International Congress on Environmental Modelling and Software

4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain - July 2008

Jul 1st, 12:00 AM

Prediction of Significant Wave Height Based on Regression Trees

A. Etemad-Shahidi

J. Mahjoobi

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Prediction of Significant Wave Height Based on Regression Trees

A.Etemad-Shahidi[a]***, J.Mahjoobi*** [b]

[a] *College of Civil engineering,Iran University of Science and Technology*
(etemad@iust.ac.ir)
[b] *College of Civil engineering,IranUniversity of Science and Technology*
(jmahjoobi@gmail.com)

Abstract: One of the most important factors in design of coastal and offshore structures is significant wave height. Thus, an accurate prediction of wave height is of great importance. In this paper, an alternative approach based on regression trees was applied for prediction of significant wave height. The data set used in this study comprises of wave and wind data gathered from deep water location in Lake Michigan, from 15 September to 10 December, 2002. In this study the data set was divided into two groups. The first one that comprises of 58 days (1392 data point) wind and wave measurement was used as training data to develop the regression tree. The second one that comprises of 29 days (686 data point) wind and wave measurement was used as testing data to verify the model. Wind speeds belonging up to six previous hours were given as input variables, while the significant wave height *(H_s)* was the output parameter. *CART* algorithm was employed for building and evaluating regression trees and outputs of models with different lags were compared. Result showed that regression trees can be used successfully for prediction of H_s. In addition it was found that error statistics of the models for prediction of H_s decrease as wind speed lag increases. Finally, the results of *CART*-based model, was compared with artificial neural networks, Results indicated that error statistics of neural networks were marginally more accurate than regression trees.

Keywords: Wave Prediction; Decision trees; Regression Trees; CART Algorithm, neural networks

1. Introduction

The estimation of significant wave height is essential for almost any engineering activity in the ocean. Different methods such as empirical, numerical and soft computing approaches have been proposed for significant wave height prediction. Recently, artificial neural networks have been widely used to predict wave parameters [e.g. Makarynskyy et al. 2005, Agrawal and Deo 2002, Makarynskyy 2004]. A review of neural network applications in ocean engineering is given in Jain and Deo, [2006]. Recently, other soft computing techniques such as Fuzzy Inference System (FIS) and Adaptive-Network-based Fuzzy Inference System (ANFIS) have been used to develop wave prediction models (e. g. Kazeminezhad et al. 2005, Ozger et al. 2007). These studies have shown that the wind speed is the most important parameter in wave parameters prediction. Prediction of significant wave height is basically an uncertain and random process and hence is difficult to accomplish by using deterministic equations. Therefore, it is ideally suited to regression trees since it is primarily aimed at recognition of a random pattern in a given set of input values. Regression trees are useful to model a random input with the corresponding random output and their application does not require knowledge of the underlying physical process as a precondition. In this paper, regression trees were invoked for prediction of significant wave height using different lags. For this purpose, *CART* algorithm was employed for building and evaluating regression trees.

2. Regression Trees

Decision trees are powerful and popular tools for classification and prediction. The advantage of decision trees is due to the fact that, decision trees represent rules. Rules can readily be expressed so that humans can understand them. A decision tree is a tree in which each non-leaf node is labelled with an attribute or a question of some sort, and in which the branches at that node correspond to the possible values of the attribute, or answers to the question. Leaf nodes are labelled with a class. Decision trees are used for classifying instances. One starts at the root of the tree, and taking appropriate branches according to the attribute or question asked about at each branch node, one eventually comes to a leaf node. The label on that leaf node is the class for that instance. The Classification and Regression Trees (*CART*) method of Breiman et al. [1984] generates binary decision trees. Regression tree building centers on three major components: (1) a set of questions of the form, Is $X \le d$? Where *X* is a variable and *d* is a constant. The response to such questions is yes or no; (2) goodness of split criteria for choosing the best split on a variable; and (3) the generation of summary statistics for terminal nodes. Here, the least squared deviation (*LSD*) impurity measure is used for splitting rules and goodness of fit criteria**.** The *LSD* measure *R(t)* is simply the weighted within node variance for node *t*, and it is equal to the resubstitution estimate of risk for the node. It is defined as:

$$R(t) = \frac{1}{N_W(t)} \sum_{i \in t} \omega_i f_i \left(y_i - \bar{y}(t)\right)^2 \tag{1}$$

where $N_W(t)$ is the weighted number of records in node *t*, ω_i is the value of the weighting field for record *i* (if any), f_i is the value of the frequency field (if any), y_i is the value of the target field, and $\bar{y}(t)$ is the mean of the dependent variable (target field) at node *t*. The *LSD* criterion function for split s at node t is defined as:

$$Q(s,t) = R(t) - R(t_L) - R(t_R) \tag{2}$$

where $R(t_R)$ is the sum of squares of the right child node, and $R(t_L)$ is the sum of squares of the left child node. The split s is chosen to maximize the value of *Q(s, t)*.

Stopping rules control how the algorithm decides when to stop splitting nodes in the tree. Tree growth proceeds until every leaf node in the tree triggers at least one stopping rule. Any of the following conditions will prevent a node from being split:

a. All records in the node have the same value for all predictor fields used by the model.
b. The number of records in the node is less than the minimum parent node size.
c. If the number of records in any of the child nodes resulting from the node's best split is less than the minimum child node size.
d. The best split for the node yields a decrease in impurity that is less than the minimum change in impurity.

In regression trees, each terminal node's predicted category is the weighted mean of the target values for records in the node. This weighted mean is calculated as:

$$\bar{y}(t) = \frac{1}{N_W(t)} \sum_{i \in t} \omega_i f_i y_i \tag{3}$$

where $N_W(t)$ is defined as:

$$N_W(t) = \sum_{i \in t} \omega_i f_i \tag{4}$$

3. Employed Data

The data set used in this study comprises of wave and wind data gathered from deep water location in Lake Michigan, from 15 September to 10 December, 2002. The data set was collected by National Data Buoy Center (NDBC) in station 45007 at 42° 40´ 30´´ N and 87° 01´ 30´´ W (Figure 1), where water depth is 176.4 m. Wind and wave data were collected using 3-meter discus buoy at 1-hour intervals. The wind speed at buoy was measured at a height of 5 meter above the mean sea level. Tables 1 and 2 show ranges and average values of different parameters of training and testing data sets.

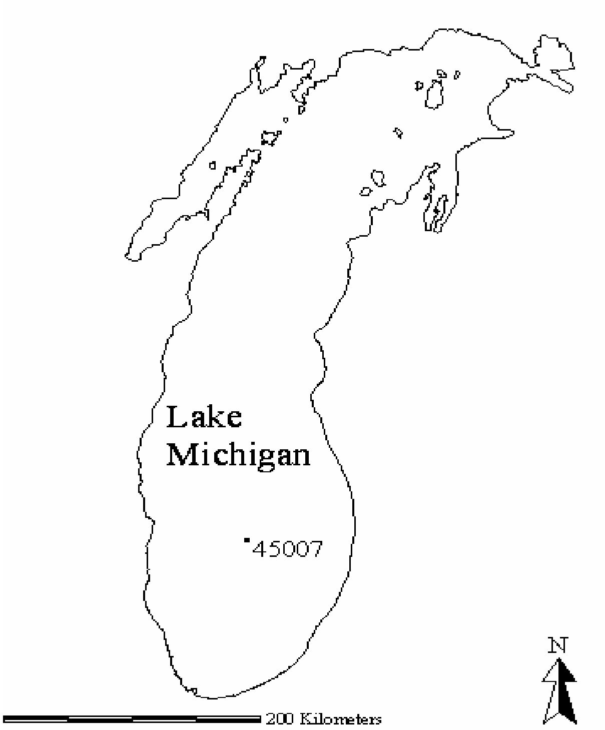

Figure 1. Lake Michigan bathymetry and location of *NDBC* buoy 45007 located at 43°37´09´´ N and 77°24´18´´ W.

Table 1: Ranges and average values of different parameters in training data

Parameter	Range	Average
Wind speed (m/s)	0.1-16.6	7.19
Significant wave height (m)	0.15-3.43	1

Table 2: Ranges and average values of different parameters in testing data

Parameter	Range	Average
Wind speed (m/s)	0.7-16.5	7.84
Significant wave height (m)	0.17-3.36	1.19

4. Building and Evaluating Regression Trees

Here the data set was divided into two groups. The first one that comprises of 58 days (1392 data point) wind and wave measurement was used as training data to develop the regression tree. The second one that comprises of 29 days (686 data point) wind and wave measurement was used as testing data to verify the model. *CART* algorithm was employed to build the regression tree. Wind speeds belonging up to six previous hours were given as input variables, while the significant wave height (H_s) was the output parameter. SPSS Clementine software was used to apply *CART (http://www.spss.com/clementine/).*

The performances of models were evaluated using three statistical measures: (1) bias, which shows the mean error;

$$bias = \bar{y} - \bar{x} \qquad (5)$$

(2) Scatter index (*SI*), is the root mean square error normalized by the mean of observed values of the reference quantity;

$$SI = \frac{\sqrt{\frac{1}{n}\sum\left((y_i - \bar{y}) - (x_i - \bar{x})\right)^2}}{\bar{x}} \qquad (6)$$

and (3) Coefficient of correlation (*R*), which is a measure of strength of the linear relationship developed by a model.

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (7)$$

where x_i is an observed value, y_i is a predicted value and n is the number of observations, $\bar{x}$ is the mean of x and $\bar{y}$ is the mean of y.

Time series plots of the significant wave height obtained from observations and predicted by the regression trees are displayed in Figures 2 and 3. As seen, the model prediction of H_s matched very well with the observed data.

Table 3 shows the error statistics of models with different wind speed lags (for testing data). As can be seen, the models slightly underestimate significant wave heights (bias=-0. 1 m) in the studied case. Results also indicate that error statistics of the models for prediction of H_s decrease as wind speed lag increases.

Table 3: Error statistics of prediction significant wave height by regression trees

Wave parameter/ model	bias	*SI* (%)	*R*
H_s (m) / 6hr- time lag	-0.08	22.56	0.926
H_s (m) / 5hr- time lag	-0.095	24.71	0.922
H_s (m) / 4hr- time lag	-0.092	24.72	0.922
H_s (m) / 3hr- time lag	-0.1	25.23	0.92
H_s (m) / 2hr- time lag	-0.1	25.75	0.916
H_s (m) / 1hr- time lag	-0.106	27.8	0.9
H_s (m) / no- time lag	-0.109	30.94	0.86

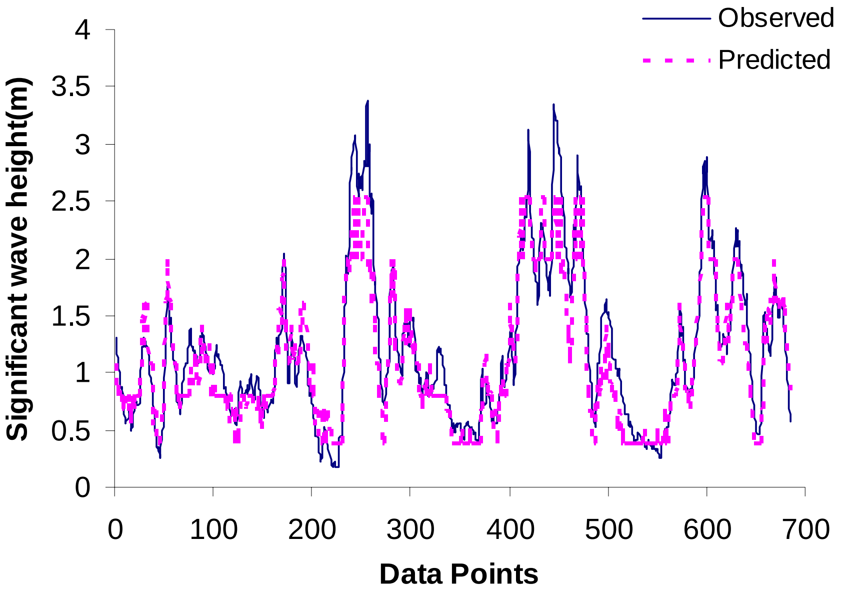

Figure 2. Time series of measured and predicted (6 hour wind speed lag) significant wave heights.

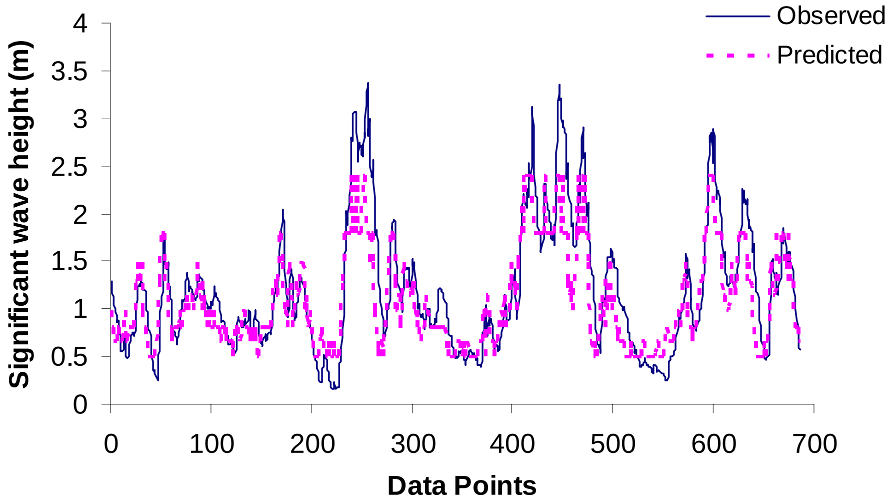

Figure 3. Time series of measured and predicted (0 hour wind speed lag) significant wave heights.

In addition, for comparison between regression trees and artificial neural networks, a three-layer feed-forward neural network (the back-propagation network) [Haykin, 1999] with the sigmoid transfer functions was used. Wind speeds belonging up to six previous hours were given as input variables. From 2 to 20 nodes for the hidden layer were examined. The best topology was found to be 7×15×1 (neurons in the input × hidden × output layers). After testing network, Coefficient of correlation, scatter index and bias were obtained 0.94, 20.96 and -.092 m, respectively. Results indicated that artificial neural networks were marginally more accurate than regression trees.

5. Summary and Conclusions

Significant wave height is an important parameter in the design of coastal and offshore structures. In this study, regression trees were used successfully for prediction of significant wave height variation responding to wind forcing. *CART* algorithm was employed for building and evaluating regression trees. The data set used in this paper comprises of wave and wind data gathered from deep water location in Lake Michigan. Wind speeds belonging up to six previous hours were given as input variables. Result show that *CART* algorithm is skilful in prediction of significant wave heights in the studied case. Furthermore, it was found that the error statistics of the models for prediction of H_s decrease as wind speed lag increases. Also, results of regression trees were compared with those of artificial neural networks (ANNs). Results show that ANNs models were marginally more accurate than regression trees. Therefore, the regression trees can be used as a cost effective and easy to use tool for engineers and scientists with much less effort required for the implementation of process-based models.

References

Agrawal J.D., Deo M.C. On-line wave prediction, Marine Structure, 15, 57–74, 2002.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees. Belmont, CA: Wadsworth Statistical Press, 1984.

Haykin S. Neural Networks: a Comprehensive Foundation. 2th Ed. Prentice-Hall, NJ, 842pp, 1999.

Jain, P and M. C. Deo, M.C. Neural networks in ocean engineering, International Journal of Ships and Offshore Structures, 1, 25-35, 2006.

Kazeminezhad M.H., Etemad-Shahidi A., Mousavi S.J. Application of fuzzy inference system in the prediction wave parameters. Ocean Engineering. 32, 1709-1725,2005.

Makarynskyy O. Improving wave predictions with artificial neural networks. Ocean Engineering, 31, 5–6, 709–724, 2004.

Makarynskyy O., Pires-Silva A.A., Makarynska D., Ventura-Soares, C. Artificial neural networks in wave predictions at the west coast of Portugal, Computers & Geosciences,; 31, 4, 415-424,2005.

Ozger M., Sen Z. Prediction of wave parameters by using fuzzy logic approach. Ocean Engineering., 34, 460-469,2007.