Local Modelling by SOM partitioning and linear regression for Ecological Modelling

Whigham, P.A.

Information Science Dept., University of Otago, New Zealand, E-Mail: pwhigham@infoscience.otago.ac.nz

Keywords: *Local Model, Self Organising map, SOM, regression, sensitivity analysis.*

EXTENDED ABSTRACT

Global modelling approaches construct a single model that covers all of the training data or data used to describe a system. However, for many problems, especially in the natural sciences, the system under consideration is best understood in terms of a number of distinct states, with different patterns and processes operating for each state. The issue of state-based modelling, where different models are constructed for different states, brings in several issues that need to be addressed, including: how are the states determined, how are they represented, how does the modelling of the system determine which states are appropriate during some phase of modelling, and how are state transitions determined.

The self-organising map (SOM) is a topologically based unsupervised clustering algorithm (Kohonen 1982) that constructs prototypical descriptions of the dataset in a spatially structured format. The resulting clusters are traditionally positioned on a two-dimensional map, where neighbors on the map are similar prototypes representing the cluster centers for a subset of the data. Visualizing the prototype values on the map may then be used to indicate relationships between variables on the SOM – variable values that cluster together are associated with each other. The SOM has been successfully applied to a variety of ecological data sets for visualization of relationships and representations of state change in ecological systems (Giraudel and Lek 2003).

Traditional modelling of ecological systems have used global models, such as artificial neural networks or multivariate linear regressions, to produce predictive models that allow the important variables and processes of a system to be studied. Although they have been generally successful, a single global model may not allow an understanding of how the system changes in state and which variables are important under different conditions.

This paper addresses the problem of modelling and knowledge elucidation for multivariate, real-valued data sets using a SOM to cluster the training data, and then construct local linear regression models for each SOM neuron (best matching unit, bmu). The minimum number of examples to be used by each local regression is used to specify the degree of generalization for each local model. The basic framework of the model, named SOM-MLR, is shown in Figure 1.

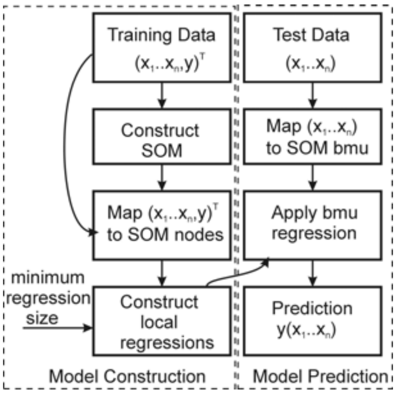

Figure 1. The SOM-MLR Framework

The minimum regression size is determined using either a cross-validation procedure, or the Akaike information criteria (AIC). This allows either global or local tuning of each regression model to occur. Since the final models are all linear regressions a variety of sensitivity analyses are possible. This paper will demonstrate the use of two approaches: the local correlation coefficient for each regression to indicate the relationship of each independent variable under a variety of states, and a local response curve based on using the prototype values from the SOM.

The properties of SOM-MLR are demonstrated by modelling the bloom dynamics of *Microcystis aeruginosa* for a regulated river system. The predictive performance and sensitivity analyses of the model highlight the benefits of using a local modelling approach to complex ecological systems.

1. INTRODUCTION

The Self-Organising Map (Kohonen 1982, Kohonen 2001) is a topologically-based unsupervised clustering algorithm that has been successfully applied to a variety of ecological data sets for visualisation of relationships and representation of state changes in ecological systems (Giraudel and Lek 2003). The SOM has been applied to such problems as aquatic insect richness (Park et al. 2003), fish assemblages (Brosse et al. 2001), community structure (Chon et al. 1996, Giraudel and Lek 2001) and stream ecosystems (Schleiter et al. 2002). The popularity of SOMs suggests that ecological problems can often be described by a set of local states, with potentially different processes and contexts operating for each state.

Traditional modelling of ecological systems have used global models, such as artificial neural networks or multivariate linear regressions, to produce predictive models that allow the important variables and processes of a system to be studied. Although they have been generally successful, a single global model may not allow an understanding of how the system changes in state and which variables are important under different conditions. In particular, analysis of correlation with a global model can often lead to a poor or misleading understanding of how a variable interacts with the dependent variable. For example, if a variable is negatively correlated with the dependent variable for some states of the system, and positively correlated for other states, the resulting global model may determine that there is no strong correlation one way or the other (Piras and Germond 1998). Since understanding an ecological system often involves the boundaries of behaviour (such as when algal blooms are high or low), a global model is not always going to allow an understanding of this change in state, even though the model has good predictive behaviour. The advantage of constructing local versus global regression models has been clearly argued by Fotheringham (2002), where spatially varying regression models are described. Although a global model can be used for overall trends, when a system behaves differently under different states a set of local models, associated with each state, may produce more valid interpretations of the underlying system behaviour.

This paper addresses the problem of modelling and knowledge elucidation for multivariate, real-valued data sets using a SOM to initially cluster the training data, and then constructing local linear regression models for each SOM neuron. A generalisation operator will be described that allows the local linear models to avoid over-fitting by exploiting the topological nature of the SOM. The use of local linear models also allows a number of approaches to interpreting the important variables of the system, their correlation with the dependent variable under a variety of states of the system, and their contribution to the dependent variable response as the state of the system changes. The system will be referred to as SOM-MLR.

The remainder of this paper is organised as follows: §2 gives background on previous related work and local modelling approaches, §3 describes the general algorithm for local model construction, §3.1 outlines the generalization operators, §3.2 describes the modelling process, §3.3 presents the local analyses tools, and §4 applies SOM-MLR to the modelling and interpretation of *M.aeruginosa* for a regulated river systems. Finally, §5 summarizes the work and addresses future research.

2. RELATED WORK

This section describes previous local approaches to modelling using a SOM. Although the approach is not new, the application of the general concept to ecological systems has not been considered.

Early work (Walter et al. 1990) focused on predicting highly non-linear time sequence data using a set of linear regressive models and a current state vector derived from a SOM. This approach did not take into account the topological characteristics of the SOM for generalization, and was specifically designed for time series trajectory analysis. A second time series approach was described by Principe and Wang where the SOM neurons and their local neighbourhood were used to construct local linear predictors (Principe and Wang 1995). In this work each model was based on a local neighbourhood of neuron vectors, however the training data were discarded once the SOM vectors were determined.

Vesanto (1997) used a SOM to partition and preserve the training data in a similar manner to SOM-MLR. A linear regression model was then constructed for each local training set. Although Vesanto mentions the issue of a minimum local data set size to ensure the stability of the regression, no further comments are made regarding how this might influence the generalization of the overall model. A similar use of a SOM for partitioning time series data (Lendasse et al. 1998) was used where the average of each data item within a cluster was used to construct an initial matrix of values.

A local SOM approach for selecting relevant input variables for non-linear regression is described in Piras (1998). Here a linear correlation coefficient is constructed over the local data that has been partitioned by the SOM. This work demonstrated that a local modelling approach combined with linear regressions can produce interpretations of variable interaction that cannot be easily distinguished by a global modelling approach.

To date there is little work on applying local modelling approaches to ecological systems. In particular, the use of a SOM as a clustering tool that allows local linear models to be constructed for ecological visualization and prediction has not been considered, and hence this paper presents some of the properties of such an approach.

3. THE SOM-MLR FRAMEWORK

This section assumes knowledge of the principles behind self-organising maps. A more formal and complete description can be found in Hautaniemi (2003) or Kohonen (2001).

Multiple linear regression assumes a continuous random dependent variable, y, and *n* independent variables $x_1, x_2, \ldots x_n$. The values of the independent variables are known quantities and hence the model is defined as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \quad (1)$$

where ε is a normally distributed random variable with mean zero and unknown standard deviation. The estimates for each β and ε is performed using a least squares minimization based on a set of data (or patterns) $p_i(x_1, x_2, \ldots, x_n, y)$, to produce the regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n \quad (2)$$

The two-dimensional SOM is defined as a lattice of points $m_i(x,y)$. Associated with each point or neuron $\mathbf{m}_i$ is an associated set of weights (also called a codebook or prototype), $\mathbf{w}_i \in \Re^{n+1}$ which are trained to match a generalised description of the pattern data. Define the function d(**a**,**b**) as the Euclidean distance between two *n*-dimensional components, **a** and **b**. Hence the distance between two neurons, $\mathbf{m}_i$, $\mathbf{m}_j$, on the lattice is $d(\mathbf{m}_i,\mathbf{m}_j)$, and the distance between a pattern $\mathbf{p}_i$ and neuron weight $\mathbf{w}_j$ is $d(\mathbf{p}_i,\mathbf{w}_j)$. Training of the SOM occurs by repeatedly presenting the set of patterns, and for each pattern **p** selecting the best matching neuron or unit (bmu) $\mathbf{m}_c$, defined as the neuron with the minimum distance between the presented pattern **p** and the current SOM neuron weights $\mathbf{w}_c$:

$$d(\mathbf{p},\mathbf{m}_c) = \min_i(d(\mathbf{p},\mathbf{w}_i)) \quad (3)$$

The best matching neuron ($\mathbf{m}_c$) weights, and those within the current neighbourhood, are then updated, based on the current learning rate. During training the learning rate and neighbourhood size are gradually reduced to allow convergence of the prototype weights. At the completion of training each pattern vector $\mathbf{w}_i$ associated with $\mathbf{m}_i$ represents a prototype vector of the training data. The topological training of the SOM implies that neighboring neurons generally having similar weight vectors, and therefore represent similar states. For SOM-MLR, the training (pattern) data is now partitioned to the SOM, where for each pattern **p** the best matching unit $\mathbf{m}_c$ is determined, and the pattern stored with this neuron. This set of patterns is referred to as $\mathbf{P_i}$. The final partition set, $\mathbf{R_i}$, used to construct the local regression equations, requires a minimum partition size for each $\mathbf{m}_i$. The process of constructing R_i will now be described.

3.1. Generalizing SOM-MLR

The concept of generalization is fundamental to producing a model with good prediction behavior with previously unseen data. In the case of SOM-MLR this involves determining either a global minimum regression size, applied to all local partitions of the training data, or a local minimum regression size, applied independently to each SOM neuron.

At the start of this process, each $\mathbf{R}_i$ is set to the corresponding $\mathbf{P_i}$ for each neuron $\mathbf{m}_i$. Increasing the number of patterns in $\mathbf{R_i}$ involves adding patterns $\mathbf{P_j}$ from the neighborhood of $\mathbf{m}_i$, where each additional partition list is selected based on their neighborhood and their neuron distance $d(\mathbf{m}_i, \mathbf{m}_j)$. This process is repeated with increasing neighborhood distance until the desired minimum number of patterns in $\mathbf{R_i}$ has been achieved.

The global minimum regression size may be determined in two ways: using a cross-validation of the training data to find the optimal regression size by minimizing the mean squared error over the validation set, and by applying a modified Akaike Information criteria (AIC) (Hurvich and Simonoff 1998), which measures a tradeoff between model complexity and model accuracy. In this case the average AIC is calculated over all local regression, and this figure is used to represent the global measure of model suitability. Since there are only a finite number of possible increments to the minimum regression size an exhaustive search of all increments of training size is possible. The cross-validation result, or AIC, can be recorded for each increase in minimum regression size, and the performance of the best resulting model used as the global minimum regression size for the model.

Local generalizing is performed using the AIC measure. Each SOM partition is gradually increased using the neighborhood training examples, and the minimum AIC measure over all possible partition sizes used for each SOM node. Hence the minimum regression size may be different for each SOM node.

3.2. The SOM-MLR Prediction Process

Once the final partitions **R** have been formed and the associated regression equations constructed for each $\mathbf{R_i}$, predictive modelling may occur. Given a test pattern $p_i(x_1,x_2,\ldots,x_n)$, the best matching SOM unit $\mathbf{m}_c$ is determined based on the minimum distance between $\mathbf{p}_i$ and the SOM neurons. Note that this does not include the independent variable y in the distance measurement. The regression equation associated with $\mathbf{m}_c$ is then used to predict the independent variable y for the given pattern p_i. Hence for each pattern presented for prediction, the SOM is used as the gating mechanism to select the local model, as shown in Figure 1.

3.3. SOM-MLR Sensitivity Analysis

Ecological modelling aims to produce both robust models for prediction and to understand how the underlying processes in a system interact. This exploration of how the response of a model can be apportioned to different sources of variation and how the model depends on the information presented to it is generally termed sensitivity analysis (Saltelli 2000).

The local models used with SOM-MLR are linear equations of the form of Eqn. (2), and therefore a number of simple methods for exploring the underlying properties of the modelled system are possible. One approach to interpreting the local model contributions is to determine the correlation coefficient for each R_i (Piras and Germond 1998), and mapping this for each state of the SOM. This has been previously shown to allow changes in the contribution of each independent variable to be assigned to particular states of the system. The plot of correlation coefficient for each $\mathbf{m}_i$ is ordered based on increasing dependent variable prototype value, and therefore allows both the positive and negative correlations in the variable for different states to be described.

A common form of sensitivity analysis is to model the response of the dependent variable as one independent variable is varied (Jeong et al. 2001, Scardi and Harding 1999). Normally the remaining independent variables are held constant (typically set to the mean value for each variable) while the selected variable is altered. This allows a general measure of the response of the dependent variable to the selected independent variable, however it does not consider the possible interactions between independent variables due to their mutual correlations. Since the control of correlation within a sample is important if meaningful results are to be obtained, a method for producing sampling values that go beyond a simple mean value is desirable (Helton and Davis 2000). This is achieved in SOM-MLR by using the prototype vectors associated with each neuron as typical values of the independent variables. Hence, as a given independent variable is varied from its lowest to highest measured value by incremental steps, the best matching neuron is selected for each step, and the associated regression equation used for prediction, with all other independent variables taking on the prototype values associated with the matching neuron. Although this does not explicitly build a model of correlation between variables, the use of prototype values for the “fixed” independent variables means that the combination of input values to each local regression is meaningful. A plot showing the dependent variable response for a range of values of the selected independent variable can now be used to interpret the dependencies between the selected variable and the modeled system.

4. INVESTIGATING BLOOM DYNAMICS FOR A RIVER SYSTEM

This section will demonstrate the use of SOM-MLR for modelling the dynamics *Microcystis aeruginosa* in the lower Nakdong River, South Korea. The dataset has been previously described and modelled by (Jeong et al. 2003) and is interesting since the system is externally driven by regulation controls and has a complex ecological response due to the changes in flow throughout the year. Since the majority of limnological data for the lower Nakdong exhibits distinct inter-annual variation, and there is a strong seasonal weather response, it is expected that a local state-based approach should be a useful tool for exploring the properties of the system. In this section the entire dataset will be used to construct a model, and some sensitivity analysis performed to explore the underlying properties of the system. Subsequently in §4.3 a SOM-MLR model will be constructed based on training data for 1995-1998 and the predictive behaviour for 1994 presented.

4.1. Experimental Setup

The SOM dimension was determined by the ratio of the two largest eigenvalues of the eigenvectors of the training data (Kohonen 2001). Using this ratio the dimensions were set based on the heuristic that the number of neurons should be approximately the square root of the number of training examples. This gave a SOM of dimensions 10x21. The SOM was then linearly initialized to allow quicker convergence, although (unpublished) experiments showed that the system was quite robust to parameter settings. The number of training epochs was set at 1000, with an initial learning rate of 0.1. All data were normalized to a mean of zero and standard deviation of 1. The AIC measure was used to locally determine the minimum regression size for each neuron. The river data were lagged by one day between independent and dependent variables, so that the inferred model was predicting one-day-ahead prediction.

4.2. Modelling the complete dataset

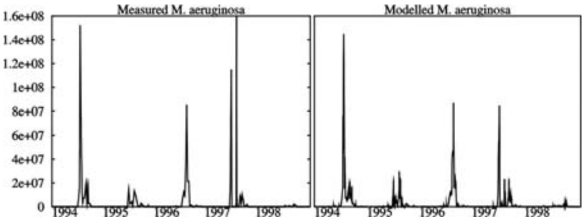

Figure 2. Model Prediction for *M.aeruginosa (cell count abundance)*

The resulting model for all years had an r^2 of 0.8, which showed that the model was not over fitting the data. The resulting prediction for *M.aeruginosa* is shown in Figure 2. The main point to note regarding the dynamics of the system are that there are bloom peaks interspersed with very low values, perhaps suggesting that the system moves between different states.

A comparison of this model with a global regression highlights some of the advantages of a local modelling approach. To enable this comparison a second model using SOM-MLR was constructed, using a SOM architecture of 2x1 neurons. This gave just two possible states to represent the system, and therefore approximated a global model. This global model had an r^2 of 0.36, and although this model predicted the timing of the events, the magnitude of each bloom was well underestimated (data not shown). As will be shown, the global model also had some limitations when interpreting the system response.

The correlation coefficient for the range of *M.aeruginosa* values is shown in Figure 4 for the local and global SOM-MLR models. The global model implied that many of the independent variables, such as secci depth, silica, ammonia-N, dissolved oxygen, and so on had almost no correlation with *M.aeruginosa* abundance. However, the full local SOM-MLR model shows that these and other variables have a more complex relationship to the dependent variable, and change their correlation particularly from low to high *M.aeruginosa* abundance. For example, secci depth showed a positive correlation for low abundance, and a negative correlation for high abundance. Similar changes in relationship based on different abundance states can be seen for many of the variables, highlighting that the interactions of the system are more complex than indicated by the global model.

The response curves for increasing values of the independent variables are shown in Figure 5, along with the global model response curves. Once again the local model shows significantly different behavior from the global model, and indicates some general non-linear trends, even though the underlying local models are linear in form. For example, the responses for Cladocera and O.limosa suggest high abundance for low values of these independent variables, which decay non-linearly as their values increase. The local model responses also suggest more complex dynamics over the complete range of independent variables than the global model, and could be used to suggest hypotheses regarding the interactions of the measured elements of the system.

4.3. Prediction Accuracy on unseen data

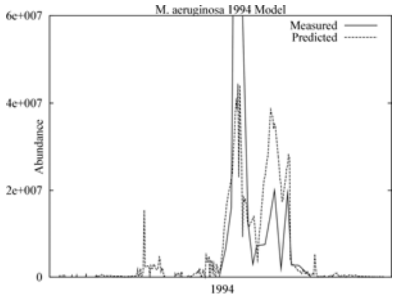

Figure 3. 1994 Abundance Prediction

To demonstrate that the SOM-MLR approach allows reasonable predictive accuracy on unseen data, the previous setup was trained using the years

1995-1998, and withholding the 1994 data. The final model prediction for 1994 had an r^2 of 0.46, which is comparable to the genetic programming approach described in (Jeong et al., 2003). The resulting behaviour is shown in Figure 3, where it can be seen that the timing of blooms are well modelled, although the magnitudes are not always correct.

5. CONCLUSIONS

This paper introduces the benefits of a local modelling approach for ecological data analysis and exploration. The use of a SOM for clustering and gating of the local models offers the benefits of a SOM for visualization of the dataset, as well as a topological neighborhood to control the generalization of the local models. The sensitivity analyses presented here show just some of the possible approaches that could be used with this clustering framework, and demonstrates that the use of local versus global models allows more information regarding the state-based behavior of ecological systems to be presented and analyzed. There are a number of important issues to be studied with this framework, including how robust the system is to changing SOM architectures, the use of other information metrics for model generalization, extensions to the basic linear regression formalism of the models and a formal justification for the construction of the regressions based on the topological neighborhood of the SOM. Clearly, however, the approach of local modelling should become a standard tool for complex system modelling and prediction.

6. ACKNOWLEDGMENTS

The author would like to thank Prof. Joo and Dr. Jeong of Busan National University for access to the Nakdong River data.

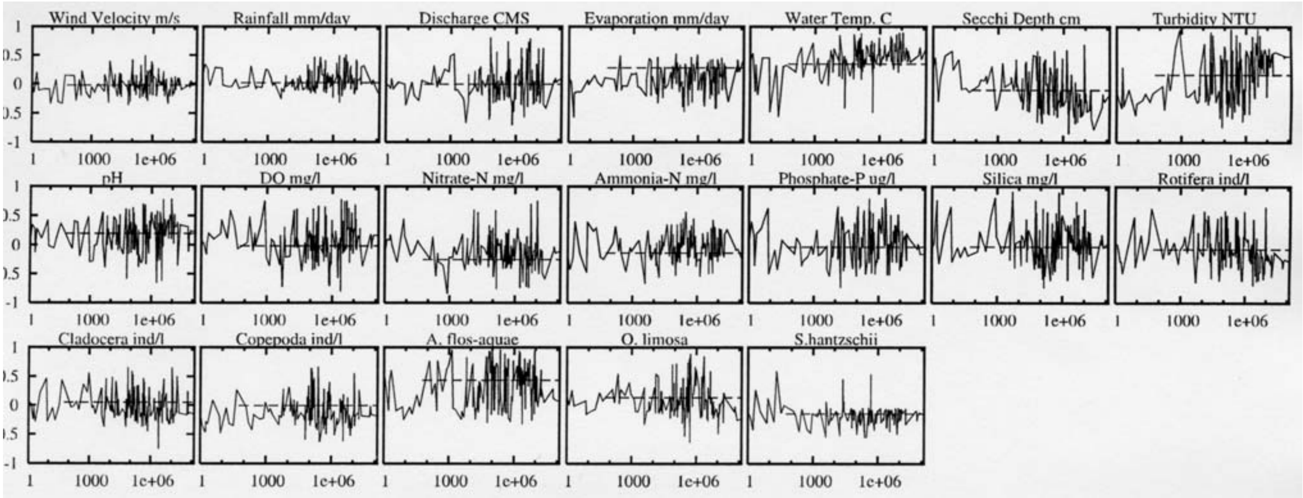

Figure 4. Correlation coefficients for local and global (dashed) models. X axis is for increasing abundance.

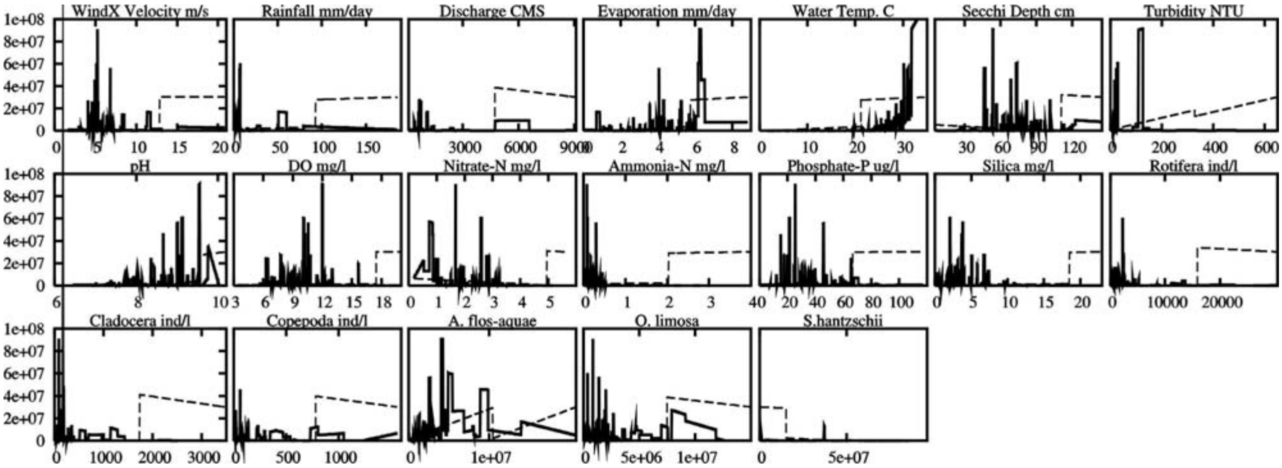

Figure 5. Response curves for local and global (dashed) models. Y axis is M.aeruginosa abundance. X axis varies based on the variable under consideration.

7. REFERENCES

Brosse, S., Giraudel, J.L. and Lek, S., (2001), Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. Ecol. Modelling, 146(1-3): 159-166.

Chon, T.S., Park, Y.S., Moon, K.H. and Cha, E.Y., (1996), Patternizing communities by using an artificial neural network. Ecol. Modelling, 90: 69-78.

Fotheringham, A.S., Brunsdon, C. and Charlton, M., (2002), Geographically Weighted Regression: the analysis of spatially varying relationships. John Wiley & Sons, Ltd.

Giraudel, J. and Lek, S., (2003), Ecological Applications of Unsupervised Artificial Neural Networks. In: F. Recknagel (Editor), Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation. Springer, Berlin Heidelberg, pp. 15-33.

Giraudel, J.L. and Lek, S., (2001), A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol. Modelling, 146(1-3): 329-339.

Hautaniemi, S., Yli-Harja, O. and Astola, J., (2003), Analysis and Visualization of Gene Expression Microarray Data in Human Cancer using Self-Organizing Maps. Machine Learning, 52: 45-66.

Helton, J. and Davis, F., (2000), Sampling-Based Methods. In: A. Saltelli, K. Chan and E.M. Scott (Editors), Sensitivity Analysis. John Wiley & Sons, Ltd, pp. 101-153.

Hurvich, C.M. and Simonoff, J.S., (1998), Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J.R. Statist. Soc. B, 60(2): 271-293.

Jeong, K., Joo, G., Kim, H., Ha, K. and Recknagel, F., (2001), Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecol. Modelling, 146: 115-129.

Jeong, K., Kim, D., Whigham, P.A. and Joo, G., (2003), Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol. Modelling, 161(1-2): 67-78.

Kohonen, T., (1982), Self-organised formation of topographically correct feature maps. Biol. Cybern., 43: 59-69.

Kohonen, T., (2001), Self-Organizing Maps. Springer.

Lendasse, A., Verleysen, M., de Bodt, E., Cottrell, M. and Gregoire, P., (1998), Forecasting time-series by Kohonen classification, ESANN'98 Proc. of the European Symposium on Artificial Neural Networks. D-Facto public, Bruges, Belgium, pp. 221-226.

Park, Y., Cereghino, R., Compin, A. and Lek, S., (2003), Applications of artificial neural netwoks for patterning and predicting aquatic insect species richness in running waters. Ecol. Modelling, 160(3): 265-280.

Piras, A. and Germond, A., (1998), Local linear correlation analysis with the SOM. Neurocomputing, 21: 79-90.

Principe, J.C. and Wang, L., (1995), Non-linear time series modelling with self-organizing feature maps, Proc. NNSP'95, IEEE Workshop on Neural Networks for Signal Processing. IEEE Service Centre, Cambridge, MA, USA, pp. 11-20.

Saltelli, A., (2000), What is Sensitivity Analysis? In: A. Saltelli, C. K. and E.M. Scott (Editors), Sensitivity Analysis. John Wiley & Sons, Ltd., pp. 3-13.

Scardi, M. and Harding, L.W., (1999), Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Modelling, 120: 213-223.

Schleiter, I.M., Obach, M., Wagner, R., Werner, H. and Schmidt, H., (2002), Modelling Ecological Interrelations in Running Water Ecosystems with Artificial Neural Networks. In: F. Recknagel (Editor), Ecological Informatics: Understanding Ecology by Biologically-inspired computation. Springer-Verlag, Berlin.

Vesanto, J., (1997), Using the SOM and local models in time-series prediction, Proc. of WSOM'97, Workshop on Self-Organizing Maps. Helsinki Univ. of Technology, Neural Networks Research Centre, Finland, Espoo, Finland.

Walter, J., Ritter, H. and Schulten, K., (1990), Non-linear prediction with self-organizing maps, Int. Joint Conf. on Neural Networks. IEEE Service Centre, Piscataway NJ, IJCNN-90-San Diego.