NEURAL NETWORKS IN RECONSTRUCTING MISSING WAVE DATA IN SEDIMENTATION MODELLING

Bhattacharya B.1, Shrestha D. L.2 and Solomatine D. P.1
1International Institute for Infrastructural, Hydraulic and Environmental Engineering (IHE), P.O. Box 3015, 2601 DA Delft, The Netherlands.
T: +31 15 215 1841 F: +31 15 212 2921 Email: bha@ihe.nl, sol@ihe.nl
2Department of Hydrology & Meteorology, Babarmahal, Kathmandu, Nepal.

Abstract: Decision support systems for efficient water management and control require the measured time series data such as rainfall, runoff, water level, waves, wind etc. In practical situations these time series often contain gaps that might originate from failures of apparatus or some other reasons. For example the wave data measured at the wave measuring station Europlatform in the North Sea consists of gaps in the wave data time series. This data is of paramount importance for understanding the sediment dynamics in the North Sea along the Dutch coast in general and for assessing the sedimentation in the nearby approach channel of the Port of Rotterdam in particular. An artificial neural network (ANN) has been used for filling in the gaps of the wave data time series from the wind data time series. Average mutual information has been used for determining the lag effect of wind. The direction of wind data has been encoded with a circular coding algorithm. The accuracy of the ANN model in estimating wave heights was found to be reasonable and the model was used to fill in the gaps of the wave data time series for the purpose of subsequent sedimentation modelling, also based on an ANN.

Keywords: missing data, wave, sedimentation, artificial neural network, average mutual information

1. INTRODUCTION

Efficient water management and control depends upon the quality of the measured data. For this purpose data is recorded in time and referred as time series data. This historical data is utilised in decision-making and future planning of water resources systems. Examples include rainfall, runoff, water level, wind, waves, temperature etc. Often these time series are characterised by short breaks in records due to several reasons such as malfunctioning of monitoring equipments, absence of the observer, holidays, power failure, communication line breakdown, disasters (earthquakes, landslides, cyclone) etc. These discontinuities lead to problems in future when a modelling system or a decision support system requires to make use of this measured data. This necessitates filling in the gaps.

Filling in missing data is of high practical motivation and normally the relevant books give an advise to trying to preserve some statistical properties of the original distribution, not providing details on how to do it in practical situations. Abebe et al (1999, 2000) used a fuzzy rule based system for filling in missing precipitation data using available data from nearby gauging stations of the same catchment. The predictive accuracy of the fuzzy rule based model was found to be slightly better than an artificial neural network (ANN) model and a naïve statistical model built with the same data. Wang (2001) used an ANN for predicting missing precipitation data for a catchment in the Netherlands. Khalil et al (1998) presents a
review of several available techniques currently in use for filling in gaps of time series data. Khalil et al (2001) suggested using groups in data based on seasons for building ANN models to predict missing runoff data.

In the present research a wave data time series that is measured in the North Sea offshore of the Dutch coastline is considered. This data is of high importance for understanding the sediment dynamics in the North Sea along the Dutch coast in general and for assessing the sedimentation in the nearby approach channel of the Port of Rotterdam in particular (Vuurens, 2001). Unfortunately, the wave data time series contains gaps and causes difficulty for using wave data as an input parameter in any model for morphological assessment of the fairway leading to the port. Accordingly, there is an urgent need to build a model that can be used for filling in the missing wave data.

An ANN has been used to build a model for predicting wave height when it is unavailable. The input parameters to the model have been selected using average mutual information. The reconstructed wave data time series has been used in developing a sedimentation model of the approach channel for the Port of Rotterdam using an ANN and will be reported elsewhere.

2. DATA AVAILABILITY
Waves are measured every 10 minutes at the wave measuring station Europlatform in the North Sea (see Fig. 1) by a directional wave buoy, called WAVEC, which contains 3-dimensional magnetic acceleration meters. Wave heights, wave period and wave directions are recorded. From this data significant wave heights, average time periods and mean wave directions were computed. During the time period 1991-2000 missing data constitutes about 3% of the total data under consideration. These gaps span mostly for short time duration (several hours) and occasionally for substantial time durations (up to a week).

![Figure 1: Location plan showing the wave and wind measuring station](image-url)
Hourly wind data is collected along the Dutch coast at several locations including the Europlatform. Wind is described by two characteristics: wind-speed and wind-direction. Wind-speed is measured by an anemometer and wind direction is measured by a wind vane in degrees from the North; both at a standard distance of 10 meters above the ground surface. Fortunately, there was no missing data in the wind data time series. Table-1 provides the overview of the available wind and wave data at the Europlatform.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Period</th>
<th>Sampling frequency</th>
<th>Missing data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waves</td>
<td>1991-2000</td>
<td>10 minutes</td>
<td>≈ 3 %</td>
<td>National Institute for Coastal and Marine Management, the Netherlands</td>
</tr>
<tr>
<td>Wind</td>
<td>1991-2000</td>
<td>1 hour</td>
<td>None</td>
<td>Royal Dutch Meteorological Institute</td>
</tr>
</tbody>
</table>

3. RECONSTRUCTING THE WAVE DATA TIME SERIES

The recent advances in the use of machine learning methods such as ANNs, fuzzy logic approach etc. suggested that these methods can be utilised in building a model to predict wave data from other available information (Abebe et al, 2000; Khalil et al, 2001; Bhattacharya and Solomatine, 2000). Once an acceptable model can be developed, it can be used in estimating the missing data. Considering several advantages of an ANN such as its noise handling capabilities etc. we have chosen an ANN for this purpose.

3.1. ARTIFICIAL NEURAL NETWORKS

The development of ANNs was inspired by the studies of the ability of the brain to learn from experience without predefined knowledge of underlying physical relationships. An ANN is a broad term covering a large variety of network architectures, the most common of which is a Multi Layered Perceptron. Such a network is trained by the so-called error-backpropagation method which is a specialised version of a gradient-based optimisation algorithm.

Given a set of input vectors and the associated target (output) vectors, the objective of an ANN is to learn a functional relationship between the input vectors and the target vectors. Each target vector \( z \) is an unknown function \( f \) of the input vector \( x \):

\[
z = f(x)
\]

The task of the network is to learn the function \( f \). The network includes a set of parameters (weights vector), the values of which are varied to modify function \( f' \), which is computed by the network and should be as close as possible to \( f \). The weight parameters are determined by training (calibrating) the ANN based on the training data set. More details about ANNs can be found in Schalkoff (1997) and Haykin (1999).

3.2. SELECTION OF MODEL PARAMETERS

Selecting the right input variables in a data driven model constitutes the most significant part of building a model. One of the possibilities in building a model to predict wave data could be to analyse it as a chaotic time series and to use previous wave data to predict future wave data (for details see Abarbanel, 1996; Solomatine et al, 2000). However, we took another approach of building a model that uses other available information to predict the wave data as well.

At a 30-km distance from the Europlatform there is another wave-measuring station Lichteiland Goere. Unfortunately, in most occasion gaps in data from the two stations coincided. The determining factor for waves is wind and wind data was available for all occasions when wave data was missing. Accordingly, the wind data was chosen as the
model’s input variable. However, the time resolution of the two time series is different – 10 minutes for waves vs 1 hour for the wind. To bring them to the same time frame the wave time series was converted to hourly time resolution by computing hourly average wave data. Thus the task of the model to be built would be to estimate the hourly wave height.

3.3. AVERAGE MUTUAL INFORMATION

For selecting the right input and output variables in a data driven model most often correlation coefficient is used to investigate the linear dependency between two variables. We adopted another approach known as Average Mutual Information (AMI) (see Abebe and Price, 2002). AMI is based on Shannon’s entropy theory and is a measure of information available from one set of data having the knowledge of another set of data (Shannon, 1948). The AMI between two measurements \( a_i \) and \( b_j \) drawn from sets A and B is defined by

\[
I_{AB} = \sum_{a_i,b_j} P_{AB}(a_i,b_j) \log_2 \left( \frac{P_{AB}(a_i,b_j)}{P_A(a_i)P_B(b_j)} \right)
\]

where \( P_{AB}(a_i,b_j) \) is the joint probability density of for the measurements A and B resulting in values \( a \) and \( b \) and \( P_A(a_i) \) and \( P_B(b_j) \) are the individual probability density for the measurements of A and B. If the measurements of a value from A resulting in \( a_i \) completely independent of the measurement of a value from B resulting in \( b_j \) then the average mutual information \( I_{AB} \) is zero.

As a measure of information, the advantage of the AMI measure compared to other approaches such as cross-correlation is that it is independent of any pre-defined function. For discrete measurements the actual AMI-values depend on the number of class intervals used to calculate the probability densities (Abarbanel, 1996).

The AMI measures of waves at the Europlatform with itself and with wind speed at the Europlatform were computed at varying lag times (Fig. 2). AMI helps to find out how much information about the future wave is available from the past wave and wind data. The lag time corresponding to the maximum AMI (0.93) between wind and wave data is 4 hours. The AMI between wind and wave is close to its maximum value (> 0.8) for lag times between 2 to 7 hours. Accordingly, wind data with lag times 2 to 7 hours was used in building the model.

![Figure 2: Average mutual information between wave and wind. The AMI between wind and wave is highest (0.93) for a 4-hour time lag](image-url)
3.4. CIRCULAR CODING OF WIND DIRECTION

Wind is a vector as it has direction as well as magnitude. Special care needs to be taken for coding direction of wind. Intuitively a direction of 3° North is quite close to a direction 357° North but the numbers expressing them are not adjacent. Direction of wind could be coded like a season which rolls endlessly round without a discontinuity between 1° North and 360° North (Pyle, 1999). This can be done by setting the variable to represent direction, say WindDir, as 1 for 0° North and 0 for 180° North. For directions 0° to 180° the variable WindDir may change from 1 toward 0 whereas for the other half, for directions between 180° to 360° from 0 toward 1. The situation is illustrated in Fig. 3 where the value of the variable WindDir is shown for 45°, 135°, 225° and 315° North. All other positions in between these quadrants have a unique value for WindDir. With the help of this diagram the direction of wind was coded without the problem of encountering discontinuities.

3.5. EXPERIMENTAL SET UP

Based upon the AMI measure the following input variables were selected for building the model (referred to as Model-1) to estimate the current wave height:

\[
\text{Inputs : } WS_{t-2}, WS_{t-3}, WS_{t-4}, WS_{t-5}, WS_{t-6}, WS_{t-7}, \text{WindDir}
\]

\[
\text{Output: } W_t
\]

where:
- \(WS\) = wind speed at the Europlatform,
- \(\text{WindDir}\) = average of the wind direction codes (as per Fig. 3) for time span \(t-4\) to \(t\),
- \(W\) = Significant wave height at the Europlatform.

The software NeuroSolutions (www.nd.com) was used for the ANN modelling. Multi-Layered Perceptron (MLP) network trained by back-propagation algorithm was used because of its simplicity and capability to learn non-linear relationship. The hyperbolic tangent function was used for the hidden layer with linear transfer function at the output layer. For ANN training, we used the backpropagation algorithm with momentum rule for 5000 epochs and training was stopped when the mean squared error (MSE) reached its minimum on the cross-validation data set. The first 30000 data points were used for training, the next 5000 data points were used for cross-validation followed by 5000 data points for verification.

The AMI measure in Fig. 2 also suggested that substantial information about the current wave data \((W_t)\) is available in the wave data observed at the immediately previous time step \((W_{t-1})\). This stimulated us to take up another experiment with an additional input as \(W_{t-1}\). The list of
input and output variables considered for building this ANN model (referred to as Model-2) with the same data and training rules is listed below:

Inputs: $W_{t-1}, WS_{t-2}, WS_{t-3}, WS_{t-4}, WS_{t-5}, WS_{t-6}, WS_{t-7}, WindDir$

Output: $W_t$

It is noteworthy that Model-2 was trained with the measured $W_{t-1}$ which is not available in most occasions when the model is used to fill in the data missing during considerable time. To explore this situation, we have built yet another model (referred to as Model-3) by using Model-2 but ensuring that the model uses the predicted value of $W_t$ as $W_{t-1}$ in the next time step.

4. RESULTS AND DISCUSSIONS

Fig. 4 shows the performance of the Model-1 for a part of the verification data. It is seen that the estimated wave height mostly follows the measured one though certain errors can be observed. Table-2 provides the information about the performance of the models in training and verification. It can be observed that the RMSE for Model-1 is about 27 cm whereas the median value of waves is 300 cm. So on an average the estimation error of Model-1 would be about 10% of the median value which though not negligible but was found acceptable for the purpose of subsequent sedimentation modelling.

As can be anticipated from Fig. 2, it is seen from Table-2 that the performance of Model-2 is far better than Model-1. The RMSE is only about 10 cm (compared to 27 cm for Model-1). Analysis of observed data showed that in many cases the gaps in data was for a week. If the Model-2 is used then the estimated value of the wave height has to be used as an input in the next time step. This may lead to accumulation of error that may not remain bounded. This can be observed when Model-2 was used with predicted wave height as an input (explained with the Model-3). The error in this case is much higher. Analysis of error in estimating wave heights showed (Fig. 5) that estimation error for Model-1 and Model-2 remained more or less the same within the prediction time horizon but it blew up with time for the Model-3. Accordingly, Model-1 was selected as the final model for filling in the missing wave data.
Table 2: Training and verification statistics of the three models

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>NMSE</td>
</tr>
<tr>
<td>Model-1</td>
<td>28.9</td>
<td>0.144</td>
</tr>
<tr>
<td>Model-2</td>
<td>11.7</td>
<td>0.023</td>
</tr>
<tr>
<td>Model-3</td>
<td>11.7</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Figure 5: Average absolute error in estimating wave height by the different models (on verification data). The error in Model-3 blows up after certain time whereas it remains around the RMSE for the Model-1 and Model-2

5. CONCLUSIONS
Reconstructing time series data from incomplete data set constitutes the subject of this paper. A data-driven model based on an artificial neural network (ANN) was built for recovering the missing wave heights. The performance of the model was found to be reasonable and the model was used for filling in the missing wave data.

The input variables for building the ANN model were selected using Average Mutual Information (AMI) which was found to be very useful for discovering information content about the target variable from the associated variables.

The ANN model uses wind as its principal information for reproducing waves. The direction of wind was found to be an important factor influencing the wave height. The circular coding method was used to represent the direction of wind in the neural network model.

It was observed that reconstructing time series data is problematic when the antecedent value of a variable that is an input to the model, contains substantial information about the present value of this variable that is an output of the model. If the gaps of considerable length are to be filled in, the value estimated by the model needs to be used as an input continuously for several prediction steps and this causes the accumulation of the error. It would be safer to use a model that uses other available information and not the antecedent value of the output unless the errors are believed to be low enough.
The presented model was used as a component in an ANN model for predicting sedimentation which is being built for the Rotterdam port authorities. The results of that research are to be reported shortly.

ACKNOWLEDGEMENT
Part of this work was performed in the framework of the project "Data mining, knowledge discovery and data-driven modelling" of the Delft Cluster research programme supported by the Dutch government.

List of abbreviations
- r : Correlation coefficient
- RMSE : Root mean square error
- NRMSE : Normalised root mean square error

REFERENCES
Wang, L. (2001) Artificial neural network and model trees for reconstruction of missing data and runoff forecasting – application to catchments in Salland, the Netherlands, MSc Thesis (No. HH 421), IHE-Delft, the Netherlands.