neighborhood outlier detection*. Expert Systems with Applications, 2010, 37 (12), pp. 8745-8749.

13. Breunig, M. M., Kriegel, H.-P., Ng, R. T., *et al.*: LOF: *Identifying density-based local outliers.* In W. Chen, J. F. Naughton, & P. A. Bernstein (Eds.), Proceedings of the ACM SIGMOD international conference on management of data, ACM Press , Dallas, Texas , 2000, pp. 93–104.

14. Troncoso, A., Salcedo-Sanz, Casanova-Mateo, S., Riquelme, J.C, C., Prieto, L. : *Local models-based regression trees for very short-term wind speed prediction*. Renewable Energy, 2015, 81, pp. 589-598.

CONTACT WITH THE AUTHORS

Hassan Ghassemi
e-mail: gasemi@aut.ac.ir
Department of Maritime Engineering
Amirkabir University of Technology
Hafez avenue, 14717 Tehran
Iran

Kumars Mahmoodi
e-mail: kumarsmahmoodi@aut.ac.ir
Department of Maritime Engineering
Amirkabir University of Technology
Hafez avenue, 14717 Tehran
Iran