Applying neural networks with active neurons to sea-water quality measurements

Evaggelos V Hatzikos

Technological Educational Institute of Thessaloniki, Greece

Leonidas Anastasakis

Loughborough University, Leicestershire, United Kingdom

Nick Bassiliades and Ioannis Vlahavas

Aristotle University of Thessaloniki, Greece

Abstract: *This study examines the presence of either linear or nonlinear relationships between a number of popular sea-water quality indicators such as water temperature, pH, amount of dissolved oxygen and turbidity. The data are obtained from a set of sensors in an underwater measurement station. The neural networks with active neurons are applied to the prediction of each one of the above four indicators and their performance is compared against a benchmark prediction method known as the random walk model. The random walk model is the simpler prediction method, which accepts as the best prediction for a variable its current value. The neural network with active neurons is a black box method, which contrary to neural networks with passive neurons does not require a long set of training data. The results show that for daily predictions the neural network with active neurons is able to beat the random walk model with regard to directional accuracy, namely the direction (upward or downwards) of the modelling object in the next day.*

Keywords: *Neural Networks, System Modeling, Active Neurons*

1. INTRODUCTION

The objective of this report is to investigate if it is possible to predict a number of water quality variables produced by an under-water measurement set-up. A set of sensors is used to record the sea-water temperature, pH, conductivity, salinity, amount of dissolved oxygen and turbidity and their measurements are stored in a database. The data form time series and a number of modelling methods could be used to reveal the information hidden within the data. However, this study will focus only on the development of prediction models for the temperature, pH, amount of dissolved oxygen and turbidity due to their higher importance in terms of commercial exploitation.

A variety of linear and nonlinear modelling techniques could be applied but in this study we focus on applying neural networks with active neurons since it is believed to be a more appropriate prediction algorithm for noisy and short time series. Their prediction ability is shown by comparing their performance against the random walk model, which serves as a benchmark model in prediction tasks. The random walk model simple states that the future value of a variable will be equal to its current value supporting in that way the unpredictability of the modelling object. However, due to the correlation and interactions between the water quality variables it is interesting to investigate if there is an underlying mechanism governing the data and therefore prove the predictability of these variables. The identification of such models is particular useful for ecologists and environmentalists since they will be able to predict in advance the pollution levels in the sea water and therefore instruct all the necessary precaution measures.

2. DATA ANALYSIS

The data in this study are produced by the Andromeda-analyser, for more details check [1] and [2], which measures water temperature, pH, conductivity, salinity, amount of oxygen and turbidity in sea-water. The original data were collected on July and early August of 2004 at an hourly basis with a sampling interval of 9 seconds. This study focus on the interactions between four of the most important water quality measurements, the temperature, pH, amount of dissolved oxygen and turbidity and therefore the rest of the measured variables are excluded from this study. The data are characterised by some outliers and a number of missing values due to temporary inefficiency of the analyser as well as problems in the transmission of the data. Furthermore, it is observed that the variation of the measurements during the day is very small and therefore it was decided to average the data over one day. As a result the corresponding models will perform the one step ahead prediction of the four water quality variables. The existence of missing values was overcome by using a linear interpolation method to replicate their values. Despite the limitation of such approach it was decided that since the variation of the variables is very small the linear interpolated values could be seen as close enough to its real value.

The data are split into two sets: one for estimating the model unknown coefficients and the second for validating the model performance. The first set corresponds to the 80% of the initial data set producing 25 observations (from 03/07/04 to 27/07/04) and are used to the subsequent data analysis methods and the model estimation procedure. The second set (the remaining 20% of the initial data set) is known as the testing set and consists of the last 8 observations of the initial data set (from 28/07/04 to 04/08/04). The testing set does not take part at any stage of the modelling approach and thus can be seen as a valid set for testing the model generalisation performance.

Table 1 shows the cross-correlations together with their significance between the four variables. The correlations are calculated from the one-month data set collected from the 03rd of July 2004 to the 27th of July 2004 (i.e. training set). It is expected that all four variables will be strongly correlated to each other since it is believed that changes in water temperature and water clarity can affect the amount of oxygen in water. In addition the water of low clarity could contain organisms, which can also affect the acidity of the water. However, as it can be seen in table 1, the turbidity is not significant correlated with any of the other water quality factors and although the presence of a significant correlation does not show causality it is rational to exclude turbidity from the modelling procedure of the remaining three variables.

Tab. 1: Correlation coefficients amongst the four sea-water quality variables.

		Water Temperature	pH	Oxygen	Turbidity
Water Temperature	Correlation	1	0.904(**)	0.678(**)	0.208
	Sig. (2-tailed)	.	0.000	0.000	0.318
pH	Correlation	0.904(**)	1	0.794(**)	0.105
	Sig. (2-tailed)	0.000	.	0.000	0.616
Oxygen	Correlation	0.678(**)	0.794(**)	1	0.234
	Sig. (2-tailed)	0.000	0.000	.	0.259
Turbidity	Correlation	0.208	0.105	0.234	1
	Sig. (2-tailed)	0.318	0.616	0.259	.

** Correlation is significant at the 0.01 level (2-tailed).

The above table shows the correlation between the variables at time t and hence it is not particular useful for showing the effect of past values of the variables at time t.

Such information can be obtained by examining the autocorrelation (ACF) and cross correlation function (CCF) for all four variables. Figures 1, 2, 3, 4, 5 and 6 demonstrate that temperature, pH and oxygen are correlated to each other while reconfirm that turbidity is not an influential factor to the other three variables. Additionally the autocorrelation function of all the variables is also plotted in figures 7, 8, 9 and 10 respectively. It is shown that past values of temperature and pH can be considered as potential inputs to an auto regression model for each one of these variables. Oxygen and turbidity seems to be uncorrelated, which implies that it could be very difficult to perform one-day ahead predictions for these two variables.

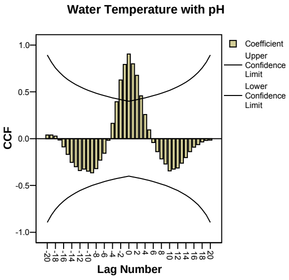

Fig. 1: Cross-correlation between water temperature and pH

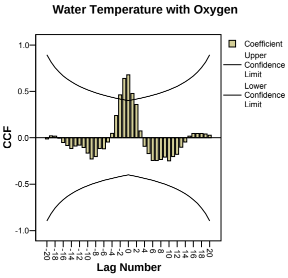

Fig. 2: Cross-correlation between water temperature and oxygen

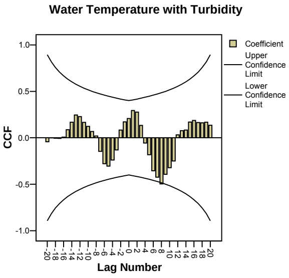

Fig. 3: Cross-correlation between water temperature and turbidity

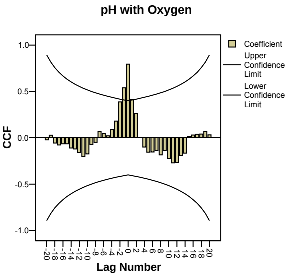

Fig. 4: Cross-correlation between pH and oxygen

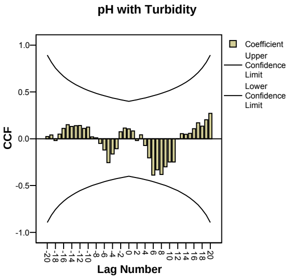

Fig. 5: Cross-correlation between pH and turbidity

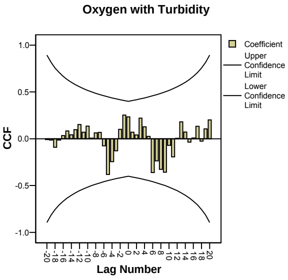

Fig. 6: Cross-correlation between oxygen and turbidity

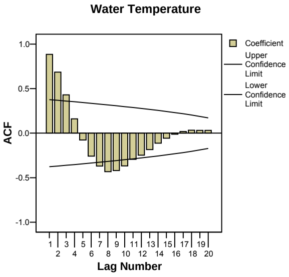

Fig. 7: Autocorrelation for water temperature

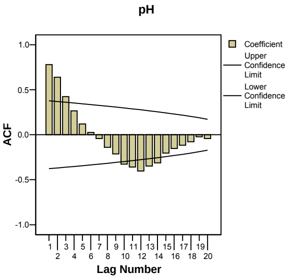

Fig. 8: Autocorrelation for pH

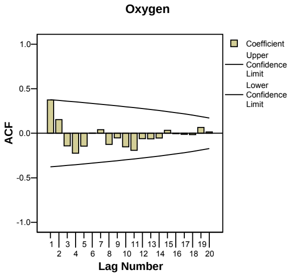

Fig. 9: Autocorrelation for oxygen

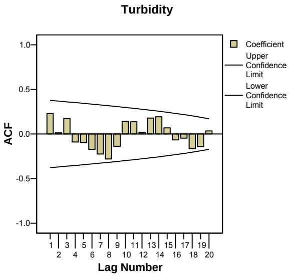

Fig. 10: Autocorrelation for turbidity

3. NEURAL NETWORKS WITH ACTIVE NEURONS

Neural networks with active neurons were developed as an appropriate modelling method for noisy and short time series, check [3] and [4]. Such neural networks follow the principles of self-organisation and induction and let the data themselves decide on the number of hidden layers, neurons as well as input variables. On the contrary neural networks with passive neurons make subjective decisions for the number of hidden neurons and layers, while the algorithms applied to identify the optimum set of input variables are affected by the initial learning conditions. Finally, they require long time series to sufficiently identify the underlying laws governing the data and therefore are not appropriate for modelling objects characterised by noisy and short time series.

For every pair of inputs an individual neuron is created. However, its activation function (instead of fixed) is optimally found amongst a variety of different functions starting with the simpler linear function of the two inputs and evolving until the complete second order activation function in equation (1) is estimated. The final activation function is chosen as the one, which minimises the accuracy criterion on an external set of data.

(1) $$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2 + \alpha_4 x_1^2 + \alpha_5 x_2^2$$

After the activation functions in every neuron for every pair of inputs are computed the neurons are classified according to their performance on the external set of data. A pre-specified number of the best performing neurons is selected to continue in the modelling procedure while the remaining are deleted. Then the outputs of the new active neurons form the input vector for the second layer of the neural network and the above procedure is repeated. The optimum number of layers is decided by finding the minimum in the plane complexity (i.e. number of layers) vs. performance criterion on the external set of data. A variety of external criteria exist but in this study the Prediction Error Sum of Squares (PESS) is used. In addition, due to the small number of data the cross-validation method is used to compute the neurons' prediction performance. According to that every single datum is used to compute the model prediction performance and therefore the data are not divided into training and validation data.

4. RESULTS

The conclusions driven from the correlation and autocorrelation analysis are used to create the input vector for the models. For each one of the four water quality variables three different models are considered. The first one is the naïve random walk model, which assumes that the model prediction is equal to its current value as equation (2) shows, where ε_t is white noise.

(2) $$y_t=y_{t-1}+\varepsilon_t$$

The second is an autoregressive model where only past values of the variable itself are used in the input vector (AR-model). The number of past values is derived from the autocorrelation function. In this study since the neural networks with active neurons are able to identify the optimum number of inputs it is decided to extend the set of potential inputs beyond those indicated by the autocorrelation function and hence for each one of the four quality variables three past values are considered. Finally, in the third model the number of inputs is increased by considering past values of the other water quality variables as well (MI-model). The number of past values is again three and based on the cross-correlation analysis it is proposed to exclude turbidity from the set of potential predictors. As it is shown in table 1 as well as figures 3, 5 and 6 turbidity is not correlated with the other three variables so it is preferred to exclude it from the input vector. Furthermore, tests have shown that its inclusion is not significantly improving model prediction performance.

The model performance is estimated according to three different performance measures. First, the R squared is computed which is a measure of accuracy that illustrates the percentage of data variance explained by the model. It takes values between 0 and 1 where a value of 1 shows that the derived model is able to explain 100% of the data variance. Second, the normalised mean square is computed, which is also a measure of accuracy but it is perhaps better than R squared for comparing model performance because it takes into account the bias and error variance. Finally the percentage of correctly predicting the direction in the daily movements of the variables is computed. From table 2 where the final set of input variables for every model is presented it is seen that the autoregressive models show high similarity with the random walk model. In all four variables the current value of the variables is chosen as the most significant influential factor to determine its future value. The similarity is also obvious by looking at table 3 where the model performance on the testing set is presented. With regard to accuracy measures the random walk model seems to be either better are at least as good as its autoregressive counterpart. However, looking at the model's ability to predict the direction of the future movements (i.e. the variable's value moving either upwards or downwards) the autoregressive models is significant better.

In case that a multivariate model (i.e. MI-model) is considered table 2 shows that past values of the other variables are also important factors to determine the model's future state. Table 3 shows that the R^2 is higher for the pH and oxygen models than it is at the corresponding autoregressive model while the model ability to correctly identify the direction in the movement of the variable future values remains either same (oxygen) or better than that of either the random walk model or the autoregressive models (water temperature and pH). It is clear that despite the low marginal accuracy of the model, their ability to predict the trend of the future values is significantly better than that of the random walk model. Finally, figures 11, 12 and 13 shows the in-sample (1 to 22 days) and out-of-sample (23 to 30 days) performance of all three models for the water temperature, pH and oxygen, while figure 14 shows the model performance for the turbidity. In the latter only the random walk and autoregressive model are shown.

Tab. 2: The set of initial inputs to take part in the prediction models for the four water quality variables.

Model \ Inputs	Temperature	pH	Oxygen	Turbidity
AR-model	Temp.: t-1, t-3	pH: t-1	Oxygen: t-1	Turbidity: t-1
MI-model	Temp.: t-1, t-3	Temp.: t-1, t-2, t-3 Oxygen: t-2, t-3	Oxygen: t-3 Temp.: t-1, t-2, t-3 pH: t-1, t-2, t-3	-

Tab. 3: The marginal and directional accuracy of the models estimated on the testing set of data.

	Temperature			pH			Oxygen			Turbidity		
Model	NMSE	R^2	%	NMSE	R^2	%	NMSE	R^2	%	NMSE	R^2	%
RW	0.0391	0.8164	25	0.0577	0.7872	25	0.1044	-0.050	50	0.1137	0.0737	37.5
AR	0.0508	0.6887	50	0.0545	0.7872	50	0.0949	-0.050	50	0.1261	0.0737	75
MI	0.0508	0.6887	50	0.1372	0.8618	75	0.2712	0.2222	50	-	-	-

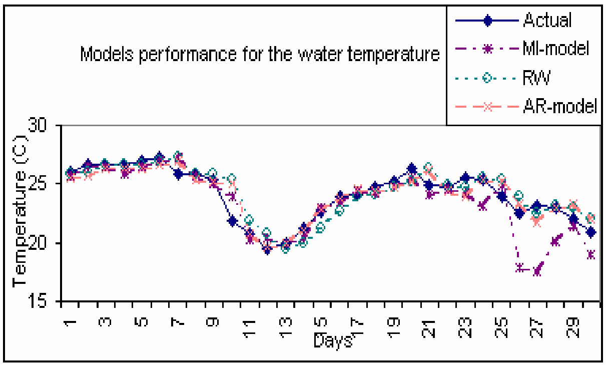

Fig. 11: In- and out-of-sample performance for water temperature

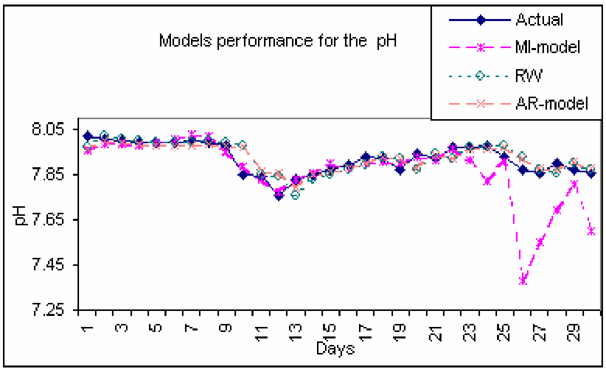

Fig. 12: In- and out-of-sample performance for pH

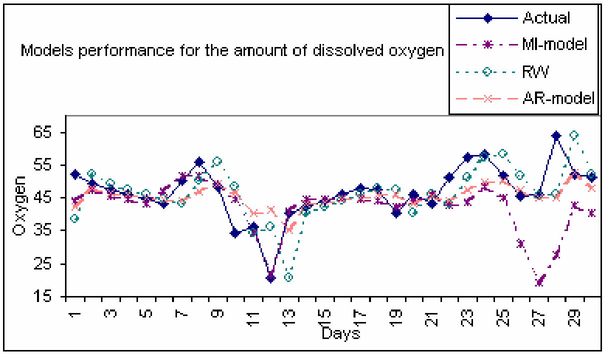

Fig. 13: In- and out-of-sample performance for dissolved oxygen

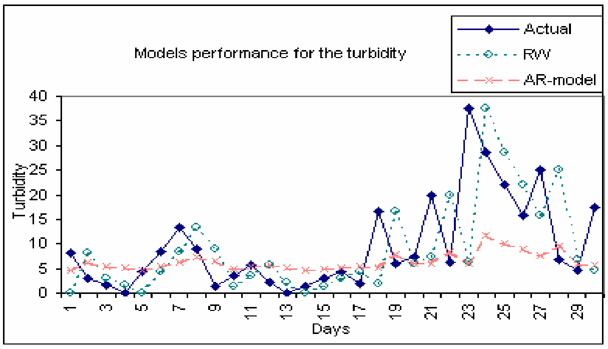

Fig. 14: In- and out-of-sample performance for turbidity

5. CONCLUSIONS AND FUTURE WORK

This study has focused on building both autoregressive and multivariate one step ahead prediction models for four main sea-water quality variables, namely the water temperature, pH, amount of dissolved oxygen and turbidity. Neural networks with active neurons were chosen since they do not require a large number of training data and they are also thought to perform better for under determined and noisy tasks such as the measurement of water quality variables. Their models' performances were compared against that of the random walk model for both marginal and directional accuracy prediction measures. With regard to marginal accuracy measures it was found that the random walk model is better but this shouldn't surprise us since the daily variation of all variables is extremely low. However, looking at the prediction of the future movement of the variables it was found that neural network models with active neurons are performing better than the random walk model. For water temperature, pH and turbidity the improvement in sign prediction is significant while no change is found for the case of the oxygen. The ability to correctly identify the direction of future changes is important for scientists since it would allow them to take the appropriate measures to eliminate the effects of higher pollution levels in sea-water.

Future study will concentrate on improving the data acquisition and producing more accurate measurements. By doing that it will be possible to create a longer time series, which will allow the experimentation with more, modelling techniques. Having a longer time series will allow us to find data with similar characteristics to those close in the prediction period and using these data in the training of the neural network with active neurons could improve further the model prediction performance. The variation of the daily measurements is very small so it would be also interesting to increase the sampling period to either weekly or monthly data and then estimate the performance of the neural networks in such volatile data.

6. REFERENCES

[1] Hatzikos, E.V. (1998) The Andromeda network for monitoring the quality of water and air elements. *Proc. 2nd Conference on Technology and Automation*, October 1998, Thessaloniki, GREECE.

[2] Hatzikos, E.V. (2002) A fully automated control network for monitoring polluted water elements. *Proc. 4th Conference on Technology and Automation*, October 2002, Thessaloniki, GREECE.

[3] Ivakhnenko, A.G., Móller, J.A. (1995) Self-organization of nets of active neurons. *Systems Analysis Modelling Simulation* 20(1-2), 93-106.

[4] Móller, J.A., Lemke, F. (1999) *Self-Organising Data Mining: An Intelligent Approach to Extract Knowledge from Data*. Berlin, Germany.