0304-3800
368 (1999)

ECOLOGICAL MODELLING

INTERNATIONAL JOURNAL ON ECOLOGICAL MODELLING AND SYSTEMS ECOLOGY

I S E M

The Journal of the
International Society of Ecological Modelling

SPECIAL ISSUE:
Application of Artificial neural networks in Ecological Modelling

GUEST EDITORS:
Sovan Lek
Jean Francois Guegan

ELSEVIER

ECOLOGICAL MODELLING

International Journal on Ecological Modelling and Systems Ecology

Volume 120/2-3 (1999)

I S E M

The Journal of the
International Society of Ecological Modelling

SPECIAL ISSUE:
Application of Artificial neural networks in Ecological Modelling

GUEST EDITORS:
Sovan Lek
Jean Francois Guegan

ELSEVIER

Ecological Modelling 120 (1999) 65–73

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Artificial neural networks as a tool in ecological modelling, an introduction

Sovan Lek [a,*], J.F. Guégan [b]

[a] *CNRS, UMR 5576, CESAC-Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, France*
[b] *Centre d'Etude sur le Polymorphisme des Micro-organismes, Centre I.R.D. de Montpellier, U.M.R C.N.R.S.- I.R.D. 9926, 911 avenue du Val de Montferrand, Parc Agropolis, F-34032 Montpellier cedex 1, France*

Abstract

Artificial neural networks (ANNs) are non-linear mapping structures based on the function of the human brain. They have been shown to be universal and highly flexible function approximators for any data. These make powerful tools for models, especially when the underlying data relationships are unknown. In this reason, the international workshop on the applications of ANNs to ecological modelling was organized in Toulouse, France (December 1998). During this meeting, we discussed different methods, and their reliability to deal with ecological data. The special issue of this ecological modelling journal begins with the state-of-the-art with emphasis on the development of structural dynamic models presented by S.E. Jorgensen (DK). Then, to illustrate the ecological applications of ANNs, examples are drawn from several fields, e.g. terrestrial and aquatic ecosystems, remote sensing and evolutionary ecology. In this paper, we present some of the most important papers of the first workshop about ANNs in ecological modelling. We briefly introduce here two algorithms frequently used; (i) one supervised network, the backpropagation algorithm; and (ii) one unsupervised network, the Kohonen self-organizing mapping algorithm. The future development of ANNs is discussed in the present work. Several examples of modelling of ANNs in various areas of ecology are presented in this special issue.

Keywords: Backpropagation; Kohonen neural network; Self-organizing maps; Ecology; Modelling; ANN Workshop

1. Introduction

Ecological modelling has grown rapidly in the last three decades. To build his models, an ecologist disposes a lot of methods, ranging from numerical, mathematical, and statistical methods to techniques originating from artificial intelligence (Ackley et al., 1985), like expert systems (Bradshaw et al., 1991; Recknagel et al., 1994), genetic algorithms (d'Angelo et al., 1995; Golikov et al., 1995) and artificial neural networks, i.e. ANN (Colasanti, 1991; Edwards and Morse, 1995).

ANNs were developed initially to model biological functions. They are intelligent, thinking machines, working in the same way as the animal brain. They learn from experience in a way that no conventional computer can and they can rapidly solve hard computational problems. With

* Corresponding author. Tel.: +33-561-558687; fax: +33-561-556096.
E-mail address: lek@cict.fr (S. Lek)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00092-7

the spread of computers, these models were simulated and later research was also directed at exploring the possibilities of using and improving them for performing specific tasks.

In the last decade, research into ANNs has shown explosive growth. They are often applied in physics research like speech recognition (Rahim et al., 1993; Chu and Bose, 1998) and image recognition (Dekruger and Hunt, 1994; Cosatto and Graf, 1995; Kung and Taur, 1995) and in chemical research (Kvasnicka, 1990; Wythoff et al., 1990; Smits et al., 1992). In biology, most applications of ANNs have been in medecine and molecular biology (Lerner et al., 1994; Albiol et al., 1995; Faraggi and Simon, 1995; Lo et al., 1995). Nevertheless, a few applications of this method were reported in ecological and environmental sciences at the beginning of the 90's. For instance, Colasanti (1991) found similarities between ANNs and ecosystems and recommended the utilization of this tool in ecological modelling. In a review of computer-aided research in biodiversity, Edwards and Morse (1995) underlined that ANNs have an important potential. Relevant examples are found in very different fields in applied ecology, such as modelling the greenhouse effect (Seginer et al., 1994), predicting various parameters in brown trout management (Baran et al., 1996; Lek et al., 1996a,b), modelling spatial dynamics of fish (Giske et al., 1998), predicting phytoplankton production (Scardi, 1996; Recknagel et al., 1997), predicting fish diversity (Guégan et al., 1998), predicting production/biomass (P/B) ratio of animal populations (Brey et al., 1996), predicting farmer risk preferences (Kastens and Featherstone, 1996), etc. Most of these works showed that ANNs performed better than more classical modelling methods.

2. Scope of this particular issue

The pressures to understand and manage the natural environment are far greater now than could ever have been conceived even 50 years ago, with the loss of biodiversity on an unprecedented scale, fragmentation of landscapes, and addition of pollutants with the potential of altering climates and poisoning environments on a global scale. In addition, many ecological systems present complex spatial and temporal patterns and behaviours.

Recent achievements in computer science provide unrivaled power for the advancement of ecology research. This power is not merely computational: parallel computers, having hierarchical organization as their architectural principle, also provide metaphors for understanding complex systems. In this sense, in sciences of ecological complexity, they might play a role like equilibrium-based metaphors had in the development of dynamic systems ecology (Villa, 1992).

ANNs have recently become the focus of much attention, largely because of their wide range of applicability and the case with which they can treat complicated problems. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the output of new independent input data. ANNs imitate the learning process of the animal brain and can process problems involving very non-linear and complex data even if the data are imprecise and noisy. Thus they are ideally suited for the modelling of ecological data which are known to be very complex and often non-linear.

For this reason, we organized the first workshop on the applications of ANNs in ecological modelling in Toulouse in December of 1998. This special volume gathers some of the papers presented.

3. What is an artificial neural network

An ANN is a 'black box' approach which has great capacity in predictive modelling, i.e. all the characters describing the unknown situation must be presented to the trained ANN, and the identification (prediction) is then given.

Research into ANNs has led to the development of various types of neural networks, suitable to solve different kinds of problems: auto-associative memory, generalization, opti-

mization, data reduction, control and prediction tasks in various scenarios, architectures etc. Chronologically, we can cite the Perceptron (Rosenblatt, 1958), ADALINE, i.e. Adaptive linear element (Widrow and Hoff, 1960), Hopfield network (Hopfield, 1982), Kohonen network (Kohonen, 1982, 1984), Boltzmann machine (Ackley et al., 1985), multi-layer feed-forward neural networks learned by backpropagation algorithm (Rumelhart et al., 1986). The descriptions of these methods can be found in various books such as Freeman and Skapura (1992), Gallant (1993), Smith (1994), Ripley (1994), Bishop (1995), etc. The choice of the type of network depends on the nature of the problem to be solved. At present, two popular ANNs are (i) multi-layer feed-forward neural networks trained by backpropagation algorithm, i.e. backpropagation network (BPN), and (ii) Kohonen self-organizing mapping, i.e. Kohonen network (SOM). The BPN is most often used, but other networks has also gained popularity.

3.1. Multi-layer feed-forward neural network

The BPN, also called multi-layer feed-forward neural network or multi-layer perceptron, is very popular and is used more than other neural network types for a wide variety of tasks. The BPN is based on the supervised procedure, i.e. the network constructs a model based on examples of data with known outputs. It has to build the model up solely from the examples presented, which are together assumed to implicitly contain the information necessary to establish the relation. A connection between problem and solution may be quite general, e.g. the simulation of species richness (where the problem is defined by the characteristics of the environment and the solution by the value of species richness) or the abundance of animals expressed by the quality of habitat. A BPN is a powerful system, often capable of modelling complex relationships between variables. It allows prediction of an output object for a given input object.

The architecture of the BPN is a layered feed-forward neural network, in which the non-linear elements (neurons) are arranged in successive layers, and the information flows unidirectionally, from input layer to output layer, through the hidden layer(s) (Fig. 1). As can be seen in Fig. 1, nodes from one layer are connected (using interconnections or links) to all nodes in the adjacent layer(s), but no lateral connections within any layer, nor feed-back connections are possible. This is in contrast with recurrent net-

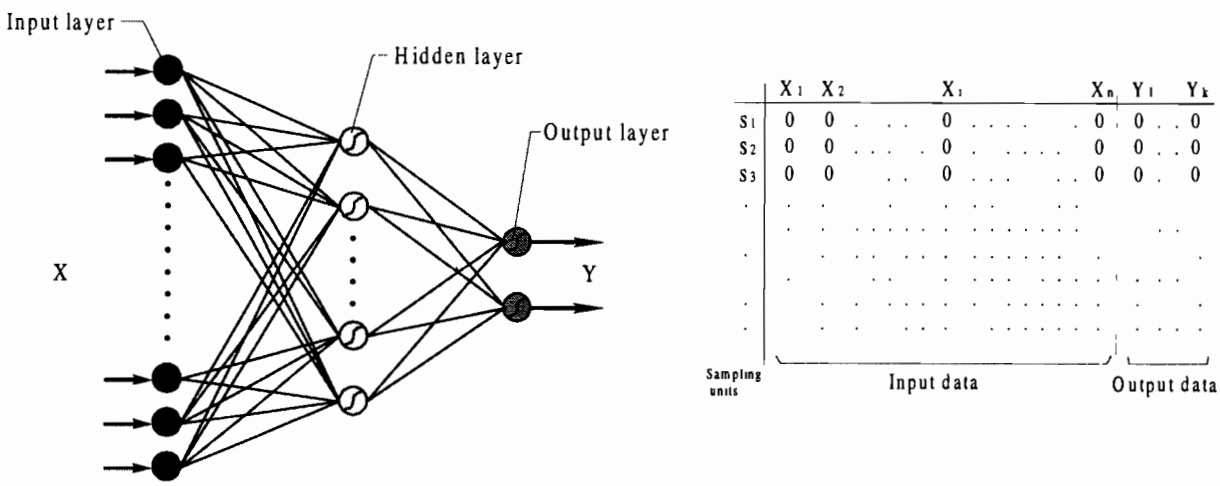

Fig. 1. Schematic illustration of a three-layered feed-forward neural network, with one input layer, one hidden layer and one output layer. The right-hand side of the figure shows the data set to be used in backpropagation network models. $X_1,\ldots, X_n$ are the input variables, $Y_1, \ldots, Y_k$ are the output variables, S_1, S_2, S_3, ... are the observation data.

works where feed-back connections are also permitted. The number of input and output units depends on the representations of the input and the output objects, respectively. The hidden layer(s) is (are) an important parameters in the network. BPNs with an arbitrary number of hidden units have been shown to be universal approximators (Cybenko, 1989; Hornick et al., 1989) for continuous maps and can therefore be used to implement any function defined in these terms.

The BPN is one of the easiest networks to understand. Its learning and update procedure is based on a relatively simple concept: if the network gives the wrong answer, then the weights are corrected so the error is lessened so future responses of the network are more likely to be correct. The conceptual basis of the backpropagation algorithm was first presented in by Webos (1974), then independently reinvented by Parker (1982), and presented to a wide readership by Rumelhart et al. (1986).

In a training phase, a set of input/target pattern pairs is used for training and presented to the network many times. After the training is stopped, the performance of the network is tested. The BPN learning algorithm involves a forward-propagating step followed by a backward-propagating step. A training set must have enough examples of data to be representative for the overall problem. However, the training phase can be time consuming depending on the network structure (number of input and output variables, number of hidden layers and number of nodes in the hidden layer), the number of examples in the training set, the number of iterations (see Box 1).

Typically, for a BPN to be applied, both a training and a test set of data are required. Both training and test sets contain input/output pattern pairs taken from real data. The first is used to train the network, and the second to assess the performance of the network after training. In the testing phase, the input patterns are fed into the network and the desired output patterns are compared with those given by the neural network. The agreement or disagreement of these two sets gives an indication of the performance of the neural network model.

Box 1. A brief algorithm of backpropagation in neural networks

(1) Initialize the number of hidden nodes
(2) Initialize the maximum number of iterations and the learning rate (η). Set all weights and thresholds to small random numbers. Thresholds are weights with corresponding inputs always equal to 1.
(3) For each training vector (input $X_p = (x_1,x_2,\ldots,x_n)$, output Y) repeat steps 4–7.
(4) Present the input vector to the input nodes and the output to the output node;
(5) Calculate the input to the hidden nodes: $a_j^h = \Sigma_{i=1}^{n} W_{ij}^h x_i$. Calculate the output from the hidden nodes: $x_j^h = f(a_j^h) = \frac{1}{1+e^{-a_j^h}}$.
Calculate the inputs to the output nodes: $a_k = \Sigma_{j=1}^{L} W_{jk} x_j^h$ and the corresponding outputs: $\hat{Y}_k = f(a_k) = \frac{1}{1+e^{-a_k}}$.
Notice that $k=1$ and $\hat{Y}_k = \hat{Y}$, L is the number of hidden nodes.
(6) Calculate the error term for the output node: $\delta_k = (Y - \hat{Y}) f'(a_k)$ and for the hidden nodes: $\delta_j^h = f'(a_j^h) \Sigma_k \delta_k W_{jk}$
(7) Update weights on the output layer: $W_{jk}(t+1) = W_{jk}(t) + \eta \delta_k x_j^h$ and on the hidden layer: $W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j^h x_i$

As long as the network errors are larger than a predefined threshold or the number of iterations is smaller than the maximum number of iterations envisaged, repeat steps 4–7.

Another decision that has to be taken is the subdivision of the data set into different sub-sets which are used for training and testing the BPN. The best solution is to have separate data bases, and to use the first set for training and testing the model, and the second independent set for validation of the model (Mastrorillo et al., 1998). This situation is rarely observed in ecology studies, and partitioning the data set may be applied for testing the validity of the model. We present here two partitioning procedures:

1. if enough examples of data sets are available, the data may be divided randomly into two parts: the training and test sets. The proportion may be 1:1, 2:1, 3:1, etc. for these two sets. However, the training set still has to be large enough to be representative of the problem and the test set has to be large enough to allow correct validation of the network. This procedure of partitioning the data is called k-fold cross-validation, sometimes named the hold-out procedure (Utans and Moody, 1991; Geman et al., 1992; Efron and Tibshirani, 1995; Kohavi, 1995; Kohavi and Wolpert, 1996; Friedman, 1997).
2. if there are not enough examples available to permit the data set to be split into representative training and test sets, other strategies may be used, like cross-validation. In this case, the data set is divided into n parts usually small, i.e. containing few examples of data. The BPN may now be trained with $n-1$ parts, and tested with the remaining part. The same network structure may be repeated to use every part once in a test set in once of the n procedures. The result of these tests together allow the performance of the model to be determined. Sometimes, in extreme cases, the test set can have only one example, and this is called the leave-one-out or sometime Jacknife procedure (Efron, 1983; Kohavi, 1995). The procedure is often used in ecology when either the available database is small or each observation is unique information and different to the others.

3.2. Kohonen self-organizing mapping (SOM)

Kohonen SOM falls into the category of unsupervised learning methodology, in which the relevant multivariate algorithms seek clusters in the data (Everitt, 1993). Conventionally, at least in ecology, reduction of multivariate data is normally carried out using principal components analysis or hierarchical clustering analysis (Jongman et al., 1995). Unsupervised learning allows the investigator to group objects together on the basis of their perceived closeness in n dimensional hyperspace (where n is the number of variables or observations made on each object).

Formally, a Kohonen network consists of two types of units: an input layer and an output layer (Fig. 2). The array of input units operates simply as a flow-through layer for the input vectors and has no further significance. In the output layer, SOM often consist of a two-dimensional network of neurons arranged in a square (or other geometrical form) grid or lattice. Each neuron is connected to its n nearest neighbours on the grid. The neurons store a set of weights (weight vector) each of which corresponds to one of the inputs in the data. The SOM algorithm can be characterized by several steps (see Box 2).

Box 2. A brief algorithm of self-organizing mapping neural networksLet a data set of observations with n-dimensional vectors:

Initialise the time parameter t: $t=0$.

(1) Initialise weights W_{ij} of each neuron j in the Kohonen map to random values (for example, random observations).

(2) Present a training sample$x(t)=[x_1(t),\ldots,x_n(t)]$ randomly selected from the observations.

(3) Compute the distances d_j between x and all mapping array neurons j according to: $d_j(t)=\Sigma_{i=1}^{n}[x_i(t)-W_{ij}(t)]^2$ where $x_i(t)$ is the i^{th} component of the N dimensional input vector and $W_{ij}(t)$ is the connection strength between input neuron i and map array neuron j at time t expressed as a Euclidean distance.

(4) Choose the mapping array neuron j^* with minimal distance d_{j^*}: $d_{j^*}(t)=\min[d_j(t)]$.

(5) Update all weights, restricted to the actual topological neighbourhood $NE_{j^*}(t)$: $W_{ij}(t+1)=W_{ij}(t)+\eta(t)(x_i(t)-W_{ij}(t))$ for $j\in NE_{j^*}(t)$ and $1\leq i\leq n$. Here $NE_{j^*}(t)$ is a decreasing function of time, as is the gain parameter $\eta(t)$.

(6) Increase the time parameter t

(7) If $t<t_{\max}$ return to step 2

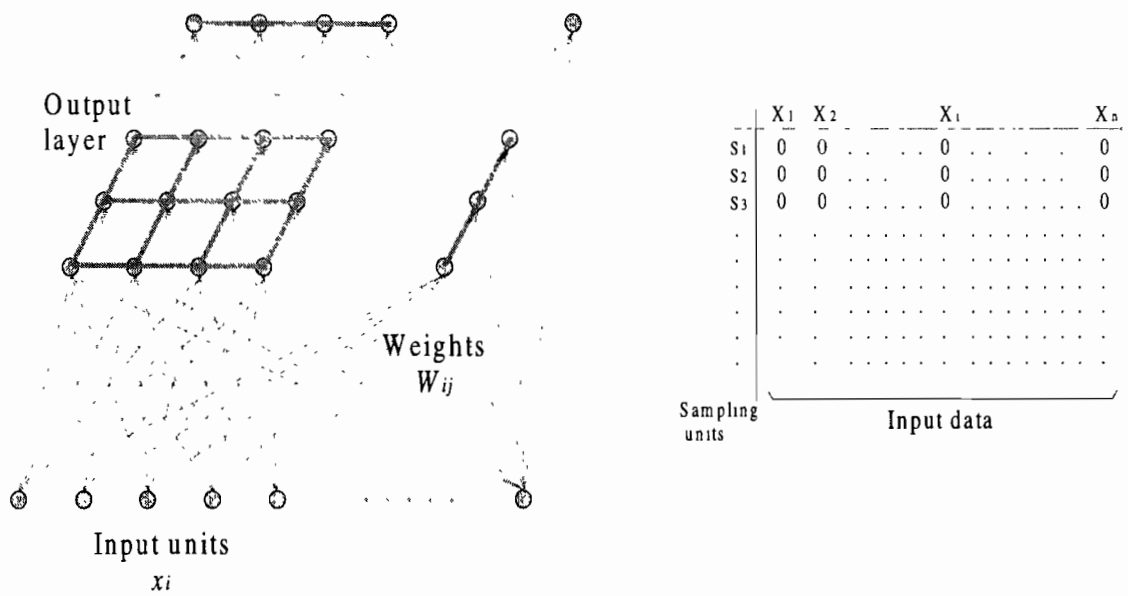

Fig 2. A two-dimensional Kohonen self-organizing feature map network. The right-hand side shows the data set to be used in Kohonen self-organizing mapping models. X_1, ..., X_n are the input variables, S_1, S_2, S_3, . are the observation data

Since the introduction of the Kohonen neural network (Kohonen, 1982, 1984), several training strategies have been proposed (see e.g. Lippmann, 1987; Hecht-Nielsen, 1990; Freeman and Skapura, 1992) which deal with different aspects of the use of the Kohonen network. In this section, we will restrict the study to the original algorithm proposed by Kohonen (1984).

4. Overview of the presented papers

During the three days of the workshop on ANN applications in ecology, 45 oral communications and posters were presented. They were thoroughly discussed by 100 or so participants coming from 24 countries. The session started with the general review 'state-of-the-art of ecological modelling with emphasis on development of structural dynamic models' (Jørgensen, see paper in the next chapter). Then applications of ANNs in several fields of ecology were presented: primary production in freshwater and marine ecosystems (seven papers), remote sensing data (six papers), population and community ecology and ecosystems (six papers), global change and ecosystem sensitivity (six papers), fishery research in freshwater and marine ecosystems (four papers), evolutionary ecology and epidemiology (three papers), population genetics (two papers) and seven remaining papers which rather concerned the methodological aspects, i.e. improvement of ANN models in ecological modelling. Some of these papers have been selected for publication in this special issue. The aim of this special issue, as well as of this first workshop, was both to contribute to an improvement of methodology in ecological modelling and to stimulate the integration of ANNs in ecological studies.

Most of the papers propose the use of a backpropagation algorithm in ANN models. Certain papers suggest improvement by including the Bayesian (see Vila et al.' paper) or radial base functions (see Morlini's paper). Only a few papers used unsupervised learning to model remote sensing data, microsatellite data, or marine ecology data (see Foody's paper).

5. Future developments of ANNs in ecological modelling

In 1992, during the first international conference on mathematical modelling in limnology (Innsbruck, Austria), Jørgensen (1995) presented a review on ecological modelling in limnology. He noted the rapid growth of ecological modelling and proposed a chronological development in four generations of models. The first models covered the oxygen balance in streams and the prey-predator relationships (the Lotka-Volterra model) in the early 1920s. The second phase of modelling (in the 1950s and 1960s) was particularly concerned with population dynamics. The third generation started from the 70's with more complicated models and rapidly became tools in environment management, e.g. eutrophication models. In the fourth generation, more recent models are becoming increasingly able to take the complexity, adaptability and flexibility of ecosystems into account.

As the modelling techniques available in the fourth generation of ecological models, researchers have a lot of methods ranging from numerical, mathematical and statistical methods to techniques based on artificial intelligence, particularly ANNs. During the last 2 decades of the current century, the growing development of computer-aided analysis, easily accessible to all researchers has facilitated the applications of ANNs in ecological modelling.

To use ANN programmes, ecologists can obtain freeware or shareware using different web sites in the World. Users interested could find these programmes by filling in 'neural network' as a keyword in the search procedure of the web explorer. Thus, they can obtain many computer ANN programmes functioning with all operating systems (Windows, Apple, Unix stations, etc.). Moreover, increasingly specialized ANN packages are proposed at acceptable prices for personal computers and most professional statistical software now proposes ANN procedures included (e.g. SAS, Splus, Matlab, etc.).

The development of computers and ANN software must allow ecologists to apply ANN methods more easily to resolve the complexity of relationships between variables in ecological data. A lot of reports, and especially the papers presented in this first workshop on the applications of ANNs in ecology, demonstrate the importance of these methods in ecological modelling. The second workshop on this subject is programmed for November 2000 in Adelaide University (Australia), and is being organized by F. Recknagel (Email: frecknag@waite.adelaide.edu.au) and S. Lek (Email: lek@cict.fr). You are cordially invited to participate in this meeting.

Acknowledgements

We would like to express our cordial thanks to Elsevier Science B.V. and to Professor S.E. Jørgensen for accepting to publish these Proceedings in a special volume of Ecological Modelling. Special thanks are due to the different agencies which have supported the ANN workshop (Centre National de Recherche Scientifique, Paul Sabatier University, Electricité De France, Agence de l'eau d'Adour-Garonne, Caisse d'épargne Midi-Pyrénées, French ministry of Foreign Affairs, the regional council of Midi-Pyrénées, OKTOS).

References

Ackley, D.H., Hinton, G.E., Sejnowski, T.J., 1985. A learning algorithm for Boltzmann machines. Cogn. Sci. 9, 147–169.

Albiol, J., Campmajo, C., Casas, C., Poch, M., 1995. Biomass estimation in plant cell cultures: a neural network approach. Biotechn. Progr. 11, 88–92.

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996. Stochastic models that predict trouts population densities or biomass on macrohabitat scale. Hydrobiologia 337, 1–9.

Bishop, M.C., 1995. Neural Networks for Pattern Recognition Clarendon Press, Oxford, UK, p. 482

Bradshaw, J.A., Carden, K.J., Riordan, D., 1991. Ecological Applications Using a Novel Expert System Shell. Comp. Appl. Biosci. 7, 79–83.

Brey, T., Jarre-Teichmann, A., Borlich, O., 1996 Artificial neural network versus multiple linear regression: predicting P/B ratios from empirical data. Marine Ecol Progr. Series 140, 251–256.

Chu, W.C., Bose, N.K., 1998. Speech Signal Prediction Using Feedforward Neural-Network. Electr Lett 34, 999–1001

Colasanti, R.L., 1991. Discussions of the possible use of neural network algorithms in ecological modelling. Binary 3, 13–15.

Cosatto, E., Graf, H.P., 1995. A Neural-Network Accelerator for Image-Analysis. IEEE Micro 15, 32–38.

Cybenko, G, 1989. Approximations by superpositions of a sigmoidal function, Mathematics of Control Signals and Systems 2, 303–314.

d'Angelo, D.J, Howard, L.M., Meyer, J.L., Gregory, S.V., Ashkenas, L.R., 1995. Ecological use for genetic algorithms: predicting fish distributions in complex physical habitats. Can. J Fish. Aquat. Sc. 52, 1893–1908.

Dekruger, D., Hunt, B.R., 1994. Image-Processing and Neural Networks for Recognition of Cartographic Area Features. Pattern Recogn. 27, 461–483

Edwards, M., Morse, D.R, 1995. The potential for computer-aided identification in biodiversity research. Trends Ecol. Evol. 10, 153–158

Efron, B, 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Statist. Assoc. 78 (382), 316–330.

Efron, B., Tibshirani, R.J. 1995. Cross-validation and the bootstrap: estimating the error rate of the prediction rule Rep Tech. Univ. Toronto.

Everitt, B.S. 1993 Cluster analysis. Edward Arnold, London.

Faraggi, D., Simon, R., 1995. A neural network model for survival data. Stat. Med. 14, 73–82.

Freeman, J.A., Skapura, D.M., 1992. Neural networks. algorithms, applications and programming techniques. Addison-Wesley Publishing Company, Reading, Massachusetts, USA.

Friedman, J.H., 1997. On bias, variance, 0/1-loss and the curse-of-dimensionality Data Mining and Knowledge Discovery 1, 55–77.

Gallant, S.I., 1993. Neural network learning and expert systems. The MIT Press, Massachusetts, USA, p. 365.

Geman, S., Bienenstock, E, Doursat, R., 1992. Neural networks and the bias/variance dilemma'. Neural computation 4, 1–58.

Giske, J., Huse, G, Fiksen, O., 1998. Modelling spatial dynamics of fish. Rev. Fish. Biol. Fish. 8, 57–91.

Golikov, S.Y., Sakuramoto, K., Kitahara, T, Harada, Y., 1995. Length-Frequency Analysis Using the Genetic Algorithms. Fisheries Sci. 61, 37–42.

Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.

Hecht-Nielsen, R., 1990. Neurocomputing. Addison-Wesley, Massachusetts, USA.

Hopfield, J.J., 1982. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad. Sci. USA 79, 2554–2558.

Hornick, K, Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural networks 2, 359–366.

Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R., 1995. Data analysis in community and landscape ecology. Cambridge University Press, England.

Jørgensen, S.E., 1995. State-of-the-art of ecological modelling in limnology. Ecol. Model. 78, 101–115.

Kastens, T.L., Featherstone, A.M., 1996 Feedforward backpropagation neural networks in prediction of farmer risk preference. Am. J. Agri. Econ. 78, 400–415.

Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection. Proc. of the 14th Int. Joint Conf. on Artificial Intelligence, Morgan Kaufmann Publishers, 1137–1143.

Kohavi, R., Wolpert, D.H., 1996. Bias plus variance decomposition for zero-one loss functions. In: Saitta, L. (Ed.), Machine learning· Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Bari, Italy, pp. 275–283.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69.

Kohonen, T., 1984 Self-organization and associative memory. Springer-Verlag, Berlin (Germany).

Kung, S.Y., Taur, J.S., 1995. Decision-Based Neural Networks with Signal Image Classification Applications. IEEE Trans. on Neural Networks 6, 170–181.

Kvasnicka, V., 1990. An application of neural networks in chemistry. Chem. papers, 44(6): 775–792

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquatic Living Res. 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996b. Application of neural networks to modelling nonlinear relationships in ecology Ecol Model. 90, 39–52.

Lerner, B., Levinstein, M, Rosenberg, B., Guterman, H, Dinstein, I., Romem, Y., 1994. Feature selection and chromosomes classification using a multilayer perceptron neural network., IEEE Int Confer. on Neural Networks, Orlando (Florida), pp 3540–3545

Lippmann, R.P., 1987. An introduction to computing with neural nets. IEEE Acoust. Speech Signal Process. Mag., April: 4-22.

Lo, J Y., Baker, J.A, Kornguth, P.J., Floyd, C.E., 1995. Application of Artificial Neural Networks to Interpretation of Mammograms on the Basis of the Radiologists Impression and Optimized Image Features. Radiology 197, 242–242.

Mastrorillo, S., Dauba, F., Oberdorff, T., Guégan, J.F., Lek, S., 1998. Predicting local fish species richness in the Garonne river basin., C.R. Acad. Sci. Paris. Life Sciences 321, 423–428.

Parker, D.B., 1982. Learning logic Invention report S81-64, File 1, Office of Technology Licensing, Stanford University.

Rahim, M.G., Goodyear, C.C., Kleijn, W.B., Schroeter, J., Sondhi, M.M., 1993. On the Use of Neural Networks in Articulatory Speech Synthesis. J. Acoustical Soc Am. 93, 1109–1121.

Recknagel, F., Petzoldt, T., Jaeke, O., Krusche, F., 1994. Hybrid Expert-System Delaqua-A Toolkit for Water-Quality Control of Lakes and Reservoirs. Ecol. Model. 71, 17–36.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Ripley, B.D., 1994. Neural networks and related methods for classification. J. R. Stat. Soc., B 56 (3), 409–456.

Rosenblatt, F., 1958. The Perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecol. Progr. Series 139, 289–299.

Seginer, I., Boulard, T., Bailey, B.J., 1994. Neural network models of the greenhouse climate. J. Agric. Eng. Res. 59, 203–216.

Smith, M., 1994. Neural networks for statistical modelling. Van Nostrand Reinhold, NY, p. 235.

Smits, J.R.M., Breedveld, L.W., Derksen, M.W.J., Katerman, G., Balfoort, H.W., Snoek, J., Hofstraat, J.W., 1992. Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. Anal. Chim. Acta 258, 11–25.

Utans, J., Moody, J.E., 1991. Selecting neural network architectures via the prediction risk: application to corporate bond rating prediction. In Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street, IEEE Computer Society Press, Los Alamitos, CA.

Villa, F., 1992. New computer architectures as tools for ecological thought. Trends Ecol. Evol. 7, 179–183.

Webos, P., 1974. Beyond regression: new tools for prediction and analysis in the behavioral sciences. Thesis, Harvard University.

Widrow, B., Hoff, M.E., 1960. Adaptive switching circuits. IRE Wescon conference record, August 23–26, 4: 96–104.

Wythoff, B.J., Levine, S.P., Tomellini, S.A., 1990. Spectral peak verification and recognition using a multilayered neural network. Anal. Chem. 62 (24), 2702–2709.

ELSEVIER

Ecological Modelling 120 (1999) 75–96

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

State-of-the-art of ecological modelling with emphasis on development of structural dynamic models

Sven Erik Jørgensen *

DFH, Environmental Chemistry, University Park 2, 2100 Copenhagen, Denmark

Abstract

The paper deals with two major problems in ecological modelling today, namely how to get reliable parameters? and how to build ecosystem properties into our models? The use of new mathematical tools to answer these questions is mentioned briefly, but the main focus of the paper is on development of structural dynamic models which are models using goal functions to reflect a current change of the properties of the biological components in the models. These changes of the properties are due to the enormous adaptability of the biological components to the prevailing conditions. All species in an ecosystem attempt to obtain most biomass, i.e. to move as far away as possible from thermodynamic equilibrium which can be measured by the thermodynamic concept exergy. Consequently, exergy has been proposed as a goal function in ecological models with dynamic structure, meaning currently changed properties of the biological components and in model language currently changed parameters. An equation to compute an exergy index of a model is presented. The theoretical considerations leading to this equation are not presented here but references to literature where the basis theory can be found are given. Two case studies of structural dynamic modelling are presented: a shallow lake where the structural dynamic changes have been determined before the model was developed, and the application of biomanipulation in lake management, where the structural dynamic changes are generally known. Moreover. it is also discussed how the same idea of using exergy as a goal function in ecological modelling may be applied to facilitate the estimation of parameters.

Keywords: Structural dynamic models; Biological components; Ecological modelling

1. Introduction: state-of-the-art of ecological modelling

Ecological modelling originates from Lotka–Volterra and Streeter–Phelps in the 1920s, while the comprehensive use of models in environmental management started in the beginning of the 1970s. During the seventies we learnt that development of ecological models requires a comprehensive knowledge to the functioning of ecosystems and that it is extremely important to find a balanced complexity considering the available data, the ecosystem and the focal problem. Meanwhile many models have been developed and today we have the experience from more than 4000 ecological models which have been used as tool in research or environmental management.

* Fax: +45-35375744.
E-mail address: sej@mail dfh.dk (S E. Jørgensen)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00093-9

Recently, a book 'Environmental and Ecological Modelling' by Jørgensen et al. (1995b), Lewis Publisher, reviewed more than 400 models and gave details about the models. The idea was to give the experience from previous modelling studies to those who want to develop models of similar ecosystems or focusing on similar environmental problems. In spite of the widely gained experience in ecological modelling, we are still facing serious problems in modelling, which, however, we attempt to overcome. The main problems are in short:

1. usually/often we cannot get sufficient data to develop models which can give reliable prognoses
2. the parameter estimation is often the weakest point in modelling
3. the models do not reflect the real properties of ecosystems, particularly their adaptability and ability to meet change in forcing functions with change in species composition. Several research ideas have been pursued to solve these problems:
   - 3.1. fuzzy models are used to overcome the problem of a poor data base (Jørgensen, 1994a)
   - 3.2. use of chaos and fractal theory in modelling to improve the parameter estimation (Jørgensen, 1995, 1997)
   - 3.3. use of catastrophe theory in modelling as an attempt to model structural changes (see Jørgensen, 1997).
   - 3.4. use of artificial intelligence in parameter estimation (Kompare, 1995)
   - 3.5. recently developed parameter estimation methods (Jørgensen, 1995, 1997, 1998)
   - 3.6. data base of ecological parameters (Jørgensen et al., 1991). A parameter data base three times larger than the volume published in 1991 is under development on a CD.
   - 3.7. development of structural dynamic models by use of goal functions (Jørgensen, 1986, 1988, 1990, 1992a,b, 1994a,b,c, 1997; Jørgensen et al., 1995a; Jørgensen and Padisak, 1996; Jørgensen and de Bernardi, 1997) to account for the ecosystem properties.

This presentation will concentrate on the last development (vii), but will also briefly touch the use of chaos, fractal and catastrophe theory and the relationship between the development of structural dynamic models and the emergence of additional parameter estimation methods.

2. How can we consider ecosystem properties in our model developments?

Ecology deals with irreducible systems (Wolfram, 1984a,b; Jørgensen, 1990, 1992a,b, 1994a, 1995; Jørgensen et al., 1995a). We cannot design simple experiments which reveal a relationship that can in all detail be transferred from one ecological situation and one ecosystem to another situation in another ecosystem. That is possible for instance with Newton's laws on gravity, because the relationship between forces and acceleration is reducible. The relationship between force and acceleration is linear, but growth of living organisms is dependent on many interacting factors, which again are functions of time. Feedback mechanisms will simultaneously regulate all the factors and rates and they also interact and are functions of time, too (Straskraba, 1979).

Table 1 shows the hierarchy of regulation mechanisms, that are operating at the same time. The regulation mechanisms operate over different time scales which are indicated in the table. From this example the complexity alone clearly prohibits the reduction to simple relationships that can be used repeatedly. An ecosystem consists of so many interacting components that it is impossible ever to be able to examine all these relationships and even if we could, it would not be possible to separate one relationship and examine it carefully to reveal its details, because the relationship is different when it works in nature with interactions from the many other processes, from when we examine it in a laboratory with the relationship separated from the other ecosystem components. The observation, that it is impossible to separate and examine processes in real ecosystems, corresponds to that of the examinations of organs that are separated from the organisms in which they are working. Their functions are com-

Table 1
The hierarchy of regulating feedback mechanisms, (Jørgensen, 1994a, 1997)

Level	Explanation of regulation process	Exemplified by phytoplankton growth	Time scale
1	Rate by concentration in medium	Uptake of phosphorus in accordance with phosphorus concentration	Min–h
2	Rate by needs	Uptake of phosphorus in accordance with intracellular concentration	Min–h
3	Rate by other external factors (biochemical adaptation)	Chlorophyll concentration in accordance with previous solar radiation	Days
4	Adaptation of properties (biological adaptation)	Change of optimal temperature for growth	Days–months
5	Selection of other species	Shift to better fitted species	Weeks–years
6	Selection of other food web	Shift to better fitted food web	Months–years
7	Mutations, new sexual recombinations and other shifts of genes	Emergence of new species or shifts of species properties	10–10^5 years

pletely different when separated from their organisms and examined in for instance a laboratory from when they are placed in their right context and in 'working' condition. These observations are indeed expressed in ecosystem-ecology. A known phrase is: 'everything is linked to everything' or: 'the whole is greater than the sum of the parts' (Allen 1988). It implies that it may be possible to examine the parts by reduction to simple relationships, but when the parts are put together they will form a whole, that behaves differently from the sum of the parts. This statement requires a more detailed discussion of how an ecosystem works.

The complexity of an ecosystem is formed not only by a high number of interacting components; the complexity is far more complex. Ecosystems belong to the class of systems denoted complex adaptive systems (Brown, 1995). The number of feedbacks and regulations is extremely high and makes it possible for the living organisms and populations to survive and reproduce in spite of changes in external conditions. Numerous examples can be found in the literature. If the actual properties of the species are changed the regulation is named adaptation. These regulations correspond to level 3 and 4 in Table 1. Phytoplankton is for instance able to regulate its chlorophyll concentration according to the solar radiation. If more chlorophyll is needed because the radiation is insufficient to guarantee growth, more chlorophyll is produced by the phytoplankton. The digestion efficiency of the food for many animals depends on the abundance of the food. The same species may be of different sizes in different environments, depending on what is most beneficial for survival and growth. If nutrients are scarce, phytoplankton may become smaller and vice versa. In this latter case the change in size is a result of a selection process, which is made possible because of the distribution in size as illustrated in Fig. 1.

The feedbacks are constantly changing, i.e. the adaptation is adaptable in the sense that if a regulation is not sufficient another regulation process higher in the hierarchy of feed-backs–see Table 1–will take over. The change in size within the same species is for instance only limited. When this limitation has been reached, other spe-

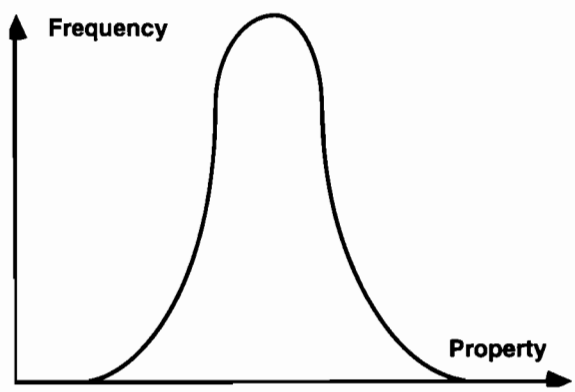

Fig. 1 Typical Gaussian frequency distribution of size within the same species.

cies will take over. It implies that not only the processes and the components, but also the feedbacks can be replaced, if it is needed to achieve a better utilisation of the available resources. The ecosystem and its properties emerge as a result of many simultaneous and parallel focal-level processes, as influenced by even more remote environmental features. This implies that the environment of a system includes historical factors as well as immediately cogent ones (Patten, 1997). The history of the ecosystem and its components is therefore important for the reactions and further development of the ecosystem. It is one of the main ideas behind Patten's indirect effect that the indirect effect accounts for the 'history,' while the direct effect only reflects the immediate effect. The importance of the history of the ecosystem and its components emphasises the need for a dynamic approach and supports the idea that we will never observe the same situation in an ecosystem twice. The history will always be 'between' two similar situations. Therefore, the equilibrium models may fail in their conclusions, particularly when we want to look into reactions on the system level.

Ecosystems show furthermore a high degree of heterogeneity in space and in time. An ecosystem is a very dynamic system. All its components and particularly the biological ones are steadily moving and their properties are steadily modified, which is why an ecosystem will never return to the same situation again. Every point is furthermore different from any other point and therefore offering different conditions for the various life forms. This enormous heterogeneity explains why there are so many species on earth. There is, so to say, an ecological niche for 'everyone' and 'everyone' may be able to find a niche where he is best fitted to utilise the resources. Ecotones, the transition zones between two ecosystems, offer a particular variability in life conditions, which often results in a particular richness of species diversity. Studies of ecotones have recently drawn much attention from ecologists, because ecotones have pronounced gradients in the external and internal variables, which give a clearer picture of the relation between external and internal variables. Margalef (1991) claims that ecosystems are anisotropic, meaning that they exhibit properties with different values, when measured along axes in different directions. It means that the ecosystem is not homogeneous in relation to properties concerning matter, energy and information, and that the entire dynamics of the ecosystem works toward increasing the differences. These variations in time and space make it particularly difficult to model ecosystems and to capture the essential features of ecosystems.

3. Ecosystems have dynamic structure

Ecosystems and their biological components, the species, develop/evolve steadily and in the long term perspective toward higher complexity. Darwin's theory describes the competition among species and states that the species, that are best fitted to the steadily changed prevailing conditions in the ecosystem will survive. The species are currently able to offer new combinations of properties due to self-organisation (Kauffman, 1996), sexual recombinations and mutations. All species in an ecosystem are confronted with the question: how is it possible to survive or even grow under the prevailing conditions? The prevailing conditions are considered as all factors influencing the species, i.e. all external and internal factors including those originating from other species. This explains the coevolution, as any change in the properties of one species will influence the evolution of the other species.

Species are generally more sensitive to stress than functional properties of ecosystems. Schindler (1988) observed in experimental acidifications of lakes that functional properties such as primary production, respiration and grazing were relatively insensitive to the effects of a continued exposure to acidification, while early signs of warning could be detected at the level of species composition and morphologies. This underlines the importance of development of models, denoted structural dynamic models, able to predict the changes in focal properties of the dominant species, included a possible shift in species composition by significant changes of external factors. All natural external and internal factors of ecosys-

tems are dynamic–the conditions are steadily changing, and there are always many species waiting in the wings, ready to take over, if they are better fitted to the emerging conditions than the species dominating under the present conditions. There is a wide spectrum of species representing different currently changed combinations of properties available for the ecosystem. The question is, which of the available combinations of properties are best able to ensure survival and growth under the present conditions and which available combinations of properties are best able to offer survival and growth under the conditions one time step further and two time steps further and so on? The necessity in Monod's sense is given by the prevailing conditions-the species must have genes or maybe rather phenotypes (meaning properties) which match these conditions, to be able to survive. But the natural external factors and the genetic pool available for the test may change randomly or by 'chance'.

Steadily new mutations (misprints are produced accidentally), sexual recombinations (the genes are mixed and shuffled) and results of self-organising processes (Kauffman, 1996) will emerge and give steadily new material to be tested toward the question: which species are best fitted under the conditions prevailing just now? These ideas are illustrated in Fig. 2. The external factors are steadily changed and some even relatively fast-partly at random, e.g. the meteorological or climatic factors. The species of the system are selected among the species available and represented by the genetic pool, which again is currently changed by mutations, new sexual recombinations and self-organising processes. What is named ecological development is the changes over time in nature caused by the dynamics of the external factors, giving the system sufficient time for the reactions, including an organisation of the network.

Evolution, on the other hand, is related to the genetic pool. It is the result of the relation between the dynamics of the external factors and the dynamics of the genetic pool. The external factors steadily change the conditions for survival and the genetic pool steadily comes up with new solutions to the problem of survival. Darwin's theory assumes that populations consist of individuals, who:

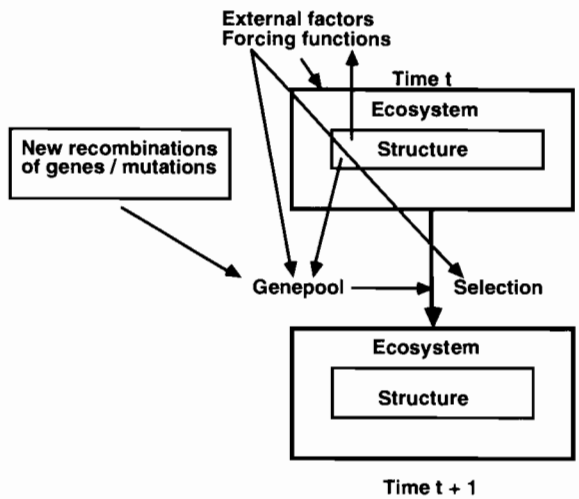

Fig. 2. Conceptualisation of how the external factors steadily change the species composition. The possible shifts in species composition are determined by the gene pool, which is steadily changed due to mutations and new sexual recombinations of genes. The development is, however, more complex. This is indicated by (1) arrows from 'structure' to 'external factors' and 'selection' to account for the possibility that the species are able to modify their own environment and thereby their own selection pressure and show self-organisation; (2) an arrow from 'structure' to 'gene pool' to account for the possibilities that the species can to a certain extent change their own gene pool.

1. On average produce more offspring than is needed to replace them upon their death–this is the property of high reproduction.
2. Have offspring which resemble their parents more than they resemble randomly chosen individuals in the population–this is the property of inheritance.
3. Vary in heritable traits influencing reproduction and survival (i.e. fitness)–this is the property of variation.

All three properties are part of the presentation in Fig. 2. The high reproduction is needed to get a change in the species composition caused by changes in external factors. The variability is represented in the short and long term changes in the genetic pool and the inheritance is needed to see an effect of the fitness test in the long run. Without the inheritance every new generation would

start from the same point and it would not be possible to maintain the result of the fitness test. The evolution is able to continue from the already obtained results. The species are continuously tested against the prevailing conditions (external as well as internal factors) and the better they are fitted, the better they are able to maintain and even increase their biomass. The specific rate of population growth may even be used as a measure for the fitness (Brown, 1995). But the property of fitness must of course be inheritable to have any effect on the species composition and the ecological structure of the ecosystem in the long run. Natural selection has been criticised for being a tautology: fitness is measured by survival and survival of the fittest therefore mean survival of the survivors. However, the entire Darwinian theory including the above mentioned three assumptions, should not be conceived as a tautology, but may be interpreted as follows: the species offer different solutions to survival under given prevailing conditions and the species that have the best combinations of properties to match the conditions, have also the highest probability of survival and growth.

Fitness is therefore a question of having the best combination of properties under the prevailing conditions and survival (growth) is the award to the organisms which have the fittest combination of properties. The formulation by Ulanowicz (1986) may also be applied: Those populations are fittest that best enhance the auto catalytic behaviour of the matter–energy loops in which they participate. Man-made changes in external factors, i.e. anthropogenic pollution have created new problems, because new genes fitted to these changes do not develop overnight, while most natural changes have occurred many times previously and the genetic pool is therefore prepared and fitted to meet the natural changes. The spectrum of genes is able to meet most natural changes, but not all of the man-made changes, because they are new and untested in the ecosystem.

The evolution moves toward increasing complexity in the long run; see Fig. 3. The fossil records have shown a steady increase of species diversity. There may be destructive forces–for instance man–made pollution or natural catastrophes–for a shorter time, but the probability that

1. new and better genes are developed; and
2. new ecological niches are utilised

will increase with time. The probability will even-again excluding the short time perspective-increase faster and faster, as the probability is roughly proportional to the amount of genetic material on which the mutations and new sexual recombinations can be developed.

It is equally important to note that a biological structure is more than an active non-linear system. In the course of its evolution, the biological structure is continuously changed in such a way that its structural map is itself modified. The overall structure thus becomes a representation of all the information received. Biological structure represents through its complexity a synthesis of the information with which it has been in communication (Schoffeniels, 1976). Evolution is maybe the most discussed topic in biology and ecology and millions of pages have been written about evolution and its ecological implications. Today the basic facts of evolution are taken for granted and the interest has shifted to more subtle classes of fitness/selection, i.e. toward an understanding of the complexity of the evolutionary processes. The coevolution explains the interactive processes

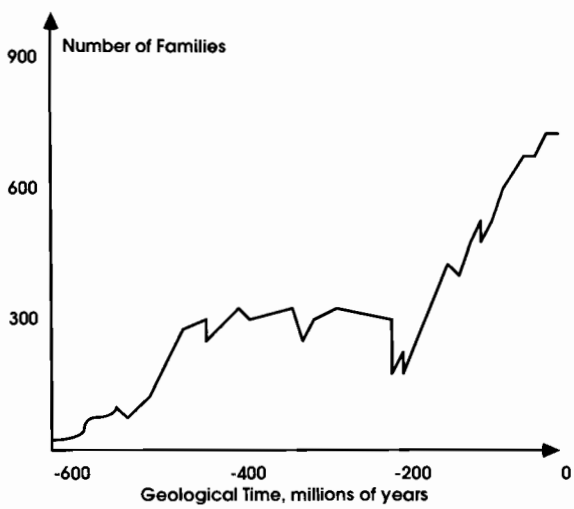

Fig. 3 Changes in species diversity over geological time (Shugart, 1998).

among species. It is difficult to observe a coevolution, but it is easy to understand that it plays a major role in the entire evolution process. The coevolution of herbivorous animals and plants is a very illustrative example. The plants will develop toward a better spreading of seeds and a better defence towards herbivorous animals. This will in the latter case create a selection of the herbivorous animals that are able to cope with the defence. Therefore the plants and the herbivorous animals will coevolve. Coevolution means that the evolution process cannot be described as reductionistic, but that the entire system is evolving. A holistic description of the evolution of the system is needed.

The Darwinian and neo-Darwinian theories have been criticised from many sides. It has for instance been questioned whether the selection of the fittest can explain the relatively high rate of the evolution. Fitness may here be measured by the ability to grow and reproduce under the prevailing conditions. It implies that the question raised according to the Darwinian theories (see the discussion above) is: 'which species have the properties that give the highest ability for growth and reproduction?' We shall not go into the discussion in this context-it is another very comprehensive theme–but just mention that the complexity of the evolution processes is often overlooked in this debate. Many interacting processes in the evolution, including self-organisation (Kauffman, 1996) may be able to explain the relatively high rate of evolution that is observed.

4. Problems associated with development of structural dynamic models

The problem associated with the development of ecological models is in short, that we base our model on an analysis of an ecosystem at a given time t, when the external factors and the species composition are given, but we would like to challenge the model to predict what is going to be the response to a given change in the external factors at a later time $t+1$, when not only the external factors but also the species composition and their adaptation processes have adapted to the new

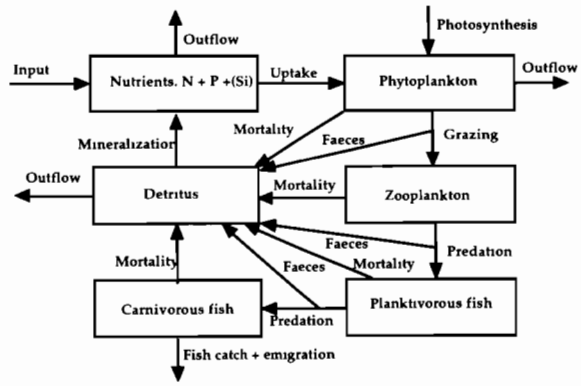

Fig. 4. The conceptual diagram of a typical eutrophication model The boxes indicate state variables and the arrows processes.

situation. Organisms develop and coevolve in the fitness landscape (Kauffman, 1996). When we analyse the ecosystem, we can presume that the organisms and their network have found if not the optimum solution then at least an excellent solution of combination of properties to obtain the highest possible survival and growth for all the organisms. The problem arises when the fitness landscape is changed due to change in the external factors, change in the properties of the organisms and even in the constraints associated with the interdependence of the organisms and their coevolution. We have with other words a different, in the best case a slightly different, fitness landscape at time $t+1$, and our model should describe that the ecosystem is able to find a new excellent solution to the challenge of the new fitness landscape. The model should therefore currently change the parameters representing the properties of the species included in the model. The organisms and species with the combination of properties offering the highest or even a very high peak in the fitness landscape should be represented as the new components in the model.

Let us describe the problem by means of an example. The model shown in Fig. 4 has been used several times to describe the development of eutrophication. A typical application of the model anticipates that we reduce the inputs of phosphorus (or other nutrients) significantly. A removal of 90% or even more is possible by the right environ-

mental management plan, but it implies that other species of phytoplankton are better fitted to the new and lower nutrient concentration. Therefore we should be able to change the parameters accordingly in the model. Furthermore, the zooplankton will change too because their food source has been changed in size and maybe even in elementary composition. This implies that also the planktivorous fish will change their properties because they will (probably) meet a food source in form of zooplankton with (probably, often) increased size. How can we determine all these changes in parameters? We know, that each species tries to get the best possible survival and fastest growth. Survival could be measured by the biomass, but an optimisation of the biomass of several species at the same time, requires that we sum up the biomass or perhaps make an addition of the weighted biomass. An addition of the biomass seems not to be an appropriate idea, as the plants and trees for instance in a forest ecosystem will be so dominant, that changes in the biomass of for instance foxes will be negligible. The possibility for a fish to find a new pathway for survival under new and emergent circumstances is furthermore much better than for phytoplankton due to the more advanced properties of a fish. A fish can move to the corner of the ecosystem where the food resources are most abundant, it can see, smell and hear in which direction it is most beneficial to move. A fish carries more information in its genes, an information which is used to obtain a better survival.

The crucial question is: 'How can we quantify the height in the fitness landscape? If we could propose a quantification of the 'size of fitness', we could be able calculate the fitness for any combination of properties for instance by an ecological model and select the combination giving the best fitness. The thermodynamic variable exergy seems to be an appropriate candidate as measure of the height of fitness. Exergy is defined as the amount of work, the system can perform when brought into thermodynamic equilibrium with a well defined reference state (for instance the same system at thermodynamic equilibrium at the same temperature and pressure as the considered ecosystem). Exergy measures therefore the distance from thermodynamic equilibrium, where there is no structure and no free energy available The amount of exergy stored in the system has the following advantages as measure of the fitness height:

1. Biomass contribute significantly to the exergy. The contribution is the free energy of biomass, approximately 18.7 kJ/g. Survival is measured by the biomass of living organisms.
2. Information has also exergy accordance to Boltzmann (1905). The free energy (work) of information is RT lnW, where W is the number of possible microstates among which one has been selected for the focal system. This contribution from information implies that the information carried by the various species will be included in the amount of exergy stored in the system. The information is applied by the species to ensure survival or even growth under the prevailing conditions.
3. Biomass and information are directly linked to the structure and order of the system in opposition to the random state at thermodynamic equilibrium. The total distance in energy unit from thermodynamic equilibrium is equal to the exergy.

It can be shown (Jørgensen et al., 1995a; Jørgensen, 1997) that the exergy of a model, Ex, may be calculated as (the system at thermodynamic equilibrium as reference state as indicated above):

$$\mathrm{Ex} = \sum_{i=0}^{i=n} \beta_i c_i \qquad (1)$$

where β_i is a weighting factor accounting for the information the species carry, while c_i is the concentration in for instance g/m^3 for aquatic ecosystems and g/m^2 for terrestrial ecosystems. How it is possible to come from the definition of exergy to Eq. (1) can be found in the references given above. Calculated β-values for various organisms are shown in Table 2 (Sources: Li and Grauer, 1991; Lewin, 1994). They are based on information about the number of non-nonsense genes in the various species. These calculations are based on the use of the same system as the one under consideration but at thermodynamic equilibrium anticipating the same temperature and pressure.

The unit applied is exergy in detritus exergy equivalents. As 1 g of detritus has approximately 18.7 kJ of free energy, it is easy to obtain the exergy content i kJ by multiplication by 18.7 of the number resulting from Eq. (1).

Application of Eq. (1) for determination of the exergy content corresponding to the model implies that the calculations determine the amount of work the system can perform entirely due to its chemical composition and its content of information, because there is no difference between the temperature, pressure and other potentials between the focal system and the reference state. Moreover, the exergy found by these computations is of course only the exergy of the model which is always a simplification of the real ecosystem. If we know the composition of the real ecosystem (which we will never be able to do in all details) then we could of course use the same equation to find the exergy of the ecosystem. It seems, however, more appropriate to use the term 'an exergy index' for Ex in accordance with Eq. (1), because we will never know the complete composition of an ecosystem and the equation (see for instance Jørgensen, 1997) is anyhow an approximation as most thermodynamic calculations are. It is assumed that the exergy index expresses the fitness (the height in the fitness landscape) and can be applied to find the combination of parameters (properties of the species) ensuring the best survival and growth.

Table 2
Approximate number of non repetitive genes

Organisms	Number of information genes	Conversion factor [a]
Detritus	0	1
Minimal cell (Morowitz, 1992)	470	2.7
Bacteria	600	3.0
Algae	850	3.9
Yeast	2000	6.4
Fungus	3000	10.2
Sponges	9000	30
Moulds	9500	32
Plants, trees	10 000–30 000	30–87
Worms	10 500	35
Insects	10 000–15 000	30–46
Jellyfish	10 000	30
Zooplankton	10 000–15 000	30–46
Fish	100 000–120 000	300–370
Birds	120 000	390
Amphibians	120 000	370
Reptiles	130 000	400
Mammals	140 000	430
Human	250 000	740

[a] Based on number of information genes and the exergy content of the organic matter in the various organisms, compared with the exergy contained in detritus. 1 g detritus has about 18.7 kJ exergy (= energy which can do work).

5. Modelling structural dynamics

If we follow the usually applied modelling procedure, we will attain a model that describes the processes in the focal ecosystem, but the parameters will represent the properties of the state variables as they are in the ecosystem during the examination period. They are not necessarily valid for another period of time, because we know that an ecosystem is able to regulate, modify and change them, if needed as response to the change in the prevailing conditions, determined by the forcing functions and the interrelations between the state variables. Our present models have rigid structures and a fixed set of parameters, reflecting that no changes or replacements of the components are possible. This may cause problems for the modeller, as he in the calibration phase attempts to find a set of parameters that is able to give an acceptable fit between model results and observations. It may be an impossible task-not because the model gives an incorrect picture of the focal processes in the ecosystem, but because the properties of the components covered by the parameters do change during the time of simulation due to seasonal and diurnal changes of the forcing functions. It may therefore be necessary to use time-varying parameters to get an acceptable model calibration. Patten (1997) has used this approach in a linear bear model. He demonstrates that it is possible to use a set of linear differential equations with time-varying parameters to get a good accordance between model and observations. He claims that the use of non-linear differential equations often is based on our attempt to get an acceptable fit by the unrealistic use of a rigid set of parameters.

We need to introduce parameters (properties) that can change according to changing forcing functions and general conditions for the state variables (components) to be able to optimise continuously the ability of the system to move away from thermodynamic equilibrium. Consequently, we may be able to hypothesise, that the level 5 and 6 in the regulation hierarchy Table 1 can be accounted for in our model by a current change of parameters according to an optimisation of exergy computed by Eq. (1). The idea is currently to test if a change of the most crucial parameters is able to produce a higher exergy of the system and, if that is the case, to use that set of parameters. Exergy is used in the modelling procedure as a so-called goal function. Thereby we obtain a better description of the regulation mechanisms in our model. If this hypothesis works, we obtain more realistic models that are able to describe more accurately our observations, and we get at least a certain support for the hypothetical fourth law of thermodynamics.

The type of models that are able to account for the change in species composition as well as for the ability of the species, i.e. the biological components of our models, to change their properties, i.e. to adapt to the prevailing conditions imposed on the species, are sometimes called structural dynamic models to indicate, that they are able to capture structural changes. They may also be called the next generation of ecological models to underline that they are radically different from previous modelling approaches and can do more, namely describe changes in species composition. It could be argued that the ability of ecosystems to replace present species with other (level 6 in Table 1), better fitted species, can be modelled by construction of models that encompass all actual species for the entire period that the model attempts to cover. This approach has, however, two essential disadvantages. The model becomes first of all very complex, as it will contain many state variables for each trophic level. It implies that the model will contain many more parameters that have to be calibrated and validated. This will introduce a high uncertainty to the model and will render the application of the model very case specific (Nielsen, 1992). In addition, the model will still be rigid and not give the model the property of the ecosystems to have continuously changing parameters even without changing the species composition (Fontaine, 1981).

Exergy has been used most widely as a goal function in ecological models, and the result of some case studies will be presented and discussed below. It should be emphasised, that we are calculating by the proposed method only an approximate and relative value of the exergy, based on statistical thermodynamic considerations. A relative value is, however, sufficient for the use of an exergy index as goal functions in models. It is obviously of theoretical as well as of environmental management interest to develop models which are able to predict changes in the species composition and/or in the ecological structure or at least to indicate the changes of the important properties of the dominant species to account for ecosystem reactions to changes in external factors. The idea of the new generation of models presented here is to find continuously a new set of parameters (limited for practical reasons to the most crucial (= most sensitive) parameters) which is better fitted for the prevailing conditions of the ecosystem. 'Fitted' is defined in the Darwinian sense by the ability of the species to survive and grow, which may as already discussed be measured by exergy (see Jørgensen, 1982, 1986, 1990; Jørgensen and Mejer, 1979; Jørgensen et al., 1995a; Mejer and Jørgensen, 1979). Fig. 5 shows the proposed modelling procedure, which has been applied in the cases presented below. The use of exergy calculations to vary continuously the parameters has only been used in 10 cases, of which two biogeochemical models will be discussed below. One of the 10 case studies has been used to support the so-called Intermediate Disturbance Hypothesis.

6. Presentation of a case study illustrating the application of the structural dynamic modelling approach

The results from Søbygaard Lake (Jeppesen et al., 1990) are particularly fitted to test the applicability of the described approach to structural

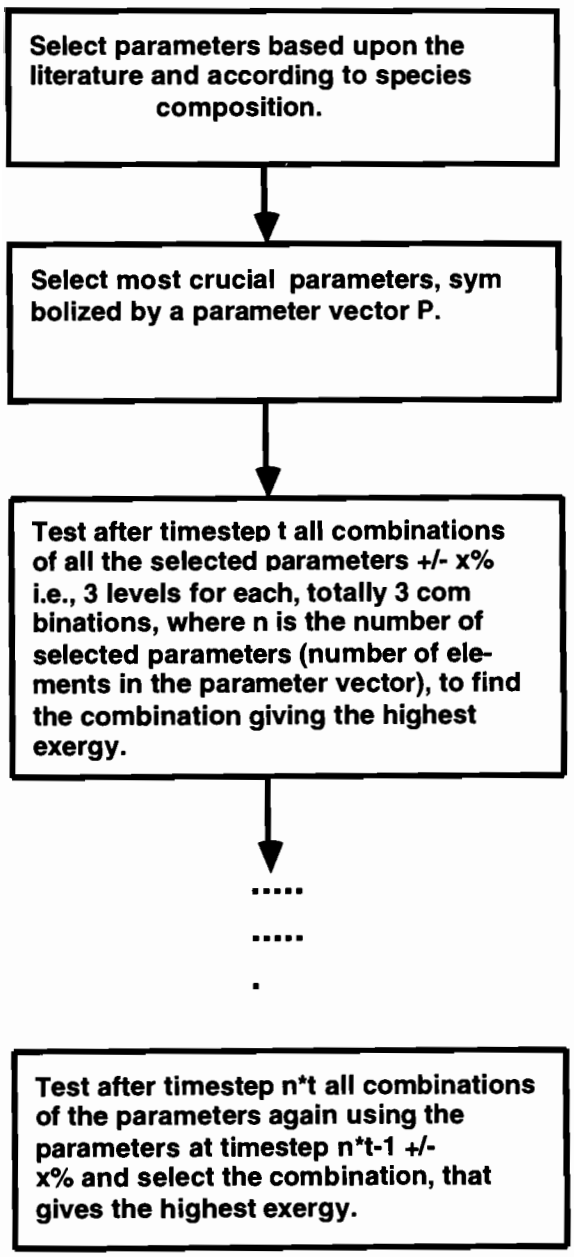

Fig. 5. The procedure used for the development of structural dynamic models.

dynamic models. As an illustration to structural dynamics of ecosystems and the possibilities to capture the flexibility of ecosystems, the case study of Søbygaard Lake will be presented in detail. Søbygaard Lake is a shallow lake (depth 1 m) with a short retention time (15–20 days). The nutrient loading was significantly reduced after 1982, namely for phosphorus from 30 to 5 g P/m2y. The reduced load did, however, not cause reduced nutrients and chlorophyll concentrations in the period 1982–1985 due to an internal loading caused by the storage of nutrients in the sediment (Jeppesen et al., 1990). However, radical changes were observed in the period 1985–1988. The recruitment of planctivorous fish was significantly reduced in the period 1984–1988 due to a very high pH caused by the eutrophication. As a result zooplankton increased and phytoplankton decreased in concentration (the summer average of chlorophyll A was reduced from 700 in 1985 to 150 μg/l in 1988). The phytoplankton population even collapsed in shorter periods due to extremely high zooplankton concentrations. Simultaneously the phytoplankton species increased in size. The growth rate decreased and a higher settling rate was observed (Jeppesen et al., 1990). The case study shows, in other words, pronounced structural changes. The primary production was, however, not higher in 1985 than in 1988 due to a pronounced self-shading by the smaller algae in 1985. It was therefore very important to include the self-shading effect in the model, which was not the case in the first model version, which therefore gave wrong figures for the primary production. Simultaneously a more sloppy feeding of the zooplankton was observed, as zooplankton was shifted from Bosmina to Daphnia.

The model applied has six state variables: N in fish, N in zooplankton, N in phytoplankton, N in detritus, N as soluble nitrogen and N in sediment. The model was developed by use of the software STELLA. As seen, only the nitrogen cycle is included in the model, but as nitrogen is the nutrient controlling the eutrophication, it may be sufficient to include only this nutrient. The aim of the study is to be able to describe by use of a structural dynamic model the continuous changes in the most essential parameters using the procedure shown in Fig. 5. The data from 1984–1985 were used to calibrate the model and the two parameters that it is intended to change from 1985 to 1988, got the following values by this calibration:

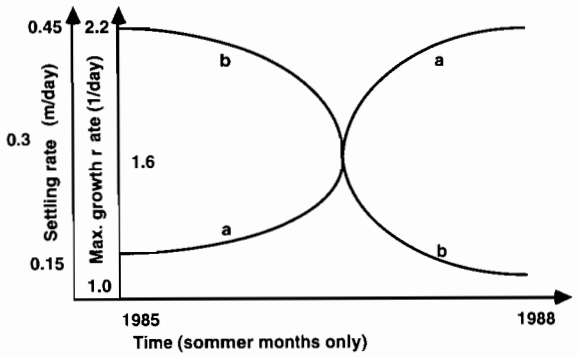

Fig. 6. The continuous changed parameters obtained from the application of a structural dynamic modelling approach on Søbygaard Lake are shown. a covers the settling rate of phytoplankton and b the maximum growth rate of phytoplankton.

Maximum growth rate of phytoplankton: 2.2 day -1

Settling rate of phytoplankton: 0.15 day -1

The state variable fish-N was kept constant $=6.0$ during the calibration period, but an increased fish mortality was introduced during the period 1985–88 to reflect the increased pH. The fish stock was thereby reduced to 0.6 mg N/l-notice the equation mort $=0.08$ if fish >6 (may be changed to 0.6) else almost 0. A time-step of $t=5$ days and $x\%=10\%$ was applied; see Fig. 5. This means that nine runs were needed for each time step to select the parameter combination that gives the highest exergy. The results are shown in Fig. 6 and the changes in parameters from 1985 to 1988 (summer situation) are summarised in Table 3. The proposed procedure is able to simulate approximately the observed change in structure. The maximum growth rate of phytoplankton is reduced by 50% from 2.2 to 1.1 day^{-1}, which is approximately according to the increase in size. It was observed that the average size was increased from a few 100 μm^3 to 500–1000 μm^3, which is a factor of about 2–3 (Jeppesen et al., 1990). It would correspond to a specific growth reduction by a factor $f=22/3-32/3$ (see Jørgensen, 1997; Peters, 1983). It means that:

$$\text{growth rate in } 1988 = \text{growth rate in } 1985/f \quad (2)$$

where f is between 1.58 and 2.08, while 2.0 is found by use of the structural dynamic modelling approach.

The settling was 0.2 m day^{-1} (range 0.02–0.4) in 1985, while it was 0.6 m day^{-1} (range 0.1–1.0) in 1988. By the structural dynamic modelling approach was found an increase from 0.15 to 0.45 day^{-1}, the factor being the same-three-but with slightly lower values. The phytoplankton concentration as chlorophyll-A was simultaneously reduced from 600 to 200 μg/l, which is approximately according to the observed reduction. All in all it may be concluded that the structural dynamic modelling approach gave an acceptable result and that the validation of the model and the procedure in relation to structural changes was positive. The structural dynamic modelling approach is of course never better than the model applied, and the presented model may be criticised for being too simple and not accounting for the structural dynamic changes of zooplankton.

For further elucidation of the importance to introduce a parameter shift, it has been tried to run the 1985 situation with the parameter combination found to fit the 1988 situation and vice versa. It was not possible to get a workable model if the parameters from 1985 was used to simulate the 1987 and 1988 data. The structural changes were so pronounced that a prediction based upon the parameters from 1985 for 1987–1988 would give completely wrong results. These exergy and stability results for this exercise are shown in Table 4. The results demonstrate that it is of great importance to apply the right parameter set to given conditions. If the parameters from 1985 are used for the 1988 conditions a lower exergy is obtained and the model to a certain extent behaves chaotically while the 1988 parameters used on the 1985 conditions give a significantly lower

Table 3
Parameter combinations giving the highest exergy

Rate (day^{-1})	Maximum Growth (m*day^{-1})	Settling Rate
1985	2.2	0.15
1988	1.1	0.45

Table 4
Exergy and stability by different combinations of parameters and conditions

Parameter	Conditions	
	1985	1988
1985	75.0 Stable	39 8 (average) Violent fluctuations. Chaos
1988	38.7 Stable	61 4 (average) Only minor fluctuations

exergy. These results are consistent with Jørgensen (1995), where it was shown that parameters may be estimated by use of the principle applied in structural dynamic models, i.e. that the parameter combination giving the highest exergy should be expected in the real ecosystems. If we have a high certainty for all the parameters except let us say two, these two missing parameters could be found as the combinations of these two parameters in possible parameter space that would give the highest exergy. If slightly too high values of the parameters, above the values giving maximum exergy, would be applied the model would behave chaotic. The parameter giving the highest exergy will therefore operate as the edge of the chaos, which is according to the results presented in Kauffman (1996).

The results of the presented case study show that it is important for ecological and environmental models to contain the property of flexibility, which we know ecosystems possess. If we account for this property in the models, we obtain models that are better able to produce reliable predictions, particularly when the forcing functions on the ecosystems change and thereby provoke changes in the properties of the important biological components of the ecosystem. In some cases we get completely different results, when we apply a continuous change of the parameters from when we use fixed parameters. In the first case we get results that are better in accordance with our observations and as we know that the parameters do actually change in the natural ecosystems, we can only recommend the application of this approach as far as possible in ecological modelling.

The property of dynamic structure and adaptable parameters is crucial in our description of ecosystems and should therefore be included in all descriptions of the system properties of ecosystems. The few examples presented here show that it is feasible to account for the adaptability of the properties in models, although a more general experience is needed before clear recommendations on the application can be given.

7. The use of structural dynamic models to understand the application of biomanipulation

The eutrophication and oligotrophication of a lacustrine environment do not proceed according to a linear relationship between nutrient load and vegetative biomass, but display rather a sigmoid trend with delay, as shown in Fig. 7. The hysteresis reaction is completely in accordance with observations (Hosper, 1989; Van Donk et al., 1989) and it can be explained by structural changes (de Bernardi, 1989; Hosper, 1989; Sas, 1989; de Bernardi and Giussani, 1995). At increasing nutrient level, a lake ecosystem shows a marked buffer capacity to variations, as only slightly higher phytoplankton concentrations are observed. It can be explained by a current increasing removal rate of phytoplankton by grazing and settling. The zooplankton concentration and the

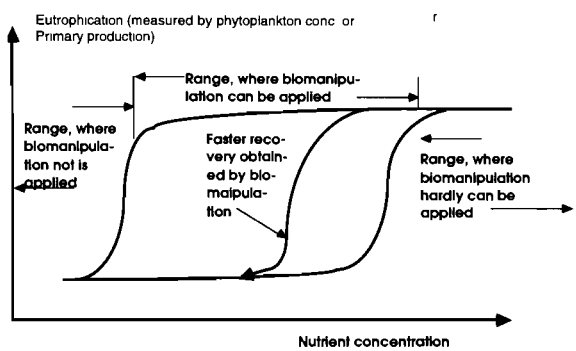

Fig. 7. The hysteresis relation between nutrient level and eutrophication measured by the phytoplankton concentration is shown. The possible effect of biomanipulation is shown. An effect of biomanipulation can hardly be expected above a certain concentration of nutrients, as indicated on the diagram.

concentration of predatory fish are maintained at relatively high level under these circumstances. At a certain level of eutrophication it is, however, not possible for zooplankton to increase the grazing rate further, and the phytoplankton concentration will increase very rapidly by slightly increasing concentrations of nutrients. When the nutrient input is decreased under these conditions a similar buffer capacity to variation is observed. The structure has now changed to a high concentration of phytoplankton and planktivorous fish which causes a resistance and delay to a change where the second and fourth trophic levels become dominant again.

Willemsen (1980) distinguishes two possible conditions:

1. A bream state characterised by turbid water, high eutrophication, low zooplankton concentration, absent of submerged vegetation, large amount of breams, while pike is hardly found at all.
2. A pike state, characterised by clear water, low eutrophication. Pike and zooplankton are abundant and there are significant fewer breams.

The presence of two possible states in a certain range of nutrient concentrations may explain why biomanipulation not always has been used successfully. According to the observations referred in the literature, success is associated with a total phosphorus concentration below 50 μg/l (Lammens, 1988) or at least below 100–200 μg/l (Jeppesen et al., 1990), while disappointing results are often associated with phosphorus concentration above this level of more than approximately 120 μg/l (Benndorf, 1987, 1990) with a difficult control of the standing stocks of planktivorous fish (Mills et al., 1987; Shapiro, 1990; Koschel et al., 1993).

Scheffer (1990) has used a mathematical model based on catastrophe theory to describe these shifts in structure. This model does however not consider the shifts in species composition, which is of particular importance for zooplankton. The zooplankton population undergoes a profound structural change when we increase the concentration of nutrients passing from a dominance of calanoid copepods to small cladocera and rotifers according to the following references: Carpenter et al., 1985, 1987; Sterner, 1989; de Bernardi and Giussani, 1995; Giussani and Galanti, 1995. It would therefore be interesting to test if structural dynamic models, i.e. models which consider the current changes in properties of the species due to changes in the conditions (the forcing functions, mainly the concentration of nutrients), could be used to give a better understanding of the relationship between concentrations of nutrients and the vegetative biomass and to explain possible results of biomanipulation. This section refers the results achieved by development of a structural dynamic models with the aim to understand the above described changes in structure and species compositions (Jørgensen and de Bernardi, 1998).

The applied model has six state variables, dissolved inorganic phosphorus, phytoplankton, phyt, zooplankton, zoopl., planktivorous fish, fish 1, predatory fish, fish 2 and detritus, detritus. The forcing functions are the input of phosphorus, in P, and the through flow of water determining the retention time. The latter forcing function determines also the outflow of detritus and phytoplankton. The conceptual diagram is similar to Fig. 4, except that only phosphorus is considered as nutrient, as it is presumed that phosphorus is the limiting nutrient.

Simulations have been carried out for phosphorus concentrations in the in flowing water of 0.02, 0.04, 0.08, 0.12, 0.16, 0.20, 0.30, 0.40, 0.60 and 0.80 mg/l. For each of these cases the model was run for any combination of a phosphorus uptake rate of 0.06, 0.05, 0.04, 0.03, 0.02, 0.01 1/24 h and a grazing rate of 0.125, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8 and 1.0 1/24 h. When these two parameters were changed a simultaneous changes of phytoplankton and zooplankton mortalities were made according to allometric principles (see Peters, 1983). The parameters which are made variable to account for the dynamics in structure are therefore for phytoplankton growth rate (uptake rate of phosphorus) and mortality and for zooplankton growth rate and mortality.

The settling rate of phytoplankton was made proportional to the $(\text{length})^2$. Half of the additional sedimentation when the size of phytoplankton increases corresponding to a decrease in the

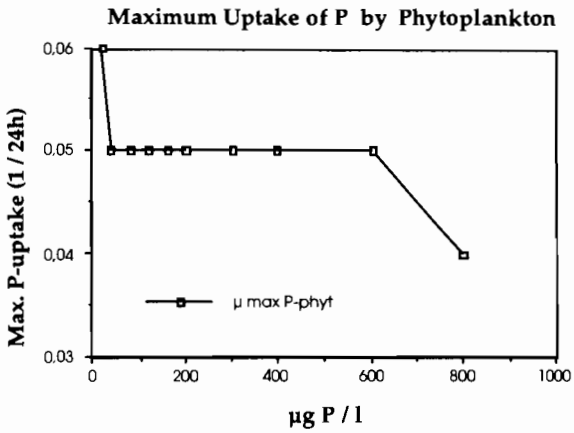

Fig. 8. The maximum growth rate of phytoplankton obtained by the structural dynamic modelling approach is plotted versus the phosphorus concentration.

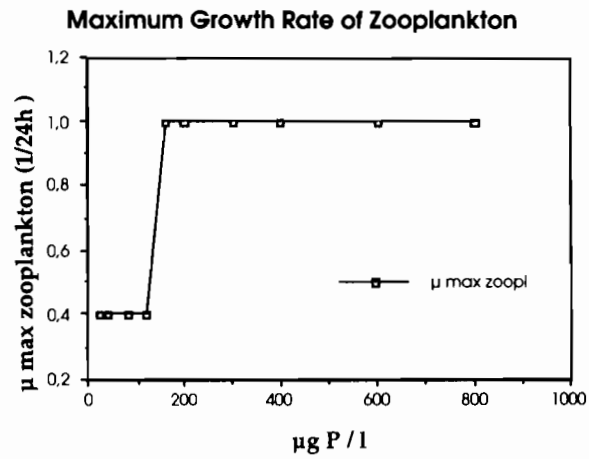

Fig. 9. The maximum growth rate of zooplankton obtained by the structural dynamic modelling approach is plotted versus the zooplankton concentration.

uptake rate, was allocated to detritus to account for resuspension or faster release from the sediment. A sensitivity analysis has revealed that exergy is most sensitive to changes in these five selected parameters which also represent the parameters which change significantly by size. As mentioned in the introduction, a change in size is observed as a response to changes in nutrient loading. The six, respectively 9 levels selected above represent approximately the range in size for phytoplankton and zooplankton.

For each phosphorus concentration 54 simulations were carried out to account for all combinations of the two key parameters. Simulations over three years, 1100 days, were applied to ensure that either steady state, limit cycles or chaotic behaviour would be attained. It was according to the above presented structural dynamic modelling approach presumed that the combination giving the highest exergy under the prevailing conditions should be selected as representing the process rates in the ecosystem. If exergy oscillates even during the last 200 days of the simulation, the average value for the last 200 days was used to decide on which parameter combination would give the highest exergy. The combinations of the two parameters, the uptake rate of phosphorus for phytoplankton and the grazing rate of zooplankton giving the highest exergy at different levels of phosphorus inputs are plotted in Figs. 8 and 9. The uptake rate of phosphorus for phytoplankton is gradually decreasing when the phosphorus concentration increases. As seen the zooplankton grazing rate changes at the phosphorus concentration 0.12 mg/l from 0.4 1/24 to 1.0 1/24 h, i.e. from larger species to smaller species, which is according to the expectations.

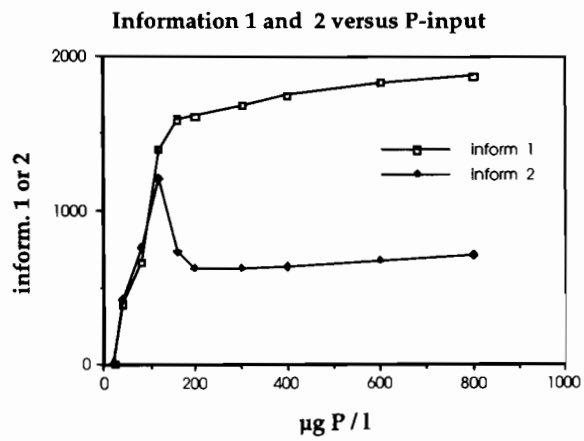

Fig. 10. The exergy is plotted versus the phosphorus concentration. Information 1 corresponds to a maximum zooplankton growth rate of 1/24 h and information 2 corresponds to a maximum zooplankton growth rate of 0.4 1/24 h. The other parameters are the same for the two plots, included the maximum phytoplankton growth rate taken from Fig. 8 as function of the phosphorus concentration.

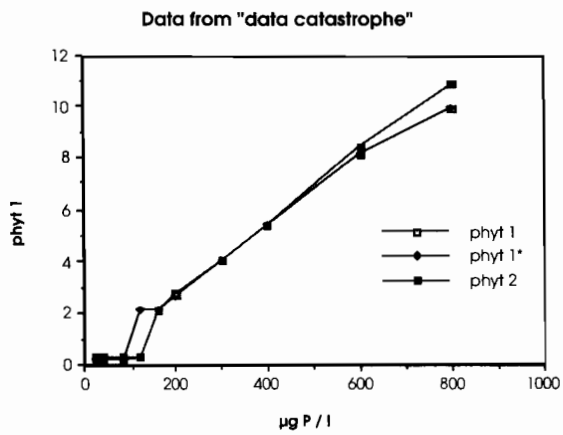

Fig. 11 The phytoplankton concentration as function of the phosphorus concentration for parameters corresponding to 'information 1' and 'information 2'; see Fig. 10. The plot named 'information 1*' coincides with 'information 1', except for a phosphorus concentration of 0.12 mg/l, where the model shows limit cycles. At this concentration, information 1* represent the higher phytoplankton concentration, while information 1 represent the lower phytoplankton concentration. Notice that the structural dynamic approach can explain the hysteresis reactions.

Fig. 10 shows the exergy, named on the diagram information, with an uptake rate according to the results in Fig. 8 and a grazing rate of 1.0 1/24 h (called information 1), respectively of 0.4 1/24 h (called information 2). Below a phosphorus concentration of 0.12 mg/l the information 2 is slightly higher, while information 1 is significantly higher above this concentration. The phytoplankton concentration increases for both parameter sets with increasing phosphorus input, as shown Fig. 11, while the planktivorous fish shows a significantly higher levels by a grazing rate of 1.0 1/24 h, when the phosphorus concentration is 0.12 mg/l (= valid for the high exergy level). Below this concentration the difference is minor. The concentration of fish 2 is higher for the case 2 corresponding to a grazing rate of 0.4 1/24 h for phosphorus concentrations below 0.12 mg/l. Above this value the differences are minor, but at a phosphorus concentration of 0.12 mg/l the level is significant higher for a grazing rate of 1.0 1/24 h, particularly for the lower exergy level, where also the zooplankton level is highest.

If it is presumed that exergy indices can be used as a goal function in ecological modelling, the results seem to be able to explain why we observe a shift in grazing rate of zooplankton at a phosphorus concentration in the range of 0.1–0.15 mg/l. The ecosystem selects the smaller species of zooplankton above this level of phosphorus because it means a higher level of the exergy index, which can be translated to a higher rate of survival and growth. It is interesting that this shift in grazing rate only gives a little higher level of zooplankton, while the exergy index level gets significantly higher by this shift, which may be translated as survival and growth for the entire ecosystem. Simultaneously, a shift from a zooplankton, predatory fish dominated system to a system dominated by phytoplankton and particularly by planktivorous fish takes place.

It is interesting that the levels of exergy indices and the four biological components of the model for phosphorus concentrations at or below 0.12 mg/l parameter combinations are only slightly different for the two parameter combinations. It can explain why biomanipulation is easy in this concentration range. Above 0.12 mg/l the differences are much more pronounced and the exergy index level is clearly higher for a grazing rate of 1.0 1/24 h. It should therefore be expected that the ecosystem after the use of biomanipulation easily fall back to the dominance of planktivorous fish and phytoplankton. These observations are consistent with the general experience of success and failure of biomanipulation; see above.

If the concentrations of zooplankton and fish 2 is low, and high for fish 1 and phytoplankton, i.e. we are coming from higher phosphorus concentrations, the simulation gives with high probability also a low concentration of zooplankton and fish 2. When we are coming from high concentrations of zooplankton and of fish 2, the simulation gives with high probability also a high concentration of zooplankton and fish 2, which correspond to an exergy index level slightly lower than obtained by a grazing rate of 0.4 1/24 h. This grazing rate will therefore still be prevailing. As it also takes time to recover the population of zooplankton and particularly of fish 2 and in the other direction of fish 1, these observations ex-

plain the presence of hysteresis reactions. An interpretation of the results points toward a shift at 0.12 mg/l, where a grazing rate of 1.0 1/24 h yields limit cycles. It indicates an instability and a probably easy shift to a grazing rate of 0.4 1/24, although the exergy level is in average highest for the higher grazing rate. A preference for a grazing rate of 1.0 1/24 h at this phosphorus concentration should therefore be expected, but a lower or higher level of zooplankton is dependent on the initial conditions.

The model is considered to have general applicability and has been used to discuss the general relationship between nutrient level and vegetative biomass and the general experiences by application of biomanipulation. When the model is used in specific cases, it may however be necessary to include more details and change some of the process descriptions to account for the site specific properties, which is according to general modelling strategy. It could be considered to include two state variables to cover zooplankton, one for the bigger and one for the smaller species. Both zooplankton state variables should of course have a current change of the grazing rate according to the maximum value of the goal function. The model could probably also be improved by introduction of size preference for the grazing and the two predation processes which is in accordance with numerous observations. In spite of these shortcomings of the applied model, it has been possible to give a right qualitative description of the reaction to changed nutrient level and biomanipulation, and even to indicate an approximately correct phosphorus concentration, where the structural changes may occur. This may be due to an increased robustness by the structural dynamic modelling approach.

8. Further support to the hypothesis

The hypothesis which we have applied to develop structural dynamic models-to describe adaptation and/or shifts to better fitted species with other properties-may be formulated as follows: If a system receives an inflow of energy (for ecosystems solar radiation), it will be able to utilise this energy flow to move away from thermodynamic equilibrium. If more combinations of processes and components are available to utilise this energy flow, the combination which can obtain the highest storage of exergy (= provide the biggest distance from thermodynamic equilibrium) will win. It has been possible in addition to the structural modelling studies, to find a few case studies (see below) where several pathways are available to utilise the flow of exergy and where the exergy gained by the system can be calculated directly (Jørgensen, 1997). These (few) case studies support the presented hypothesis, as the selected pathways give the system the highest (stored) exergy.

The sequence of oxidation of organic matter (see for instance Schlesinger, 1997) is as follows: by oxygen, by nitrate, by manganese dioxide, by iron (III), by sulphate and by carbon dioxide. It means that oxygen will always out-compete nitrate which will out-compete manganese dioxide and so on. The amount of exergy stored in ATPs (ATP represents a storage of 42 kJ exergy per mole) in the microorganisms winning the competition as a result of the oxidation processes decreases in the same sequence, as it should be expected if the hypothesis was valid; see Table 5. Numerous experiments have been performed to imitate the formation of organic matter in the primeval atmosphere on earth 4×10^9 years ago (see for instance Jørgensen, 1997). Various

Table 5
kJ/equiv available to build ATP for various oxidation processes of organic matter at pH 7.0 and 25°C

Reaction	Available (kJ/equiv)
$CH_2O + O_2 \rightarrow CO_2 + H_2O$	125
$CH_2O + 0.8\ NO_3^- + 0.8H^+ \rightarrow CO_2 + 0.4\ N_2 + 1.4\ H_2O$	119
$CH_2O + 2MnO_2 + H^+ \rightarrow CO_2 + 2\ Mn^{2+} + 3H_2O$	85
$CH_2O + 4FeOOH + 8H^+ \rightarrow CO_2 + 7H_2O + Fe^{2+}$	27
$CH_2O + 0.5SO_4^{2-} + 0.5H^+ \rightarrow CO_2 + 0.5HS^- + H_2O$	26
$CH_2O + 0\ 5CO_2 \rightarrow CO_2 + 0.5CH_4$	23

sources of energy have been sent through a gas mixture of carbon dioxide, ammonia and methane. Analyses have shown that a wide spectrum of various compounds included amino acids is formed under these circumstances, but generally only compounds with rather large negative free energy (i.e. high exergy storage) will form an appreciable part of the mixture (Morowitz, 1968).

At the biochemical level, we find that different plants operate three different biochemical pathways for the process of photosynthesis: (a) the C3 or Calvin Benson cycle; (b) the C4 pathway; and (c) the crassulacean acid metabolism (CAM) pathway. The latter pathway is the less efficient than the two other possible pathways, measured as g plant biomass formed per unit of energy received. Plants using CAM pathway can, however, survive in harsh, arid environment, but the photosynthesis will switch to C3 as soon as the availability of water is sufficient (see Shugart, 1998). Givnish and Vermelj (1976) made the assumption leaves optimise the payoff of having leaves of a given size versus maintaining leaves of a given size. They can by this assumption which corresponds to optimisation of exergy storage, explain the size of leaves in a given environment dependent on the solar radiation and the humidity. The entire evolution has been towards organisms with an increasing number of non-nonsense genes and more types of cells, i.e. towards storage of more exergy due to the increased information content.

9. The application of the exergy index to assess the value of unknown parameters

A detailed examination of the relationship between the behaviour and the value of a specific parameter, in this case the maximum growth rate of zooplankton, has been made in Jørgensen (1995). Fig. 12 shows the results of simulations with a model with the following state variables: nutrients, phytoplankton, zooplankton and detritus. The maximum growth rate of zooplankton has been varied. The model is run to steady state, if a steady state can be obtained. The exergy index expressed as 'exergy of g organic matter'/l is plotted versus the maximum growth rate of

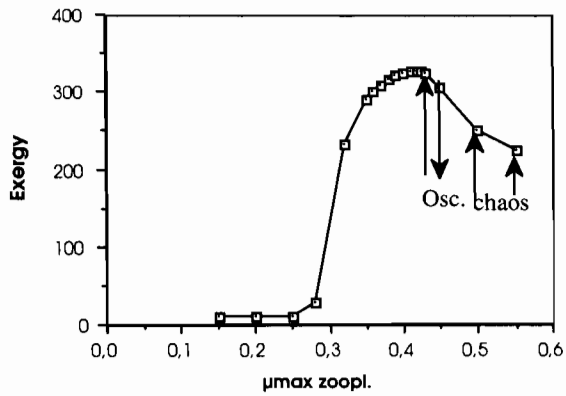

Fig. 12. Exergy as mg detritus/l is plotted versus the maximum growth rate of zooplankton for a model with nutrients, detritus, phytoplankton and zooplankton as state variables.

zooplankton. On the figure is indicated, whether a steady state can be obtained, or whether fluctuations occur. If regular oscillations occur, the average of the exergy for one oscillation is used. At a maximum growth rate of 0.5 1/day regular oscillations occur, and the average level of exergy is slightly lower than for a maximum specific growth rate of 0.425 1/day. At a maximum specific growth rate of 0.6 1/day an even lower average exergy is obtained and the regularity is smaller. At higher growth rate the exergy and the state variables exhibit violent and irregular changes.

The highest level of exergy is obtained for maximum growth rate of zooplankton slightly lower than the values, that give exhibit chaotic behaviour; see Fig. 12. The highest exergy is therefore for this particular model obtained at the 'edge of chaos'. The maximum growth rate obtained at the highest level of exergy can furthermore be considered realistic, i.e. according to the range found in the literature for the maximum specific growth rate of zooplankton; see Jørgensen et al. (1991). Fig. 13 shows the same plot as Fig. 12, but with introduction of fish in the model. Lower specific growth rate means that zooplankton get bigger in size following general allometric relationships (see Peters, 1983). This behaviour of the model is entirely following several observations in nature: predation by fish yields zooplankton, that often has bigger size (provided that fish doesn't have any size preference which, however,

may be the case) and has slower growth rates (see Peters, 1983). The maximum specific growth rate found at maximum exergy for the model run with fish is also within the range of values found in nature: approximately 0.15/0.5 1/day; see Jørgensen et al. (1991), and Jørgensen (1988), Jørgensen (1994a).

If the fish is removed from the model again. The level of exergy decreases drastically due to the loss of the information embodied in the fish, but by increasing the maximum growth rate of zooplankton the exergy increases again. The results from Fig. 12 is reproduced again with the highest level of information at a maximum growth rate of 0.425 1/day, and at the edge of chaos.

The fractal dimension may be considered a measure of the chaotic behaviour. The fractal dimensions obtained for the plots of exergy versus the time for various levels of the maximum zooplankton growth rate for the model runs without fish are shown on the Fig. 14. As seen the fractal dimension increases with increasing maximum growth rate of zooplankton as expected due to the more and more violent fluctuations of the state variables and thereby the exergy. When the maximum growth rate increases more and more violent fluctuations result with higher and higher

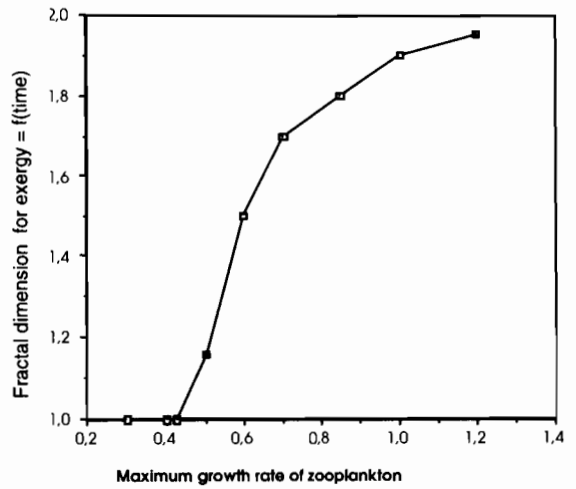

Fig. 14. The fractal dimension obtained for the plots exergy = f(time) for various values of the maximum growth rate of zooplankton is shown.

maximum values, smaller and smaller minimum values and increasing occurrence of the smaller values, resulting in decreasing average values of the exergy. It is illustrated Fig. 15, where the average exergy is plotted versus the fractal dimension.

In this case is obtained a fractal dimension of 1.0 for the values of the maximum growth rate of

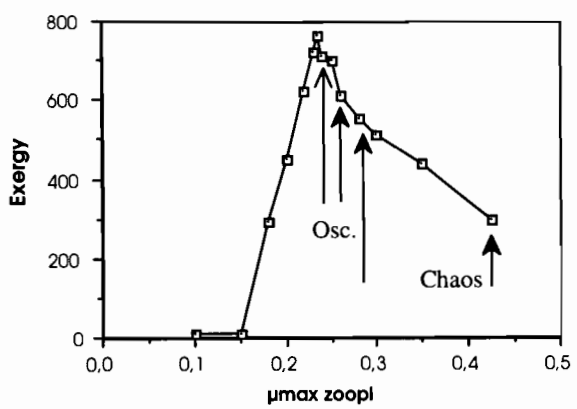

Fig. 13. Exergy as mg detritus/l is plotted versus maximum growth rate of zooplankton with nutrients, detritus, phytoplankton, zooplankton and fish as state variables. Notice that the exergy is higher than in Fig. 12 due to the presence of fish, and that the maximum growth rate at maximum exergy level is lower. No size preference is assumed for the zooplankton predated by the fish.

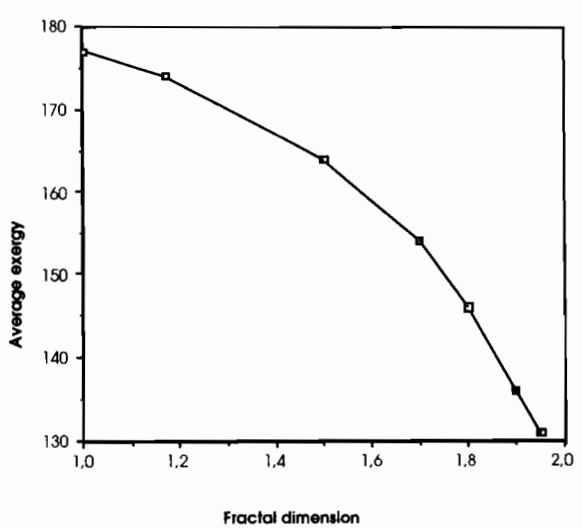

Fig. 15. The average exergy is plotted versus the fractal dimension based on the case study used in Figs. 12–14.

zooplankton ≤ 0.425 1/day, because the model considers a steady state situation where no fluctuations in the phytoplankton due to variations in temperature and solar radiation are considered. If normal diurnal and seasonal changes are considered these parameter values will exhibit a fractal dimension slightly more than 1, but the fractal dimension will still increase when the maximum growth rate is > than the maximum growth rate at maximum exergy index.

The values of exergy and the fractal dimensions in the here illustrated case study are of course dependent on all the selected parameter values. The shown tendency is, however, general: the highest exergy is obtained by a parameter value above which chaotic behaviour, increasing fractal dimension for the state variables and the exergy index as function of time and decreasing average exergy index occur. The parameter estimation is as mentioned in the introduction often the weakest point for many of our ecological models. Reasons are:

- an insufficient number of observations to enable the modeller to calibrate the number of more or less unknown parameters,
- no or only little literature information can be found,
- ecological parameters are generally not known with sufficient accuracy
- the structure shows dynamical behaviour, i.e., the parameters are continuously changing to achieve a better adaptation to the ever changing conditions; see also Jørgensen (1988, 1992a),
- or a combination of two or more of these points.

The above-mentioned results seem to reduce these difficulties by imposing the ecological facts that all the species in an ecosystem have the properties (described by the parameter set) that are best fitted for survival under the prevailing conditions. The property of survival can currently be tested by use of exergy, since it is survival translated into thermodynamics. Coevolution, i.e. when the species have adjusted their properties to each other, is considered by application of exergy for the entire system. Application of the ecological law of thermodynamics as constraint on our ecological models enable us to reduce the feasible parameter range, which can be utilised to facilitate our parameter estimation significantly.

10. Conclusions

Ecosystems are very different from physical systems mainly due to their enormous adaptability. It is therefore crucial to develop models that are able to account for this property, if we want to get reliable model results. The use of exergy as goal functions to cover the concept of fitness seems to offer a good possibility to develop a new generation of models, which is able to consider the adaptability of ecosystems and to describe shifts in species composition. The latter advantage is probably the most important, because a description of the dominant species in an ecosystem is often more essential than to assess the level of the focal state variables.

It is possible to model a competition between a few species with quite different properties, but the structural dynamic modelling approach makes it feasible to include more species even with only slightly different properties, which is impossible by the usual modelling approach; see also the unsuccessful attempt to do so by Nielsen (1992). The rigid parameters of the various species make it difficult for the species to survive under changing circumstances. After some time only a few species will still be present in the model, opposite what is the case in reality, where more species survive because they are able to adapt to the changing circumstances. It is therefore important to capture this feature in our models. The structural dynamic models seem promising in this respect, although more experience is needed before a final conclusion on their applicability can be made.

It is interesting that the ranges of growth rate actually found in nature (see for instance Jørgensen et al., (1991)) are those, which give stable, i.e. non-chaotic conditions. All in all it seems possible to conclude that the parameters that we can find in nature today, are in most cases those which assure a high probability of survival and growth in all situations; chaotic situations are thereby avoided. The parameters that could give

possibilities for chaotic situations, have simply been excluded by selection processes. They may give high exergy in some periods, but later the exergy becomes very low due to the violent fluctuations and it is under such circumstances that the selection process excludes the parameters (properties), that cause the chaotic behaviour.

References

Allen, P.M., 1988. Evolution: Why the whole is greater than the sum of the parts. In: W.Wolff, C.J. Soeder and F.R. Drepper (editors), Ecodynamics: Contribution to Theoretical Ecology. Part 1: Evolution. Proceedings of an International Workshop, 19–20. October, 1987, Jülich, Germany, Springer Verlag, Berlin, pp. 2–30.

Benndorf, J., 1987 Food-web manipulation without nutrient control: A useful strategy in lake restoration? Schweiz Z Hydrol. 49, 237–248.

Benndorf, J., 1990 Conditions for effective biomanipulation. Conclusions derived from whole-lake experiments in Europe. Hydrobiologia 200/201, 187–203

Boltzmann, L., 1905. The Second Law of thermodynamics. Populare Schriften. Essay No 3. (address to Imperial Academy of Science in 1886). Reprinted in English in Theoretical Physics and Philosophical Problems, Selected Writings of L. Boltzmann. D. Reidel, Dordrecht.

Brown, J.H., 1995. Macroecology The University of Chicago Press, Chicago, IL, p. 269.

Carpenter, S.R., Kitchell, J.F., Hodgson, J.R., 1985. Cascading trophic interactions and lake productivity. BioScience 35, 639–643.

Carpenter, S.R., Kitchell, J.F., Hodgson, J.R., Cochran, P.A., Elser, J J., Elser, M M., Lodge, D.M., Kretchmer, D., He, X., von Ende, C.N., 1987. Regulation of lake primary productivity by food web structure Ecology 68, 1863–1876.

de Bernardi, R. and Giussani, G. 1995. Biomanipulation: Bases for a Top-down Control 1-14. In Guidelines of Lake Management, Volume 7. Biomanipulation in Lakes and Reservoirs, edited by De Bernardi, R. and Giussani, G. ILEC and UNEP. p. 211.

de Bernardi, R., 1989. Biomanipulation of aquatic food chains to improve water quality in eutrophic lakes 195-215 In: Ravera, O. (Ed), Ecological Assessment of Environmental Degradation, Pollution and Recovery. Elsevier, Amsterdam, p. 356.

Fontaine, T.D., 1981. A selfdesigning model for testing hypotheses of ecosystem development 281-291. In: D. Dubois (editor), Progress in ecological engineering and management by mathematical modelling, Proc. 2nd Int. Conf., State-of-the-Art of Ecological Modelling, 18–24 April 1980, Liege, Belgium. p. 720

Giussani, G., Galanti, G., 1995 Case Study: lake Candia (Northern Italy) 135-146. In. De Bernardi, R. and Giussani (editors). Guidelines of Lake Management, Volume 7 Biomanipulation in Lakes and Reservoirs, G. ILEC and UNEP. 211 pp

Givnish, T J , Vermelj, G J., 1976. Sizes and shapes of liana leaves. Am Natur 110, 743–778.

Hosper, S.H., 1989. Biomanipulation, new perspective for restoring shallow, eutrophic lakes in The Netherlands Hydrobiol Bull 73, 11–18.

Jørgensen, S.E. 1982 A holistic approach to ecological modelling by application of thermodynamics 72–86. In: Mitsch, Bosserman and Dillon (editors). Systems and energy., Ann Arbor, p. 176.

Jørgensen, S.E., 1986. Structural Dynamics model. Ecol. Modelling 31, 1–9.

Jørgensen, S.E., 1988. Use of models as an experimental tool to show the structural changes are accompanied by increased exergy Ecol. Model. 41, 117–126.

Jørgensen, S.E., 1990. Ecosystem theory, ecological buffer capacity, uncertainty and complexity. Ecol. Model. 52, 125–133.

Jørgensen, S.E , 1992a Development of models able to account for changes in species composition Ecol. Model. 62, 195–208.

Jørgensen, S.E., 1992b. Parameters, ecological constraints and exergy Ecol Model. 62, 163–170.

Jørgensen, S.E., 1994a. Fundamentals of Ecological Modelling, 2nd edition Elsevier, Amsterdam, p. 630.

Jørgensen, S.E., 1994b. Models as instruments for combination of ecological theory and environmental practice Ecol. Model 75/76, 5–20.

Jørgensen, S E , 1994c. Review and comparison of goal functions in system ecology. Vie et Milieu 44, 11–20

Jørgensen, S.E , 1995 The growth rate of zooplankton at the edge of chaos. J. Theor. Biol. 175, 13–21.

Jørgensen, S.E., 1997. Integration of Ecosystem Theories. A Pattern, 2nd edition. Kluwer, Dordrecht, p. 400.

Jørgensen, S.E., 1998. An improved parameter estimation procedure in lake modelling. Lakes and Reservoirs: Res. Manag. 3, 139–142.

Jørgensen, S E., Mejer, H.F.., 1979. A holistic approach to ecological modelling Ecol Model 3, 39–61.

Jørgensen, S.E., Padisak, J., 1996. Does the intermediate disturbance hypothesis comply with thermodynamics? Hydrobiologia 323, 9–21.

Jørgensen, S.E., de Bernardi, R., 1997. The application of a model with dynamic structure to simulate the effect of mass fish mortality on zooplankton structure in Lago de Annone. Hydrbiologia 356, 87–96.

Jørgensen, S.E , de Bernardi, R., 1998. The use of structural dynamic models to explain the success and failure of biomanipulation. Hydrobiologia 379, 147–158.

Jørgensen, S.E., Nielsen, S.N., Jørgensen, L.A., 1991 Handbook of Ecological Parameters and Ecotoxicology Elsevier. Amsterdam, pp. 1320

Jørgensen, S.E., Nielsen, S.N., Mejer, H.F., 1995a. Emergy, environ, exergy and ecological modelling. Ecol. Modelling 77, 99–109

Jørgensen, S.E., Halling-Sørensen, B., Nielsen, S.N., 1995b. Handbook of Environmental and Ecological Modelling. CRC Press, Boca Raton, FL, US, p. 672

Jeppesen, E.J., Mortensen, E., Sortkjaer, O., Kristensen, P., Bidstrup, J., Timmermann, M., Jensen, J P., Hansen, A.M., Søndergaard, M., Müller, J.P., Jensen, J., Riemann, B., Lindegaard-Petersen, C., Bosselmann, S., Christoffersen, K., Dall, E., Andersen, J.M., 1990. Fish manipulation as a lake restoration tool in shallow, eutrophic temperate lakes. Cross-analysis of three Danish case studies Hydrobiologia 200/201, 205–218.

Kauffman, S., 1996. At Home in the Universe The Search for Laws of Complexity Penguin Books, Oxford University Press, Oxford, p. 320.

Kompare, B., 1995. The Use of Artificial Intelligence in Ecological Modelling, Ph.D Thesis at DFH, University Park 2, Copenhagen, p. 360.

Koschel, R., Kasprzak, P., Krienitz, L., Ronneberger, D., 1993. Long term effects of reduced nutrient loading and food-web manipulation on plankton in a stratified Baltic hard water lake. Verh int. ver Limnol. 25, 647–651

Lammens, E.H.R.R., 1988. Trophic interactions in the hypertrophic Lake Tjeukemeer: Top-down and bottom-up effects in relation to hydrology, predation and bioturbation, during the period 1974–1988. Limnologica (Berlin) 19, 81–85

Lewin, B., 1994. Genes V. Oxford University Press, Oxford, p. 1272

Li, W.H., Grauer, D., 1991. Fundamentals of Molecular Evolution. Sinauer, Sunderland, Massachusetts, p. 660.

Margalef, R., 1991. Networks in ecology. In: Higashi, M. and Burns, T P. (editors). Theoretical Studies of Ecosystems: the Network Perspectives. Cambridge University Press, Cambridge, pp. 41–57

Mejer, H F. and Jørgensen, S.E., 1979. Energy and ecological buffer capacity 829–846 In State-of-the-art of ecological modelling. ISEM, Copenhagen. p. 866.

Mills, E.L., Forney, J.L., Wagner, K.J. 1987. Fish predation and its cascading effect on the Oneida Lake food chain 118–131. In: Kerfoot, R. and Sih, S. (editors), Predation-direct and indirect impacts on aquatic communities. University Press, New England, Hanover & London. p. 324.

Morowitz, H I., 1968 Energy flow in biology. Academic Press, New York, p. 180.

Morowitz, H I., 1992 Beginnings of Cellular life. Yale University Press, New Haven and London, p. 260.

Nielsen, S.N., 1992 Application of maximum exergy in structural dynamic models. Ph.D. Thesis, DFH, Institute A, Section of Environmental Chemistry, Copenhagen, Denmark. p 51.

Patten, B.C., 1997. Bear Model for Aironduck National Park. Ecol Model. 100, 11–42.

Peters, R.H., 1983. The Ecological Implication of Body Size. Cambridge University Press, Cambridge, p. 286

Sas, H. (coordination) 1989. Lake restoration by reduction of nutrient loading Expectations, experiences, extrapolations. St. Augustin. Academia Verl. Richarz., p. 497.

Scheffer, M., 1990. Simple models as useful tools for ecologists. Elsevier, Amsterdam, p 192

Schindler, D.W., 1988. Effects of acid rain on freshwater ecosystems. Science 239, 149–157

Schlesinger, W H., 1997. Biogeochemistry. An Analysis of Global Change. Academic Press, New York, p 588

Schoffeniels, E., 1976. Anti-Chance. Pergamon Press, New York, p. 198.

Shapiro, J., 1990. Biomanipulation. the next phase-making it stable. Hydrobiologia 200/210, 13–27

Shugart, H.H., 1998. Terrestrial Ecosystems in Changing Environments. Cambridge University Press, Cambridge, p 537

Sterner, R.W., 1989. The role of grazers in phytoplankton succession 107–140 In: Sommer, U. (Ed.), Plankton Ecology. Springer Verlag, Germany, p. 476.

Straskraba, M., 1979. Natural control mechanisms in models of aquatic ecosystems. Ecol Model. 6, 305–322.

Ulanowicz, R.E., 1986. Growth and Development, Ecosystem Phenomenology Springer Verlag, New York, p. 203.

Van Donk, E., Gulati, R.D., Grimm, M.P., 1989. Food web manipulation in lake Zwemlust: positive and negative effects during the first 2 years. Hydrobiol. Bull. 23, 19–35.

Willemsen, J., 1980 Fishery aspects of eutrophication. Hydrobiol. Bull. 14, 12–21.

Wolfram, S., 1984a Cellular automata as models of complexity. Nature 311, 419–424.

Wolfram, S., 1984b Computer software in science and mathematics. Sci. Am. 251, 140–151.

ELSEVIER

Ecological Modelling 120 (1999) 97 - 107

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Applications of the self-organising feature map neural network in community data analysis

Giles M. Foody *

Department of Geography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Abstract

Freedom from restrictive assumptions that underlie many quantitative techniques make neural networks attractive for ecological investigations. The potential of the self organising feature map (SOFM) neural network for the classification, and to a lesser extent, ordination of vegetation data was investigated. The SOFM output was shown to correspond closely to classifications obtained from three alternative clustering algorithms, with similar samples located close together in the SOFM output space. Moreover, the classes were distributed spatially in the SOFM output by their relative similarity. This was evident with comparison against classifications derived at various levels of a hierarchical classification that revealed that the classes aggregated during each step of the hierarchical classification also tended to lie close together in the SOFM output space. As a consequence, the spatial distribution of classes in the SOFM output may represent the data in a manner similar to an ordination analysis. Some evidence for this inference is provided by comparison with the results of a standard ordination analysis.

Keywords: Kohonen SOFM; Vegetation classification; Ordination

1. Introduction

Quantitative methods have been used increasingly for ecological investigations since the 1950s (Greig-Smith, 1980). Many of the methods used are based on conventional statistics. Consequently, the analyses are often based on a set of often untenable assumptions (Potvin and Roff, 1993). For example, the general linear model that underlies many methods of community classification and ordination, two major areas of quantitative ecology, requires the satisfaction of ecologically unrealistic assumptions. It is, for example, assumed that the relationships between variables are linear when often they may be non-linear and even non-monotonic (Greig-Smith, 1980, 1996; Terborgh et al., 1997). Furthermore, the underlying assumption of normally distributed data is often not satisfied with ecological data (Tong, 1992; Potvin and Roff, 1993) and emphasis is placed on typical rather than the sometimes more important extreme values in the data set (Gaines and Denny, 1993). Deviations from the assumptions of a particular quantitative method are not always important, particularly if emphasis

* Fax: +44-1703-593295.
E-mail address: gmf@soton.ac.uk (G.M. Foody)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00094-0

is on low level data exploration (Greig-Smith, 1980) but can result in major misuses and misinterpretations. Attention has, therefore, turned to the refinement of the techniques so they may be appropriately applied or to adoption of alternative methods. Thus, for example, non-linear ordination analyses or use of classification methods based on fuzzy sets have attracted interest (e.g. Bosserman and Ragade, 1982; Bradfield and Kenkel, 1987; Ludwig and Reynolds, 1988; Equihua, 1990; Tong, 1992; Foody, 1996). Further, possibilities are offered by alternative paradigms such as neural networks that are free of constraining assumptions.

A variety of neural networks have been used in ecological research. Much attention has focused on feedforward neural networks (e.g. Lek et al., 1995; Mastrorillo et al., 1997; Maier and Dandy, 1998) with considerably less directed at exploiting the potential of the Kohonen or self organizing feature map (SOFM) networks (e.g. Chon et al., 1996). This type of neural network organises the data by similarity. The output of the SOFM is a low, typically two-dimensional, array in which similar samples are clustered together. The aim of this paper was to evaluate the use of the SOFM in community data analysis, particularly in relation to classification but with some reference to ordination. Despite some similarities, these two types of analyses have different aims and applications. Classification, for example, seeks to form groups of samples with similar attributes whereas ordination aims to arrange samples such that their similarity is reflected in their relative position or order (Goldsmith et al., 1976; Greig-Smith, 1980). The main focus of this paper was on the spatial arrangement of samples in the SOFM output and its relationship to outputs from alternative classification and ordination approaches.

2. Self organizing feature map (SOFM)

This section aims to provide a brief overview of the salient features of the SOFM neural network. Unlike other widely used neural networks, the SOFM uses unsupervised learning and produces a topologically ordered output that displays the similarity between the samples presented to it (Davalo and Naim, 1991; Schalkoff, 1992). The

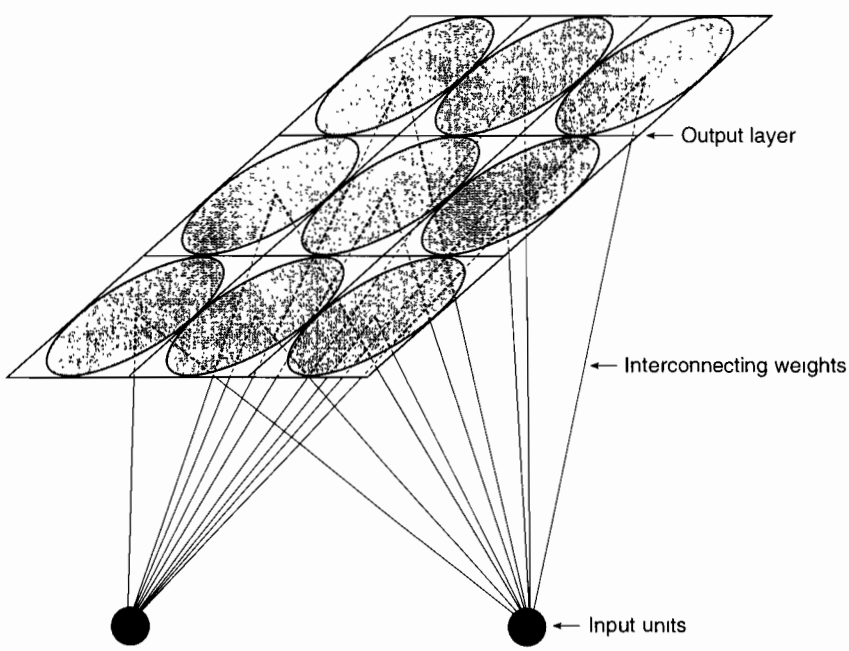

Fig. 1. A basic SOFM Note that each unit in the input layer is connected to every unit in the output layer by a weighted connection.

network consists of only two layers (Fig. 1). The input layer contains a node or unit for each variable (e.g. species) in the ecological data set. The input units operate in a similar fashion to those in other neural networks, effectively acting as a means of presenting the data for each sample to the network in an appropriate format. Unlike the more widely used feedforward neural network, there is no hidden layer and the input units are instead connected directly to units in the output layer or Kohonen layer. This output layer is also typically, but not necessarily, a two-dimensional array of units and each of these units is connected to every unit in the input layer by a weighted connection. Lateral interaction between units in the output layer also ensures that learning is a competitive process in which the network adapts to respond in different locations for inputs that differ. Consequently, samples that are similar should be associated with units that are close together in the output layer while a dissimilar sample would be associated with a distant unit elsewhere in the output layer. While the rows and columns on the output layer can be interpreted as co-ordinate axes to locate units and upon which the output of the SOFM may be interpreted they need not have an explicit meaning or relation to the ecological variables of the input data set. The projection depicted by the SOFM output is also non-linear. The distance between output units is, therefore, difficult to evaluate objectively but does, however, provide information on the similarity of samples associated with the units (Blayo and Demartines, 1991; Goodacre et al., 1994).

As with other neural networks the analysis is based on the solution of a large number of simple operations that can be performed in parallel. Since each of the n input units is linked to every output unit by a weighted connection, each output unit has the same number of weights associated with it as the dimensions of the input data vectors. Each output unit, i, is fed the input data vector, $I = (I_1, I_2 \ldots I_n)$, for each sample in parallel and has an associated weight vector, $W_i = (W_{i1}, W_{i2} \ldots, W_{in})$. At the outset the weights are set randomly but adjusted on each training iteration t. This adjustment is often based on the Euclidean distance measurement D_i made for each output unit with

$$D_i = \sqrt{\sum_{j=1}^{n} (I_j - W_{ij})^2} \quad (1)$$

The output unit with the lowest distance is the closest to the particular input sample and is taken to be the 'winning' or best matching unit, b. Once b has been identified, the weights connecting the input and output units may then be subject to adjustment. Weight adjustment is, however, constrained to include only those weights associated with output units close to b with all other weights unaffected. Weight up-dating is, therefore, undertaken only within a defined neighbourhood, N, of b and the size of this neighbourhood is generally reduced during the learning phase. Although all units, including the winning unit, within the neighbourhood are included in the weight up-dating process the magnitude of change made is also a function of distance from the winning unit. The weights associated with units close to the winning unit are subject to a larger change than those associated with units further from the winning unit. Typically the weight up-dating is achieved by a function such as,

$$\Delta W_{ij} = \alpha(I_j - W_{ij})(\sin d_{bi}/2d_{bi}) \quad (2)$$

where α is the learning rate, d_{bi} is a measure of the distance between units i and b, and the term (sin $d_{bi}/2d_{bi}$) acts to reduce the magnitude of weight changes with increasing distance from b. The magnitude of α is typically defined as a decreasing function of iteration. The final output of the SOFM is dependent on the selected network parameter values, notably the size and shape of the output layer and the number of iterations together with N and α and their associated 'shrinkage' terms (Schalkoff, 1992). Typically these parameters are defined subjectively on the basis of trial investigations.

3. Test site and data

Attention focused on data from surveys of vegetation acquired from sites in Exmoor National Park, UK (Fig. 2). Large tracts of the Park are covered with moor and heath, much of high conservational value, and these were the subject of the field surveys. The analyses were undertaken

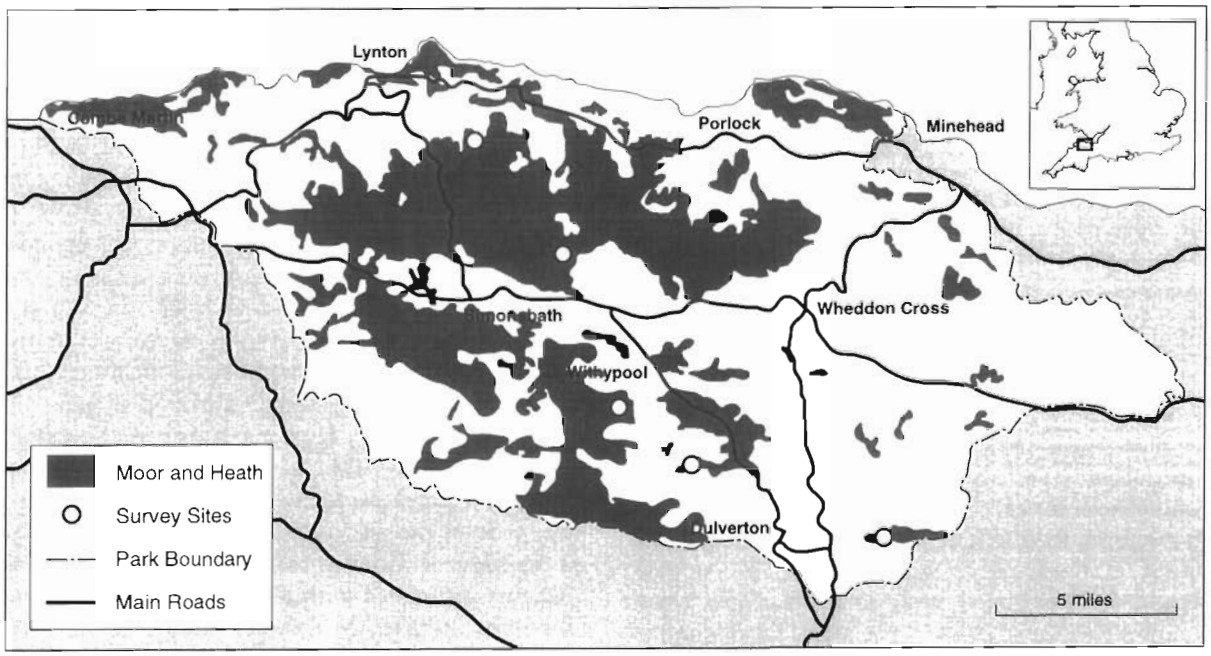

Fig. 2. Location of the test sites within Exmoor National Park.

on a data set that comprised the presence/absence of 10 vegetation types in 1 m quadrats sampled along transects located at five sites in the Park (Fig. 2). These vegetation data were acquired typically at 50 or 100 m intervals along the transects, the sampling approach was defined to support a vegetation mapping project. Here the data from 211 quadrats acquired during a ~6 week period in mid-summer were used.

4. Methods

The nature of the SOFM output is dependent on the settings of the various network parameters. Here the parameters were selected on the basis of trial runs and the literature, with the study directed mainly at the exploration of the spatial arrangement of samples in the SOFM's output space. The main focus of the investigation was a SOFM comprising 25 output units arranged in a square layer and 10 input units. The weights between the input and output units were initially set randomly and then adjusted during the course of network learning. The parameter defining the size of the neighbourhood around the winning unit, identifying the weights to be included in the up-dating phase of network learning, was set at 0.65 and was reduced with iteration by the function

$$N_t = N_o(1 - t/T) \tag{3}$$

where N_o is the initial neighbourhood size, N_t the neighbourhood size at iteration t and T the total number of iterations to be performed. The learning rate was initially set at 0.30 and also declined with iteration by the function:

$$\alpha_t = \alpha_0(1 - t/T) \tag{4}$$

where α_0 and α_t represent the initial setting and that at iteration t, respectively. The total number of learning iterations, T, was 50 000.

For comparative purposes the results of the SOFM analysis were evaluated against outputs from a set of widely used alternative techniques. The main focus was on three other forms of unsupervised classification. The methods used were a basic k-means clustering algorithm, a hi-

erarchical clustering algorithm (using between group linkages based on Euclidean distance measurements) and the fuzzy c-means (FCM) algorithm. Of these classifiers, the FCM is perhaps the least encountered but has been used for the classification of ecological data (e.g. Equihua, 1990; Foody, 1996). The FCM is a non-hierarchical clustering algorithm that may be used to subdivide a data set into c clusters or classes. In a fuzzy c-partition of a data set, the membership functions characterise the membership of each sample in all classes. Memberships close to unity indicate a high degree of similarity between a sample and a class whereas memberships close to zero indicate little similarity between a samples and a class. The algorithm used to derive these membership values was that described by Bezdek et al. (1994) using Euclidean distance measurements and with the parameter $m = 2$. A conventional hard classification was achieved by allocating each sample to the class with which it has the highest fuzzy membership value.

All classifications have a large subjective component and so comparison of the groupings derived from different algorithms, or even from the same algorithm with a different set of parameters, is likely to reveal differences. Furthermore, the evaluation of the differences is difficult and the identification of the most appropriate classification contentious. Nonetheless, the concern here is that the SOFM offers a largely assumption-free and flexible method of classification that may sometimes be more applicable than other methods. The evaluation of the SOFM classification is, therefore, made relative to the other methods with particular emphasis on how classes derived from the other algorithms are located in the space defined by SOFM output layer. As the potential of the SOFM was to be evaluated relative to the other classifications in the absence of a means for absolute evaluation, the classifications were not optimised. Other studies have sought to compare the accuracy of SOFM classifications with a range of other classifications (e.g. Waller et al., 1998). Here the focus was on the spatial arrangement of the vegetation samples in the SOFM output space with particular emphasis on the potential to organise the data into classes. If the SOFM is providing a realistic classification of the data and the other techniques are reasonably applicable, it would be expected that samples belonging to the same class, as assessed by the other classification algorithms, would cluster together in the SOFM output layer. Furthermore, it would be expected that classes that are aggregated in the hierarchical classification would be relatively close in the SOFM outputs. The classifications derived from the other algorithms, therefore, provide the backdrop for the evaluation of the SOFM classification.

5. Results and discussion

After the iterative learning phase in the SOFM analysis, each of the 211 samples was associated with an output unit. Each output unit contained some of the samples and there was no obvious discrete grouping of cases (Fig. 3a). This is likely to be a function of both the nature of the vegetation at the site and low level of measurement precision of the vegetation data, but it is possible that different network parameters, particularly a larger output layer, could have produced a sharper classification. Nonetheless, mapping the vegetation presence/absence data for each sample into the SOFM output reveals that some classes are associated with different parts of the SOFM output layer (Fig. 4), with some occurring across much of the output layer while others were very concentrated into a limited region. For example, the samples containing members of the *Cyperaceae* were associated with the units along the base of the SOFM outputs while gorse was associated almost exclusively with the upper right corner of the SOFM output layer. The SOFM, therefore, appeared to have organized the samples such that the various output units were associated with different vegetation groupings. To assess this more rigorously, the SOFM output was compared with the other classifications.

As an initial step, various three and four class classifications were derived using the three alternative classification algorithms, these numbers selected subjectively on the basis of the nature of the vegetation and dimensions of the SOFM.

These three classifications (k-means, hierarchical clustering and FCM) differed markedly. For instance, the pairwise correspondence between the three 4-class classifications revealed no more than 58% agreement. The main concern here, however, was that classes defined by each of the alternative classifications should be associated with different parts of the SOFM output. That is, the samples of any class should be clustered together in the SOFM output. For each of the alternative classification outputs, the class associated with each sample was mapped onto the location of the sample in the SOFM output. Each SOFM output unit was then associated with the class that domi-

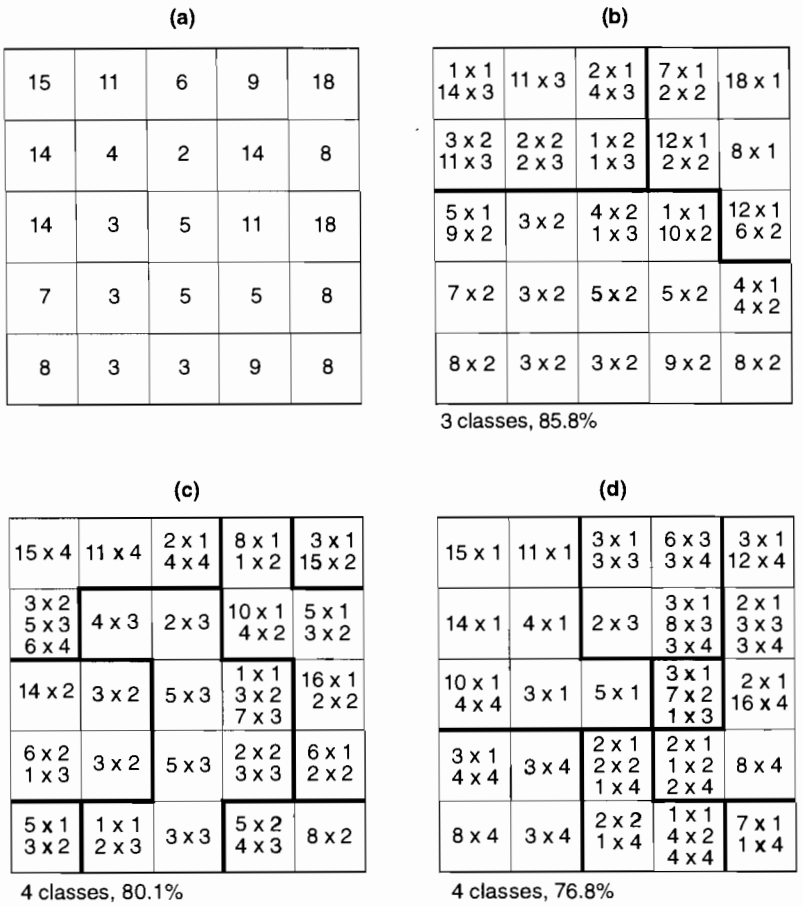

Fig. 3. Summary of the class allocations made by a selection of classifications and their relationship with the SOFM output. (a) a representation of the SOFM output layer showing the number of samples in the data set associated with each unit; (b) relationship between the class allocations derived from the 3 class k-means classification and the SOFM output; (c) relationship between the class allocations derived from the 4 class k-means classification and the SOFM output; (d) relationship between the class allocations derived from the 4 class fuzzy c-means classification and the SOFM output. Note that in (b)-(d) the number of samples ε allocated to a class λ defined by one of the three alternative classification algorithms is indicated as $\varepsilon \times \lambda$; the class codes used are those derived from the alternative algorithm and are not comparable between the various alternative algorithms. Each SOFM output unit was labelled with the code of the class that dominated it or, if there was a tie, by the code of the co-dominant class that also dominated surrounding units. The SOFM output space was then partitioned to show the location of the classes defined by the alternative classifications. The boundaries of the classes defined are indicated by the bold lines between output units. The level of agreement between the partitioned SOFM output space and the alternative classification was then expressed as the percentage of samples with a class label defined from the alternative classification algorithm that occupied the region associated with the same label in the partitioned SOFM space.

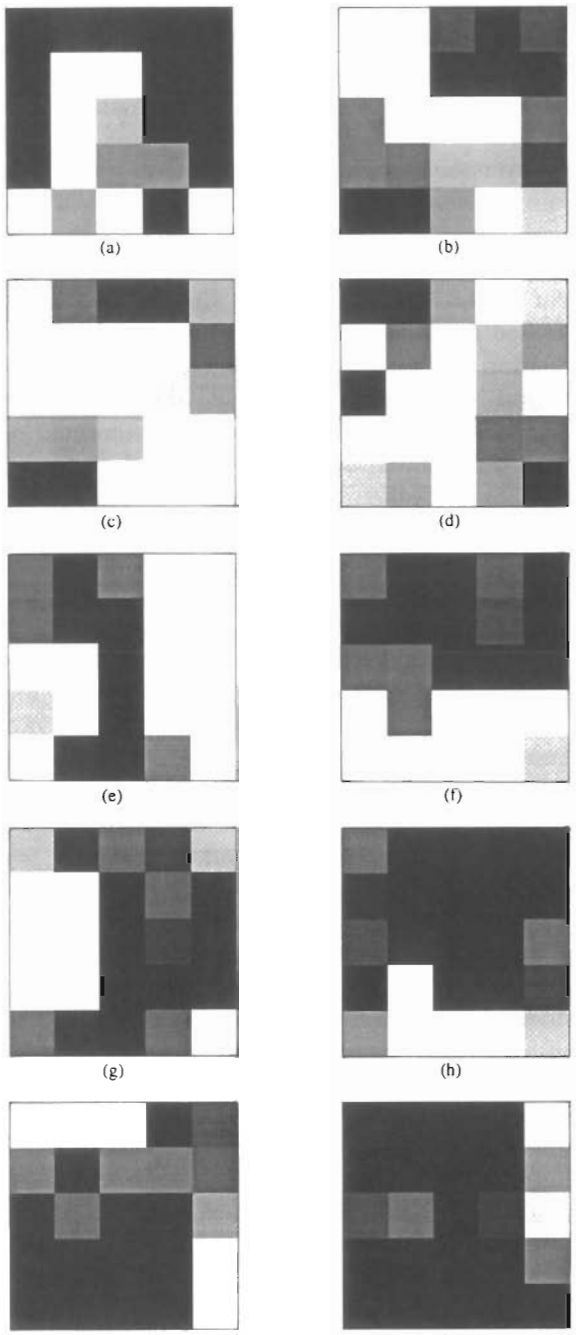

Fig. 4. Distribution of the ten vegetation types in the SOFM output. The grey level of each output unit indicates the percentage of samples associated with that unit (Fig. 3a) for which the specified vegetation type was present (white = 100%, black = 0%). The classes, as defined in the ground data, were (a) *Juncaceae*; (b) broad leaved grasses; (c) fine leaved grasses; (d) *Molinia*; (e) *Ericaceae*; (f) *Cyperaceae*; (g) flowering plants; (h) mosses; (i) bracken; and (j) gorse.

nated the samples it contained and the SOFM output space thereby partitioned by the classes defined by the alternative classification algorithm. To derive an index of the level of agreement between the derived partitioning of the SOFM output space and the allocations from an alternative classification, the percentage of samples with the same class label was computed. Fig. 3 shows the distribution of the samples of the classes defined in some of the classifications. In each, it was apparent that the samples of each class are clustered in the SOFM output layer, with a relatively high degree of agreement (> 76%). In comparison against the three class classification with the k-means algorithm (Fig. 3b), for example, the classes occupied the upper right corner, upper left corner and lower portion of the SOFM output. Moreover, disagreements (samples associated with a SOFM output unit more strongly associated with a class other than the one allocated by the alternative classification algorithm) were spatially concentrated around the (arbitrary) boundaries of the classes depicted on the SOFM output. Overall, it was apparent that samples of each class derived from the other alternative classifications were clustered in the SOFM output with the classes occupying different regions of the SOFM output space (Fig. 3).

Since the number of classes permitted can have a significant influence on the nature and quality of an unsupervised classification or clustering analysis, the hierarchical classification was designed to produce a range of classifications, with between 2 and 10 classes. The nature of the class aggregation in the hierarchical classification is shown in Fig. 5. As with the evaluations relative to the other classifications, the allocated class label of each sample was mapped onto its location in the SOFM output layer at each level in the hierarchy defined. As previously, samples of the same class were found to cluster and the classes appeared to be associated with different parts of the SOFM output layer (Fig. 6). In addition to the classes occupying different locations, it was apparent that the samples and hence classes were distributed by relative similarity. This was evident in comparison against the classifications derived at different levels of the hierarchical classification. The samples of classes

that were aggregated at any stage of the hierarchical classification tended to lie within neighbouring output units. Note, for instance, that the samples of classes 3 and 5, joined at the first step of the hierarchical classification, and 2 and 10, joined at the second step of the hierarchical classification, lie close together in the SOFM output (Figs. 6a and 6b). Similar trends are apparent throughout the hierarchy, with the classes merged into the final two class scheme (Fig. 5) occupying approximately the corresponding area in the SOFM output (Figs. 6a and 6f). Following the class aggregation through the hierarchy, the SOFM, therefore, appeared to have provided a classification in which the classes corresponded closely to those defined from the other algorithm and display them in a two dimensional array in which similar classes are located close to each other.

The SOFM output appears, therefore, to locate samples in terms of their relative similarity. This is clearly desirable for a classification, and the analyst could seek to group together samples located in neighbouring SOFM output units into

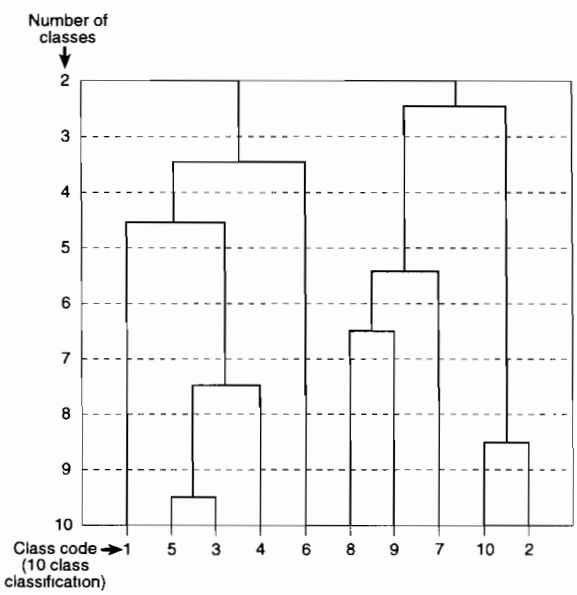

Fig. 5. Summary of class aggregation with the hierarchical classification. The class codes used were those defined in the initial, 10 class, classification

classes. However, the SOFM is also providing a representation of the data that may yield some information on sample similarity analogous to an ordination.

Ordination aims essentially to arrange the samples spatially in a manner that reflects their similarity (Goldsmith et al., 1976) which may, therefore, have some correspondence to the output of an SOFM. Although the data set is too limited for a rigorous assessment, a crude investigation of the potential of the SOFM for ordination was undertaken. For this, the mean vegetation vector of the samples allocated to each SOFM output unit was derived and input to a basic ordination analysis, a principal components algorithm (PCA). The PCA, a widely used form of ordination, recast the data and the location of the SOFM output units in the space defined by the first two components is shown in Fig. 7. While care is required in the interpretation, particularly as the data are not ideal and the main components leave much of the variance unaccounted for, some trends are apparent. For example, the SOFM output units in columns A–E and row E–Y (Fig. 7b) are arranged in order along the axis representing second principal component and, slightly less clearly, columns K–O and P–T are ordered along that representing the first principal component. The SOFM, therefore, appears to have ordered the data in a manner related to the axes of the PCA. Unlike the SOFM, however, the PCA allows the distance between samples to be measured in a well-defined and consistent manner along the derived axes and evaluation of the importance of each axis. The selection of which method to use will, therefore, depend on the objectives of the study (Greig-Smith, 1980) with perhaps the SOFM most suited to situations in which the analysis needs to be free from assumptions about the data and/or for low-level data exploration. Although the comparison of the PCA and SOFM analyses is difficult (Blayo and Demartines, 1991) the results do indicate the potential for some interpretation of the SOFM output like an ordination but a more detailed and rigorous analysis is required.

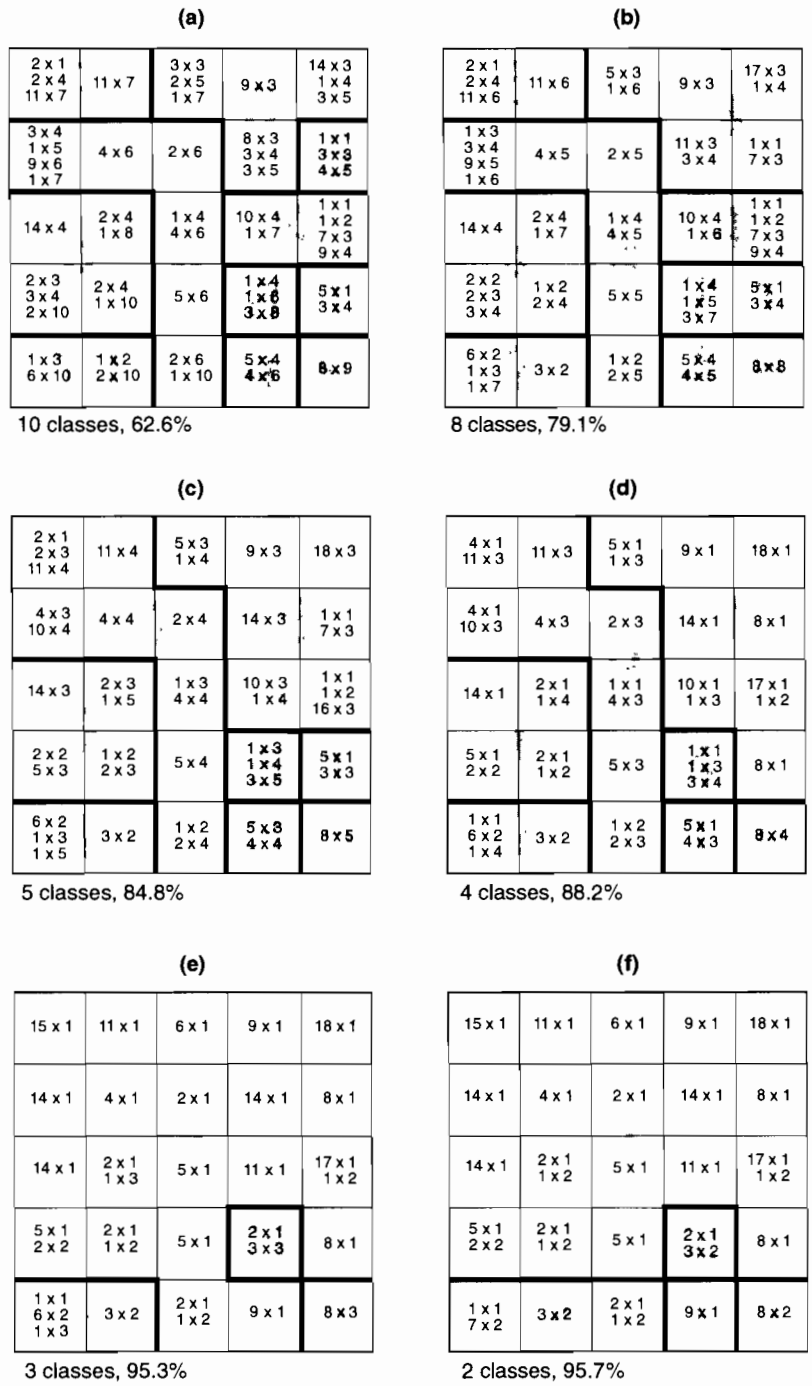

Fig. 6. Location of the classes defined at various steps in the hierarchical classification in the SOFM output space The representation is similar to that used in Fig. 3 and the class labels correspond to those derived from the hierarchical classification. (a) 10 class classification, with the class labels corresponding to those depicted in Fig. 5; (b) 8 class classification; (c) 5 class classification; (d) 4 class classification (relate to Figs. 3c and 3d); (e) 3 class classification and (f) 2 class classification. Note in (f) the regions of the SOFM output space associated with the two classes corresponds closely to the regions associated with the classes defined at the initial 10 class classification from which they were derived (Figs. 5 and 6a).

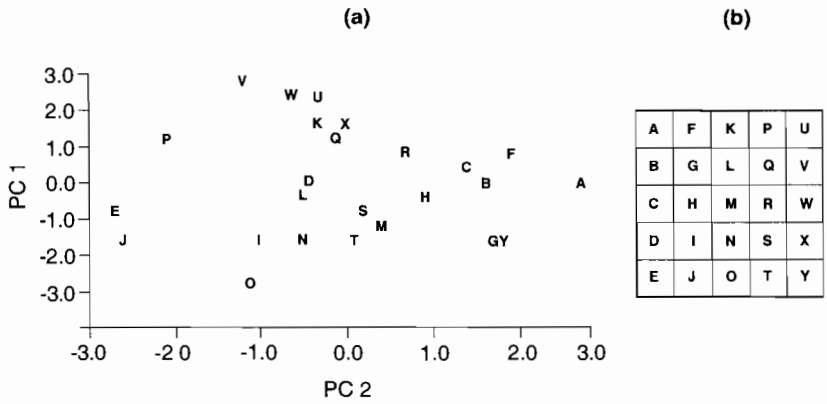

Fig. 7 Comparison of the SOFM output with the results of a principal components analysis. (a) the location of each SOFM output unit, defined in (b), is shown within the space defined by the first two principal components (which accounted for 42.4% of the variance).

6. Conclusions

Neural networks are attractive for ecological studies. Analyses of vegetation survey data with a SOFM neural network showed that samples of classes, defined by other classifications, clustered in the SOFM output space. Moreover, these samples and classes were arranged by similarity, with the class aggregation of a hierarchical classification corresponding essentially to the merging of neighbouring regions in the SOFM output layer. This indicates that the SOFM has potential as a tool for ecological classification (grouping similar samples) and ordination (arranging samples in an ordered manner). Further investigations are, however, required to fully evaluate the use of SOFM in ecological studies. It must be stressed, however, that while offering a simple and largely assumption-free approach the use of the SOFM is not without its problems and limitations. Neural networks such as the SOFM are not a panacea and have a large subjective component with significant analyst input required (e.g. in specification of the network parameters or in group identification).

Acknowledgements

I am grateful to the late Dr L.F. Curtis who, as the Exmoor National Park Officer, generously provided the data set used and the referees for their comments on the paper. The SOFM analyses were conducted with the NCS NeuFrame package.

References

Bezdek, J.C., Ehrlich, R., Full, W., 1994. FCM: the fuzzy c-means clustering algorithm. Computers and Geosci. 10, 191–203.

Blayo, F., Demartines, P., 1991. Data analysis: how to compare Kohonen neural networks to other techniques? In: Prieto, A. (Ed.), Artificial Neural Networks Springer-Verlag, Germany, pp 469–475.

Bosserman, R W., Ragade, R.K., 1982. Ecosystem analysis using fuzzy set theory. Ecol. Model. 16, 191–208.

Bradfield, G E., Kenkel, N C , 1987. Non-linear ordination using flexible shortest-path adjustment of ecological distances. Ecology 68, 750–775.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996 Patternizing communities by using an artificial neural network Ecol. Model. 90, 69–78.

Davalo, E , Naim, P., 1991 Neural Networks. Macmillan, Basingstoke, p. 145.

Equihua, M , 1990. Fuzzy clustering of ecological data J Ecol. 78, 519–534

Foody, G.M., 1996. Fuzzy modelling of vegetation from remotely sensed imagery. Ecol. Model. 85, 3–12.

Gaines, S D , Denny, M.W., 1993. The largest, smallest, highest, lowest, longest and shortest: extremes in ecology. Ecology 74, 1677–1692.

Goldsmith, F.B., Harrison, C.M., Morton, A.J., 1976 Description and analysis of vegetation. In: Moore, P D., Chapman, S B (Eds.), Plant Ecology, 2nd edition Blackwell, Oxford, pp. 437–524.

Goodacre, R., Neal, M J., Kell, D.B., 1994. Rapid identification using prolysis mass spectrometry and artificial neural networks of Propionibacterium acnes isolated from dogs J. Appl. Bacteriol. 76, 124–134.

Greig-Smith, P., 1980 The development of numerical classification and ordination. Vegetatio 42, 1–9.

Greig-Smith, P , 1996. Applications of numerical methods in rain forest. In: P.W. Richards (Ed.), The Tropical Rain Forest, 2nd edition, Cambridge University Press, Cambridge, appendix 2, pp. 497–502.

Lek, S., Belaud, A., Dimpoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Marine and Freshwater Res. 46, 1229–1236

Ludwig, J.A., Reynolds, J.F., 1988. Statistical Ecology. In: A Primer on Methods and Computing. Wiley, NY, p. 337.

Maier, H.R., Dandy, G.C., 1998. The effect of internal parameters and geometry on the performance of backpropagation neural networks: an empirical study'. Environ. Model. Software 13, 193–209.

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river Freshwater Biol 38, 237 246.

Potvin, C., Roff, D A., 1993. Distribution-free and robust statistical methods. viable alternatives to parametric statistics? Ecology 74, 1617–1628.

Schalkoff, R., 1992 Pattern Recognition: Statistical Structural and Neural Approaches. Wiley, NY, p. 364.

Terborgh, J., Flores, C., Mueller, N.P., Davenport, L., 1997. Estimating the ages of successional stands of tropical trees from growth measurements. J. Tropic. Ecol. 14, 833–856.

Tong, S.T.Y , 1992 The use of non-metric multidimensional scaling as an ordination technique in resource survey and evaluation: a case study from southeast Spain. Appl. Geogr. 12, 243–260.

Waller, N.G , Kaiser, H.A., Illian, J.B., Manry, M , 1998 A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and 3 hierarchical cluster analysis algorithms. Psychometrika 63, 5–22

ELSEVIER

Ecological Modelling 120 (1999) 109–118

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Radial basis function networks with partially classified data

Isabella Morlini *

Istituto di Statistica, Università degli Studi di Parma, Via J.F. Kennedy 6, I-43100 Parma, Italy

Abstract

The problem of estimating a classification rule with partially classified observations, which often occurs in biological and ecological modelling, and which is of major interest in pattern recognition, is discussed. Radial basis function networks for classification problems are presented and compared with the discriminant analysis with partially classified data, in situations where some observations in the training set are unclassified. An application on a set of morphometric data obtained from the skulls of 288 specimens of *Microtus subterraneus* and *Microtus multiplex* is performed. This example illustrates how the use of both classified and unclassified observations in the estimate of the hidden layer parameters has the potential to greatly improve the network performances.

Keywords: Classification; Discriminant analysis; Mixture analysis, Radial basis function networks

1. Introduction

One of the major problems related to practical applications in pattern recognition is the presence of partially classified data. In these situations the population from which the sample is taken consists itself of a number of several homogeneous sub-populations, but the group membership of the training data is known only for some input vectors. If the quantity of data available is sufficiently large, and the proportion of unclassified observations is small, then the simplest solution is to discard those patterns from the data set. This approach, however, is implicitly assuming that the cause of the omission of the group membership is independent of the data itself. If the reason of the omission of the group membership depends on the data, then this approach will modify the effective data distribution (Bishop, 1995). When there is too little data to discard the unclassified one, or when the proportion of unclassified observations is high, it becomes important to use all the information which is potentially available from the incomplete patterns. It is intuitively clear, in fact, that the unclassified observations, as well as the classified ones, contain some knowledge about the distribution of the measured variables in the different groups.

The purpose of this work is to show the benefits of using the information contained in a partially classified data set to the maximum extent. Radial basis function networks are introduced and demonstrated to be a suitable method in situations where some observations in the training data are unclassified. An application on an ecological

* Fax: +39-0521-902375.

E-mail address: morlini@economia.econ.unipr.it (I. Morlini)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00095-2

problem, which illustrates how to include unclassified observations in the network training, and which compares the network performances with those reached by conventional discriminant analysis and by discriminant analysis with partially classified observations, is presented. The network performances are measured in terms of classification error rate and generalisation to unobserved patterns.

2. Radial basis function networks

Radial basis function (RBF) networks provide a powerful technique for generating multivariate, non-linear mappings (Broomhead and Lowe, 1988). Unlike the widely used multi-layer perceptron, that is based on units which compute a non-linear function of the scalar product of the input vector and a weight vector, the activation of a RBF hidden neuron is determined by the distance between the input vector and a prototype vector. The RBF network mapping from a d-dimensional input space x to a c-dimensional target space t is a linear combination of a set of M basis functions, which take the form:

$$y_k(x) = \sum_{j=1}^{M} w_{kj}\phi_j(\|x-\mu_j\|) + w_{k_0} \quad k = 1,\ldots, c \qquad (1)$$

where x is the d-dimensional input vector with elements x_i and μ_j is the vector determining the centre of basis function ϕ_j and has elements μ_{ji}. The basis functions can be normalised (Moody and Darken, 1989) through lateral connections between different hidden units in the network diagram, so that the output becomes:

$$y_k(x) = \sum_{j=1}^{M} w_{kj} \frac{\phi_j(\|x-\mu_j\|)}{\sum_{j=1}^{M} \phi_j(\|x-\mu_j\|)} \quad k = 1,\ldots, c \qquad (2)$$

Usually the distance $\|x-\mu_j\|$ is taken to be Euclidean and several form of basis functions can be considered, the most common being the Gaussian:

$$\phi_j(\|x-\mu_j\|) = \exp\left(-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right) \qquad (3)$$

where the standard deviation σ_j, also called smoothing parameter, determines the width of the hidden unit. If the basis functions are Gaussians, then the hidden units assume a localised nature: the network forms a representation in the space of hidden units which is local with respect to the input space, because, for a given input vector, only few hidden units will have significant activations. The use of radial basis functions can be motivated from a number of different concepts as function approximation, noisy interpolation, density estimation and optimal classification theory (Bishop, 1995). In this work we are considering the use of such networks for a classification problem. A multilayer perceptron can separate classes by using hidden units, which form hyperplanes, or hypersurfaces in the input space, and for this reason can be related to discriminant analysis. A RBF network is able to model each class distribution by local kernel functions, and so can be rather compared with the kernel discriminant analysis. If, in a classification problem, the goal is to model the posterior probabilities $p(C_k|x)$ for each of the classes C_k, $(k = 1,\ldots, c)$, then these probabilities can be obtained through Bayes' theorem, using prior probabilities $p(C_k)$ as follows:

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)} = \frac{p(x|C_k)P(C_k)}{\sum_{k=1}^{c} p(x|C_j)P(C_j)} \qquad (4)$$

where $P(\cdot)$ indicates a probability and $p(\cdot)$ a probability density function. If the class-conditional distributions are obtained by using not a single kernel function, but a mixture model constituted by a common pool of M basis functions, labelled by an index j and equal for every density, then the probabilities $p(x/C_k)$ and $p(x)$ can be written as

$$p(x|C_k) = \sum_{j=1}^{M} p(x|j)P(j|C_k) \qquad (5)$$

and

$$p(x) = \sum_{k=1}^{c} p(x|C_k)P(C_k) = \sum_{j=1}^{M} p(x|j)P(j) \qquad (6)$$

where priors $P(j)$ are given by

$$P(j) = \sum_{k=1}^{c} P(j|C_k)P(C_k) \qquad (7)$$

The posterior probabilities can be obtained by substituting Eqs. (5) and (6) into Bayes' theorem (4) and adding an extra factor of $1 = P(j)/P(j)$ to give:

$$P(C_k|x) = \frac{\sum_{j=1}^{M} P(j|C_k)p(x|j)P(C_k)P(j)}{\sum_{j'=1}^{M} p(x|j')P(j')P(j')} = \sum_{j=1}^{M} w_{kj}\phi_j(x) \quad (8)$$

This expression represents a radial basis function network (Bishop, 1995), in which the normalised basis functions are given by

$$\phi_j(x) = \frac{P(x|j)P(j)}{\sum_{j'=1}^{M} p(x|j')P(j')} = P(j|x) \quad (9)$$

and the second layer weights are given by

$$w_{kj} = \frac{P(j|C_k)P(C_k)}{P(j)} = P(C_k|j) \quad (10)$$

After training, for a particular partition of the data into c groups, the value of each k output neuron, $(k = 1, \ldots, c)$ can be interpreted as the posterior probability of corresponding class membership. Thus, following the optimal classification rule (Anderson, 1984), in a two class problem an observation should be classified as belonging to group k if the value of the corresponding output unit is bigger than 0.5. In practice, when least squares are used to set the second layer parameters and the target values are coded with the 1-of-c coding scheme (so that they sum to unity), the output values are forced to sum to unity but they are not forced to lie in the range [0, 1]. If the output values do not lie in this range, they should be normalised.

The major problems related to a RBF network are the determination of the number of basis functions and the choice of the parameters. The faster and simplest procedure is to create a Probabilistic Neural Network (Specht, 1990) which has N localised hidden units centred on each input vector. In these networks the parameters σ_j are usually heuristically determined. One approach is to choose all σ_j to be equal and to be given by some multiple of the average distance between the basis function centres. This ensures that the basis functions overlap to some degree and hence give a relatively smooth representation of the distribution of the training data. In order to determine the number of basis functions by the complexity of the data, rather by the size of the data set, a subset of the input vectors can be chosen by forward selection or orthogonal least squares to serve as centres. A different approach is to choose the number of basis functions and determine the parameters by supervised or unsupervised methods. An exhaustive list of these methods, together with their theoretical issues, is in Bishop (1995). A k-means procedure is adopted in the example of section 4. This procedure proposed by Moody and Darken (1988), sets the centres of basis functions equal to the cluster centres found by the k-means clustering algorithm, and the standard deviations σ_j equal to the average distances to the z-nearest clusters. Moody and Darken (1988) report good empirical results for using this procedure. The main drawback of this method is that the number of basis functions must be defined a priori. This leads to similar problems as the 'number of hidden units' dilemma in the multi layer perceptron, since it is very difficult to estimate an appropriate number of basis functions. In Section 4 we determine the optimal number of clusters (and, therefore, the optimal number of basis functions in the RBF network) on the basis of the within-groups and between-groups deviances, for different number of groups. Once the parameters of the hidden layer are determined, the network has to be trained to produce the optimal values of the second layer weights. When the error function is a quadratic function of these weights, its minimum can be found in terms of the solution of a set of linear equations. In fact, if we indicate with N the number of training cases and with $t_k(x_n)$ the target value for output unit k when network is presented with input vector x_n $(n = 1, \ldots, N;$ $k = 1, \ldots, c)$, then the sum of squares error function is given by

$$E = \sum_{n=1}^{N} \sum_{k=1}^{c} \{y_k(x) - t_k(x)\}^2 \quad (11)$$

where y_k is defined in Eqs. (1) and (2). Training is then very fast and does not have the problem of local minima.

3. Estimating group membership with partially classified observations

In real applications, especially in biological and ecological modelling, it sometimes happens that group membership is known only for a subset of the original sample. This can arise, for example, when the exact determination of group membership requires high laboratory costs. In these situations, classical supervised methods, like the discriminant analysis or the multi-layer perceptron, are often applied. Classified observations are used to estimate the discrimination rule and this rule is then applied to unclassified observations, to determine the corresponding group membership. Evidently, this procedure does not use the information contained in the data to the maximum extent, since it is clear that the unclassified observations contain some information about the distribution of the measured variables in the groups, as well. There is also some theoretical literature on the benefits of using unclassified observations for estimation (O'Neill, 1978; McLachlan and Basford, 1988). On the other hand, using an unsupervised procedure (like mixture analysis, cluster analysis or the Kohonen network) over the entire data set means ignoring group membership of classified observations and, therefore, discarding important available information. Airoldi et al. (1995) found that mixture analysis, compared with discriminant analysis on a data set with partially classified observations, reveals highly unstable estimates. They conclude that ignoring group membership is a bad idea. In statistics, an iterative method that uses the information contained in both classified and unclassified observations in the parameter estimation is fairly well developed under the name of discriminant analysis with partially classified data (*discrimix*). This method (McLachlan and Basford, 1988; Airoldi et al., 1995) has the potential to greatly improve the estimation of the classification rule. However, it is a re-estimation procedure which may involve some technical problems in the solution of the equation system. These drawbacks are the computational time and costs, the eventual convergence to a singular estimate of the covariance matrix (that will cause the algorithm to fail), the absence of convergence or the convergence to a local maximum. Some of these problems can be overcome with a constrained maximum solution and the availability of good computer programs. Therefore, the main drawback of this method seems to be the assumption of multivariate normality of the density function in each group. This assumption is indispensable in discriminant analysis with partially classified data, since the density function appears explicitly in one equation of the system. This is also a crucial difference to discriminant analysis, where calculus can be justified without assuming normality or any other particular distribution.

RBF networks in which the basis functions parameters are estimated by unsupervised procedures are particularly advantageous for applications with partially classified observations, since the hidden layer parameters can be determined using both labelled and unlabelled data, leaving a relatively small number of parameters in the second layer to be determined using the classified data. It must be remarked that using unsupervised methods for determining the hidden units parameters, doesn't mean ignoring group memberships in the entire procedure, since the second layer parameters are determined by the solutions of a set of linear equations, which includes target values. One advantage of RBF networks, over discriminant analysis with partially classified data, is that they do not require iterative procedures in the estimate of the second layer parameters. Moreover, they do not need the assumption of multivariate normality or any other particular distribution of the density function of the input variables in each group.

Next section illustrates how the use of unsupervised procedures for the determination of the basis function parameters and, consequently, the use of unlabelled data in the estimate of the classification rule in a problem with partially classified observation, can improve the performances of a RBF network. RBF networks are

also compared with discriminant analysis and discriminant analysis with partially classified observations.

4. Real data set example

4.1. The microtus data

This example is based on the classification of two species of voles (Flury, 1997, pp. 333–339). The two species, *Microtus multiplex* and *Microtus subterraneus*, differ in the number of chromosomes, but are morphometrically difficult to distinguish. The geographic ranges of distribution of the two species overlap to some extent in the Alps of southern Switzerland and northern Italy (Krapp, 1982; Niethammer, 1982). *M. subterraneus* is smaller than *M. multiplex* in most measurements. It usually occurs at elevations from 1000 m to over 2000 m, but it is also found at lower elevations. *M. multiplex* is found at similar elevations, and also at latitudes from 200 to 300 m (South of the Alps). Much of the data available are in form of skull remains, either fossilised or from owl pellets. Till now, no reliable criteria based on cranial morphology have been found to distinguish the two species. The data set consists of eight variables measured on the skulls of 288 specimens found at various places in central Europe: X1 = width of upper left molar 1; X2 = width of upper left molar 2; X3 = width of upper left molar 3; X4 = length of incisive foramen; X5 = length of palatal bone; X6 = condylo incisive length or skull length; X7 = skull height above bullae; X8 = skull width across rostrum. Variables X1 to X5 are measured in mm/1000; variables X6 to X8 are in mm/100. These cranial measurements are relatively inexpensive to carry out, since they can be measured with a measurescope (accurancy 1/1000 mm) and dial calipers (accuracy i/100 mm). Nevertheless, the exact determination of the species requires a costly chromosomal investigation. For this reason, only 89 of the skulls were analysed to identify their species: 43 specimens were from *M. multiplex* and 46 from *M. subterraneus*. The chromosomes were not analysed and species was not determined for the remaining 199 observations.

Airoldi et al. (1995) report a discriminant analysis, a finite mixture analysis and a discriminant analysis with partially classified observations (which they call *Discrimix*) of this data set. Here, we seek to analyse the data with RBF networks and to compare the classification capabilities of different models. The analysis is first performed using both classified and unclassified observation in the optimisation of the basis function parameters. In order to reach better generalisation capabilities, a pre-processing stage is then applied to the network. Results are finally compared with those reached by a RBF network with parameters determined using the sole 89 classified specimens and with those reached by other statistical models.

In the RBF networks considered in the following the input variables are combined via the Euclidean distance function, so that the contribution of an input variables depends heavily on its variability relative to other inputs. In order to give the same importance to every input variable, variables are standardised to zero mean and unit variance before every process.

4.2. Computation of the error rates

Two types of error rates are used to assess the performance of classification procedure. The first, the simplest and most popular error, is the *plug-in error rate:* it is the proportion of observations misclassified when the classification rule is applied to the data in the training sample. The second, the *cross-validation error* (Stone, 1974), is obtained as follows. The sample is divided in k subsets of equal size. The network is trained k times, each time leaving out one of the subsets from training, and using the omitted subset to compute the error rate. If k equal the sample size, and only one observation is used each time to compute the proportion of observation misclassified, than cross validation reduces to the *leave-one-out* error rate. The plug-in error rate is very fast to compute and, since it uses the entire sample to train the network, it is very advantageous when only a little sample is available. The main drawback of the plug-in error rate is that it tends to be overly optimistic, that is, it tends to underestimate the

Table 1
ANOVA table for different number of clusters

Number of clusters	Deviance between	Degree of freedom	Deviance between	Degree of freedom	Deviance total	Degree of freedom	R^2
2	1153.229	1	1142.771	286	2296	1	0.5023
3	1447 640	2	848.360	285	2296	1	0 6305
4	1565.199	3	730.801	284	2296	1	0.6817
5	1636.804	4	659.196	283	2296	1	0.7129
6	1681.827	5	614.173	282	2296	1	0.7325
7	1715.849	6	580 151	281	2296	1	0.7473

probability of misclassifying future observations, since the error is calculated over the same data employed during training. Cross validation gives a better estimate of the generalisation error, namely, the average misclassification rate over the entire space of possible inputs. For this reason, cross validation is often preferred, but if k gets too small, the error estimate is pessimistically biased because of the difference in sample size between the full-sample analysis and the cross-validation analyses. For this reason, a value of $k = 10$ is chosen, since it is shown to offer good empirical results in literature.

4.3. *Using both classified and unclassified observations in a RBF network*

Eq. (9) points out that the basis functions depend solely on the input data and ignore any target information. In particularly, the basis function parameters should be chosen to form a representation of the probability density of the input data and the centre μ_j should be regarded as *prototype* of the input vectors. This justifies the use of unsupervised procedures to determine the basis function parameters, which are usually very fast and can be run a number of time, in order to test the robustness of the results, with low computational costs. Following Moody and Darken (1989), the *k-means* clustering algorithm is performed to optimise both the basis function centres and the widths. The optimal number of clusters is heuristically chosen comparing the within-groups and between-groups deviances, for different values of k. Due to an increase in the number of clusters, the deviance between groups (which indicates the share of total deviance 'explained' by the aggregation of the observations in clusters) increases, while the deviance within (which indicates the error minimised by the algorithm) decreases. As long as the increase in the deviance between groups is considerable, we think it justifies the increase in the complexity of the grouping structure (due to the addition of new groups). We stop adding clusters when this increase becomes poor, in order to reach a good compromise between the proportion of the total deviance 'explained' by the aggregation in groups and a parsimonious number of clusters (which means a clearer and simpler representation of the data set). The ANOVA table obtained running the k-means cluster analysis for the 288 observations, for different values of k (using the package SPSS for Windows, release 7.5), is reported in Table 1. The coefficient R^2 is the ratio between the deviance between groups and the total deviance. The increase in R^2 from 2 to 3 clusters is considerable. From 3 to 4 groups it is still fairly great, while from 4 to 5 clusters it becomes poor. From 5 to 6 and from 6 to 7 groups the increase in R^2 is nearly negligible. The 'optimal' grouping structure, the one which appears to lead to the best *trade off* between number of clusters and variance in each cluster, seems therefore to be associated with $k = 4$.

In a RBF network with eight input nodes (one for each variable), four hidden nodes with centres determined by the cluster means and widths determined by the minima distance between all the other clusters, and second layer weights determined by linear regression, the plug in error rate is 5.62%, while the cross validation error rate is 2.28%.

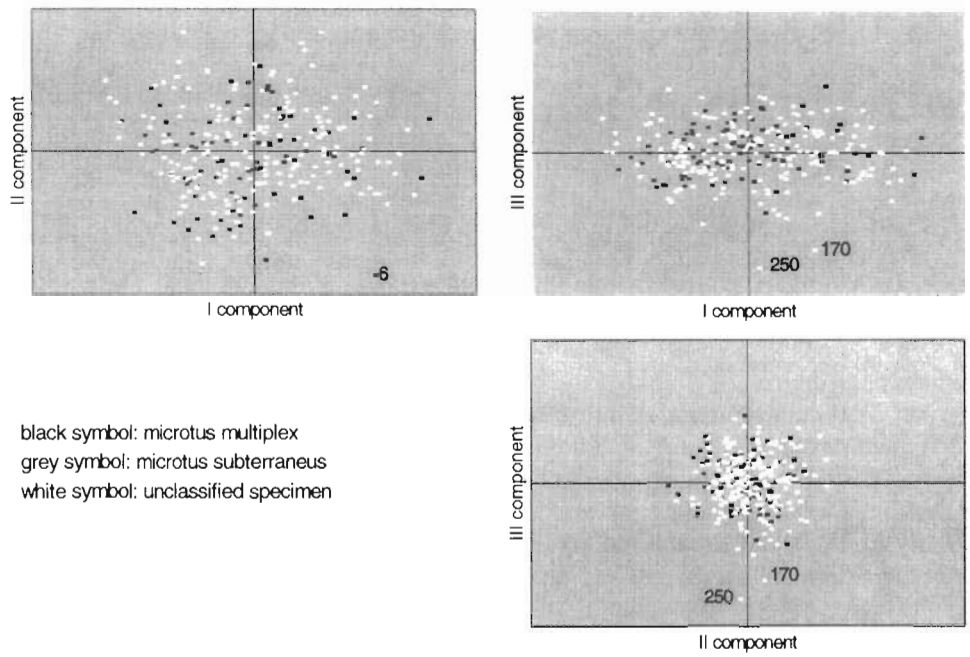

Fig. 1. Scatter plot matrix of the first three principal components.

Analysing the correlation matrix of the input data, it can be noted that the eight variables are highly correlated and the information related to many of these variables is therefore redundant. When input variables are highly correlated, a subset of these variables or a linear transformation of these into new, fewer variables may describe the data equally well and, in accordance with the principle of parsimony (or 'Occam's Razor') the simplest model, the one with fewer variable, should be preferred. Moreover, the network performances may improve with a reduction of the input vector dimensionality (and the related loss of information), since a network with fewer inputs has fewer adaptive parameters to be determined. These parameters are more likely to be properly constrained by a data set of limited size, leading to a network with better generalisation properties. As a pre-processing stage, a principal component analysis is performed in order to form linear combinations of the original variables and generate new (less) input variables for the network. The scatter plot diagram of the first three principal components is reported in Fig. 1. Using the scores of the first n principal components as input variables, the proportion of original information that is preserved can be measured. Since the first three principal components retain the 88% of the original variance, only the 12% of original information is lost using these scores as input variables. The scatter plot diagram of Fig. 1 also reveals the presence of possible multivariate outliers, since observations 6, 170 and 250 clearly stand aside from the cloud of points. In order to determine the basis function parameters, cluster analysis is then performed with $k = 4$ and without this three possible outliers. The second layer parameters are determined by least squares, with a training set of 88 observations (unit 6 is discarded also for linear regression). With this pre-processing step, the plug in error rate of the RBF network is 3.37%, while the cross validation one is 4.49%.

An alternative pre-processing concerning discard of six (redundant) input variables and elimination of the three possible outliers is also applied. Performing the analysis with the sole variables X1 and X4, in which the two groups are

Table 2
Error rates of a radial basis function networks with parameters determined using both classified and unclassified observations

Error rate	RBF network with eight input variables (%)	RBF network with three input variables	RBF network with two input variables (%)
Plug-in	5.62	3.37	3.37
Cross validation	2.28	4 49	3.37

well separated (see Airoldi et al., 1995), the plug in and the cross validation error rates are both 3.37%.

Table 2 summarises the results obtained in the different analysis. Particularly attention must be paid to the first analysis, since the cross validation error rate of the RBF network is less than the plug-in one.

This is a fairly unusual and unexpected result, even if it is not impossible in theory. The explanation of this phenomenon can be related to the normalisation of the basis functions. Normalisation is desirable for a classification problem, since at every point in the input space the sum of the basis function is forced to sum to unity so that, in mixture underlying model, the activation of each basis function can be interpreted as the posterior probability of the presence of corresponding feature in input space (see Eq. (9)) and the network outputs can be interpreted as Bayesian posterior probabilities of group memberships (Bishop, 1995). However, normalisation leads to a number of side effects which are described in Murray-Smith (1994). Some of these side effects should be considered here, in order to motivate the better performances of the network in the test set rather then in the training set. The first one is that when the basis functions are Gaussians, the normalisation results the whole of the input space being covered and not just the region of the input space defined by the training data. The second one is that basis functions with different widths (which are used in the application) can become multimodal, meaning that their activations increase as the distance function between the input vector and the centre decreases (this phenomenon is called 'reactivation' of the basis functions). A final side effect, which also concerns basis functions with different widths, is that the maxima may no longer be at their centres. These three normalisation effects, which are more pronounced as the input dimension increases, due to the increased number of neighbouring units in higher dimensions, justify results reported in the first column of Table 2. From a heuristic point of view, we have noted that, performing the analysis with an unnormalised RBF network, the plug is error rate is less than the cross-validation one.

4.4. Using only classified observations in a RBF network

In a classical set of a probabilistic neural network, the 89 specimens with known group membership should constitute the training sample and, in a subsequent stage, the trained neural network should be used to assign the remaining 199 specimens to either the *M. multiplex* or the *M. subterraneus* group. Using a probabilistic neural network with eight input nodes, one for each explanatory variable and 89 hidden nodes with equal width parameters and centres determined by the input vectors, the following numerical results are obtained. The plug in error rate is 1.12% and the cross validation error rate is 10.1%. Using the first three principal components as input variable, the plug in and the cross validation error rates are both 6.82. Performing the analysis with the two variables X1 and X4, the misclassified observations in the training set are 5 and the plug in error rate is therefore 5.62%. The cross validation error is 8.99%. The reduction of the input vector dimensionality improves the generalisation properties of the network, but these numerical results are still remarkably worse that those previously obtained. The advantage of using a RBF network with basis function parameters determined using both classified and unclassified observations is there-

fore apparent, since generalisation of a result obtained from a particular data set is one of the most important concerns in quit every real applications.

4.5. Comparisons with other concurrent methods

For the 89 classified observations and using the discriminant analysis the following numerical results are obtained for all eight variables (for theoretical and empirical comparisons between discriminant analysis and other classification tools see, for example, Hand, 1981; Ripley, 1994). With prior probabilities given by the relative frequencies of observations in each group, the plug in error rate is 5.62% and the cross validation one is 6.74%. With equal prior probabilities the error rates are, respectively, 4.49 and 6.74%. Using variables X1 and X4, only, the plug in error rate is 4.49% and the cross validation is 5.62%, both for equal and different prior probabilities. Numerical results and parameter estimations obtained from discriminant analysis with partially classified observations are reported in Airoldi et al. (1995). Here it should be noted that error rates obtained with two input variables are remarkably similar to those obtained by conventional discriminant analysis. The advantage of *discrimix* over discriminant analysis is apparent performing bootstrap analysis, since it reveals that the estimates from *discrimix* are typically much smaller. From a numerical point of view, RBF network with basis function parameters given by k-means cluster analysis outperforms procedure *discrimix*. However, comparison between *discrimix* and RBF network should be more detailed, since the purposes of these two methods are different. Discriminant analysis with partially classified observations (like conventional discriminant analysis and mixture analysis) attempts to estimate the parameters of a population which is known to be composed of a fixed number of homogeneous sub-populations. It directly models the class distributions by Gaussian mixtures in the sampling paradigm. The outputs of a RBF network represent, in an underlying mixture model, the posterior probabilities of class memberships. However, procedure k-means partition a data set deterministically into subgroups and the number of these sub-populations is heuristically determined. The hidden layer of a RBF network is used to learn bout the class distributions and to estimate the number of sub-clusters in the training data, when this number is unknown. Procedure k-means can be seen as a particular limit of the expectation-maximisation (EM) algorithm used in *discrimix*. It can be shown that in case of Gaussian basis functions with a common width parameter σ and in the limit $\sigma \to 0$, the EM update formula for a basis function centre reduces to the k-means update formula (Dempster et al., 1977). However, means and variances of the k-clusters are not in general considered as estimators of the parameters of the component densities. Similarly, the mixing coefficient w_{kj}, which are determined by the EM algorithm in *discrimix*, are given by least squares in the RBF network and should be motivated from a geometrical point of view rather than from the principle of maximum likelihood. A final observation relates to the assumption of multivariate normality of the density function in each group. In procedure *discrimix* this density function appears explicitly in the update formula. On the contrary, calculus performed by a RBF network can be justified without assuming normality or any other particular distribution.

If the classification rules found by *discrimix* and RBF network are applied to the observations with unknown group membership, results are remarkably similar. Of the 199 unclassified specimens, 100 are classified as *M. multiplex*, 75 as *M. subterraneus*, and 24 observations are near the classification boundary, giving rise to considerable uncertainty in allocating them in one of the two groups both with *discrimix* and RBF network.

The CPU time is not a real problem, for the *Microtus* data, in any case. Running *Discrimix* takes about 10 s of CPU time on a 486PC, using the Gauss software (Airoldi et al., 1995). Running the principal components for the pre processing stage in the neural network set-up takes about 3 s of CPU time on a pentium PC, using the SPSS for Windows release 7.5. It takes less then 3 s for each run of the k-means cluster analysis and for the solution of the linear equations, to determine the network parameters. However, for very large data

sets, the computational costs are usually higher in *discrimix*. A further technical problem of *discrimix* if that the re-estimation formula must not deterministically converge, while convergence is demonstrated for the k-means algorithm.

5. Discussion

The idea of using RBF networks to process inclomplete data is not new (see Bishop, 1995, p. 184). This work is an attempt to explain and illustrate the use of RBF networks in situations where partially classified data sets occur and to show the differences between this methodology and other competitive methods which are often used in these situations. The goal of this paper is to make RBF networks more popular, since they appear to be rather less well known than the classical multi-layer perceptron, in the neural networks field, and than discriminant analysis and discriminant analysis with partially classified observations, in statistics. The application on the *Microtus* data demonstrates that RBF networks are a suitable methodological tool for ecological modelling, since the example is a rather typical case. The benefits of using RBF networks with partially classified observations is that no information is wasted and if very few observations are labelled the only alternative to estimate a classification rule is procedure *discrimix*. On the other hand, procedure *discrimix* is not a suitable tool in situations where the normality of the density function in each group is not verified and, for very large data sets, can lead to some technical problems in the solution of the equation systems. These problems are overcome in a RBF network in which the basis functions are trained with the k-means algorithm and the second-layer weights are given by least squares.

References

Anderson, T W., 1984. An Introduction to Multivariate Statistical Analysis. Wiley, NY, p 374.

Airoldi, J.P., Flury, B., Salvioni, M., 1995. Discrimination between two species of Microtus using both classified and unclassified observations'. J. Theor Biol. 177, 247–262.

Bishop, M.C , 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, UK, p. 482.

Broomhead, D.S., Lowe, D., 1988. Multi-variable functional interpolation and adaptive networks Complex Syst. 2, 321–335.

Dempster, A.P , Laird, N.M., Rubin, D B , 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat. Soc., B 39 (1), 1–38

Flury, B., 1997. A First Course in Multivariate Statistics. Springer-Verlag, NY, p. 713

Hand, D.J., 1981. Discrimination and Classification Wiley, NY, p. 218

Krapp, F., 1982. *Microtus multiplex* (Fatio, 1905) Alpen-Kleinwuhlmaus. In Niethammer, J. and Krapp, F., Handbuch der Säugetier Europas, Band 2/I, Nagetiere II, Akademische Verlagsgesellschaft, pp. 319–428.

McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and application to Clustering Marcel Dekker, NY, p. 272.

Moody, J , Darken, C.J., 1988. Learning with localised receptive fields. In· Touretzky, D., Hinton, G., Sejnowsky, T. (Eds.), Proceedings of the 1988 Connectionist Models Summer School. Morgan and Kauffman, San Mateo, pp. 133–143.

Moody, J., Darken, C.J , 1989. Fast learning in networks of locally-tuned processing units. Neural Comput. 1 (2), 281–294

Murray-Smith, R., 1994. A Local Model Network Approach to Nonlinear Modelling. Ph.D. Thesis, Department of Computer Science, University of Strathclyde, Glasgow, Scotland, Nov.1994, pp. 71–79.

Niethammer, J., 1982. *Microtus subterraneus* (de Sélys-Longchamps, 1836). In Niethammer, J. and Krapp, F., Handbuch der Säugetier Europas, Band 2/I, Nagetiere II, Akademische Verlagsgesellschaft, pp. 397–418.

O'Neill, T.J., 1978. Normal discrimination with unclassified observations. J Am. Stat. Assoc. 73, 821–826.

Ripley, B.D., 1994. Neural networks and related methods for classification. J. R. Stat. Soc., B 56 (3), 409–456.

Specht, D.F., 1990. Probabilistic Neural Networks. Neural Networks 3, 110–118

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictor. J R. Stat. Soc., B 36, 111–147.

ELSEVIER

Ecological Modelling 120 (1999) 119–130

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Neural network architecture selection: new Bayesian perspectives in predictive modelling Application to a soil hydrology problem

Jean-Pierre Vila [a,*], Vérène Wagner [a], Pascal Neveu [a], Marc Voltz [b], Philippe Lagacherie [b]

[a] *INRA, Laboratoire de Biométrie, 2 Place Viala, F-34060 Montpellier, France*
[b] *INRA, Station de Sciences des Sols, 2 Place Viala, F-34060 Montpellier, France*

Abstract

The aim of this paper is to present to the community of ecologists concerned with predictive modelling by feedforward neural network, a new statistical approach to select the best neural network architecture (number of layers, number of neurons per layer and connectivity) in a set of several candidate networks. The interest of this approach is demonstrated on a soil hydrology problem. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Ecological modelling; Soil sciences; Neural networks; Non-linear regression; Bayesian model selection

1. Introduction

In a large number of ecological studies predictive modelling approaches are employed. Some of them are mechanistic but many are empirical. Among the latter linear regression is often used. But in many instances, the relationships between the predicting and predicted variables are not truly linear, and linear regression is then used only because the non-linear form of the relationships are not known. In that case, a neural network modelling can be a well-adapted alternative non-parametric solution to this modelling problem.

For example, this is the case in soil hydrology, where large efforts have been made to predict by multilinear regressions the soil hydraulic properties from easy-to-measure soil variables like clay, silt, sand and organic matter contents (e.g. Wösten and Van Genuchten, 1988; Williams et al., 1992; Kern, 1995; Bastet et al., 1998). In effect the knowledge of soil hydraulic properties, i.e. water retention characteristics and hydraulic conductivity of the soil, are essential for modelling transport of water and solutes through the soil and evaluating the availability of soil water to plants. However, it must be recognized that the performance of the predictions of soil hydraulic properties by multilinear regression equations remains unsatisfying for site-specific applications. One reason that can explain this is the fact that

* Corresponding author. Fax: +33-4-67521427
E-mail address: vila@ensam.inra.fr (J.-P. Vila)

PII: S0304-3800(99)00096-4

the relationships between the soil hydraulic properties and the basic soil variables are complex and basically non-linear (Hillel, 1980). Consequently, one way of improving the prediction of soil hydraulic properties is to look for multivariate non-linear relationships. In this respect artificial neural networks are interesting tools as they do not require to specify a priori the shape of the non-linear relationships, and can easily take into account qualitative variables as predictors. Tamari et al. (1996) already experienced this approach for predicting soil hydraulic conductivity from particle size distribution, organic matter content and bulk density. In their work and in other applications of neural networks one major problem is the selection of the best neural network architecture (number of layers, number of neurons per layer and connectivity) in a set of several candidate networks.

In this paper we present a new statistical approach to this general problem and illustrate it with a case study in which we use feedforward neural networks for predicting the water retention properties of the soils of a region in southern France.

Besides the fundamental question of choosing the right set of input variables, defining an adequate neural topology for approximating a function by a feedforward neural network still remains an unsatisfactorily solved question. The search for a satisfactory compromise between good data fitting on one side and good generalization performance on the other side, has oriented the design of several on-line and off-line building procedures. Among on-line techniques, constructive algorithms, as for example cascade correlation (Fahlman and Lebière, 1990; Kwok and Yeung, 1997), follow an incremental approach by starting with a small network and trying to increase it step by step. On the other hand pruning algorithms as optimal brain damage (Le Cun et al., 1990) or optimal brain surgeon (Hassibi et al., 1994), follow a decremental approach by starting with a large network and trying to eliminate unnecessary connections (Reed, 1993). But both types of step-by-step evolution do not ensure the reaching of the best topology. Moreover, most of the used termination criteria lack clear statistical meaning. Among off-line selection procedures cross-validation (Golub et al., 1979) is one of the most favored, because of its simplicity and apparent objectivity, but it is pointwisely dependent on the dataset and cannot take into account any probabilistic information.

Statistically-based comparison techniques divide themselves into two main groups:

- asymptotic comparison tests (Wald test, likelihood ratio test) and procedures (Akaïke criteria, . .) (Seber and Wild, 1989)
- comparison procedures based on an approximate Bayesian analysis (MacKay, 1992; Thodberg, 1996).

Neither of these two approaches is definitely satisfying. The first one relies upon samples of sufficiently large size and is often restricted to the comparison of embedded networks. The second one assumes the disposal of a pertinent prior probability distribution for the network parameters (weights and biases) and even if this prior is available this approach can suffer from several controversial points and drawbacks: treatments of the hyperparameters introduced by the prior distribution (integrated out or not (Wolpert, 1993; MacKay, 1995)), debatable estimation of the complex posterior weight distributions by questionable Gaussian approximation (MacKay, 1992) or by heavy Monte Carlo procedures (Neal, 1996), to cite a few of them.

However, the Bayesian approach could appear as the most promising: when no reliable subjective prior distribution is available the modern Bayesian theory can provide nevertheless powerful solutions, and among them efficient procedures for the identification of pertinent prior distributions allowing exact posterior distribution calculation (Berger, 1985; Bernardo and Smith, 1994). We then decided to consider the problem of the selection of a neural network architecture in this renewed Bayesian framework.

Following Bernardo and Smith (1994), we based our neural network model selection procedure on the maximization of an expected utility criterion defined from a predictive sample reuse procedure. By comparison with the cross-validation procedure based on pointwise predictions, this criterion uses a predictive probability distri-

bution determined for each candidate model. It then selects the model under which some predictive probability-based internal consistency of the training dataset is maximized. For a given candidate model this predictive distribution is asymptotically estimated from the assumed Gaussian likelihood of the data and the corresponding conjugate prior density of the model parameters. The heart of this approach is in the determination of this particular prior density of the model parameters which offers the advantage of allowing analytic calculations of parameter and network response posteriors.

The main ideas of this Bayesian network architecture selection approach will be now briefly described. We refer to Vila et al. (1998) for a full theoretical description. All the necessary Bayesian prerequisites are relegated in Appendices A and B at the end of the paper.

2. A general Bayesian non-linear regression model selection procedure

Let us first introduce the general principles of the model comparison procedure that we shall apply to our neural network architecture selection problem. This procedure is based on the maximization of an expected utility criterion and is described at length in Appendix B.

Let $\{M_j\}_{j\in J}$ be $N = \text{card}\ (J)$ models to be compared from $Z_n = ((x_1, y_1),\ldots,(x_n, y_n))$, n independently and identically distributed (i.i.d.) pairs of observations.

Model M_j (for example a neural network) is given by:

$$y_i = f_j(x_i, \theta_j) + \varepsilon_i \quad 1 \le i \le n \tag{1}$$

where $(x_{i_1}, \ldots x_{i_l}) \in \mathbb{R}^l$, $1 \le i \le n$; the $\{\varepsilon_i\}$ are independently and identically normally distributed with mean 0 and variance denoted $1/\lambda_j$; $\theta_j \in \mathbb{R}^{q_j}$.

We randomly select $K \le n$ observations $\{(x_{l_1}, y_{l_1}),\ldots,(x_{l_K}, y_{l_K})\}$ from the dataset and for each model M_j we compute the expected utility criterion U_j according to Eq. (15) in Appendix B:

$$U_j = \frac{1}{K}\sum_{k=1}^{K} \log p_j(y_{l_k}|x_{l_k}, Z_{n-1}[l_k]) \tag{2}$$

where $Z_{n-1}[l_k]$ denotes Z_n with observation $z_{l_k} = (x_{l_k}, y_{l_k})$ deleted and $P_j(y_{l_k}|x_{l_k}, Z_{n-1}[l_k])$ is the posterior predictive density of model M_j for the observation z_{l_k}, having observed $Z_{n-1}[l_k]$ (see Appendix B).

Among the N candidate models we shall retain the one for which the quantity U_j is maximal.

In order to compute U_j we need to know the posterior predictive density p_j, which results from a Bayesian analysis described in Section 3.

3. A Bayesian analysis of non-linear regression

Let us suppose that model M_j, as described by Eq. (1) is the true model of the data Z_n. To alleviate the notations we shall drop index j from all relevant quantities since only one model, M_j is considered in all this section.

Let $Y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$ and $X = (x_{ik})_{1 \le k \le l}^{1 \le i \le n}$

Conditional to X, the probability density of Y is multinormal with mean $F(X, \theta) = (f(x_1, \theta),\ldots, f(x_n, \theta))^t$ and variance $\frac{1}{\lambda} I_n$. We shall denote it:

$$p(Y|X, \theta, \lambda) = N_n(Y|F(X, \theta), \lambda I_n) \tag{3}$$

In this section we shall determine a posterior predictive density $p(y|x, Z_n)$ which will be used in the selection procedure. To do that, we shall first choose a particular parameter prior density: a density member of the family of conjugate prior densities related to the postulated gaussian likelihood (Bernardo and Smith, 1994). The form of this density will greatly simplify the calculation of the relevant parameter posterior density and then that of the posterior predictive density.

3.1. Prior density of (θ, λ)

It can be easily shown that $p(Y|X, \theta, \lambda)$ belongs to a $2n$-parameter exponential family of probability distributions (see definition A2) with sufficient statistics:

$$h_i(Y) = \begin{cases} y_i & 1 \le i \le n \\ y_{i-n}^2 & n < i \le 2n \end{cases} \tag{4}$$

Using proposition A1 it can be shown under regular assumptions (Vila et al., 1998) that the conjugate prior density of (θ, λ) is consistently asymptotically given, as n tends to infinity, by:

$$p(\theta, \lambda|\mathscr{T}) \asymp N_q\left(\theta|\theta^0, \lambda\sum\nolimits_0\right)\mathrm{Ga}(\lambda|\alpha, \beta) \quad (5)$$

(we use the symbol $\asymp$ to denote asymptotic equality) with:

$$\theta^0 = \operatorname{argmin}\,[\mathscr{T}_{\cdot}^0 - F(X, \theta^0)]'[\mathscr{T}_{\cdot}^0 - F(X, \theta^0)]/2 \quad (6)$$

$$\alpha = \frac{n\,\tau_0 - q}{2} + 1$$

$$\beta = \tau_0[\mathscr{T}_{\cdot}^0 - F(X, \theta^0)]'[\mathscr{T}_{\cdot}^0 - F(X, \theta^0)]/2$$

$$\sum\nolimits_0 = \tau_0\,\dot{F}_{\theta^0}^t\,\dot{F}_{\theta^0} \quad (7)$$

where $\mathscr{T} = (\tau_0, \ldots, \tau_n, \ldots, \tau_{2n})'$ is the vector of the parameters of the conjugate prior density; $\mathscr{T}_{\cdot}^0 = (\tau_1/\tau_0, \ldots, \tau_n/\tau_0)'$; $\dot{F}_{\theta^0}$, the Jacobian matrix of $F(X, \theta)$ at θ^0; $N_q(\theta|\theta^0, \lambda\Sigma_0)$, density of the multivariate gaussian distribution of dimension q, with mean θ^0 and variance matrix $(\lambda\Sigma_0)^{-1}$; $\mathrm{Ga}(\lambda|\alpha, \beta)$, density of the gamma distribution with parameters α and β, with mean $\alpha\beta^{-1}$ and variance $\alpha\beta^{-2}$.

Without any prior information on the hyperparameters $\mathscr{T}$ we followed the 'empirical Bayes' point of view (Maritz and Lwin, 1989) which sets these quantities to values maximizing the marginal density of the observations Y. We showed (Vila et al., 1998) that a very good compromise between optimality and tractability of computation is simply given by:

$$\tau_i = \begin{cases} y_i & 1 \le i \le n \\ y_{i-n}^2 & n \le i \le 2n \end{cases}$$

$$\tau_0 = 1 \quad (8)$$

The procedure which consists of building the prior distribution from the data themselves and their likelihood, rather than from a subjective approach of some a priori information, is at the border between Bayesian and frequentist statistics and does not belong to the classic Bayesian paradigm. However it has proved its high practical value when no reliable prior is available and constitutes one important chapter of modern Bayesian analysis (Robert, 1995) since the first introduction of the conjugate prior approach by Raiffa and Schlaifer (Raiffa and Schlaifer, 1961). Moreover and above all, conjugate priors have been designed to lead very easily to the corresponding posterior distributions.

3.2. Posterior density of (θ, λ)

According to Eq. (5) and proposition A2 and after some algebra, the posterior density of (θ, λ) conditional to Z_n is consistently asymptotically given as n tends to infinity, by:

$$p(\theta, \lambda|Z_n) \asymp N_q\left(\theta\middle|\theta_n, 2\sum\nolimits_0\lambda\right)Ga\left(\lambda|\alpha + \frac{n}{2}, \beta_n\right) \quad (9)$$

where:

$$\theta_n = (1/2)(2\theta^0 + (\dot{F}_{\theta^0}^t\,\dot{F}_{\theta^0})^{-1})\dot{F}_{\theta^0}^t(Y - F(X, \theta^0))$$

$$\beta_n = \beta + \frac{1}{2}(Y - F(X, \theta^0) - \dot{F}_{\theta^0}(\theta_n - \theta^0))'(Y - F(X, \theta^0))$$

3.3. Predictive posterior density $p(y|x, Z_n)$ where y and x satisfy model (1)

$$p(y|x, Z_n) = \int p(y|x, \theta, \lambda)\,p(\theta, \lambda|Z_n)\mathrm{d}\theta\,\mathrm{d}\lambda$$

where $p(y|x,\theta,\lambda) = N(y|f(x,\theta),\lambda)$ by assumption.

After some algebra:

$$p(y|x, Z_n) \asymp \mathrm{St}\left(y|f(x, \theta_n), g_n(\dot{f}_{\theta^0})\left(\alpha + \frac{n}{2}\right)\beta_n^{-1}, 2\alpha + n\right) \quad (10)$$

with:

$$\dot{f}_{\theta^0} = \left(\frac{\partial f(x, \theta)}{\partial\theta_1}, \ldots, \frac{\partial f(x, \theta)}{\partial\theta_q}\right)_{\theta = \theta^0},$$

$$g(\dot{f}_{\theta^0}) = 1 - \dot{f}_{\theta^0}\left(\dot{f}_{\theta^0}^t\dot{f}_{\theta^0} + 2\sum\nolimits_0\right)^{-1}\dot{f}_{\theta^0}^t$$

and

$$\mathrm{St}\left(y|f(x,\theta_n), g_n(\dot{f}_{\theta^0})\left(\alpha+\frac{n}{2}\right)\beta_n^{-1}, 2\alpha+n\right)$$

is the Student density with mean $f(x,\theta_n)$ and variance $\beta_n/(g_n(\dot{f}_{\theta^0})(\alpha+n/2-1))$

4. Application to neural networks

Let us suppose now that model $\{M_j\}$, the true model, is a feedforward neural network model. Again, we shall drop index j from all relevant quantities since only this model will be considered in this section.

With the empirical Bayes setting (Eq. (8)) we are using for the hyperparameters $\{\tau_i\}$, θ^0 in Eq. (6) is the maximum likelihood estimate of θ (and $\beta=[Y-F(X,\ \theta^0)]'[Y-F(X,\ \theta^0)]/2$). The whole previous procedure relies on the implicit assumption of the unimodality of the likelihood function (Eq. (3)) for each of the N candidate models. If one of these likelihood functions is multimodal, which occurs frequently for non-linear regression models, the procedure can be impaired. Multimodal likelihood is the rule for a general network mapping function $f(x,\ \theta)$ such that of a multilayer perceptron: there are several families of local optima in the parameter space (Vila et al., 1998). For a given network topology with H hidden layers and m_h units in layer h, there are $\mathrm{SF}=\prod_{h=1}^{h=H} m_h!\,2^{m_h}$ parameter settings of equal likelihood, which are obtained from each other by sign changes of the biases and input and output weights of the units and by unit interchanges. Then, each local mode of the likelihood function will in fact belong to a class of SF equivalent optima of the likelihood. Moreover, several classes of such local optima can coexist. Let NC be the total number of such classes. The relevant formulae for the prior and posterior distributions in the multimodal case are given in (Vila et al., 1998) under general assumptions (negligible overlaps between the NC × SF specific priors):

4.1. Parameter prior density of $(\theta,\ \lambda)$

$$p(\theta,\lambda)=\frac{1}{\mathrm{NC}\times\mathrm{SF}}\sum_{c=1}^{\mathrm{NC}}\sum_{s=1}^{\mathrm{SF}} p(\theta,\lambda|\theta^0_{c,s})$$

where $\theta^0_{c,s}$ is the location of the sth likelihood local optima belonging to the cth class, with $1\le c\le \mathrm{NC}$ and $1\le s\le \mathrm{SF}$ and $p(\theta,\ \lambda|\theta^0_{c,s})$ denotes the density of the conjugate parameter prior distribution computed from Eq. (5) for $\theta^0=\theta^0_{c,s}$.

4.2. Posterior density of $(\theta,\ \lambda)$

$$p(\theta,\lambda|Z_n)=\left(1/\sum_{c=1}^{\mathrm{NC}}K_c\right)\sum_{c=1}^{\mathrm{NC}}(K_c/\mathrm{SF})\sum_{s=1}^{\mathrm{SF}}p(\theta,\lambda|Z_n,\theta^0_{c,s}).$$

where:

$$K_c=\frac{\Gamma(\alpha+n/2)}{\beta_{n,c}^{\alpha+n/2}}\frac{(2\pi)^{q,2}}{\sqrt{\det(2\dot{F}'_{\theta^0_c}\dot{F}_{\theta^0_c})}} \tag{11}$$

and $p(\theta,\lambda|Z_n,\theta^0_{c,s})$ denotes the density of the conjugate parameter posterior distribution computed from Eq. (9), for $\theta^0=\theta^0_{c,s}$.

$\beta_{n,c}=(Y-F(X,\ \theta^0_c))'(Y-F(X,\ \theta^0_c))$, with θ^0_c any of the SF likelihood modes $\theta^0_{c,s}$, belonging to the cth class, $1\le c\le \mathrm{NC}$.

4.3. Posterior predictive density $p(y|x,\ Z_n)$ where y and x satisfy model (1)

$$p(y|x,Z_n)\asymp\left(1/\sum_{c=1}^{\mathrm{NC}}K_c\right)\sum_{c=1}^{\mathrm{NC}}(K_c)\,\mathrm{St}\left(y|f(x,\theta_{n,c}), g_n(\dot{f}_{\theta^0_c})\left(\alpha+\frac{n}{2}\right)\beta_{n,c}^{-1}, 2\alpha+n\right) \tag{12}$$

where $\theta_{n,c}=\theta^0_c$.

Remark 1. Practical limitation of NC:

We can notice that according to Eq. (12) $p(y|x,\ Z_n)$ is the weighted average of the predictive posterior densities related to each of the NC classes of local likelihood optima. In this weighted sum the weights K_c, as shown by Eq. (11), are in

inverse proportion to the corresponding error sum of squares $\beta_{n,c}$ and to the associated information matrix determinant, $\det(\dot{F}'_{\theta_c^0}\dot{F}_{\theta_c^0})$. This permits us to neglect in Eq. (12), classes with relatively low and sharp optima which are hopefully the most difficult to detect by numerical algorithms.

Remark 2. Posterior prediction of y given (x, Z):

The mean of the posterior predictive density (Eq. (12)) gives the minimum squared error loss prediction of y:

$$\hat{y}\Big|_{x,Z} = \left(1/\sum_{c=1}^{NC} K_c\right) \sum_{c=1}^{NC} K_c f(x, \theta_{n,c}) \qquad (13)$$

5. A soil science case study

This case study is part of a more general research program which attempts to evaluate the influence of global change on crop yields and available land ressources at the scale of several european regions (Loveland, 1996). For simulating crop yields and estimating land ressources, the water retention properties of the soils must be known. They are usually determined by measuring the soil water retention curve, namely the relationship between the metric soil water potential and the soil water content. But, since this soil property is expensive to measure, it is rarely measured for all soils in a region. As pointed out in the introduction, one way for solving this problem is to seek functions or models relating the soil water retention curves to soil properties that are largely available over the region of interest. This is what this case study aims at in the case of the Plain of Languedoc in southern France. In the following we describe the sampling and data and the results we obtained by using the statistical method described above for selecting the best predictive neural networks.

5.1. Sampling and measurements

The major soil classes of the Languedoc Plain were sampled at 138 locations (Moulènes, 1993; Leenhardt et al., 1994). At each sampling site, we dug a pit, observed the soil and distinguished the soil layers of different morphology, texture and origin. Among the 138 pits, 372 soil layers were distinguished. Each soil layer was sampled for determining both its soil water retention curve and basic soil properties that are currently determined in soil survey and can be used as predictors of the former.

For the determination of the soil water retention curve, undisturbed soil clods of about 30 cm^3 were taken. Their water contents at five metric water potentials: 3, 10, 30, 100 and 300 kPa were measured using the pressure plate extractor (Smith and Mullins, 1991). These are the variables that we consider hereafter as the output variables of the predictive neural networks to be built. They correspond to a set of points of the retention curve. In principle, for use in soil water flow models, the entire retention curve has to be know. Therefore, after prediction of the set of points, the whole curve is generally reconstructed by fitting a parametric model to the predicted points. This step will not be done here since we are only interested in evaluating the proposed neural network selection approach.

The basic soil variables that we measured were the bulk density (bd), the particle size distribution and the organic carbon content of the soil. These variables are considered as the input variables of the neural networks. The bulk density of the horizon was measured in situ with a surface gamma densimeter (Troxler 3411). Disturbed soil samples were taken for measuring in the laboratory the proportions of five particle size classes, namely clay (cl) (0–2 μm), fine silt (fs) (2–20 μm), coarse silt (cs) (20–50 μm), fine sand (fs′) (50–200 pm) and coarse sand (cs′) (200–2000 μm), and of organic carbon content (oc).

5.2. The neural network model selection

We shall restrict the presentation to the modelling of the third metric water potential ($wp30$) from the seven independent variables bd, oc, cl, fs, cs, fs′, cs′.

In order to assess the efficiency of our Bayesian selection procedure in selecting the 'best' predic-

Table 1
Learning RSS, U and CV values for the candidate architectures

Model	RSS	CV	U
A_{19}	0 136	0.1512	2.260
A_{27}	0.118	0.141	2.355
A_{34}	0 114	0 134	2.313
A_{37}	0.113	0.140	2.324
A_{46}	0.110	0 146	2.339

tive neural network architecture, we compared it with the frequently used standard cross-validation procedure. For a given dataset, and a set of candidate models, the cross-validation procedure selects the model which minimizes $CV = \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$, in which $\hat{y}_i$ is the model prediction at x_i, after adjustment on the dataset without the ith pair (x_i, y_i)

The family of one hidden layer feedforward neural networks was considered (its ability to uniformly approximate any continuous function is well-known (Cybenko, 1989). The number of units (neurons) in the layer (each with a logistic transfer function) was varied from two to six. All five resultant fully-connected networks were fitted on a same learning basis (LB) of 272 observations randomly chosen among the 372 observations of the initial dataset (DS). We then used well-known statistical pruning techniques (asymptotic t-tests) to detect and suppress some non-significantly non-null connections in each of the five fully-connected networks. The resulting architectures were then refitted and successively compared by the cross-validation procedure and the Bayesian procedure (with $K = n = 272$ in Eq. (2) and NC = 1 in Eq. (12) after application of remark 1). The results are given in Table 1.

5.2.1. *The architecture selected by the Bayesian procedure (U-criterion of Eq. (2))*

The architecture which maximizes the U-criterion is A_{27}, displayed in Fig. 1. It is made of three hidden neurons, partially connected with the seven input variables (one link is missing), which leads to 27 parameters (23 weights + four biases).

5.2.2. *The architecture selected by cross-validation procedure (CV-criterion)*

The architecture which minimizes the CV-criterion is A_{34}, displayed in Fig. 2. It is made of four hidden neurons partially connected with the seven input variables (three links are missing) which leads to 34 parameters (29 weights + five biases).

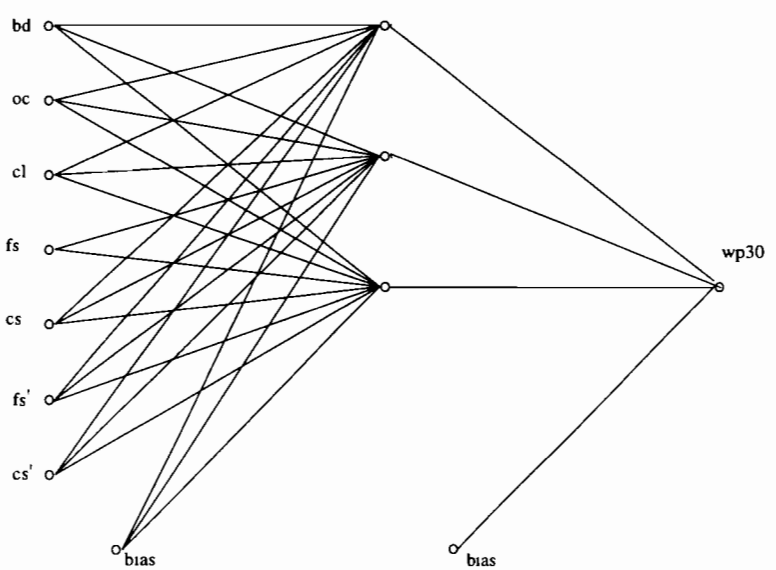

Fig. 1 Network architecture selected by the Bayesian procedure for predicting $wp30$.

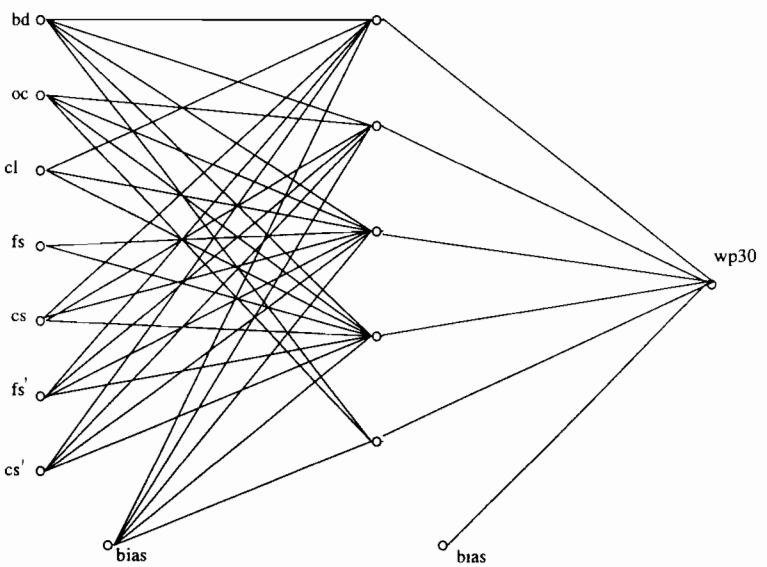

Fig. 2. Network architecture selected by the cross-validation procedure for predicting $wp30$.

The linear regression of $wp30$ on the same seven independent variables was also adjusted to the 272 observations of LB, with a residual sum of squares of 0.173.

The better fit of the learning basis LB by the CV-selected network is not surprising since this network has the highest number of connections and parameters.

5.2.3. Predictive efficiency comparison

5.2.3.1. First comparison. We computed the respective standard errors of prediction (S.E.P.) of the five networks: A_{19}, A_{27}, A_{34}, A_{37}, A_{46} plus that of the linear regression model (lrm), on the subdataset TB = DS − LB, made of the 100 remaining observations of the initial dataset. The results are given in Table 2.

The U-selected network A_{27} ranks first and the linear model lrm second. We note an increase of 23% between the S.E.P. of the U-selected network A_{27} and that of the CV-selected one A_{34}, and an increase of 6% between the S.E.P. of A27 and that of the linear model lrm. We note also the relatively bad behaviour of A_{34}, in fourth position among the six models.

5.2.3.2. Second comparison. We applied the three models previously selected and adjusted on LB: A_{27}, A_{34}, lrm, on 500 test subsamples, each made of 50 observations randomly chosen from the 100 observations of TB. The respective pointwise dependencies of both U and CV procedures on the selection data LB, was then taken into account.

In 435 cases (87%), the S.E.P. of the U-selected network, A_{27}, was lower than that of the CV-selected one, A_{34}. Among these cases, differences of more than 5% (of the U-S.E.P.) between both S.E.P. occurred in 86% of the cases.

In 350 cases (70%), the S.E.P. of the U-selected network was lower than that of the multiple linear regression model. Among these cases, differences of more than 5% between both S.E.P. occurred again in 78% of the cases.

Table 2
S.E.P. of the six models on the TB dataset

Model	S.E.P
A_{19}	0.0877
A_{27}	0.0755
A_{34}	0.0927
A_{37}	0.0959
A_{46}	0.0952
lrm	0.0801

5.2.3.3. Third comparison. As in the second one, we applied the three models A_{27}, A_{34} and lrm on 500 test subsamples, each made of 50 observations randomly chosen from TB. But now for each of the 500 tests, the three selected models were previously readjusted on 322 observations: the 272 observations of LB plus the remaining 50 non-selected observations of TB. The robustness of the two selection procedures with respect to discrepancy between the selection data and the learning data was then taken into account.

In 411 cases (82%) the S.E.P. of the *U*-selected network was lower than that of the CV-selected one. Among these cases, differences of more than 5% between both S.E.P. occurred in 84% of the cases.

In 354 cases (71%) the S.E.P. of the *U*-selected network was lower than that of the multiple linear regression model. Among these cases, differences of more than 5% between both S.E.P. occurred in 70% of the cases.

The overall superiority of the U-selected network, and in particular its increased superiority with respect to the CV-selected one between the third and the second comparison, can be explained by the density-based structure of the *U*-criterion, less data-dependent than the individual-point-based structure of the CV-criterion.

These results are in accordance with the expected greater robustness of the Bayesian selection procedure with respect to the standard cross-validation one.

6. Conclusion

Neural network predictive modelling is one method among several modern non-mechanistic modelling techniques, which can be compared with the so-called non-parametric modelling methods. They all rely on the principle 'let us have the data speak by themselves' in order to characterize the functional links between the independent and dependent variables. This approach is particularly well-suited in the analysis of complex systems, as for example bio-physical and ecological systems, where complete knowledge of all the interacting mechanisms is most often unreachable. However, as flexible modelling tools as they may be, feedforward neural networks need to be duly calibrated to prevent bad predictive performances in case of oversized architectures. Current statistical approaches of this problem, can compare advantageously with the empirically-based most favored ones. However these approaches frequently suffer of lack of generality. But Bayesian statistical analysis of the problem can offer a larger degree of applicability, as for example, the possibility to compare non-embedded networks and even to compare neural networks with models of other types.

We adapted Bayesian analysis to non-linear regression and neural network models selection. This led us to a classic predictive sample reuse procedure, based on two Bayesian concepts, i.e. conjugate prior densities and empirical Bayes setting of hyperparameters, which allow analytic characterization of posterior and predictive densities, while limiting the introduction of a priori information.

we applied this 'least false' neural model selection procedure to several case studies in biological and bio-physical complex systems, as soil hydrology, for which feedforward neural networks appear as competitive modelling tools. The study of soil hydraulic conductivity tackled in this paper, reveals the relative improved efficiency of this Bayesian selection procedure with respect to the more classic cross-validation procedure. Let us note that, as interesting as the present results appear to be, this study points out a desirable improvement of our Bayesian selection procedure: the possibility to compare multioutput neural network models (and more generally, multiresponse non-linear regression models). For our present application, this improvement would allow to select the neural architecture which simultaneously predicts at best, the five water potentials of interest. This multiresponse extension of our Bayesian procedure is presently in project.

Appendix A. Elements of Bayesian analysis

Given a random sample $Z=(z_1, \ldots, z_m)$ of m i.i.d observations, with likelihood $p(Z|\phi)$ where ϕ is an unknown vector of parameters, and a prior density $p(\phi)$, the Z-conditional posterior density is defined by:

$$p(\phi|Z)=\frac{p(Z|\phi)p(\phi)}{\int p(Z|\phi)p(\phi)d\phi}$$

Multidimensional integration is required to calculate this posterior density. In the case of general likelihood and prior, the exact analytical or numerical evaluation of these integrals is most often untractable. To perform integration over the range of ϕ, simultaneous analytic or numerical approximations are then necessary. However exact calculations can be done for a sufficiently rich class of particular priors: the family of conjugate priors.

Definition A1 (conjugate prior family). *The conjugate family of prior densities for $\phi\in\Phi$, with respect to the likelihood $p(Z|\phi)$ with sufficient statistic $t_m=t_m(Z)=\{m, s(Z)\}$ of dimension k, is defined by:*

$$\{p(\phi|\mathcal{T}), \mathcal{T}=(\tau_0, \tau_1, \ldots, \tau_k)'\in T\},$$

where:

$$T=\left\{\mathcal{T}; \int_\Phi p(s=(\tau_1, \ldots, \tau_k)|\phi, m=\tau_0)d\phi<\infty\right\}$$

and

$$p(\phi|\mathcal{T})=\frac{p(s=(\tau_1, \ldots, \tau_k)|\phi, m=\tau_0)}{\int_\Phi p(s=(\tau_1, \ldots, \tau_k)|\phi, m=\tau_0)d\phi}$$

Definition A2 (k-dimensional exponential family). *A probability density $p(z|\phi)$ where $z\in\mathcal{Z}$ and $\phi\in\Phi\subseteq\mathbb{R}^q$ belongs to a k-dimensional parameter exponential family if it can be written:*

$$p(z|\phi)=c(z)g(\phi)\exp\left\{\sum_{i=1}^{k}\psi_i(\phi)h_i(z)\right\}$$

with:

$$\frac{1}{g(\phi)}=\int_z c(z)\exp\left\{\sum_{i=1}^{k}c_i\psi_i(\phi)h_i(z)\right\}dz<\infty$$

The exponential family is said to be regular if the set of possible values of $\mathcal{Z}$ does not depend on ϕ.

Proposition A1 (Conjugate families for regular exponential families of distributions). *Let $Z=(z_1,\ldots, z_m)$ be a random sample from a k-dimensional regular exponential family distribution. Its likelihood is given by:*

$$p(Z|\phi)=\prod_{j=1}^{m}c(z_j)[g(\phi)]^m \left\{\exp\sum_{i=1}^{k}\psi_i(\phi)\left(\sum_{j=1}^{m}h_i(z_j)\right)\right\}$$

The corresponding conjugate family of prior distributions of ϕ has then the form:

$$p(\phi|\mathcal{T})=[K(\mathcal{T})]^{-1}[g(\phi)]^{\tau_0}\exp\left\{\sum_{i=1}^{k}\psi_i(\phi)\tau_i\right\},$$

$$\phi\in\Phi$$

where $\mathcal{T}$ is such that:

$$K(\mathcal{T})=\int_\Phi[g(\phi)]^{\tau_0}\exp\left\{\sum_{i=1}^{k}\psi_i(\phi)\tau_i\right\}d\phi<\infty$$

Proposition A2 (corresponding posterior and predictive densities). *Under the assumptions of proposition A1 and for the corresponding conjugate prior density for ϕ:*

(i) the posterior density for ϕ is given by:

$$p(\phi|Z, \mathcal{T})=p(\phi|\mathcal{T}+t_m(Z))$$

where:

$$\mathcal{T}+t_m(Z)=\left(\tau_0+m, \tau_1+\sum_{j=1}^{m}h_1(z_j), \ldots, \tau_k+\sum_{j=1}^{m}h_k(z_j)\right)$$

(ii) the predictive density for future observations $\tilde{Z}=(\tilde{z}_1, \ldots z_l)$ is:

$$p(\tilde{Z}|Z, \mathcal{T})=p(\tilde{Z}|\mathcal{T}+t_m(Z))$$

$$=\prod_{j=1}^{l}c(\tilde{z}_j)\frac{K(\mathcal{T}+t_m(Z)+t_l(\tilde{Z}))}{K(\mathcal{T}+t_m(Z))}$$

where:

$$t_l(\tilde{Z}) = \left[l, \quad \sum_{j=1}^{l} h_1(\tilde{z}_j), \ldots, \sum_{j=1}^{l} h_k(\tilde{z}_j) \right]$$

Appendix B. Bayesian model comparison

Let $\{M_j\}_{j\in J}$ be $N = \text{card}\ (J)$ models to be compared from $Z_n = ((x_1, y_1), \ldots, (x_n, y_n))$ n independently and identically distributed (i.i.d.) pairs of observations. Model M_j is given by Eq. (1).

Among the N candidate models we shall retain, the one for which the expectation of a given utility function $u(M_j, z, Z_n)$ (defined in the following), is maximum:

$$\bar{u}(M_j|Z) = \int u(M_j, z, Z_n) p(z|Z_n) dz \quad j\in J \qquad (14)$$

where $z = (x, y)$ is a future observation for which the predictive distribution of y at x is wanted and $p(z|Z_n)$ is a density, usually unknown, representing actual beliefs about z having observed Z_n

Let us consider the n partitions of Z_n: $Z_n = [Z_{n-1}[l_k], z_{l_k}]$, $1 \le k \le n$ where $Z_{n-1}[l_k]$ denotes Z_n with observation z_{l_k} deleted. Let us randomly choose $K \le n$ of these partitions. We have by the strong law of large numbers as n and K tend to infinity:

$$\left| \int u(M_j, z, Z_n) p(z|Z_n) dz - \frac{1}{K} \sum_{k=1}^{K} u(M_j, z_{l_k}, |Z_{n-1}[l_k]) \right| \xrightarrow{a.s} 0$$

So, the respective expected utilities of the different candidate models M_j, $j\in J$, can be compared on the basis of the quantities:

$$\frac{1}{K} \sum_{k=1}^{K} u(M_j, z_{l_k}, Z_{n-1}[l_k]) \quad j\in J$$

As we are rather interested in comparing models from a predictive distribution point of view, we shall take as utility function a logarithmic score:

$$u(M_j, z, Z_n) = \log p_j(y|x, Z_n)$$

where $p_j(y|x, Z_n)$ is the predictive posterior density, knowing Z_n, of a response y for the dependent variable observed at x, for model M_j. We then select over $j\in J$ the model M_j for which the following quantity is maximum:

$$U_j = \frac{1}{K} \sum_{k=1}^{K} \log p_j(y_{l_k}|x_{l_k}, Z_{n-1}[l_k]) \qquad (15)$$

This procedure can be considered as a Bayesian cross-validation-like process, which brings us to select the model under which the data achieve the highest level of some kind of 'internal consistency'

References

Bastet, G , Bruand, A., Quétin, P., Cousin, I., 1998. Estimation des propriétés de rétention en eau des sols à l'aide de fonctions de pédotransfert. une analyse bibliographique Etude et Gestion des sols 5 (1), 7–28.

Berger, J.O , 1985 Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory John Wiley, New York

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. 2, 303–314.

Fahlman, S.E., Lebière, C., 1990. The cascade correlation learning architecture. In: Advances in Neural Information Processing Systems 2 Morgan Kaufmann, San Mateo, CA, pp. 524–532.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21 (2), 215–223.

Hassibi, B., Stork, D.G., Wolff, G., Watanabe, T., 1994. Optimal Brain Surgeon: extensions and performance comparisons. In: Advances in Neural Information Processing Systems 6. Morgan Kaufmann, San Mateo, CA, pp. 263–270.

Hillel, D., 1980. Fundamentals of soil physics. Academic Press, New York, 413 p.

Kern, J.S., 1995. Evaluation of soil water retention models based on basic soil physical properties. Soil Sci. Soc. Am. J. 59, 1134–1141.

Kwok, T.Y , Yeung, D.Y., 1997. Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems. IEEE Trans. Neural Networks 8 (3), 630–645.

Le Cun, Y., Denker, J.S., Solla, S.A., 1990. Optimal Brain Damage. In: Advances in Neural Information Processing Systems 2. Morgan Kaufmann, San Mateo, CA, pp. 598–605.

Leenhardt, D., Voltz, M., Bornand, M., Webster, R., 1994. Evaluating soil maps for prediction of soil water properties Eur. J. Soil Sci. 45 (3), 293–301.

Loveland, P.J., 1996 The ACCESS project: agro-climatic change and European soil suitability—a spatially-distributed soil, agro-climatic and soil hydrological model. Int. Agrophys 10, 145–154.

MacKay, D.J.C., 1992. A practical Bayesian framework for backpropagation networks. Neural Comput. 4, 448–472.

MacKay, D.J.C., 1995. Hyperparameters· Optimize or integrate out? In: Maximum Entropy and Bayesian Methods. Kluwer, The Netherlands.

Maritz, J.S., Lwin, T., 1989. Empirical Bayes Methods, 2nd edition Chapman and Hall, London.

Moulènes, D., 1993. Caracterisation hydrodynamique des sols du Languedoc-Roussillon. Recherche de fonctions de pedotransfert. Msc thesis, Ecole Nationale Superieure Agronomique de Montpellier, 45 p.

Neal, R.M., 1996. Bayesian Learning for Neural Networks. Springer, New York

Raiffa, H., Schlaifer, R., 1961. Applied Statistical Decision Theory. Division of Research, Graduate School of Business Administration, Harvard University.

Reed, R., 1993. Pruning algorithms—A review. IEEE Trans. Neural Networks 4, 740–747.

Robert, C.P., 1995. The Bayesian Choice. Springer-Verlag, New York, p. 436

Seber, G.A.F., Wild, C.J., 1989. Non-linear Regression. Wiley, New York, p. 768.

Smith, K.A., Mullins, C.E., 1991. Soil Analysis: Physical methods. Marcel Dekker, New York, p. 620.

Tamari, S., Wösten, J.H.M., Ruiz-Suarez, J.C., 1996. Testing an artificial neural network for predicting soil hydraulic conductivity Soil Sci Soc. Am. J. 60, 1732–1741.

Thodberg, H.H.T., 1996. A review of Bayesian neural networks with an application to near infrared spectroscopy. IEEE Trans. Neural Networks 7 (1), 56–72.

Vila, J.P., Wagner, V., Neveu, P, 1998. Bayesian non-linear model selection and neural networks. Rapport Technique, Laboratoire de Biometrie INRA Montpellier, 22 p.

Williams, R.D., Ahuja, L.R., Naney, J.W., 1992. Comparison of methods to estimate soil water characteristics from soil texture, bulk density and carbon content. Soil Sci. 148 (6), 389–403.

Wolpert, D.H., 1993. On the use of evidence neural networks. In: Advances in Neural Information Processing Systems 5. Morgan Kaufmann, San Mateo, CA, pp. 539–546

Wösten, J.H.M., Van Genuchten, M.T., 1988. Using texture and others soil properties to predict the unsaturated soil hydraulic functions Soil Sci. Soc. Am. J. 52, 1762–1770.

ELSEVIER

Ecological Modelling 120 (1999) 131–139

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Software sensor design based on empirical data

Marie H. Masson [a,*], Stéphane Canu [b], Yves Grandvalet [a], Anders Lynggaard-Jensen [c]

[a] *Heudiasyc, UMR CRNS 6599, UTC, France*
[b] *PSI, INSA de Rouen, Rouen, France*
[c] *VKI Water Quality Institute, Århus, Denmark*

Abstract

Software sensor design consists of building an estimate of some quantity of interest. This estimate can be used either to replace a physical measurement, or to validate an existing one. This paper provides some general guidelines for the design of software sensors based on empirical data. When the model is a priori unknown, the problem can be stated in terms of non-parametric regression or black-box modelling. Complexity control is the main difficulty in this setting. A trade-off must be achieved between two antagonist goals: the model should not be too simple, and model identification should not be too variable. We propose to address this issue by a penalization algorithm, which also estimates the relevance of input features in the identification process. A data-driven software sensor should also provide accuracy and validity indexes of its prediction. We show how these indexes can be estimated for complex non-parametric methods, such as neural networks. An application in environmental monitoring, the design of an ammonia software sensor, illustrates each step of the approach.

Keywords: Software sensor; Black-box modelling; Neural networks; Complexity control; Feature selection; Ammonia prediction

1. Introduction

Nowadays SCADA (supervision control and data acquisition) systems are widely used for environmental monitoring, thus creating large databases. This paper presents a methodology exploiting the redundancy arising in those databases to replace missing measurements, or to cross-check available ones. This methodology is illustrated on a case study: the design of a 'software' or 'virtual' sensor of ammonia.

Ammonia is one of the main indicators of water pollution. Its monitoring is thus needed to assess the water quality in rivers and in waste water treatment plants (WWTP). Some important decisions are derived from this monitoring: in a river, the intake of a downstream drinking water production plant should be closed to avoid a pollution peak. In a WWTP, an efficient feedback control requires the real-time estimation of the ammonia concentration. The availability of the measurements is therefore crucial. Meanwhile, although commercial available ammonia sensors

* Corresponding author. Fax: +33-03-44234477; no: 122331.
E-mail address: mmasson@hds.utc fr (M.H Masson)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00097-6

are improving, they are still expensive to buy and to maintain. Furthermore, as pointed out by Lynggaard-Jensen (1995), they are not reliable in the long-term. and their long response time (10–25 min) causes delays in rising alarms, and difficulties in standard feedback control. It is thus essential to estimate the ammonia concentration by other means than the physical sensor, i.e. by a software sensor.

A software sensor computes an estimate of some quantity of interest, based on a mathematical model and other (faithful) measurements. The computed estimate may be used in place of the measurement when the latter is missing, or as a tool to validate an unreliable physical measurement. In most real world applications, the software sensor estimate will not be as accurate as a carefully tuned physical sensor. If it is designed to replace a physical sensor, the user should be ready to encounter an accuracy loss. But the software sensor has other purposes. It may give predictions of laboratory data, estimates when the measure is missing, and provide a sensor diagnosis when the measure is available.

There are two kinds of software sensors: model-based and data-driven. Model-based, or deterministic software sensors (e.g. Lynggaard-Jensen, 1997) can be built when the physical, biological and chemical relations between the measurements are known up to some constants, and that these constants can be identified. The model is derived from the problem analysis, and the software sensor is build thanks to the estimation of the model parameters. Data-driven, also known as black-box or statistical software sensors are to be used when no accurate model is known. Data-driven methods include kernel and spline smoothers, additive models, projection pursuit and neural networks. These methods estimate the statistical dependence between measurements. For this purpose they require a 'training set' of valid past measures, including the quantity of interest. Hence, the software sensor does not learn the physics of the process, but the behavior of the physical sensor, which had to be installed to provide the training examples.

The first part of this paper gives the general outline of the methodology for the design of a data-driven software sensor, with a special emphasis on neural networks modelling. This methodology is then illustrated by the design of an ammonia software sensor on a real application developed within the EM2S (environmental management and monitoring systems) Esprit project P-22442, which involves the following partners: Suez–Lyonnaise des Eaux (France), VKI Water Quality Institut (Denmark), Danfoss System Control (Denmark), Hitec (Norway), Computas (Norway) and Heudiasyc CNRS (France).

2. Methodology

2.1. Sample selection and data splitting

Before building a data-driven software sensor, a preliminary data validation step has to be carried out. Indeed, it is likely that some sensor failure happened during the data collection. As software sensors mimic physical sensors, sensor failures should be eliminated from the database.

This 'cleaning' step can be carried out by an expert or by automatic validation procedures based on standard signal processing methods (filtering, sequential hypothesis testing). The description of these validation procedures is out of the scope of this paper. The reader will find a review of these methods, together with their use within the EM2S project, in the report of the Diagnosis group (1996).

Once a data set has been selected, it has to be divided into learning and test sets. The learning set is used for the calibration of the software sensor, the test set for its validation. These sets should be independent for the test error to be an unbiased estimate of the modelling error. When there is no time dependency, this condition is ensured by any randomized splitting. In time series, contiguous samples are correlated and should thus appear in the same set. A usual approach is to split the set by taking the first part for training and the last part for testing. However, since environmental time-series present non-stationarities, such that seasonalities and trends, this coarse scheme can not be used. Instead, the data set is divided in blocks of sequential data that are alter-

natively allocated to test and training sets. The number of blocks results from a compromise. Each block should be large enough to ensure small dependency between blocks, and short enough so that the whole phenomenon is represented in the two sets.

2.2. Black-box modelling

In deterministic approaches, a software sensor estimates a function, which is known up to some parameters, and learning refers to the parameter calibration. The accuracy of the results depends on the amount of data available for tuning the parameters and on the appropriateness of the function. For example, a well-tuned linear model will do badly if the dependency is truly non-linear.

In the machine learning framework, the goal of black-box techniques is not explicitly stated as a function approximation problem, but as an inference problem. The aim is to approach the explained variable y for any plausible value of the explicative variables $\boldsymbol{x}$ by some function $g(\boldsymbol{x})$. To achieve this, a data set $S_l = \{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$ and a loss function l are given. The data set is the learning set from which inference is carried out, and the loss function gives a quantitative objective: how much should be paid for a given error? In this context, generalizing means achieving a small average loss on future predictions, i.e. for $\boldsymbol{x} \neq \boldsymbol{x}_i$, $i = 1,\ldots, l$. As predictions are supposed to be done on examples drawn from the distribution of the learning set p, generalization is measured by the mean loss, or prediction error PE:

$$\mathrm{PE}(g) = \int l(y, g(\boldsymbol{x}))dp(x, y) \tag{1}$$

The distribution p being unknown, the prediction error can not be computed. To minimize Eq. (1) the empirical risk minimization principle proposes to minimize:

$$R_{\mathrm{emp}}(g) = \frac{1}{\ell}\sum_{i=1}^{\ell} l(y, g(\boldsymbol{x}_i)) \tag{2}$$

constructed on the basis of the training set. If the loss function l is chosen to be quadratic, one obtains the least squares method:

$$R_{\mathrm{emp}}(g) = \frac{1}{\ell}\sum_{i=1}^{\ell} (y_i - g(\boldsymbol{x}_i))^2 \tag{3}$$

Minimizing Eq. (3) amounts to estimate the expected value of Y given $\boldsymbol{x}$, i.e the regression function $f(\boldsymbol{x}) = \mathrm{IE}[Y|X = \boldsymbol{x}]$. As f is unknown, the function g should be flexible enough to be able to approximate a large class of functions. Well-tried examples include kernel or spline regression, additive models, and artificial neural networks (see e.g. Venables and Ripley, 1994). All these regression methods are able to propose a valid solution to many problems, whereas parametric models propose a precise or invalid solution, whether the model is well specified or not. The choice of one particular method is motivated by some characteristics of the problem. Kernel or spline smoothing are used if there are only a few explicative variables (typically one or two). Additive models provide easily interpretable solutions if there are no interaction effects between explicative variables. If the dimension of the input space is high, and the size of the training set is large, neural networks are well adapted.

2.3. Complexity control

The major pitfall of flexible methods is to misuse their flexibility. The more flexible the model is, the greater is its ability to approach any function, but the more instable is the estimation problem from a finite amount of data. This is known as the approximation/estimation or bias/variance tradeoff. This issue is addressed by the control of flexibility or complexity, which is a crucial step in building an estimate of the regression function. Usually, some parsimony or smoothness conditions are imposed as means to provide this control.

In neural networks, there are two archetypal ways to control complexity: by setting the net architecture, or by setting constraints on the parameters of the net. In the first case, the network is fitted by least mean squares. Complexity is defined by the number of parameters. The optimal number of parameters, or number of hidden units is estimated by constructive or pruning methods reviewed in Reed (1993). In the second case, an oversized network is chosen, and

its parameters are estimated by penalized least mean squares (e.g. weight decay of Plaut et al., 1986). Complexity is defined by the strength of the penalization applied. Here, like in kernel or spline smoothing, the notion of parameters is no more relevant. It is replaced by the number of effective parameters introduced by Moody (1994), or degrees of freedom, as defined by Hastie and Tibshirani (1990). These two methods are respectively equivalent to subset selection and ridge regression in linear regression.

Complexity tuning is usually carried out by estimating the prediction or generalization error PE Eq. (1), and minimizing this estimate. The empirical risk Eq. (2) is a down-biased estimate of PE. The error on an independent validation set is an unbiased estimate of prediction error, but its computation requires to put aside a part of the training set for complexity tuning. Note that the validation set is a part of the training set used to provide an estimate of the optimal complexity. It should not be confused with the test set which role is to estimate the generalization ability of the software sensor on unseen data.

Analytic estimates of the prediction error exist (cf. Krieger and Zhang, 1997), but they are either loose upper bounds, either based on strong assumptions on the data distribution. Some of these estimates have been tested on neural nets, but Tibshirani (1996) experimentally showed that their reliability is much lower than resampling estimates.

Roughly speaking, resampling techniques provide a large validation set for tuning complexity, while the whole training set can still be used to calibrate the software sensor. This intensive use of the training set is done at the expense of intensive computation.

The two main resampling schemes are cross-validation (leave-one-out or leave-many-out), and bootstrap (cf. Efron and Tibshirani, 1993). These methods should be used with care for time-series, as the examples in the training set are correlated. In black-box modelling, this dependence is simulated by sampling blocks of contiguous data. Block-bootstrap requires choosing a relevant block length. It is simpler to use K-fold cross-validation, using large blocks of contiguous data. Breiman (1996) recommends the number of blocks to be between 5 and 10.

The training set is divided in K blocks of contiguous data: $(K-1)$ blocks are used for training, and one block for validating the estimate. This is repeated K times for all possible choices of validating block. For each complexity index (such as the number of parameters in subset selection), K estimates of the regression function are thus computed. The generalization error corresponding to the complexity index is then estimated by the average error on the validating blocks, which minimum estimates the optimal complexity tuning. Finally, the whole training set is used to estimate the net parameters for this tuning.

2.4. Prediction accuracy and validity domain

For practical use, a software sensor should not only provide a pointwise estimate of the quantity of interest, but also an accuracy index, such as a confidence interval. In regression, a confidence interval is usually defined as a band centered on the regression estimate, where the true regression function should lie with some confidence level. Here, the confidence interval should be understood as a band including the regression estimate, where the explained variable should lie with some confidence level. We are interested in the difference $(\hat{f}(\boldsymbol{x})-y)$ between the prediction and the true value, not in the difference $(\hat{f}(\boldsymbol{x})-f(\boldsymbol{x}))$ between the actual and the best possible predication $f(\boldsymbol{x})=\mathrm{IE}[Y|X]=\boldsymbol{x}]$.

When the software sensor is used to cross-check measurements, the confidence interval is necessary to assess the similarity of the two quantities. For missing measurements, the prediction uncertainty is needed to evaluate the risks of decisions based on the prediction.

Another interesting feature of a software sensor is its ability to provide to the end user a self-diagnosis, such as 'prediction unlikely to be valid'. This diagnosis should be given when there is some evidence showing that the operating conditions of the software sensor have changed. For example, if temperature is a feature used by the software sensor, a model trained on summer months data should not be extrapolated to winter months. As

faithful extrapolation can not be guaranteed in data-driven methods, prediction should not be assumed valid far from previously seen cases.

Actually, accuracy and validity are related issues since a very large confidence interval should be given for data far away from training data. If our only assumption about the regression function is stated as a regularity or 'smoothness' hypotheses, then we believe that a training sample $(\boldsymbol{x}_i, y_i)$ should only have a localized influence on the estimate $\hat{f}(\boldsymbol{x})$. Hence the estimate $\hat{f}(\boldsymbol{x})$ is arbitrary when $\boldsymbol{x}$ is far away from the training samples, and the confidence interval should 'blow-up'.

As stated here, assigning a confidence interval to the prediction is a difficult problem. In kernel methods (cf. Härdle, 1990), confidence intervals are determined by the interpolation of pointwise confidence interval at $\boldsymbol{x}_i$. There is no conventional means to account for the fact that $\hat{f}(\boldsymbol{x})$ is almost completely unknown for $\boldsymbol{x}$ far from training data. This may be why no results are usually given in extrapolation with kernel or spline smoothing.

The fact that almost no information on y is gained for $\boldsymbol{x}$ far from training points $\boldsymbol{x}_i$ can be easily accounted for in the Bayesian formalism. However, even if there is no theoretical hurdles, taking into account heteroscedasticity in this framework is technically very difficult.

In this paper, we propose a novel approach to estimate the confidence interval by learning from data. Compared to other estimates based on data proposed by Nix and Weigend (1995), or Heskes (1997), our method does not rely on any assumption about the conditional distribution of $(Y|\boldsymbol{x})$. It is inspired by stacking regression algorithms introduced by Wolpert (1992).

Stacking is defined as a very general technique, designed to improve the accuracy of a regression estimate. We only provide here a simple example introducing our algorithm. Stacking estimates a correction term which should be applied on the top of the predictor for test examples. This correction takes into account some properties of the test example with respect to the training set. First, the training sample is partitioned in several training and validation sets, as in K-fold cross-validation. For each learning set, a predictor $\hat{f}_k$, $k = 1,\ldots, K$ is built. For each $\hat{f}_k$, and each sample of the validation set, the error $\varepsilon_i = \hat{f}_k((\boldsymbol{x}i) - y_i)$ is computed. Moreover, other features $\boldsymbol{c}_i$, such as a distance to the training set, or the estimate $\hat{f}_k(\boldsymbol{x}_i)$ are computed. Once this has been made for each predictor $\hat{f}_k$, the new features $\boldsymbol{c}_i$ are available on the whole training set. They are used, together with $\boldsymbol{x}_i$, to estimate ε_i by a function, the correction term $\hat{g}(\boldsymbol{c}, \boldsymbol{x})$.The whole training sample is then used to estimate $\hat{f}(\boldsymbol{x})$, which prediction is corrected by $\hat{g}(\boldsymbol{c}, \boldsymbol{x})$.

To provide a confidence interval, the only modification made in our algorithm is to estimate the absolute value of the residuals $|\varepsilon_i|$. The function $\hat{g}$ is not a correction applied to the estimate, but a standard deviation estimate. The confidence interval is a variable width band centered on the regression estimate. Its width is proportional to the estimate of standard deviation, and is chosen so that a given percentage of the validation data are in the band.

3. Application

3.1. Site and data description

This case study concerns the monitoring of the Ouche river by the French water supplier Lyonnaise des Eaux. This river is used as pouring for the sewage system and the WWTP of the city of Dijon. To quantify the impact of the WWTP on the river, physical and chemical variables like water temperature, conductivity, pH, ammonia and dissolved oxygen were measured and stored in a database. The temporary monitoring station was situated downstream the sewage and WWTP outlets. As the installation and operating costs of an ammonia sensor are high, this measurement is not available in the permanent monitoring station. The feasibility of a virtual sensor was thus studied.

The raw database is made of 25 000 measurements of each variable during a period of 5 months (April–August), sampled every 6 min. Many sensor failures were detected and the validation procedure eliminated a large amount of data. The further requirement to have blocks of valid data lasting at least 10 h reduced the num-

ber of extracted data to 3200 (during the June–July period). This constraint originated from the will to be able to take into account long-term dependencies, if they were any. Each block of contiguous data was split into two equal blocks, one for the training set, and one for the test set. The learning set and the test set are thus composed of 1600 samples.

This equal splitting may seem to be a waste of data, since the only use of the test set is to validate the software sensor, and to provide a criterion for comparing different solutions. However, each block of contiguous data in the validated database consists only of 3–4 days. The halving of these periods aims at ensuring some independence between training and test data. If the data in the two sets are too dependant, the test set does only validate the software sensor ability to learn, not its ability to generalize. Finally, this work was stated as a feasibility study. In other words, it has to ensure potential results rather than to give actually accurate results.

3.2. Ammonia prediction

Since a relatively large amount of data is available and no interpretation of the results is needed, a multilayered perceptron (MLP) with one-hidden layer was chosen as predictor. The output of such a network is given by:

$$g(\boldsymbol{x}) = \sum_{k=1}^{H} w_{0k} \tanh\left(\sum_{j=1}^{d} w_{jk} x_j + \theta_k\right) + \theta_0 \qquad (4)$$

where H is the number of hidden units, d the number of input variables, w_{jk} and w_{0k} are the weights of the input-to-hidden and hidden-to-output layer, θ_k and θ_0 are the corresponding thresholds, and tanh is the units activation function.

To control the complexity, we use here a version of adaptive ridge regression introduced in Grandvalet (1998) penalizing/pruning the input variables according to their relevance, while controlling the smoothness of the input–output mapping. The overall penalization applied to the network is controlled by a unique hyper-parameter λ. Tuning the complexity is equivalent to the estimation of λ^* minimizing the generalization error. The version of adaptive ridge regression used here is illustrated in Fig. 1. It penalizes differently $d+1$ groups of variables: d groups gather the outcoming weights of the input units $\boldsymbol{w}_j$, $j=1,\ldots,d$, and one group the incoming weights of the output unit $\boldsymbol{w}_0$, excluding all bias terms which are not penalized. The first d groups are used to penalize irrelevant features, and the last one only applies smoothness constraints on the mapping. The estimate $\hat{f}_\lambda$ is defined by

$$\begin{cases} \hat{f}_\lambda = \underset{f \in F}{\operatorname{Argmin}} \dfrac{1}{\ell} \displaystyle\sum_{i=1}^{\ell} (f(\boldsymbol{x}_i) - y_i)^2 + \sum_{j=0}^{d} \lambda_j \|\boldsymbol{w}_j\|^2 \\ \text{subject to } \dfrac{1}{d+1} \displaystyle\sum_{j=0}^{1} \frac{1}{\lambda_j} = \frac{1}{\lambda} \quad , \lambda_j > 0 \end{cases} \qquad (5)$$

where F is the set of MLPs with H hidden units. Let $\boldsymbol{w}_j^{(s)}$ and $\boldsymbol{w}_0^{(s)}$ be the value of $\boldsymbol{w}_j$ and $\boldsymbol{w}_0$ at the step s of the optimization algorithm. The values of λ_j are simply updated by

$$\lambda_j^{(s)} = \frac{\lambda}{d+1} \frac{\sum_{j=0}^{d} \sqrt{\|\boldsymbol{w}_j^{(s)}\|^2}}{\sqrt{\|\boldsymbol{w}_j^{(s)}\|^2}} \qquad (6)$$

where the value of λ is determined by estimation of generalization error by cross-validation.

As was pointed out before, the main interest of adaptive ridge regression is a control of complexity together with a selection of relevant features. The robustness of the network against useless input variables is thus increased. Hence, instead

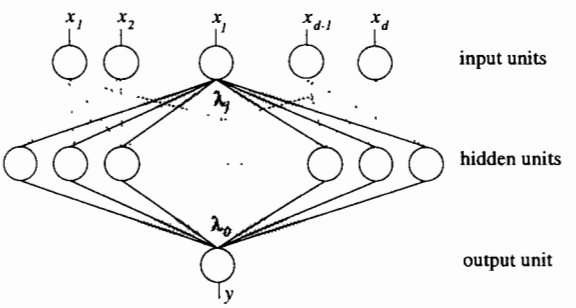

Fig 1 Weigth groups used by adaptive ridge penalization.

of using a feature selection algorithm in a separated pre-processing step, the network may be fed with a large set of potentially explanatory variables.

When dealing with time series, the input variables may include present and past variables. The role of past variables is to take into account the dynamics of the process. In our application, the input variables $(x_1,\ldots, x_d)$ were chosen to be pH, temperature, dissolved oxygen, and conductivity at present time t and previous instants $t-\alpha\Delta t$, with $\Delta t = 1$ h, and $\alpha = 1, 2,\ldots, 10$. The total number of input variables d is thus 44. This choice resulted from a compromise between a good description of short and long-term dependencies and a moderate number of input variables. We know that all these variables are not relevant, but the penalization is designed to estimate which input variables should be used by the model.

The chosen architecture is a one-hidden-layer perceptron with 20 hidden units. As there are 44 input variables, the number of free parameters is about 1000, for about 1600 points in the training sample. This number of free parameters allows a great flexibility. It is thus possible to approach a large class of functions. The number of effective parameters, hence stability of the estimate is governed by the hyperparameter λ of adaptive ridge regression.

3.3. Prediction accuracy and validity domain

The residuals obtained from the ammonia prediction support the hypothesis of heteroscedasticity. The stacking inspired confidence interval described in Section 2.4 is thus used to train a neural network estimating the prediction accuracy. As cross-validation is used to tune complexity, a major part of the computation required by stacking is already done: we have the training set for predicting the errors. We still have to compute new potentially explanatory features, taking into account the properties of test examples with respect to the training set. The feature used are different distance from $\boldsymbol{x}$ to the training sample and the prediction. We provided the distance to the training set gravity center and to the nearest neighbor in the set, with the Euclidean metric and the metric derived from the relevance index given by the adaptive ridge algorithm. The training set $\{(\boldsymbol{x}_i, \boldsymbol{c}_i, |\varepsilon_i|)\}_{i=1}^{l}$ being built, we apply the machinery used for predicting ammonia for estimating the prediction interval.

3.4. Results

Of the 44 input variables, 22 are estimated to be irrelevant during the cross-validation procedure. They are thus deleted of the training sample when estimating $\hat{f}_\lambda$, on the whole training set. The final prediction and the confidence interval on the test set is given in Fig. 2.

The confidence interval width varies in a factor 4 and is coherent with test residuals shown in Fig. 2. These results support the approach, especially the introduction of distances as input variables for accuracy, since the main explanatory variable here is the nearest neighbor distance. The predictor performance is compared for reference to three other predictors shown in Table 1. The results for MLPs are significantly improved over the ones of linear prediction, supporting the existence of nonlinearities in the dependence. The benefits from adaptive penalization in terms of prediction performance are also significant.

The second benefit of adaptive regularization is that the interpretation of results is eased by the computation of the relevance index (proportionally to $1/\lambda_j$ Eq. (5), and normalized to sum to one). This index is indicated in Table 2 for the most significant explanatory variables. Oxygen and conductivity are by far the main input variables. This result is surprising for the chemist, who would expect pH to be more important. But pH is highly correlated with oxygen, and the pH measurement is less accurate.

4. Conclusion

In this paper, we showed the feasibility of an ammonia software sensor on a real application on the site of Dijon. Since the valid data available for this study did not cover a full year, the proposed software sensor can not be implemented without further developments based on new data record-

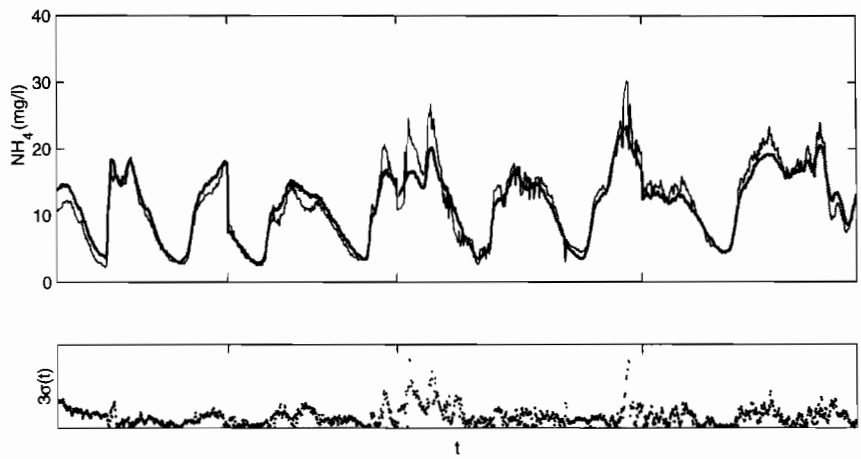

Fig. 2 Top: NH_4 measurement (thin line) and MLP output for adaptive ridge training (thick line) on test data with 90% confidence interval (shaded region); bottom: absolute errors (points) and 90% confidence interval (shaded region).

ing. However, rather than solving a particular problem, this paper aims at providing general guidelines for the design of data-driven software sensors.

When little is known about the nature of phenomenon to be modelled, black-box methods, among them neural networks, are well-adapted. They amount to estimate a quantity of interest from a set of explanatory variables on the sole basis of solved examples. A major issue in black-box modelling is to find the relevant input variables. A feature is relevant according to the dependence and to the predictor used to model this dependence. Thus feature selection should be integrated into the modelling process. The adaptive ridge algorithm is well-suited since it performs input selection while tuning the predictor complexity, which is a crucial step in black-box modelling. Thanks to this selection mechanism, non-linearities are exhibited in our application although the ratio of sample size to input dimension is low.

A usual criticism of black-box models is their failure to provide confidence interval. This confidence measure is essential for the end user to compare the predicted value with the physical measure and to diagnose a faulty behavior of the software sensor. The proposed approach allows a data-based estimation of such intervals to be built. The same supervised algorithm is applied using both the cross-validation residuals as targets and additional input variables. Among the latter, the distance between the current input and its

Table 1
Prediction error for linear regression trained with ridge regression (RR), adaptive ridge regression (ARR), and MLP with weight decay (RR), and adaptive ridge regression (ARR). The intervals are estimated from the intervals of validation set errors

Method	Prediction error
Linear+RR	4.0 ± 0.2
Linear+ARR	3.8 ± 0.2
MLP+RR	5.3 ± 0.2
MLP+ARR	3.1 ± 0.2

Table 2
Relevance (computed from λ_j) for the top six selected explicative variables

Variable	Relevance index (%)
Oxygen(t)	33
Conductivity(t)	21
Temperature(t-100)	8
pH(t-40)	5
Conductivity(t-10)	5
pH(t)	5

nearest neighbor in the training set appears to be the most relevant, justifying the use of training set characteristics as additional variables.

Acknowledgements

We would like to thanks S. Deveughèle from the Lyonnaise des Eaux for providing the data and technical support.

References

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Diagnosis group, 1996. Sensor data validation. Technical Report UTC. CNRS/EM²S/310/12-96, http://www.hds. utc fr/~em2s, p. 43

Efron, B., Tibshirani, R., 1993. An introduction to the bootstrap. In: Monographs on Statistics and Applied Probability, vol. 57. Chapman and Hall, New York, p. 436.

Grandvalet, Y., 1998. Least absolute shrinkage is equivalent to quadratic penalization. In: Niklasson, L, Bodén, M, Ziemske, T (Eds.), ICANN 1998, Perspectives in Neural Computing, vol. 1. Springer, Berlin, pp. 201–206.

Härdle, W., 1990. Applied nonparametric regression. In: Economic Society Monographs, vol 19. Cambridge University Press, New York, p. 333.

Hastie, T., Tibshirani, R, 1990 Generalized additive models. In: Monographs on Statistics and Applied Probability, vol. 43. Chapman and Hall, New York, p. 335.

Heskes, T., 1997. Practical confidence and prediction intervals. In: Mozer, M.C., Jordan, M.I., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA, pp. 176–182.

Krieger, A.M., Zhang, P, 1997 Generalized final prediction error criteria. In: Kotz, S., Read, C.B., Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Update, vol. 1 Wiley, New York, pp. 269–272.

Lynggaard-Jensen, A., 1995 Status for online sensor and automated operation of wastewater treatment plants. Proceedings Nordic Seminar· Nitrogen Removal from Municipal Wastewater, Espoo, Finland, pp. 174–186.

Lynggaard-Jensen, A., 1997. The new sensor development for wastewater treatment plants with nitrogen removal. Proceedings of Nordic Conference on Biological Nitrogen and Phosphorus Removal, Stockholm, Sweden.

Moody, J., 1994 Prediction risk and architecture selection for neural networks. In: Cherkassky, V., Friedman, J., Wechsler, H. (Eds.), From statistics to neural networks, theory and pattern recognition applications, NATO ASI series F: Computer and Systems Sciences, vol. 36. Springer, Berlin, pp. 147–165.

Nix, D., Weigend, A, 1995. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (Eds.), Advances in Neural Information Processing Systems 7. MIT Press, Cambridge, MA, pp. 489–496.

Plaut, D., Nowlan, S., Hinton, G, 1986. Experiments on learning by back propagation. Technical Report CMU-CS-86-126, Carnegie-Melon Department of Computer Science, Pittsburgh, PA, available at http.//www.cs.utoronto.ca/~hinton/backprop.ps, p. 40.

Reed, R., 1993. Pruning algorithms—a survey. IEEE Trans Neural Netw. 4 (5), 740–747.

Tibshirani, R., 1996. A comparison of some error estimates for neural networks models Neural Comput. 8 (1), 152–163.

Venables, W., Ripley, B., 1994. Modern applied statistics with S-plus. In: Statistics and Computing. Springer, New York, p. 462.

Wolpert, D., 1992. Stacked generalizations. Neural Netw. 5, 241–259.

ELSEVIER

Ecological Modelling 120 (1999) 141–156

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

pH modelling by neural networks. Application of control and validation data series in the Middle Loire river

Florentina Moatar [a,*], Françoise Fessant [b], Alain Poirel [c]

[a] *LTHE, UMR 5564, CNRS-INPG-ORSTOM-UJF, BP 53, 38041, Grenoble Cedex 9, France*
[b] *INRETS-MAIA, 2 avenue du General Malleret Joinville, 94114, Arcueil, France*
[c] *EDF-Division Technique Générale, 21, avenue de l'Europe, BP 41, 38 040, Grenoble Cedex 9, France*

Abstract

Artificial neural networks (ANNs) are applied as a new type of model to estimate the daily pH of the Middle Loire river. The model is used for pH measurement screening, error detection (abnormal values, discontinuities and recording drifts) and validating the collected data. The measured values of pH are compared with the values estimated by the ANN model using statistical tests to verify homogeneity and stationarity. River water pH is affected by numerous processes: biological, physical and geochemical. Examples are: CO_2 pressure equilibrium with the atmosphere, photosynthesis, respiration of plants, organic matter degradation, geological and mineral background, pollution etc. Inter-relationships between these processes and pH values are complex, non-linear and not well understood. As a neural network provides a non-linear function mapping of a set of input variables into the corresponding network output, without the requirement of having to specify the actual mathematical form of the relation between the input and output variables, it has the versatility for modelling a wide range of complex non-linear phenomena. For this reason the neural network approach has been selected and tested for pH modelling. We used the classical multilayer perceptron model (MLP).

River discharge and solar radiation variables are used as inputs to the MLP model. The choice of these variables is dictated by what are perceived to be the predominant processes that control pH in the Middle Loire river, which is typically eutrophic during the low-flow summer period. The influence of the previous day's flows and radiation has been evaluated in the calibration and verification test. The best network found to simulate pH was one with two input nodes and three hidden nodes. The inputs are: daily discharge and a variable called 'Index of anterior radiation', i.e. calculated as an exponential smoothing of the daily radiation variable. When calibrated over 4 years of data and tested (i.e. verified) for a one-year independent set of data, the model proved satisfactory on pH simulations, with accuracies in the order of 86%. After elaborating the pH model, the Student test and the cumulative Page–Hinkley test were applied for automatic detection of changes in the mean of the residuals from the ANN pH model. This analysis has shown that such tests are capable of detecting a measurement error occurring over a short period of time (1–4 days).

Keywords: Artificial neural network; Middle Loire river; pH; River discharge; Solar radiation

* Corresponding author. Fax: +33-476-825-286..
E-mail address: moatar@hmg.inpg.fr (F. Moatar)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00098-8

1. Introduction

French environmental regulations impose continuous monitoring of the aquatic environment at every river-site equipped with a nuclear power plant. Therefore 'Electricité de France' performs continuous data acquisition of four parameters: temperature, electrical conductivity, dissolved oxygen and pH, on an hourly basis. Field measurements do not always give a perfect view of reality. The sensor may have a bad contact due to fouling, clogging or lack of maintenance. The measurement can be influenced by external factors: humidity, temperature extremes or electromagnetic fields. The calibration of the measuring instrument may also give rise to problems. Experience has demonstrated the need to verify measurements in order to be able to distinguish between the different reasons for an anomaly: brief and unexpected though real fluctuations, systematic or progressive error in a sensor or progressive evolution of the parameter being measured. A method to critically analyse these data has been developed (Moatar, 1997). The method combines modelling and statistical evaluation. The modelling facilitates the estimation of the pH parameter values and the statistical decision tests allow the verification of the coherence of the measurements to detect inherent errors.

In this paper, the modelling of pH using neural networks and details on how to use this technique for the critical analysis of data are presented. In water, the pH is affected by the water's chemistry, particularly the concentration of some of the CO_2-system components (CO_2, $H_2CO_3^-$ and CO_3^{2-}) according to the equilibria reactions (Stumm and Morgan, 1981). The concentration of CO_2 is a function of the CO_2 pressure equilibrium with the atmosphere, as well as photosynthesis, respiration of plants and the degradation of organic matter. Under acidic conditions, where water chemistry is predominant, the pH is directly related to the flow. Several authors have modelled this relation after linearisation using regression or Box and Jenkins (1976) transfer functions (Whitehead et al., 1986; Fisher et al., 1988; Hirst, 1992). Under alkaline conditions, the CO_2 concentration which affects the pH is principally related to photosynthesis. Photosynthesis is driven predominantly by solar radiation, nutrients, temperature and algal biomass. In the eutrophic Slapy reservoir (Nesmerak and Straskraba, 1985), methods of time series analysis (Box and Jenkins, 1976) have been used to identify relationships between automatic measurements of major driving (i.e. input) variables and changes of pH as an expression of photosynthesis. These analyses have suggested that daily changes of pH are closely related to changes in solar radiation and water temperature.

The site selected for this study is the Dampierre power plant, which is located in the Middle Loire River (Fig. 1). At this location, the stream is typically eutrophic (the amount of chlorophyll-A being up to 150–250 mg/m^3) during summer low-flows. The high level of phytoplankton photosynthetic activity (>0.6 mg C/h during summer) controls the physical–chemical characteristics of the water body at this period, notably the pH. Compared with lake and reservoir studies, the strong variation in the hydrological regimes throughout the year makes river discharge a predominant parameter in determining algal biomass (Recknagel et al., 1997) and other physical and chemical variables, including pH. This was illustrated for the Dampierre site by the Principal Component Analysis run on 104 data series over 13 years (Lair and Sargos, 1993). For this site, the pH can be considered as a function of the flow and the variables characteristic of photosynthetic activity which are themselves related to the hydraulic regime and energy exchanges between the water body and the atmosphere. The purpose of the model is to quickly furnish probable pH values to validate the measured values. The calculation is based on reliable variables which are measured on a daily basis. For this reason we excluded from our model algal biomass, nutrients and carbonates which are not reliable measurements and are only measured one or twice monthly. Only the discharge and solar radiation data were used in the model. The water temperature is measured by the same monitoring system as the pH. We choose to use only those parameters which are measured independently. We did, however, test the sensitivity of the influence of temperature on the model.

A preliminary study of the daily pH-daily discharge relationship at the Dampierre station suggested that it has a non-linear and complex shape (Fig. 2). By segmenting the data after solar radiation ($S(t) < 200$ W/m^2 and $S(t) > 200$ W/m^2) we can improve the correlation of the relationship between the daily discharge and the daily pH. Moreover, the data series are nonstationary, i.e. the basic statistical characteristics such as mean and standard deviation of the process change with the time. The interannual mean and standard deviation of pH present a complex periodic behaviour (Moatar, 1997). The standard deviation of pH displays a strong annual variability not directly related to the absolute level of the pH. Transformations of the pH data usually used for modelling water resources time series (Box and Jenkins, 1976; Salas et al., 1980) do not induce complete stationarity. Discharge data series do not follow a normal distribution. In this case the Box–Cox transform (Box and Cox, 1964) is usually used to obtain normally distributed data (Lemke, 1991). For instance, the linear time series models such as ARMAX (auto-regressive moving average with exogenous inputs) models developed by Box and Jenkins (1976) are not applicable.

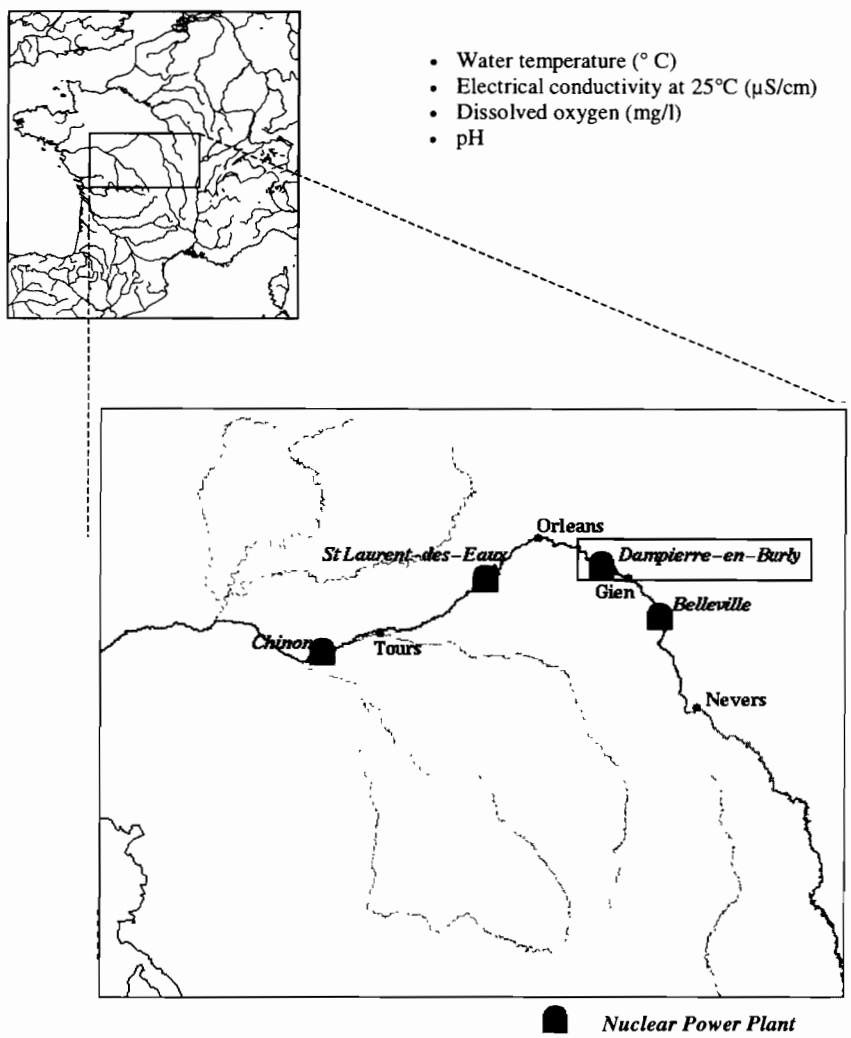

Fig. 1. Location and equipment of the Dampierre en Burly study site.

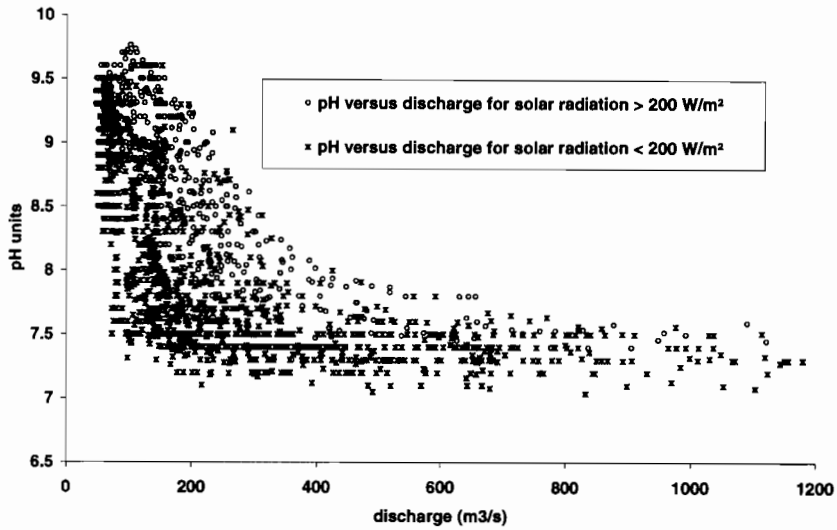

Fig. 2. Daily pH versus daily discharge.

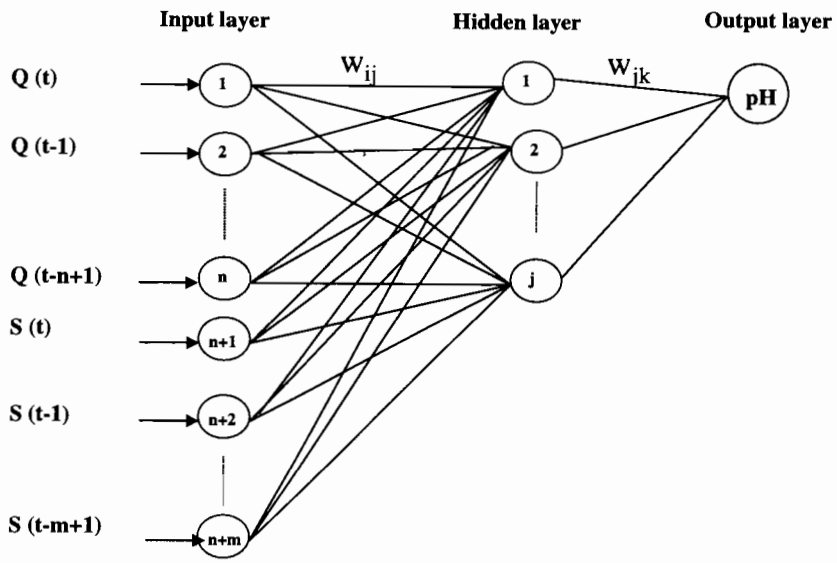

Fig. 3. Structure of the three-layer feed-forward artificial neural network used in this study.

However, when developing ANN models, the nonstationarities in the data are accounted for by the hidden layer nodes and the statistical distribution of the data does not need to be known (Maier and Dandy, 1996). Neural network models have been largely studied for the last 15 years. Although they first proceeded from physical, biological or psychological works about modelling, their use has broadly spread out to many different scientific areas. Neural networks are usually used as a particular type of non parametrical statistical model (Thiria et al., 1997). The most important and interesting characteristics shared by most neural networks models may be summarised as follows: non linear modelling capacity, generic modelling capacity, robustness to noisy data and ability to deal with high dimensional data. In the analysis of water resource phenomena, ANNs are

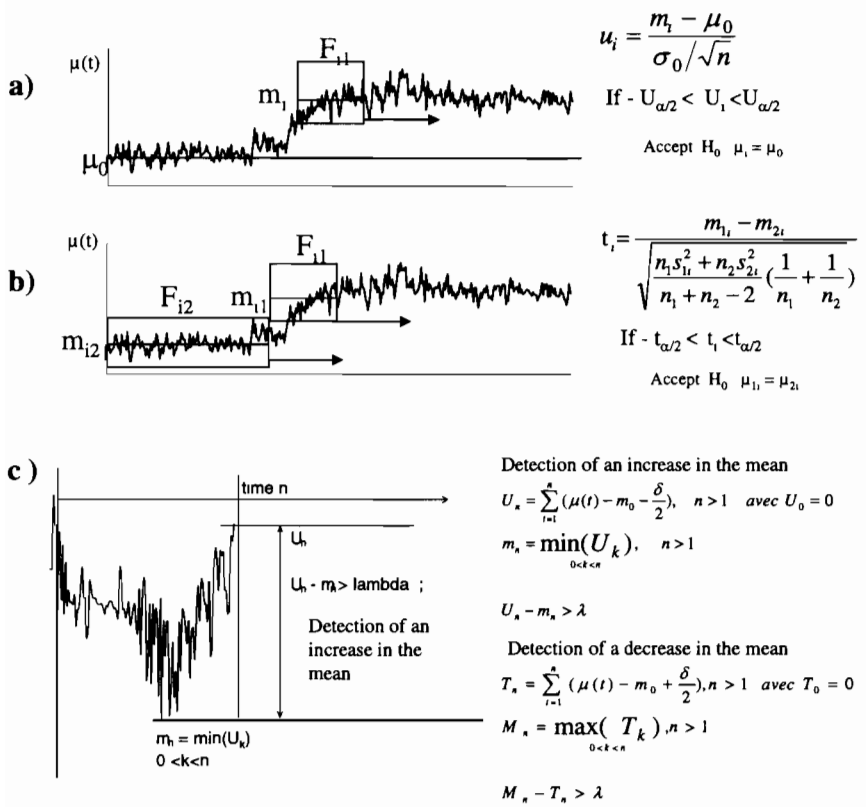

Fig. 4. Statistical tests.

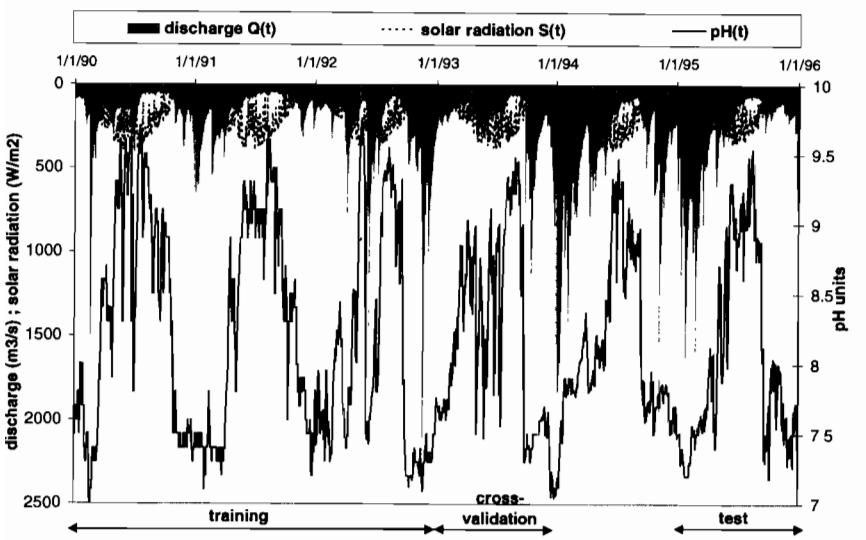

Fig. 5. Plot of the flow, solar radiation and pH time-series under study (1990–1995).

Table 1
Comparison of regression and ANN for single inputs variables

Single inputs variables	Regression		ANN	
	E criterion	S D.* (pH units)	*E* criterion	S.D.* (pH units)
$Q(t)$	0.45	0.50	0.72	0.34
Log $Q(t)$	0.69	0.39	0 71	0.34
$S(t)$	0.37	0.53	0.42	0 53
$T(t)$	0.33	0.54	0.26	0.55

* S.D. = standard deviation of residuals.

Table 2
Comparison of regression and ANN for multiple inputs variables

Multiple inputs variables	Regression		ANN	
	E criterion	S.D.* (pH units)	*E* criterion	S.D.* (pH units)
$Q(t)$ S (t)	0.60	0.41	0.77	0 30
$Q(t)$ $T(t)$	0.61	0.41	0.73	0.34
Log $Q(t)$ $S(t)$	0.73	0.31	0.73	0.33
Log $Q(t)$ $T(t)$	0.74	0.33	0.77	0.31
$Q(t)$ $S(t)$ $T(t)$	0.62	0.41	0.71	0 35
Log $Q(t)$ $S(t)$ $T(t)$	0.76	0.32	0.74	0.33

* S D. = standard deviation of residuals.

typically used to model the relation between rainfall and runoff (Dimopoulos et al., 1996; Minns and Hall, 1996). The ANN is shown to provide a better representation of the rainfall-runoff relationship than the linear ARMAX time series approach (Hsu et al., 1995; Lek et al., 1996b). In the domain of ecological modelling successful results have been obtained. For instance, Recknagel et al. (1997) studied the relationship between different species of algae and several limiting factors such as: solar radiation, nutrient concentrations, density and composition of zooplankton. Lek et al. (1996a) applied ANNs to modelling fish diversity with respect to riverine habitat characteristics.

2. The data base and the methods used in this study

2.1. Site and monitoring system description

The Loire river has a length of $\approx 1\,012$ km and a drainage area covering 115 000 km^2 of the centre and the west of France (Fig. 1). The Dampierre site, considered in this study, is situated 550 km from the source and drains 35 500 km^2 of watershed. It has the longest available record (1990–1995) of water quality parameters measurements. The monitoring system consists of a floating platform including a temperature sensor for direct measurement of water temperature (at a depth of 20 cm) in the river course and a pumping device sending a small flow of water (approx. 0.5 l/s) to the three following electrodes: pH (range: 0–14 pH unit; accuracy: $\pm 0.2\%$), Dissolved Oxygen (DO) (range: 0–20 mg/l; accuracy $\pm 1\%$) and electrical conductivity at 25°C (range: 0–1000 μS/cm; accuracy $\pm 1\%$). The pH accuracy given above is for instantaneous values and is that estimated by the manufacturer. However, the corresponding accuracy of pH (including electrode, transmission, and calibration) estimated *in situ* by comparison with laboratory measurements for the maintenance department of the Dampierre site are closer to ± 0.3 pH units. Accu-

racy is defined as two times the standard deviation (S.D.) of the check-sample readings. Furthermore, these instantaneous values, taken every 5 s are not archived as such, but as an hourly mean, which in fact is the average over 50 min (the remaining 10 min being used for the circuit cleaning cycle). In this study daily pH values were used. The hydrometeorological data used in connection with the pH data are the discharges at the Dampierre site (obtained from water-level records and a rating curve; range 46–2900 m^3/s; accuracy 8–10%) and daily solar radiation (W/m^2) measured at the meteorological station of the city of Tours located 185 km from the study site.

2.2. pH modelling by artificial neural networks

In this study, the neural network model used is the classical multilayer perceptron (MLP) with

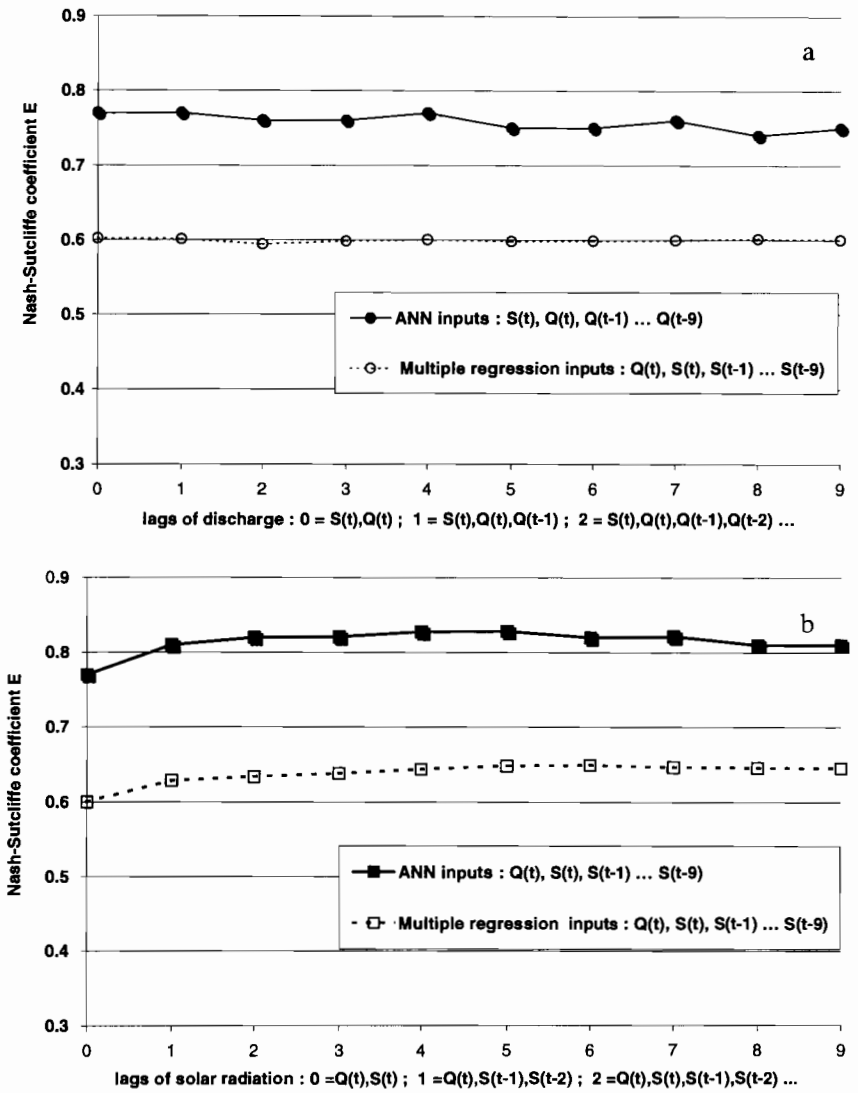

Fig. 6. Nash–Sutcliffe coefficient: (a) lags of discharge; (b) lags of solar radiation.

Table 3
Comparison of regression and ANN for multiple inputs variables: $Q(t)$ and $IS(t)$ for different values of the weighting parameter β

Multiple inputs variables	Regression		ANN	
	E criterion	S.D.* (pH units)	E criterion	S.D.* (pH units)
$Q(t)$ $IS(t)$; $\beta = 0.1$	0.61	0.41	0.79	0.29
$Q(t)$ IS (t); $\beta = 0.2$	0.61	0.41	0.79	0.28
$Q(t)$ IS (t); $\beta = 0.3$	0.62	0 40	0.80	0.28
$Q(t)$ IS (t); $\beta = 0.4$	0.63	0.40	0.81	0.28
$Q(t)$ IS (t); $\beta = 0.5$	0.63	0.39	0.82	0.27
$Q(t)$ IS (t); $\beta = 0.6$	0.64	0.39	0.82	0.27
$Q(t)$ IS (t); $\beta = 0.7$	0.65	0.38	0.83	0.26
$Q(t)$ IS (t); $\beta = 0.8$	0.64	0.38	0.79	0.29
$Q(t)$ IS (t); $\beta = 0.9$	0.63	0.39	0.79	0.29

* S.D. = standard deviation of residuals.

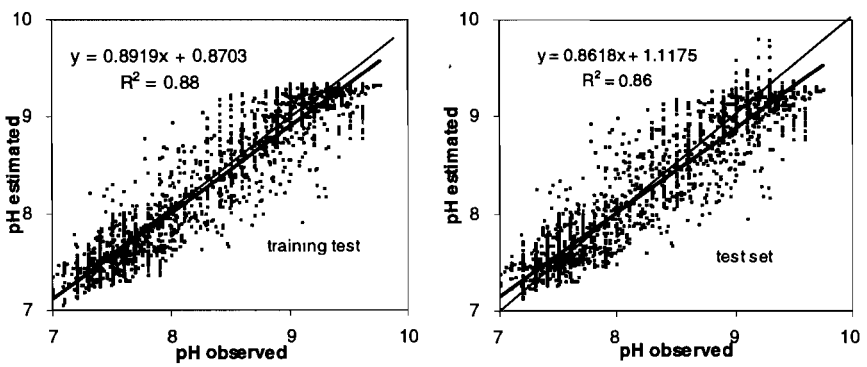

Fig. 7 Estimated versus observed pH values for the 5 years calibration period (left) and for 5 years verification (right).

one layer of hidden neurons (Fig. 3). It was developed using the commercially available software package Matlab-Neural Network Toolbox (The MathWorks Inc., 1998). The MLP consists of a large number of highly connected non-linear simple neurons. We can differentiate three types of neurons: input, output and hidden neurons. The input neurons receive information to be processed, in our case the discharge $Q(t)$ and solar radiation $S(t)$ (eventually incorporating also the discharge and solar radiation from previous days). The output neurons give the results of the neural network. In this case we have only one neuron which should return the result of the dependent variable pH (t). The hidden neurons which are neither input nor output neurons are used to keep an internal representation of the problem. The parameters associated with each of these connections are called weights. Knowledge of the network is kept in these weights. Each hidden and output unit computes its value as the weighted sum of its inputs, passed through a nonlinear function. For a given network architecture, the model calculates the weights that minimize a cost function (generally the mean square error function). Given a cost function, a network architecture and some data, the next step is to find the appropriate weights which minimize the cost function. This is usually done using an iterative procedure. The best known learning mechanism for neural networks is the backpropagation (BPA) rule of Rumelhart et al. (1986). It is a simple

gradient descent technique, which minimizes the cost function in weight space by modifying the weights in the opposite direction of the gradient error with respect to the weights. The BPA is often too slow for practical problems. Since 1986, a variety of improvements have been proposed (introduction of a momentum term, use of conjugate gradient techniques, use of second order information, etc.) (Hertz et al., 1991). We used the Levenberg–Marquardt algorithm, an alternative to the conjugate gradient techniques for fast optimization.

One of the most important features of learning systems is their ability to generalize to new situations. An early stopping procedure to stop the learning process was used for improving generalization. In this technique the available data were divided into three subsets. The first subset is the training subset which is used for computing the gradient and updating the network weights. The second subset is the validation set. The error on the validation set is monitored during the training process. The validation error will normally decrease during the initial phase of training, as does the training set error. However, when the network begins to overfit the data, the error on the validation set will typically begin to rise. When the validation error increases, the training is stopped, and the weights at the minimum of validation error are returned. The verification test subset is a set of independent data used to verify the consistency of the efficiency of the model.

The right number of hidden neurons cannot be achieved from a universal formula. Networks with too many parameters tend to memorize the

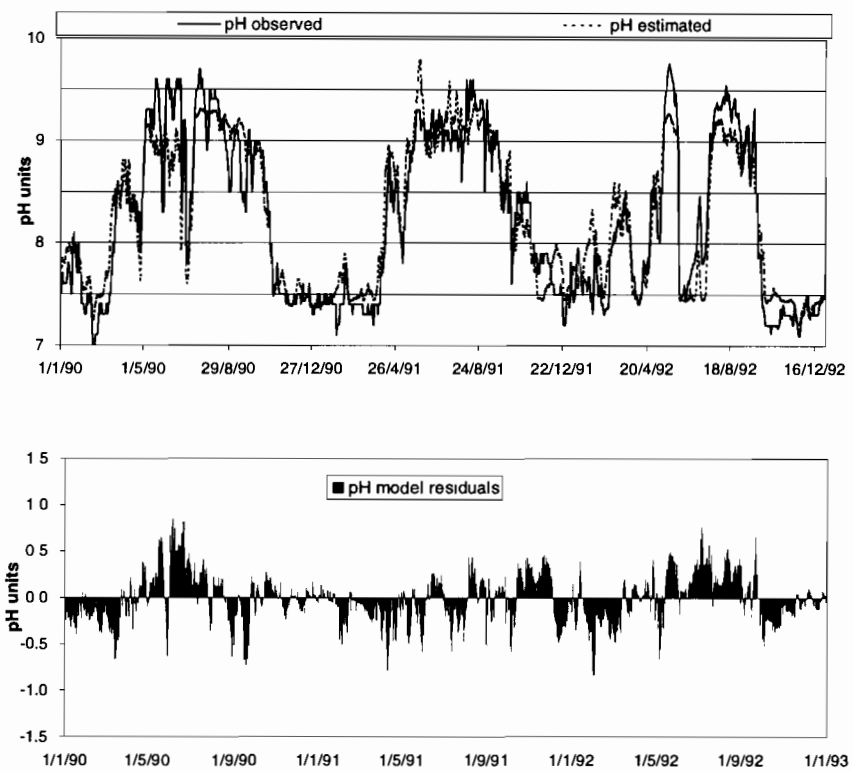

Fig. 8. Observed and estimated pH values for the period 1990–1992 inclusive (upper). Residual pH values for the same period (lower).

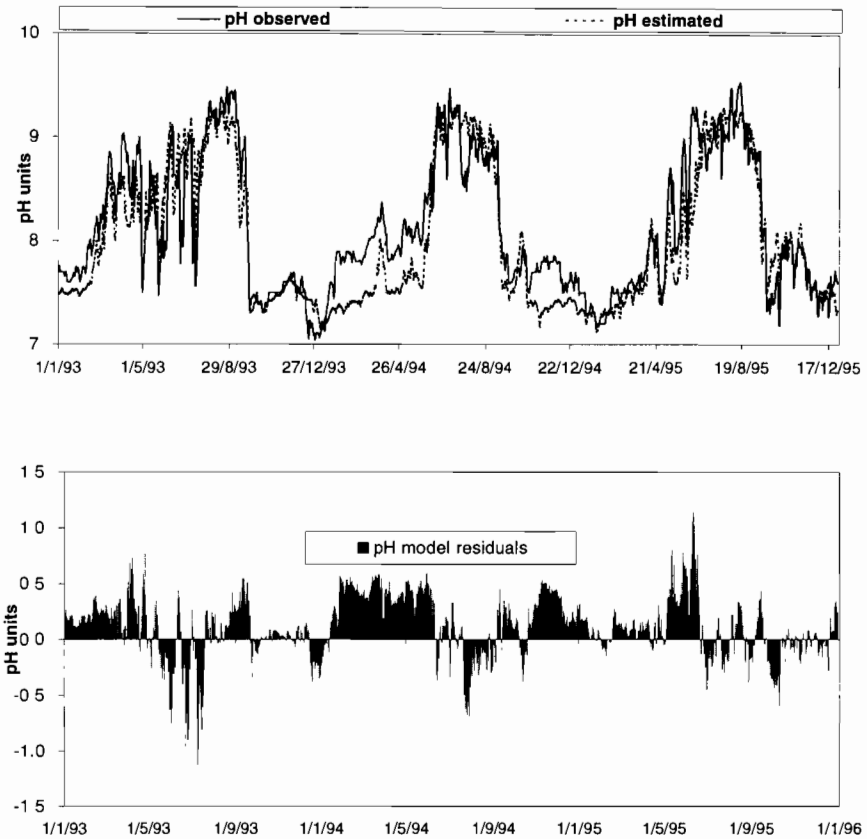

Fig 9 Observed and estimated pH values for the period 1993–1995 inclusive (upper). Residual pH values for the same period (lower).

Table 4
Statistical evaluation of estimated pH values (verification period)

Subset test year	Observed pH		Estimated pH		pH model residuals	
	Mean	S.D.*	Mean	S.D.	Mean	S.D
1990	8.35	0.78	8.36	0.67	−0.01	0.28
1991	8.27	0.74	8.33	0.73	−0.06	0.23
1992	8.04	0 77	8.04	0.65	0.00	0.27
1993	8.19	0.66	8.13	0.62	0.06	0.28
1994	8.11	0.55	7.92	0 69	0.20	0.27
1995	8.05	0.69	7.98	0.67	0.07	0.26

* S.D = standard deviation of residuals.

input patterns, while those with too few hidden parameters may not be able to simulate a complex system at all. We applied a trial-and-error approach to select the best ANN architecture. Our initial model had few parameters, we gradually added hidden neurons during learning until the optimal result is achieved in the test subset.

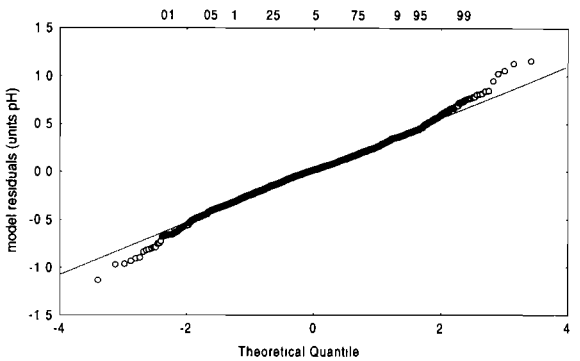

Fig. 10. Empirical distribution of residuals on Gauss paper.

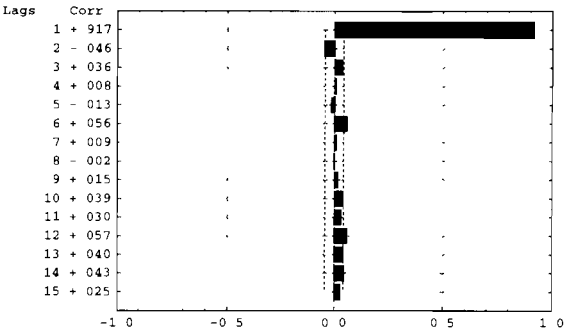

Fig. 11. Partial autocorrelation function of ANNs pH residuals.

2.3. Description of the pH data control method

A method for pH data control was developed after the pH model was built. The measured pH values were compared with those estimated by the ANN pH model using statistical tests in order to verify the homogeneity and the stationarity of the residual error series. These tests are performed for normal variables having independent observations. The series of residuals $\varepsilon(t)$ from the ANN pH model are normal (cf. Fig. 10) but have, in this case, a temporal structure (cf. Fig. 11). The modelling of this series using an autoregressive AR model allows the extraction of the independent residual series $\mu(t)$.

$$\varepsilon(t) = a_1\varepsilon(t-1) + a_2\varepsilon(t-2) + \cdots a_n\varepsilon(t-2) + \mu(t) \quad (1)$$

The order n of the AR model was estimated after analysis of the auto (ACF) and partial (PACF) autocorrelation functions. Finally, two types of statistical test are applied for automatic detection of changes in the mean of the signal $\mu(t)$:

1. The Student test comparing the mean of the values within a sliding window F_1, and either a reference mean μ_0 (Fig. 4(a)), or the mean of the values within an anterior sliding window F_2, (cf. Fig. 4(b)).
2. The Page–Hynkley test (Basseville, 1986) was performed as a cumulative sum test, where jumps in the mean occur at unknown time instants (Fig. 4(c).

The details of how the tests are applied are presented below:

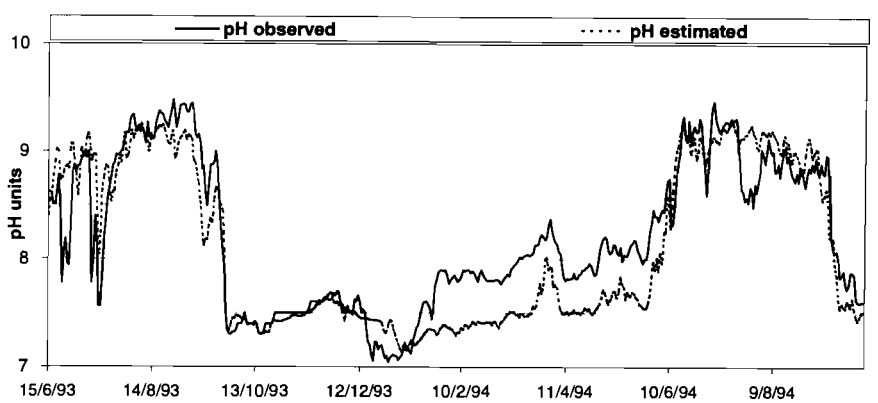

Fig. 12. Estimated and observed pH values for the period 15/06/–15/07/1993 inclusive.

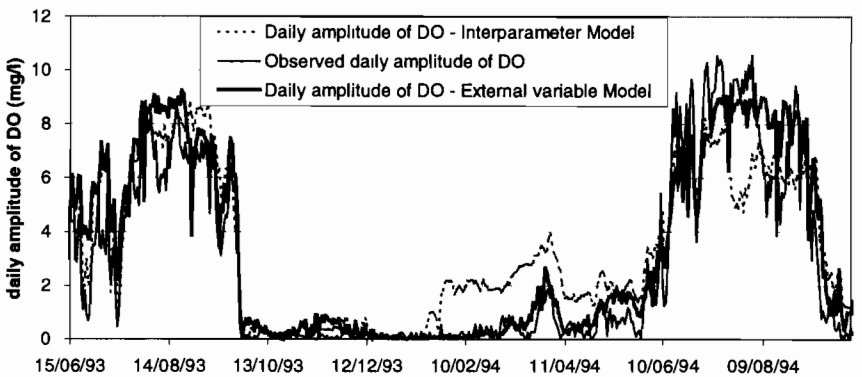

Fig. 13. Daily amplitude of DO measured and estimated with the two models: the 'Interparameter Model' and the 'External Variable Model' for the period 15/06/–15/07/1993 inclusive.

1. For each window, the statistical tests which must evolve according to known probability distribution laws (assuming the hypotheses that we are trying to prove are true) are carried out.
   1.1. To detect a change in the mean of the signal $\mu(t)$, we calculate the mean m_i within the current mobile window F_{1i} and the statistic test u_i The values μ_0 and σ_0 are the mean and S.D. calculated from independent samples known to be free of error measurements. The statistic test u_i follows, assuming no changes, a normal, centred and reduced law. The test is used to verify the hypothesis: $\mu_i = \mu_0$. If this hypothesis is confirmed, the difference between m_i and μ_0 is uniquely due to errors of estimation of the true mean population μ_i by the mean of the sample m_i
   1.2. If no reference values to test the calculated magnitudes are available, they will be calculated in two windows to allow comparison. The comparison of the mean of the two windows (m_{1i} and m_{2i}) of size n_1 and n_2, is performed for small samples ($n_1 < 30$ and/or $n_2 < 30$), sampled independently from a normal population from unknown variance but assumed to be equal to a common variance value ($\sigma_{1i}^2 = \sigma_{2i}^2 = \sigma^2$). If we assume that hypothesis H_0: $\mu_{1i} = \mu_{2i}$ is true, the statistic test follows a Student law with $n_1 + n_2 - 2$ degrees of freedom (d.f.).
2. The Page–Hinkley test (Fig. 4(c)) consists in fixing a priori a minimum jump magnitude δ to be detected, and running two tests in parallel, because the 'direction' of the jump is not known a priori (increasing or decreasing mean). The detector will set the alarm at the first time n at which $U_n - m_n > \lambda$ (cf. Eq. (2)) for detecting an increase in the mean and at the first time n at which $M_n - T_n > \lambda$ (cf. Eq. (3)), for detecting a decrease in the mean.

$$U_n = \sum_{i=1}^{n} \left(\mu(t) - m_0 - \frac{\delta}{2} \right); \quad n > 0 \text{ and } U_0 = 0$$

$$m_n = \min_{0 \le k \le n} (U_k) \tag{2}$$

$$T_n = \sum_{i=1}^{n} \left(\mu(t) - m_0 + \frac{\delta}{2} \right); \quad n > 0 \text{ and } T_0 = 0$$

$$M_n = \max_{0 < k < n} (T_k) \tag{3}$$

The limit λ is determined by learning. The initial value is calculated by the expression: $\lambda = 2*h*\sigma/\delta$ where $h = 2$ for normal distributions and σ is the standard deviation of the signal (Ragot et al., 1990).

3. Case study

3.1. Determination of appropriate ANNs model parameters

The daily pH, discharge and solar radiation values from the period 1990 to 1995 were used (cf. Section 1). For these series, data sets for the

network training, cross-validation and verification steps were prepared (Fig. 5). Data from 3 years were used for training, 1 year of data was used for cross-validation and 1 year of data was used for verification. The measurements for 1994 were not taken into account in the final calibration and the cross-validation process because of a lack of confidence in the measurements (as explained later). Each of the 5 years was chosen, one at a time, as the verification period, the other 4 years being used as the training and cross-validation data periods. The performance of the model was therefore verified using five different test samples.

For each input variable, the performance of the ANN was compared with the linear regression. The inputs variables tested was discharge $Q(t)$, natural logarithm of discharge Log $Q(t)$, solar radiation $S(t)$ and water temperature $T(t)$. Table 1 summarises the results in terms of the Nash and Sutcliffe (1970) efficiency criterion (E criterion) and the S.D. of the residuals in the verification subset. A plain improvement of regression ($E = 0.45$, S.D. $= 0.50$ for $Q(t)$ and $E = 0.69$, S.D. $= 0.36$ for Log $Q(t)$) is indicated for discharge by

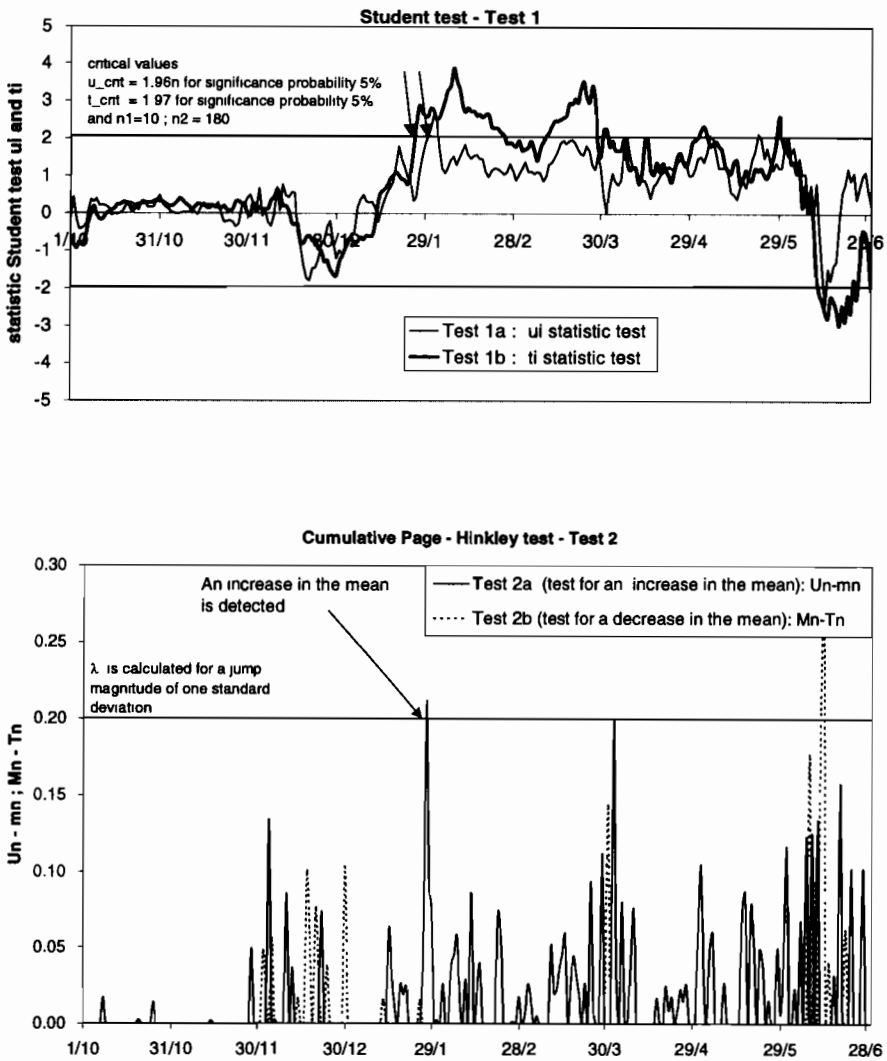

Fig. 14. Control charts: (a) Student's t test; (b) Page–Hinkley test.

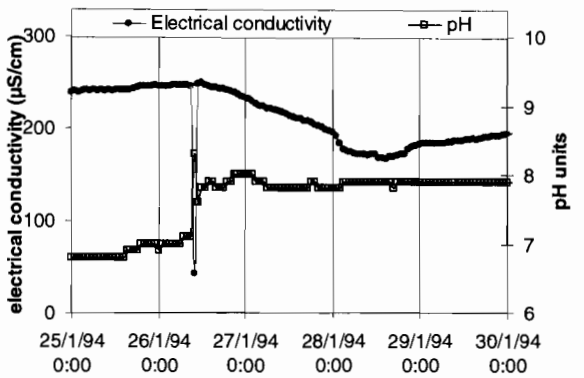

Fig. 15. pH and electrical conductivity (25/01–30/01/1994).

the ANN ($E = 0.72$, S.D. = 0.34), that confirms the nonlinear relationship between pH and discharge. In contrast, the relationship between solar radiation and pH appears to be linear, because no improvement of the regression model ($E = 0.37$, S.D. = 0.53) is obtained with an ANN model ($E = 0.42$, S.D. = 0.53). The same result is obtained if we enter as input the water temperature.

Table 2 presents the results for multiple inputs variable. Both the ANN and the regression model have a better estimation of the pH when the discharge and solar radiation and/or temperature are considered together. The best result for the multiple regression ($E = 0.76$, S.D. = 0.32) is obtained when Log Q (t), S (t), T (t) are taken into account. For the ANN, the results are similar for the diverse combination tested with a slight amelioration for the case Q (t) and S (t) ($E = 0.77$, S.D. = 0.30) or Log Q (t) and T (t) ($E = 0.77$, S.D. = 0.31). The influence of previous day's flows and previous day's radiation was then investigated. The efficiency E was calculated for the regression model and the ANN model (in the verification sets) as follows: initially the coefficient was calculated with the daily radiation S (t) and flows of N previous days Q ($t - N$), N varying from 0 to 9, as the input variables (Fig. 6(a)). It was then calculated for the case where the daily flow Q (t) and solar radiation of the N preceding days S ($t - N$) were used as the input variables (Fig. 6(b)).

Fig. 6(a) shows that in the case of the radiation, the preceding day's flows do not give better results as compared to the flow of the considered day. On the contrary, in the case of the flow, the preceding day's radiation does improve the pH estimation. Thus E, which was 0.77 with two input variables considered (flow and daily radiation), becomes 0.83 for the case of five input variables (flow and daily solar radiation and solar radiation at lag times 1, 2 . . 3, days: $t - 1$, $t - 2$, $t - 3$). To decrease the number of input variables, without losing the influence of the previous day's radiation, an exponential smoothing was applied. This variable has been called 'index of anterior radiation', IS, and has been calculated for a given day in the following manner:

$$IS(t) = \beta\, IS(t-1) + (1-\beta)\, S(t) \quad (4)$$

When applied recursively to each successive observation in the series, each new smoothed value is computed as the weighted average of the current observation S (t) and the previous smoothed observation IS ($t - 1$) depending on the value of the weighting parameter β. The optimal value of β in terms of the Nash–Sutcliffe coefficient, during both calibration and verification, was 0.7 (cf. Table 3). Finally the model has two inputs: Q (t) and IS (t).

We used the tan–sigmoid transfer function on the hidden layer and a linear transfer function on the output layer. In order to select the optimal number of hidden neurons, tests were performed by varying the number of neurons between 1 and 10. The optimal result of the test set is obtained for three neurons in the hidden layer, a choice that is justified by the absence of improvement of the model beyond this value. The data were standardised (zero mean and unity S.D.).

3.2. Results

3.2.1. pH modelling by ANNs

Finally the best model found has two inputs (Q (t) and IS (t)), three hidden neurons and one output for pH (t). The model fits the data well and explains 86% of the pH variance. The correlation coefficient is high in the calibration set ($R^2 = 0.88$) as well as in the verification set ($R^2 = 0.86$), indicating a high consistency of the model efficiency. (cf. Fig. 7). The time series of observed

and estimated values as well as the corresponding series of the residuals for the period between 1990–1992 inclusive and 1993–1995 are presented respectively in Figs. 8 and 9. The model conserves the same mean as the mean of the data. The S.D. of the estimated values is slightly smaller than for the observed values (Table 4). The mean error is zero for each year, except 1994, for which the values are underestimated. The S.D. of the errors vary between 0.23 and 0.28 pH units.

The normality and temporal structure of the residuals were analysed on the test set for 1990–1993 and 1995. Fig. 10 shows that the sample of model residuals is normal in the central part of the distribution and for more than 90% of the data. However, the partial autocorrelation function shows the existence of a temporal (i.e. persistence) structure (the autocorrelation function at lag 1 being equal to 0.9) (Fig. 11).

As shown in Fig. 12, during 5 months in 1994 (from mid-January to mid-June), the difference between the estimated and observed pH values is systematically in the order of 0.5 units. To explain this difference, another parameter measured by the monitoring system was analysed—daily amplitudes of dissolved oxygen, for which two estimation models are available. The first is a linear stochastic model, using the pH from the monitoring system (the 'Interparameter' model). The second model is based on physical principles and has variables which are not measured by the Electricité de France (EDF) monitoring system—river discharge and solar radiation. This model is called the 'External variable model'. As indicated in Fig. 13, from mid-January to mid-June, 1994, the 'External variable model' reproduces the daily amplitudes of dissolved oxygen (DO) quite well while the 'Interparameter model' systematically over-estimates them. This comparison indicates that the pH measurement for this time period is false (calibration error) or that it is significantly influenced by an external phenomenon (e.g. pollution) which cannot readily be explained. The method of critical data analysis applied to the pH and described in the following section shows that this time period is indeed suspect.

3.2.2. Control and validation of pH data

In this section, the results of the statistics tests of detection for the period 1/10/1993–30/6/1994 are presented. Using the method described in Section 2.3, the residuals $\varepsilon(t)$ from the pH model we initially decorrelated. After analysis of the autocorrelation function and the partial autocorrelation function, a first order autorregressive model was used.

$$\varepsilon(t) = 0.86\,\varepsilon(t-1) + \mu(t) \quad (5)$$

The calibration of this model as well as the calculation of the reference values (mean and S.D.), were performed for the 1991 which appears to have the most reliable measurements based on the critical analysis and validation of the other parameters. This year presents the best correlation between the modelled and measured pH values. Fig. 14 presents the test variables calculated for the series $\mu(t)$:

- u_1 for Test 1a: the mean of the values in the sliding window containing ten values compared to the mean reference $m_0 = 0$.
- t_1 for Test 1b: the mean of the values in the sliding window containing ten values compared to the mean of the window from 6 anterior months.
- $U_n - m_n$ and $M_n - T_n$ for Test 2: cumulative sum of the values from 1/10/1993. This test is re-initialised after each detection.

It is observed that in each of the three tests, the first signal is detected between 27/01/1994 and 30/01/1994, after 4 months of error free measurements (Test 1a: 30/01; Test 1b: 28/01; Test 2b: 27/01). This period corresponds with the beginning of a pH series which was already considered suspect through using the 'Interparameter' model which calculates daily amplitudes of dissolved oxygen (DO) from the pH measurements. Analysing the raw pH data, a systematic difference of 0.5 units during 6 months (previously presented in Section 3.2.1) is noted. This difference corresponds with a discontinuity observed on the 26 /01 at 10:00 (Fig. 15). For the same time step, an 'abnormal' electrical conductivity value was measured. This analysis shows that such tests are capable of detecting a measurement error occurring over a short period of time (1–4 days).

4. Conclusion

The results presented in this paper indicate that ANN clearly give satisfactory responses in the modelling of pH as a function of hydrometeorological data such as discharge and solar radiation. The best network found ($R^2 = 0.86$) to simulate pH was one with two inputs and three hidden nodes. The inputs are daily discharge Q (t) and the IS (t), 'index of anterior radiation', i.e. calculated as an exponential smoothing of the daily radiation variable. The model, which was adopted for its generality and its simplicity, and also because of the availability and reliability of the significant input variables, was integrated into our system of modelling tools which facilitate the critical analyses and validation of physical–chemical measurements. This system of modelling tools is currently in the process of being put into service on-line by EDF to allow them to follow and critically evaluate water quality parameters with respect to hydrometeorological conditions.

References

Basseville, B., 1986. On line detection of jumps in mean. Lect. Notes Contr. Inf. Sci. 77, 12–26.

Box, G., Jenkins, G., 1976. Time Series Analysis; Forecasting and Control, Holden-Day, San Francisco.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc., Ser. B 26, 211–243.

Dimopoulos, I., Lek, S., Lauga, J., 1996. Modélisation de la relation pluie-débit par les réseaux connexionnistes et le filtre de Kalman. Hydrol. Sci. J. 41 (2), 179–193.

Fisher, F., Dickson, K., Rodgers, J., Anderson, K., Slocomb, J., 1988. A statistical approach to assess factors affecting water chemistry using monitoring data. Water Resour. Bull. 24 (5), 1017–1029.

Hertz, J., Krogh, A., Palmer, R.G., 1991. Introduction to the Theory of Neural Computation, Santa Fe Institute Studies in the Sciences of Complexity, Addison Wesley, Reading, MA, 327 pp.

Hirst, D., 1992. A new technique for the analysis of continuously monitored water-quality data. J. Hydrol. 134, 95–102.

Hsu, K-L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modelling of the rainfall-runoff process. Water Resour. Res. 31 (10), 2517–2530.

Lair, N., Sargos, D., 1993. A 10-year study at four sites of the middle course of the River Loire. I-Patterns of change in hydrological, physical and chemical variables in relation to algal biomass. Hydroécol. Appl. 5 (1), 1–27.

Lek, S., Delacoste, M., Baran, Ph., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996a. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Lek, S., Dimopoulos, I., Derraz, M., Ghachtoul, Y.E., 1996b. Rainfall-runoff modelling using artificial neural networks. Rev. Sci. l'Eau 3, 319–331.

Lemke, K., 1991. Transfer function of suspended sediment concentration. Water Resour. Res. 27 (3), 293–305.

Maier, H.R., Dandy, G.C., 1996. The use of artificial neural networks for the prediction of water quality parameters. Water Resour. Res. 32 (4), 1013–1022.

Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydro. Sci. 41 (3), 399–417.

Moatar, F., 1997. Modélisations statistiques et déterministes des paramètres physico-chimiques utilisés en surveillance des eaux des rivières: Application à la validation des séries de mesures en continu (Cas de la Loire Moyenne). Ph.D. thesis, INP Grenoble, 283 pp.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, 1: a discussion of principles. J. Hydrol. 10, 282–290.

Nesmerak, I., Straskraba, M., 1985. Spectral analysis of the automatically recorded data from Slapy Reservoir, Czechoslovakia. Int Revue Hydrobiol. 70 (1), 27–46.

Ragot, J., Darouach, M., Maquin, D., Bloch, G., 1990. Validation de Données et Diagnostic, Hermès, Paris, 593 pp.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial nerural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, 1, MIT Press.

Salas, J.D., Delleur, J.W., Yevjevich, V., 1980. Applied Modelling of Hydrologic Time Series, Water Resources Publications, Book Crafters Inc., 484 pp.

Stumm, W., Morgan, J., 1981. Aquatic Chemistry, Wiley Interscience, New York, 781 pp.

The MathWorks Inc., 1998. Neural Network Toolbox User's guide version 3, The MathWorks Inc., 296 pp.

Thiria, S, Lechevalier, Y., Gascuel, O., Canu, S., 1997. Statistique et Méthodes Neuronales, Dunod, Paris, 311 pp.

Whitehead, P.G., Neal, C., Seden-Perriton, S., Christophersen, N., 1986. A time series approach to modelling stream acidity. J. Hydrol. 85, 281–303.

ELSEVIER

Ecological Modelling 120 (1999) 157–165

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece)

Ioannis Dimopoulos [a,*], J. Chronopoulos [b], A. Chronopoulou-Sereli [a], Sovan Lek [c]

[a] *Laboratory of Physics and Agricultural Meteorology, Department of General Sciences, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece*

[b] *Laboratory of Floricultural and Landscape Architecture, Department of General Sciences, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece*

[c] *CESAC UMR 5576, CNRS-Univ. Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, France*

Abstract

The aim of the present work is to propose a model for the estimation of lead concentration in grasses using urban descriptors easily accessible and to study the specific effect of each descriptor on lead concentration. Six descriptors were considered: the density of vegetation, the vegetation height, wind velocity, height of building, distance of adjacent street, traffic volume. Lead concentrations were determined in one grass species, *Cynodon dactylon* (L.) Pers, (Bermuda grass), collected from 30 different locations in Athens city. The proposed model is a multilayer perceptron (MLP) trained by backpropagation. The predictive quality of the model was judged by two cross-validation methods. The generalization ability of the model is confirmed by a determination coefficient higher than 0.91. The study of the first partial derivatives of the output of the MLP with respect to each input is used to identify of the factors influencing the lead concentration and the mode of action of each factor. Results allow to classify the environmental descriptors by their decreasing influence on lead concentration: distance of adjacent street, traffic volume, density of vegetation, wind velocity, height of building and vegetation height.

Keywords: Urban pollution; Heavy metal; Modelling; Backpropagation; Multiple regression; Sensitivity analysis

1. Introduction

In the city of Athens the constant increase of the population over the last decades has resulted in high traffic volumes and consequently high automobile emissions. Compounded by the narrowness of the roads this has caused discomfort (due to environmental conditions) to the city residents. Consequently, the air, plants and the soil are contaminated by various contaminants such as lead (Pb) (Ndiokwere, 1984; Ho and Tai, 1988; Mielke, 1991; Francek, 1992).

* Corresponding author. Fax: +30-152-94233.
E-mail address: gphy2hrk@auadec.aua.gr (I. Dimopoulos)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00099-X

In a city environment the main sources of Pb pollution are car exhausts, fumes and tyre wear, if there are no smelting sites, heavy industry or other sources of Pb contamination nearby (Akhter and Madany, 1993). In addition to the automobile emissions, the high density of large buildings amplifies pollution of plants because dispersion of the pollutants over wider areas is prevented (Capannesi et al., 1988).

Regarding the dispersion of pollutants in parks, studies suggest that the pollution burden is greater in the peripheral than in the central zones of the open areas (Shao-Lian et al., 1989; Grodzinska et al., 1990). In a previous study, the authors (Chronopoulos et al., 1997) examined the impact of traffic conditions on the vegetation and soil of two major parks in Athens and concluded that the density and composition of the peripheral vegetation has a remarkable effect on the dispersion of Pb and Cd towards the inner sites of the parks.

The concentration of pollutants in the different parts of the plants is strongly dependent on the plant species. Plant species as well as the design patterns of parks can also affect the distribution of Pb concentration in plants. A limited number of plant species that tolerate and colonize environments polluted with heavy metals are selected and used in the composition of city parks and avenue median dividers. Several plant species were studied to evaluate Pb contamination in city environments. *Cynodon dactylon* (L.) Pers. is one of the most frequently studied plant species for this purpose (Ho and Tai, 1988; Sukkop, 1990).

In order to establish realistic simulation models of Pb deposition and accumulation by plant species several inter-dependent models of environmental processes have to be linked together. Direct measurements of deposition rates using micrometeorological methods have advanced the knowledge of deposition processes. However, routine implementation of these methods for monitoring deposition rates is difficult and pollutant dispersion models for urban and industrial regions are only just beginning to be developed.

The purpose of our study is the evaluation of Pb levels in vegetation in an urban environment, using environmental parameters that are easily accessible and that strongly influence the diffusion of the pollutants which are mainly the result of high traffic (Preer, 1977; Wong, 1996). At the present study, we use and compare the predictive capacity of two statistical methods: Multiple Linear Regression (MLR) and Neural Networks (NN). Model-predicted and observed values are compared by different statistical parameters. For the NN model we propose a new simple method to study the relationship between the Pb concentrations estimated by the model and each influencing variable.

2. Materials and methods

2.1. Study area and environmental descriptors

The city centre of Athens is characterized by the presence of high densities of tall buildings and very infrequent sites covered by vegetation, such as parks. National Garden and Areos Park are the two major parks in the city centre they occupy relatively large areas of 15.8 and 24.0 hectares, respectively. These two parks are surrounded by avenues and streets, with different traffic volumes and an orientation that inhibits air circulation and dispersion of pollutants.

Squares of considerable size, covered with vegetation and able to provide comfortable environmental conditions for the citizens, are almost absent from the city of Athens. The great majority of city squares (approximately 92%) are less than 1.0 ha in size.

Samples were collected during the summer of 1995, from the plant species *Cynodon dactylon*, at 30 different locations (three public squares 1.0 ha in size, three public squares 5.0 ha in size, three public squares 10 ha in size, two parks and 14 traffic islands, Fig. 1). At each site three samples of *Cynodon* were bulked together to give a composite sample of about 5 g. A total of 140 plant samples were studied. *Cynodon dactylon* was selected for monitoring Pb contamination since it was found at all studied parks, squares and traffic islands. All plant samples were oven-dried at 70–80°C and ground to a fine powder by a microhammer mill to pass through a 1 mm mesh screen.

From each powder sample three subsamples of 1 g were weighed and metals were extracted by digestion with a 2:1 $HClO_4/HNO_3$ solution. Then the samples were filtered and diluted with deionised water to the final volume for Pb determination. Lead concentration was determined in the extracted solutions by atomic absorption spectrometry (GBC 908 FBT). The detection limit was 100 ppb for Pb with an accuracy of 1% RSD.

Every sample was described by a set of permanent descriptors (discrete and continuous).

- DENS: mean density of vegetation between the sample point and the nearest adjacent street (the values of DENS varied over the range 0–90%).
- GRAD: mean vegetation height between the sample point and the nearest adjacent street (the value of GRAD varied over the range 0–2 m).
- AIR: Wind velocity recordings were carried out with a digital measurement device at a network of 140 selected points. The measurement points were located at the plant sampling sites. The measurements were made at a height of 2.0 m above ground using a cap anemometer. A total of 38 measurement trips were conducted. After processing the data obtained, the average wind velocity was determined for the selected points. The reduction of the wind velocity for each measurement point compared with the maximum mean wind velocity was determined. Then a variable, AIR, was introduced to take into account the reduction of the wind velocity and the degree of ventilation at the measurement points. When reduction did not exceed 20%, ventilation was considered good (AIR = 3). At the points where the reduction varied between 20 and 40% ventilation was considered moderate (AIR = 2) and whenever the reduction exceeded 40% ventilation was considered poor (AIR = 1).
- BUILD: mean height of the adjacent buildings (the value of BUILD varied over the range 2–8 floors).
- DIST: distance between the sample point and the nearest adjacent street (the value of DIST varied from 0–66 m)
- TRAF: Traffic volume as expressed from the number of traffic lanes (the value of TRAF varied from 2–8 lines).

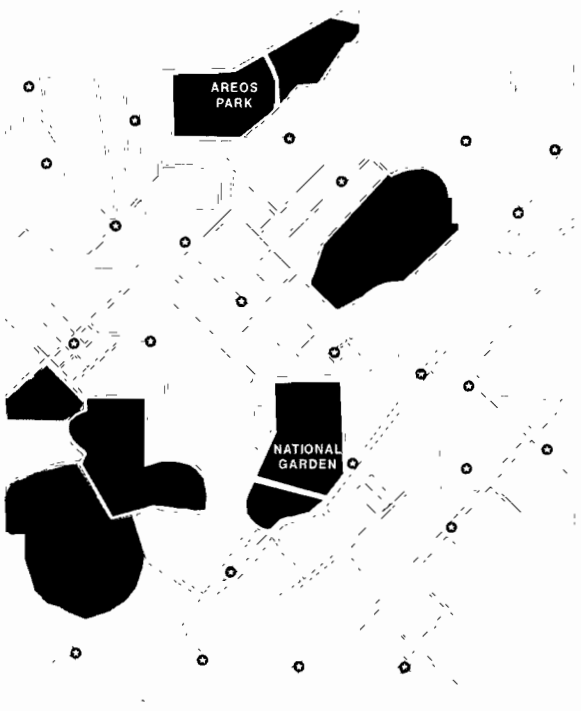

Fig. 1. Locations of measurement points in Athens city centre.

2.2. Modelling techniques

The techniques of multiple linear regression and stepwise multiple linear regression (Weisberg, 1980; Tomassone et al., 1983) were used. Calculations were done using SPSS software.

Multilayer Perceptrons (MLP), the most commonly used artificial neural networks, are general purpose, flexible, nonlinear models, $f: R^n \rightarrow R^m$, of the general form:

$$f(x) = \phi_n[W_n\phi_{n-1}[W_{n-1}\phi_{n-2}[\ldots\phi_1[W_1x]]]] \quad (1)$$

$$f^j(x) = \phi_i^L\left[\sum_{j=0}^{J_L-1} w_{ji}^L\phi_j^{L-1}\left[\sum_{u=0}^{J_{L-2}} w_{uj}^{L-1}\phi_u^{L-2}\left[\ldots\phi_l^1\left[\sum_{e=0}^{J_0} w_{el}^1 x_e\right]\right]\right]\right], \; j=1,\ldots,m \quad (2)$$

where W_i stands for the parameter matrix or weight matrix and ϕ_i stands for diagonal nonlinear operators; the elements of which are the so-called activation functions. MLP's with a nonlinear activation function are genuinely nonlinear and it has been proved (Cybenko, 1989) that, under some weak assumptions, any function can be approximated with an arbitrary accuracy by an MLP. Estimation of W is called training, learning or adaptation of the weights and regression via MLP is called supervised learning. The backpropagation algorithm is the most frequently used for training (Rumelhart et al., 1986).

A major problem in the use of MLP for model building is the determination of the optimal architecture of the network (number L of layers and J_j, $j=1...L$, where J_j is the number of node for layer j). Usually, the *trial-and-error* method is applied to test various alternative model architectures and choose the one with the optimal generalisation capability. Generalisation is defined by the ability of a model to predict data other than those on which it has been trained. A model with too many free parameters will fit the training data arbitrarily closely, but will not necessarily lead to optimal generalisation (overfitting).

Two classes of generalisation criteria are usually used for model architecture selection and model testing. The first class contains criteria based on the fitting errors (e.g. Akaike information criterion, Akaike, 1974). The second class of criteria is based on the principle of cross-validation (CV), according to which, the decisions on the model structure and predictive capacity are made on samples of data different than the sample used to estimate the parameters of the model. Usually overfitting is controlled by using a subset of the data, the validation set. This subset is not used for the computation of the weight matrix but for stopping the training process and taking decisions on the architecture parameters. The generalization ability is estimated by using another subset of the data, the test set, which neither participated in the weight estimation, nor in the architecture optimization, but only for the ultimate evaluation of the model. Separation of the data into the subsets is not straight-forward. Several questions arise concerning this method, they are discussed in Weigend et al. (1992).

One of the most efficient methods is k-fold cross-validation. The data set is divided into k approximately equal parts, and each part is used in turn as the test set for the network trained on the remainder, and the observed error rates on the k parts are averaged.

The error of a network, as a function of the weights that define it, is filled with hills and valleys. A trivial change in the training data can change the weights. Even with exactly the same training set, different random starting weights can result in dramatically different final results. Therefore, we do not dare assert that a network trained with all of the known data is essentially identical to networks trained with subsets of the data. To take into account this problem Moody and Utans (1991) propose a modification of the above cross-validation method: nonlinear k-fold cross-validation (NL K–f CV). In this work we use the two alternative kinds of CV: (1) CV with training, validation and test data sets and (2) NL K–f CV.

2.3. Preparation of data

The input data had very different orders of magnitude according to the variables. To standardize the scales of measurement, the values of the variables were converted by the relationship:

$$Z_s = \frac{X_o - \bar{X}}{\sigma_x} \tag{3}$$

with Z_s: standardized values, X_o: original values, $\bar{X}$ and σ_x the mean and standard deviation of the variable. The dependant variable Pb was also centred, reduced and converted over the interval [0...1] because the logistic function used for the NN output neuron modulates the response to values between 0 and 1.

2.4. Study of the influencing factors

In multiple linear regression, the influence of each variable can be roughly assessed by checking the final values of the regression coefficients. In mathematical terms, each coefficient of a linear model is the partial derivative of the response of the model with respect to the variable of that coefficient. The MLR partial coefficients therefore

generally give an indication of environmental reality, although it is not possible for this type of model to represent a nonlinear relationship such as that which probably exists between Pb levels and some influencing factors. On the other hand the neural network is a 'black box' type model and does not clarify the participation of each of the explanatory variables (descriptors). In this study we use a simple method based on the use of the partial derivatives of the network

response with respect to each descriptor. The link between the modification of inputs, x_j, and the variation of outputs, $y_j = f(x_j)$, is the Jacobian matrix $dy/dx^t = [\partial y/\partial x]_{mxn}$. It represents the sensitivity of the network outputs according to small input perturbations. For a network with n inputs, one hidden layer with ni nodes, and one output (i.e. $m = 1$), the gradient vector of y_j with respect to x_j is $d_j = [d_{j1},\ldots,d_{je},\ldots,d_{jn}]^T$ (Dimopoulos et al., 1995), with:

$$d_{je} = s_j \sum_{i=1}^{ni} w_{is} I_{ij} (1 - I_{ij}) w_{ei} \quad (4)$$

(under the assumption that a logistic sigmoid function is used for the activation. When s_j is the derivative of the output node with respect to its input, I_{ij} is the output of the ith hidden node for the input x_j, the scalars w_{is} and w_{ei} are the weights between the output node and the ith hidden node, and between the eth input node and the ith hidden node).

The sensitivity of the *MLP* output for the data set with respect to input x_e is:

$$SSD_e = \sum_{i=1}^{ni} (d_{je})^2 \quad (5)$$

and the derivative can be efficiently computed as a minor extension to the backpropagation algorithm used for training.

3. Results and discussion

3.1. Performance of the models

3.1.1. Multiple linear regression modelling

3.1.1.1. Complete model. With all the eight variables, the equation of the MLR model and determination coefficient became:

Pb =	−0.374DENS	+0.156GRAD	−0.033AIR	+0 097BUILD	−0.724DIST	+0 092TRAF
(*t*	−3.728	1.739	−0.617	1.646	−12.125	1.247)
(*Sig.*	0.000	0.024	0.538	0.102	0.000	0.214)

$R^2 = 0.703$ (6)

3.1.1.2. Stepwise model. Only three independent variables were retained by the model:

Pb =	−0.228DENS	+0.139 BUILD	−0.7 DIST
(*t*	−4.155	2.914	−12.650)
(Sig.	0.000	0.004	0.000)

$R^2 = 0.696$ (7)

The study of Fig. 2 shows several problems of the MLR model (Eq. (7)): an underestimation of the low values (Fig. 2a), the residuals (differences between observed and estimated values) tend to increase with estimated values (Fig. 2b). The residual distribution is far from normality (Fig. 2c).

3.1.2. Neural network

With the cross-validation approach, a good predictive model can be obtained using a network with three neurons in the hidden layer and sigmoid as activation function. In Table 1, the performance of the MLP model estimated by two CV methods is shown (MSE = Mean square error). The high value of the determination coefficient

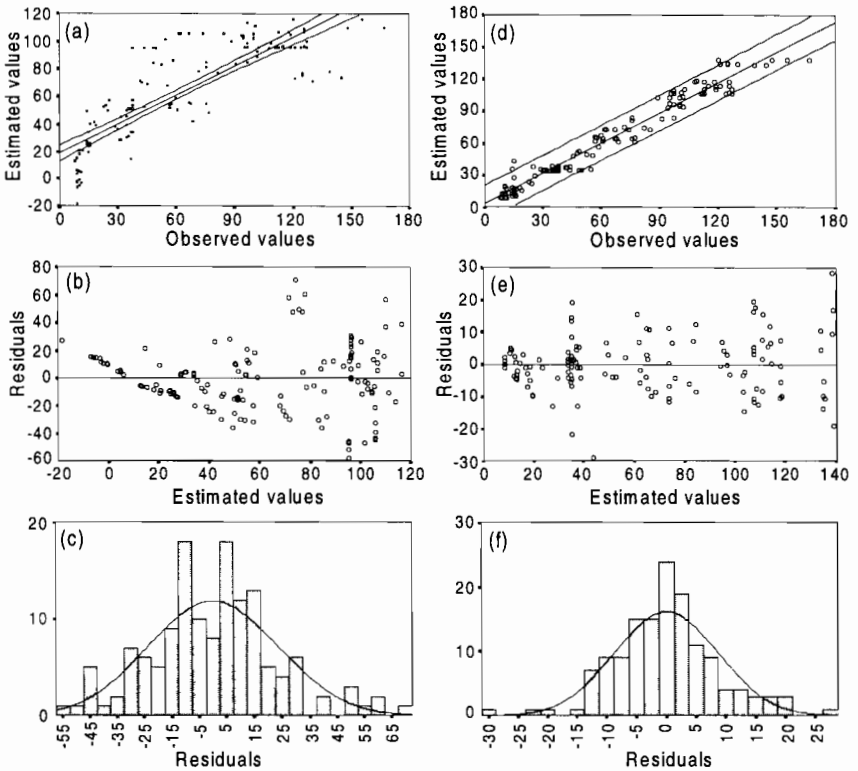

Fig. 2. Relationship between observed and estimated values of Pb; (a) MLR and (d). MLP Relationship between the residuals and the estimated values of Pb; (b) MLR and (e) MLP. Distribution of residuals (observed values-estimated values of Pb); (c) MLR and (f) MLP.

demonstrates the predictive capacity of the model (R^2 higher than 0.9). The fact that MLP provide a good predictive model was highlighted by the independence of the residuals from the variable to be predicted (Fig. 2e) and their normality (Fig. 2f). The distribution of residuals is better balanced with MLP than with MLR. Values that exceed the limits of the normal approximation are rather scarce.

Table 1
Mean square error (MSE) and determination coefficient R^2 for the NN model estimated by two alternative kinds of CV method

		MSE	R^2
CV	Training (80)	54.024	0.953
	Validation (30)	79.247	0 956
	Test (30)	107 779	0 938
NL 10-*f* CV	Training	50.37	0.972
	Test	88.547	0.911

3.2. Influence of factors

The study of MLR model (7) leads to the conclusion that the most significant factors affecting Pb diffusion are in decreasing order significance DIST, DENS and BUILD. Pb concentration decreased with DIST and DENS and increased with BUILD. The rest of the factors are either not very important or they are correlated to the three more significant factors.

The study of the MLP model, according to the method presented in Section 2.4, led to the layout of Fig. 3. Thus, for instance every point of DDENS versus DENS (Fig. 3a) resulted from Eq. (4) with $j = 1,\ldots,140$. Eq. (5) allows the variables to be classified according to their increasing influ-

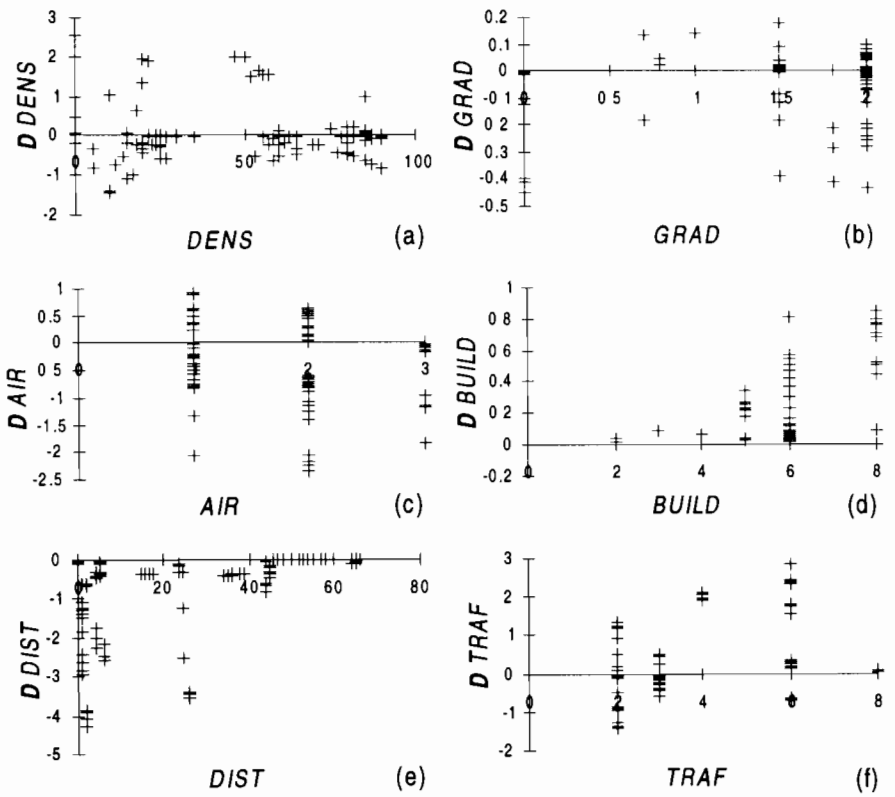

Fig. 3. Partial derivatives of the *NN* model reponse with respect to each descriptor

ence on Pb concentration: DIST (SSD_{DIST} = 339.95), TRAF (SSD_{TRAF} = 137.83), DENS (SSD_{DENS} = 100.29), AIR(SSD_{AIR} = 97.78), BUILD (SSD_{BUILD} = 12.99), GRAD (SSD_{GRAD} = 2.99). The study of Fig. 3 leads to the following remarks:

- The influence of density (DENS) on the Pb concentration is rather complicated and non-linear (Fig. 3a). The negative values of partial derivatives (DDENS) for the majority of the values of DENS show that the increase of the density contributes to the reduction of Pb concentrations.
- The influence of the height of vegetation on the reduction of Pb diffusion is shown in Fig. 3b. The negative values of partial derivatives (DGRAD) show that the height of vegetation contributes to the reduction of Pb concentration. This reduction increases with height. These results for the contribution of the density and the height of vegetation are in agreement with the remarks of Horbert et al. (1988) that plant density and structure provide an intensive decline in contamination in the central area of the parks.
- Concerning the factor AIR, two hypotheses may be made:
  1. 'Good' ventilation allows better Pb diffusion, reducing its high concentrations at points close to the emission source.
  2. 'Poor' ventilation does not facilitate Pb diffusion to distant points and can thus explain the reduction of concentrations in the centre of parks where ventilation is poor.
- The increase of the negative derivatives DAIR with AIR (Fig. 3c) shows that the first hypothesis is more positive. In a previous study (Chronopoulos et al., 1997), it has been pointed out that the dispersion of Pb depends significantly on the facility of the movement of

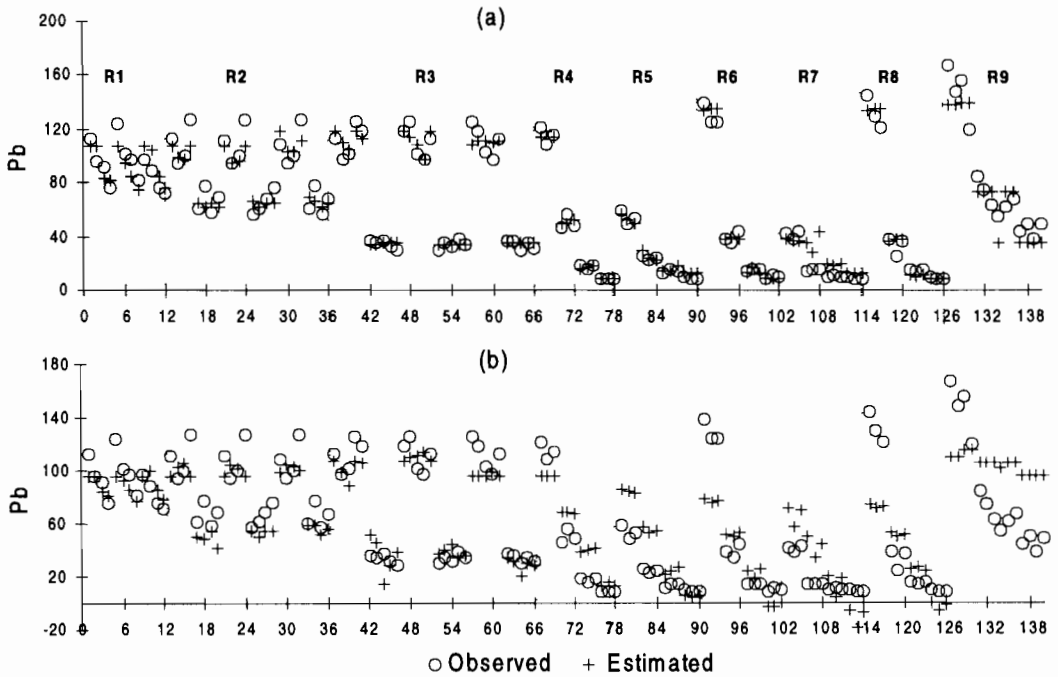

Fig. 4. Pb levels observed and estimated by the MLP model (a) and the MLR model (b) in the different regions of the study area: **R1**: three public squares (1.0 ha in size), **R2**: three public squares (5.0 ha in size), **R3**: three public squares (10.0 ha in size), **R4**. Areos Park (24.0 ha in size), **R5**: Areos Park-Mavromateon Street, **R6**: National Garden (24.0 ha in size)-Vas. Sofias Avenue, **R7**: National Garden-Irodou Attikou Street, **R8**: National Garden-Amalias, **R9**: traffic islands.

air masses, which prohibits or inhibits the dispersion of pollutants. The increase of the negative derivatives DAIR with AIR (Fig. 3c) shows that the first hypothesis is positive.

- The increase of the number of the floors of the adjacent buildings supports the increase of Pb concentrations (Fig. 3d).
- The decrease of Pb concentration with the increase of DIST is evident and nonlinear (Fig. 3e). The reduction of Pb is very intense near the emission points and becomes negligible when the distance is greater than 45 m. The MLR model without taking into consideration the traffic factor as expressed by TRAF is unable to properly estimate the Pb concentrations on traffic islands (Fig. 4b, R9). The slight reduction of the estimated values from the MLR model in that case is due to the fact that the values of the factor BUILD decrease at those points. The MLP, taking into account the factor TRAF gives much better estimations of Pb concentrations.
- The increase of the positive derivatives DTRAF with TRAF shows that Pb concentrations increase with traffic volume as expressed by the number of traffic lanes.

4. Conclusions

The basic idea behind the approach proposed here is the simulation of the system by a statistical model and the use of the resulting model to evaluate the contribution of each explanatory variable to the response of the explained variable. The comparison between the response of the model to the environmental variables on the one hand, and results from field observations on the other hand, shows similarities and indicates neural network modelling can be trusted. MLP adjusts the result of the estimations to the values actually measured. The result can be considered satisfactory since the model built up from a set of

'training data' can predict concentrations for another set of data obtained in the same geographic area.

The advantage of MLP over MLR models seems to arise from the ability of MLP to directly take into account any nonlinear relationships between the Pb concentrations and each explanatory factor. The approach proposed here can be extended to other applications in which non-linear relationships are observed.

References

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. on Automatic Control 19, 716–723.

Akhter, M.S., Madany, I.N., 1993 Heavy metals in street and house dust in Bahrain., Water. Air and Soil Pollution 66, 112–119

Capannesi, E., Caroli, S., Rosada, A., 1988. Evergreen oak leaves as natural monitor in environmental pollution. J Radioanal. Nucl. Chem. 123, 713–729

Chronopoulos, J., Haidouti, C., Chronopoulou-Sereli, A., Massas, I., 1997. Variations in plant and soil lead and cadmium content in urban parks in Athens, Greece Sci Total Environ. 196, 91–98.

Cybenko, G., 1989. Approximations by superpositions of a sigmoidal function, Math. Control. Signals and Syst. 2, 303–314.

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Process. Lett. 2 (6), 1–4.

Francek, M.A., 1992 Soil lead levels in a small town environment: a case study from Mt Pleasant, Michigan. Environ. Pollut. 76, 251–257.

Grodzinska, K., Szarek, G., Godzik, B., 1990. Heavy metal deposition in Polish National Parks -Changes during 10 years., Water. Air and Soil Pollution 49, 409–419.

Ho, Y.B., Tai, K.M., 1988 Elevated levels of lead and other metals in roadside soil and grass and their use to monitor aerial metal depositions in Hong Kong Environ. Pollution 49, 37–51.

Horbert, M., Kirchgeorg, A., Chronopoulou-Sereli, A., Chronopoulos, J., 1988. Impact of Green on the Urban Atmosphere in Athens, Scientific Series of the International Bureau KERNFORSCHUNGSANLAGE, 181 pp

Ndiokwere, C.L., 1984 A study of heavy metal pollution from motor vehicle emissions and its effect on roadside soil, vegetation and crops in Nigeria. Environ Pollution Ser. B 7, 247–254

Preer, J.R., 1977. Lead and cadmium content of urban garden vegetables, 11th Annual Conference on trace substances in environmental health, Columbia, June 7–9, pp. 179–187.

Rumelhart, D.E., Hinton, G.E, Williams, R.J., 1986. Learning representations by back-propagating error Nature 323, 533–536.

Shao-Lian, F., Hashimoto, H., Siegel, B.Z., Siegel, S.M., 1989. Period of significant source reduction., Water. Air and Soil Pollution 43, 109–118

Sukkop, H., 1990 Stadtokologie-Das Beispiel Berlin, chapter 3. Reimer-Verlag, Germany, p 425

Tomassone, R., Lesquoy, E., Miller, C., 1983 La regression, nouveaux regards sur une ancienne méthode statistique. INRA, Paris, p. 188.

Moody, J.E., Utans, J., 1991. Principled architecture selection for neural networks· Application to corporate bond rating prediction. In: Moody, J.E., Hanson, S J., Lippmann, R P. (Eds.), Advances in Neural Information Processing Systems 4. Morgan Kaufmann Publishers, San Mateo, CA, pp. 683–690.

Weigend, A.S., Huberman, B A., Rumelhart, D E., 1992. Predicting Sunspots and Exchange Rates with Connectionist Networks, In: M Casdagli and S Eubank (eds.), Nonlinear Modeling and Forecasting, Addison-Wesley, Redwood City, pp. 395–432.

Weisberg, S, 1980. Applied Linear Regression. Wiley, New York, p. 324

Wong, J.W.C., 1996. Heavy metal contents in vegetables and market garden soils in Hong Hong. Environ. Technol. 17 (4), 407–414.

ELSEVIER

Ecological Modelling 120 (1999) 167–179

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Support vector machines for optimal classification and spectral unmixing

Martin Brown *, Steve R. Gunn, Hugh G. Lewis

Image, Speech and Intelligent Systems research group, Department of Electronics and Computer Science, University of Southampton, Southampton, UK

Abstract

Mixture modelling is becoming an increasingly important tool in the remote sensing community as researchers attempt to resolve the sub-pixel, mixture information, which arises from the overlapping land cover types within the pixel's instantaneous field of view. This paper describes an approach based on a relatively new technique, support vector machines (SVMs), and contrasts this with more established algorithms such as linear spectral mixture models (LSMM) and artificial neural networks (ANN). In the simplest case, it is shown that the mixture regions formed by the linear support vector machine and the linear spectral mixture model are equivalent; however, the support vector machine automatically selects the relevant pure pixels. When non-linear algorithms are considered it can be shown that the non-linear support vector machines have model spaces which contain many of the conventional neural networks, multi-layer perceptrons and radial basis functions. However, the non-linear support vector machines automatically determine the relevant set of basis functions (nodes) from the performance constraints specified via the loss function and in doing so select only the data points which are important for making a decision. In practice, it has been found that only about 5% of the training exemplars are used to form the decision boundary region, which represents a considerable compression of the data and also means that validation effort can be concentrated on just those important data points.

Keywords: Spectral unmixing; Mixture modelling; Support vector machines

1. Introduction: mixed-pixel classification

The classification of remotely sensed data is an important use of earth observation satellite technology. In many cases, high classification accuracies are required to establish and regulate economic, social or environmental policy. Where remotely sensed data has been considered for future monitoring of the landscape, for example, it has been proposed that an acceptable accuracy limit for land cover maps derived from the classification of satellite data is 85% (Anderson et al., 1976).

Traditionally, pattern recognition has been regarded as a crisp classification process, where the algorithms are formulated for discrimination which is the practice of dividing up the feature

* Corresponding author. Present address: Unilever Research, Port Sunlight, Quarry Road, East Bebington, Wirral L63 3JW, UK. Fax: +44-151-6411825.

E-mail address: martin.q brown@unilever.com (M. Brown)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00100-3

space into a number of non-overlapping regions, and statistical pattern recognition which is the practice of modelling the posterior (or prior) distributions for a pre-defined number of discrete objects. In such problems, it is typically assumed that the process of generating observable data may be decomposed into a number of independent classes, each of which is a sub-process generating data according to a particular class conditional density for that class. When the class conditional densities do not overlap a discrimination approach is appropriate, but when they do overlap, a statistical pattern recognition approach, which models the posterior probabilities, $p(C_j|\mathbf{x})$, is more appropriate. This is generally necessary, as the chosen feature vector does not contain enough information to completely separate all the classes.

Implicit in the traditional classification process is the concept that each feature vector should be mapped into one of the classes of interest. In remote sensing however, this is unrealistic as the classes represented by a pixel's spectral features depend on the sensor's instantaneous field of view. Therefore, within each pixel multiple classes can occur, and only if every pixel is completely covered by a single class (a pure pixel) is conventional classification appropriate. However, the scale of many of the classes of interest is near pixel scale and so class mixture modelling is a fundamental part of obtaining maps from remotely sensed images. The spectral mixing can originate from overlap between different land cover classes, but also from clouds partially obscuring pure pixels. Hence, the process of mixed-pixel classification is to model the class mixing proportions (percentage ground cover area) (Horwitz et al., 1971), rather than to estimate the probability that a pixel's spectral response corresponds to a particular class label. This mixture cannot be resolved into either of the two classes, even when the input information is perfect (Foody and Cox, 1994), yet the estimate of the mixing proportions may be regarded as a statistical quantity, due to missing features and lack of training data.

This paper describes how support vector machines (SVMs) (Vapnik, 1995; Cherkassky and Mulier, 1998) can be used for mixture modelling using sets of exemplar pure pixels. SVMs are based on the concept of optimal discrimination, where the data should be correctly classified and the decision boundary should be as far away as possible from both classes. When a linear decision boundary is used, the algorithm can be shown to be equivalent to using a linear spectral mixture model (LSMM) which is subject to the sum to unity constraint. Despite the linear SVM algorithm being formulated as a discrimination technique, the two approaches are equivalent as the contour of the linear model is the discrimination boundary, so all that needs to be established is that the (unthresholded) gains of the two linear models are equivalent. In addition, the theory behind non-linear and non-separable SVM is described and this is compared with more conventional neural network approaches. An example illustrates how the SVM algorithms can be used to model the mixing problem for a remotely sensed, Landsat TM data set. To begin, the standard LSMM approach is briefly described.

2. Linear spectral mixture models

Spectral unmixing has been used as a technique for analysing the mixture of components in remotely sensed images for almost 30 years (Horwitz et al., 1971). The technique is based on the assumption that the class mixing is performed in a linear manner between so-called end-members (pure pixels). However, often little attention is given to the selection of these pure pixels, the corresponding mismatch between the model and true mixture, and the concepts that underlie the basic LSMM algorithm.

2.1. Algorithmic implementation

Assuming there exist m classes of interest and n spectral features which are used to model the class mixture, the user must specify an $(n+1)+m$ matrix $\mathbf{R}$ which contains the spectral values of the 'pure pixels', one from each class. In addition to the n spectral values within each feature vector, it is also assumed that a constant, bias value of $+1$

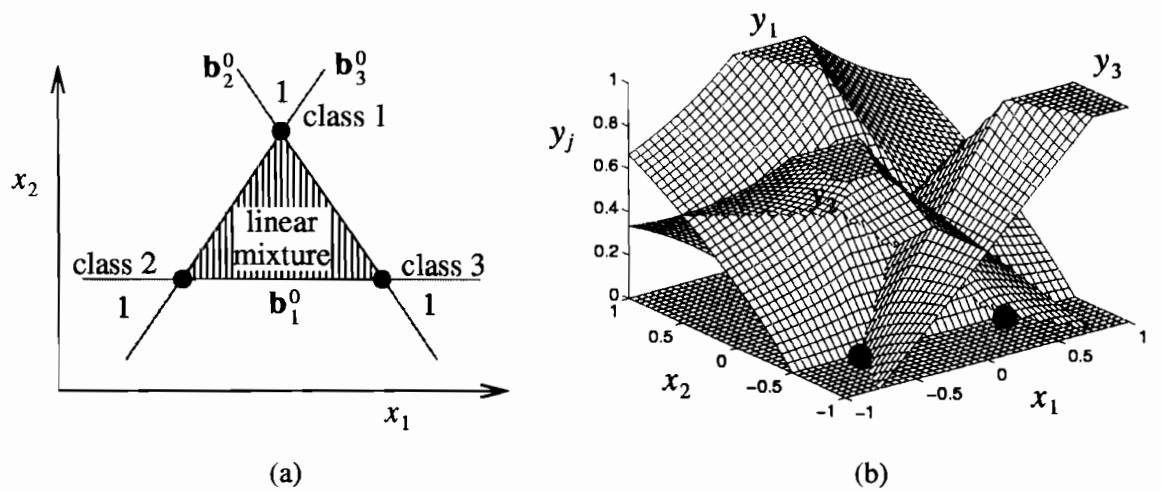

Fig. 1. A linear spectral mixture model for two inputs and three classes, in (a) two dimensions and (b) three dimensions, having linear margin boundaries, $\mathbf{b}_1^0$, $\mathbf{b}_2^0$, and $\mathbf{b}_3^0$, passing through the pure pixels occurring at (0, 0.5), (−0.5, −0.5), (0.5, −0.5), and mixing proportions y_j

has been appended to the start of each input. These pure pixel values are generally calculated to be the classes' means or else selected by the user or designer. The assumed linear model is of the form:

$$\mathbf{x} = \mathbf{R}\mathbf{y} + \varepsilon \tag{1}$$

where $\mathbf{R}$ is the matrix of pure pixels, $\mathbf{x}$ is the vector of spectral inputs, $\mathbf{y}$ is the vector of mixing proportions, and ε is a vector of errors. It is assumed that $\mathbf{R}$ is full-rank, i.e. all the pure pixels are linearly independent. Note also, that although this set of linear equations are expressed as an inverse model, the mixing proportions' estimates are usually calculated using a least-squares approach under the assumption that each class distribution can be represented by a Gaussian distribution with a common covariance matrix $\mathbf{V}$ but different means.

Letting $\mathbf{V}$ denote the covariance error matrix of the observations $\mathbf{x}$,

$$\mathbf{V} = E(\varepsilon\varepsilon^T) \tag{2}$$

then the goal of predicting the class mixtures can be formulated as minimising the weighted sum squared error:

$$J = (\hat{\mathbf{x}} - \mathbf{x})^T\mathbf{V}^{-1}(\hat{\mathbf{x}} - \mathbf{x}) \tag{3}$$

In addition, to ensure that estimated mixtures sum to unity, an additional linear constraint is introduced into the optimisation goal which states that the mixture estimates should sum to unity. This linear constraint can be combined with the quadratic error loss criteria to produce a closed-form constrained least squares (CLS) estimate (Settle and Drake, 1993) of the mixing proportions:

$$\mathbf{y} = \mathbf{C}^{-1}\mathbf{e} - \mathbf{C}^{-1}1(1^T\mathbf{C}^{-1}1)(1^T\mathbf{C}^{-1}\mathbf{e} - 1) \tag{4}$$

where $\mathbf{C} = \mathbf{R}^T\mathbf{V}^{-1}\mathbf{R}$ is the weighted auto-correlation matrix, $\mathbf{e} = \mathbf{R}^T\mathbf{V}^{-1}\mathbf{x}$ is the weighted cross correlation vector and 1 is a unity column vector. However, the partition of unity constraint in Eq. (4) does not ensure that the estimates lie in the unit interval, and in order to explicitly impose this on the solution, a common practice is to set $y_j = 0$ when $y_j < 0$ and then normalise the remaining estimates.

2.2. Interpretation

Using the CLS LSMM algorithm means that at most $n + 1$ classes can be linearly mixed, as illustrated in Fig. 1. When $m = n + 1$, each mix-

ture model has a linear margin boundary, $\mathbf{b}_j^{01}$ which passes through the remaining n pure pixels. This is uniquely determined and the least squares estimate reduces to $\mathbf{y} = \mathbf{R}^{-1}\mathbf{x}$. Within the convex region ($n+1$ dimensional simplex) enclosed by the linear margin boundaries, each model is linear and their memberships sum to unity.

When $m < n+1$, the linear margin boundaries are not uniquely determined by the remaining pure pixels, and this had led Settle and Drake to comment on the "embarrassment...among the infinity of possibilities" (Settle and Drake, 1993). In this situation, the position of the boundaries pass through the corresponding pure pixels and the orientations are determined by the properties of the covariance matrix (Brown et al., 1999).

In either case, however the selection of the pure pixel exemplars may prove to be problematic. This is especially so when there exists more than one pure pixel that could describe the class because many pure pixels could be used to separate (model the linear mixing) the classes, and expecting an expert or user to select a single one is unreasonable. In this case, end-members that are close to the classes' means are typically selected. However, these choices introduce an element of 'misclassification', as pure pixels at the edges of the classes' distributions will lie within the linear mixing margin.

3. Support vector machines

SVMs are a range of classification and regression algorithms that have been formulated from the principles of statistical learning theory developed by Vapnik (Vapnik, 1995). This theoretical framework develops a link between the empirical performance of a learning algorithm, when trained from a finite data sample, and the 'true' performance when used in practice. It has been shown that the rate of convergence of the empirical estimate to the true value is a function of the algorithm's VC-dimension. The VC-dimension of a model or classifier is, effectively, a measure of its flexibility and by minimising the model's flexibility as part of the learning process (structural risk minimisation) the risk of over-fitting the training set is reduced. SVMs therefore embody this structural risk minimisation process (Haykin, 1999).

In recent years, a number of non-linear classification and regression SVMs have been developed and these have been benchmarked against artificial neural networks (ANN). It has been found that the empirical performance of SVMs is generally as good as the best ANN solution (Hearst et al., 1998) and it has been hypothesised that this is because there are fewer model parameters to optimise in the SVM approach, reducing the possibility of over fitting the training data and thus increasing the actual performance. A major distinction between the two approaches is the training algorithm. Both SVMs and ANNs can be represented as two-layer networks (where the weights are non-linear in the first layer and linear in the second layer). However, while ANNs generally adapt all the parameters (using gradient or clustering-based approaches), SVMs choose the parameters for the first layer to be the training input vectors, because this minimises the VC-dimension (Cherkassky and Mulier, 1998). It is assumed that there are as many nodes in this layer as there are training points. A selection procedure is then used to calculate the weights in the second layer and this generally sets many of the weights to zero, which has the effect of dropping the corresponding training point from the overall calculation. Again, this selection procedure attempts to minimise the VC-dimension of the final solution. In the approaches discussed in this paper, SVMs can be considered to be sparse kernel methods.

3.1. Linear support vector machines

Consider a data set that contains two classes that are separable. For the pure pixels containing these classes, shown in Fig. 2, there are an infinite number of lines (hyperplanes) that will separate the data. The linear SVM is based on the principle

[1] A boundary $\mathbf{b}_j^{\theta}$ can be defined as $\mathbf{b}_j^{\theta} = \{\mathbf{x}: y_j(\mathbf{x}) = \theta\}$, i.e. it represents the input points for which the jth model's output is θ.

of selecting the one that maximises the minimum distance of the hyperplane from each class (the margin). This is because the VC dimension of a linear classifier is related both to the number of inputs and to the size of the calculated weights (Vapnik, 1995). Minimising the size of the weight vector produces a solution that maximally separates the classes and this is often known as the optimal separating hyperplane (OSH). As only the data points which lie on the class boundary closest to the hyperplane are involved in determining the minimum distance, effectively, all of the other data points in the training set are discarded from the calculation.

In classification approaches it is usually assumed that the labelled output lies in the unity interval, [0,1]. However, for the purposes of deriving the algorithms, it is more convenient to assume the output lies in the bi-polar interval $[-1,1]$. This can be done without loss of generality.

Maximising the margin and correctly classifying all the training data can be formulated as:

$$\min\Phi(\mathbf{w}) = 1/2\|\mathbf{w}\|_2^2 \tag{5}$$

subject to

$$(\mathbf{x}^t.\mathbf{w} + w_0)t^t \geq 1 \tag{6}$$

for the model:

$$y^t = \mathbf{w}^T.\mathbf{x}^t + w_0 \tag{7}$$

where $\{y^t\}_{t=1}^l$ are the output mixing proportions, $\mathbf{w} = (w_1,\ldots w_n)$ is the weight vector associated with the linear decision boundary, w_0 is the corresponding bias term and $\{\mathbf{x}^t, t^t\}_{t=1}^l$ is the labelled data set containing the spectral feature vector $\mathbf{x}^t$ and the target mixture proportion t^t for the tth data point. This can be formulated as a Lagrange functional producing a quadratic program (QP) with a global optimum which can be found using interior point methods (Burges, 1998; Gunn, 1998). The Lagrange multipliers effectively weight each data point according to its importance in determining the solution, and for the linearly separable two-class problem just described, only those data points that lie on the margin boundary have a non-zero Lagrange multiplier. There is just two data points for the example shown in Fig. 2. Hence, this often performs a considerable compression of the data set. Note that although the previous discussion has concentrated on a two class problem, it can be easily extended to m classes via a 1-of-m encoding of the data.

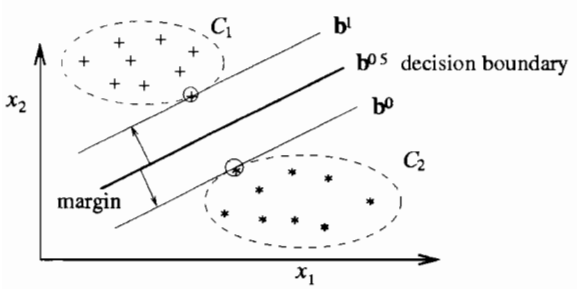

Fig. 2. A linear SVM for two inputs and two classes. The stars and crosses represent the labelled training data and the circles denote the selected support vectors which determine the linear margin's boundaries and the contours, $\mathbf{b}^{\prime\prime}$, for the second model are labelled.

It is interesting to compare this simple model with the one implicitly assumed to exist for the LSMM algorithm. In Fig. 3, there exists a convex, simplex (triangular) region where the mixture model is linear. It is also constant in the simplex (triangular) region 'behind' each prototype feature vector. Assuming that the LSMM is a true model of the mixture, that the training set contains vectors from the class' cores and that it also contains the three exemplar vectors, the result described in the next section can be established.

3.2. Linear SVM and LSMM

It can be shown that the Linear SVM and the LSMM algorithm are identical when the same information is used in their design. An outline proof of this is provided in this section and a full proof is contained in (Brown et al., 1999).

3.2.1. Theorem 1

The models formed by a linear (normalised) SVM and the LSMM are equivalent, when the same data set of pure pixels is used to train the models and the predicted mixtures are thresholded at 0 and 1.

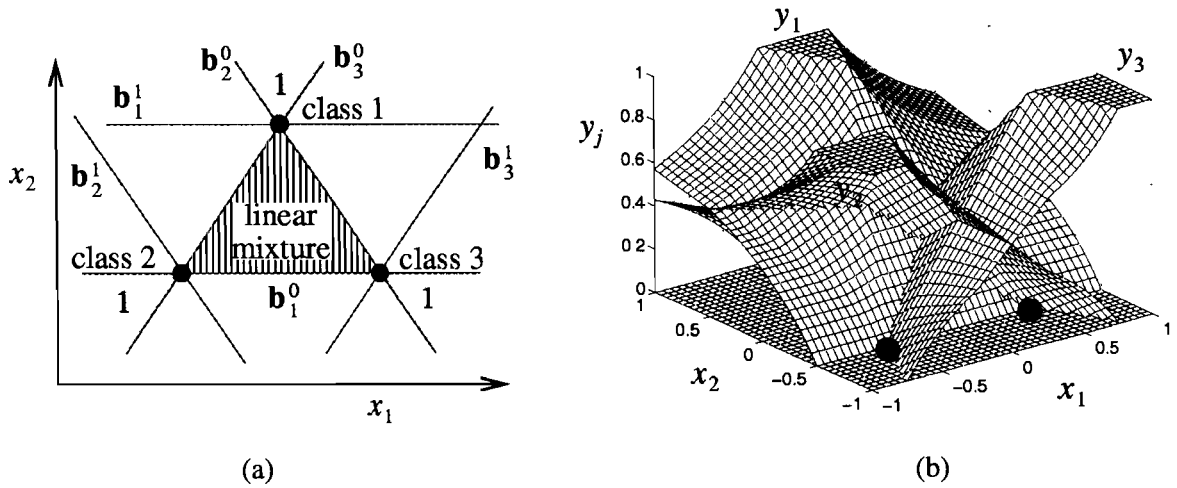

Fig. 3. Three linear SVMs and their margins which are denoted by the margins' boundaries $\mathbf{b}_j^0$ and $\mathbf{b}_j^1$ and mixing proportions y_j, for a two input problem shown in (a) two dimensions and (b) three dimensions

3.2.2. Proof

To begin with, the case when $m = n + 1$ is considered and then the problem when $m < n + 1$. When $m = n + 1$, the query point can lie either within the mixing margin of all the classes or it lies outside the margin of at least one class. When the data point lies within the margin of all the classes, all the unthresholded mixture estimates lie in the unit interval and both models are equivalent. This is easy to establish as both models contain $(n + 1)$ linear planes which pass through the same $(n + 1)$ data points, hence, the outputs must be identical. When the query point lies outside the margin of at least one class, this class estimate is thresholded at 0 or 1, and the resulting thresholded outputs are explicitly normalised to sum to unity. The motivation for thresholding the outputs of the LSMM are debatable, although it has been reported that thresholding at 0 and normalising the resulting values is a simple and quick way to ensure that the mixtures lie in the unit interval. Thresholding the output at 1 as well is natural when the mixture estimation problem is viewed as a margin maximisation process.

Now consider the case when $m < n + 1$. There are an infinite number of linear models that can fit the data set of pure pixels, and the sum to unity constraint of the CLS LSMM algorithm does not uniquely specify which one is preferable. This uniqueness is introduced when a least squares solution is used to formulate the data fitting process, even though there are fewer equations than degrees of freedom in the model, which means that the data can be interpolated exactly without the least squares solution. The least squares solution is given by choosing the weight vector that minimises its norm, where the weight vector includes the bias term (a subtle but important difference when comparing it to the SVM algorithm). The linear SVM algorithm can be slightly re-formulated for this particular data set as requiring that:

$$\min\Phi(\mathbf{w}) = 1/2\|\mathbf{w}\|_2^2 \tag{8}$$

subject to

$$(\mathbf{x}^t.\mathbf{w} + \mathbf{w}_0)t^t = 1 \tag{9}$$

Here the general inequality 'classification' constraints have been replaced by more specific equality constraints as it has been assumed that there exists only one data point for each class (the pure pixel). By exploiting well known results for quadratic programming problems with linear constraints it can be shown (Brown et al., 1999), that the formula for the weight vector (and bias term) which solves the above quadratic programming problem is equivalent to the CLS LSMM algorithm, hence the two techniques produce identical models.

The authors argue that the set of constraints given in Eq. (9) is a more fundamental way of expressing the data modelling problem, when compared to the constraints used to derive the CLS LSMM algorithm, even though they produce the same results. The data is interpolated rather than producing a least squares solution, and the mixing margin with the maximum size is produced. The properties are represented directly in the constraints, rather than being implicitly represented in the solution methodology.

The CLS LSMM algorithm also considers a noise matrix **V**, and this can be included in the linear SVM algorithm by minimising $\mathbf{w}^T\mathbf{V}\mathbf{w}$ which simply rotates the 'optimal' mixing margin. Therefore, given the same two data sets of pure pixels, where each class is represented by a single pure pixel, the two algorithms are identical. However, the linear SVM has the potential to automatically select pure pixels lying on the edge of a class core from a much larger data set. In addition, the following two sections briefly describe how non-linear and non-separable pure pixel data sets can be handled.

3.3. Linearly non-separable mixture modelling

Typically, remote-sensing data does not contain sufficient information to unambiguously assign each data point to a unique class. This situation of spectral confusion arises from the sampling of the ground surface in only a few wavebands by the satellite sensor, and occurs even if the pixel is pure. In these cases the class conditional densities overlap and the data is non-separable. The linear, separable quadratic programming problem can be modified by including a set of extra, non-negative variables $\xi_t \geq 0$ (Vapnik, 1995), which introduce a tolerance to misclassification:

$$t^t(\mathbf{w}.\mathbf{x}^t + w_0) \geq 1 - \xi_t \tag{10}$$

The modified optimisation problem becomes:

$$\min\Phi(\mathbf{w},\xi) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{t=1}^{l}\xi_t \tag{11}$$

where C is a given smoothness constraint. Therefore, the cost function is composed of a term which tries to maximise the size of the margin, $\|\mathbf{w}\|_2^2$, and a term which tries to minimise it by tolerating classification errors. The parameter C weights these two competing goals, and can be optimised using cross validation. Generally, when $C \to 0$, less emphasis is placed on the classification performance and the margin becomes wider. In contrast, when $C \to \infty$, more emphasis is placed on the classification performance and the margin becomes narrower. The task of optimising the smoothness constraint, C, in the SVM, therefore, is similar to the task of selecting the end-members
of classes for the LSMM: selecting end-members that are near to the classes' means produces a wide margin (low classification performance); whereas selecting end-members that are near to the edge of the classes' distributions produces a narrow margin (high classification performance).

3.4. Non-linear mixture modelling

Kernel-based mappings can also be used to construct more flexible, non-linear decision boundaries. In pursuing this approach, all of the previous analysis holds for forming the decision boundary, and the only change that needs to be made is to substitute a kernel function instead of the inner product between two training vectors. Common choices for kernel functions include:

Polynomial Kernel $K(\mathbf{x}^t,\mathbf{x}^j)$

$= (\mathbf{x}^t.\mathbf{x}^j + 1)^d$ where $d = 1, 2,\ldots$

Gaussian RBF $K(\mathbf{x}^t,\mathbf{x}^j) = \exp\left(\dfrac{-\|\mathbf{x}^t - \mathbf{x}^j\|_2^2}{2\sigma^2}\right)$

Ridge functions $K(\mathbf{x}^t,\mathbf{x}^j) = \tan h(b(\mathbf{x}^t.\mathbf{x}^j) - c)$

as well as various spline functions, where the mixing proportions are now given by:

$$y = \sum_{t=1}^{l}\alpha^t t^t K(\mathbf{x},\mathbf{x}^t) \tag{12}$$

Therefore, the non-linear kernels can be used to produce a wide range of decision boundaries, and the data selection procedure is equivalent to the selection of the kernel functions in the network,

i.e. only those kernels with a non-zero Lagrange multiplier, α^i, will contribute to the network's decision. It should be noted that many of the kernels that can be used within the SVM framework (any function can be used as long as it satisfies Mercer's conditions) are also widely used within artificial neural networks (ANNs), which have been widely applied within the remote sensing literature for statistical pattern recognition and area estimation. However, the 'training' procedure is very different. The weights associated with the hidden nodes in an ANN are generally trained (using a gradient method) to minimise the mean squared output error or, alternatively, some clustering-type approach is used so that they represent the mean of the local data cluster (Bishop, 1995; Ripley, 1996). Kernel methods centre the hidden nodes on unique data points and the SVM training procedure identifies those kernels (or data points) which directly influence the solution. As stated before, choosing the kernel points to be centred on the data points minimises the VC-dimension of the solution (Cherkassky and Mulier, 1998), and so reduces the risk of over-fitting the

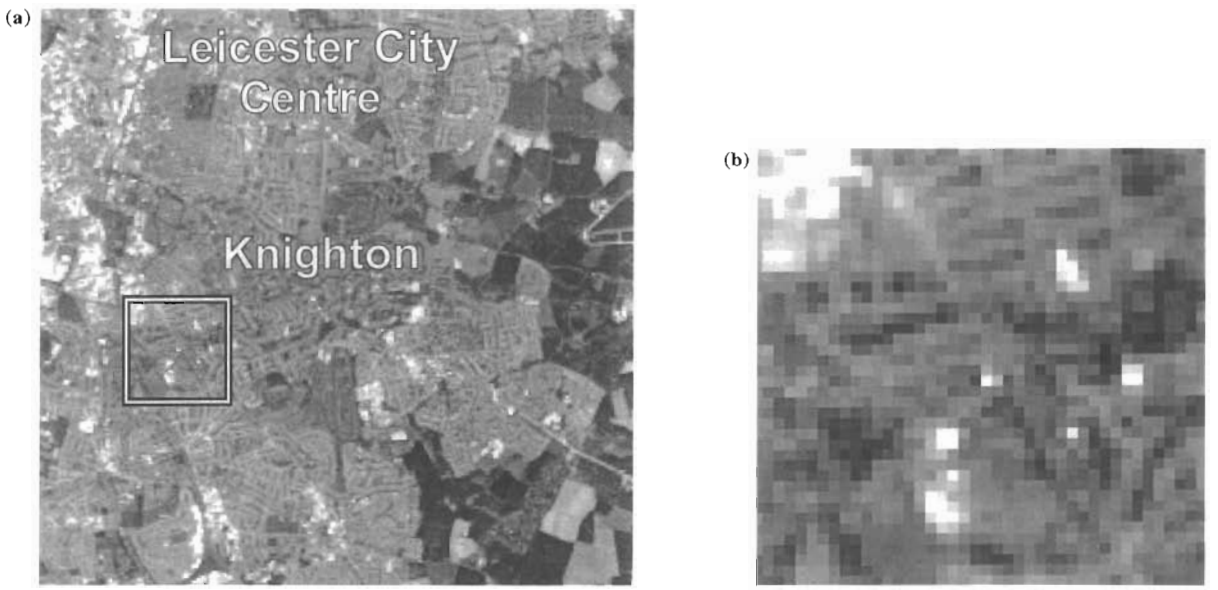

Fig. 4. Landsat TM image showing (a) the location of the data within the Knighton suburb of Leicester, UK and (b) an enlarged view of the dataset in the red channel (band 1).

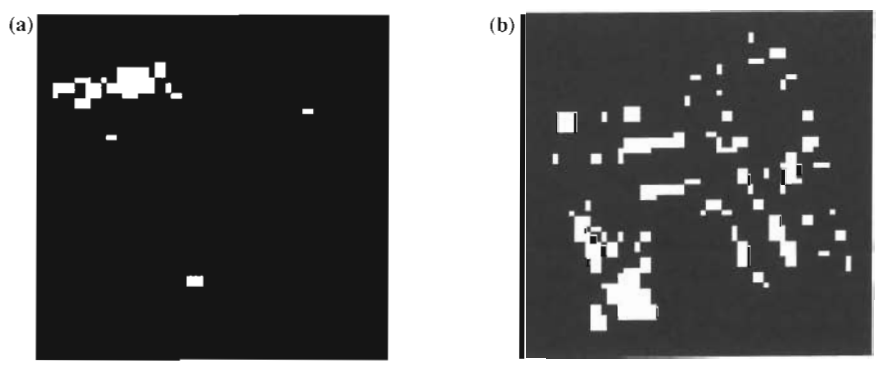

Fig. 5. Images showing the pure pixels for (a) the developed and other class and (b) the undeveloped and vegetation class.

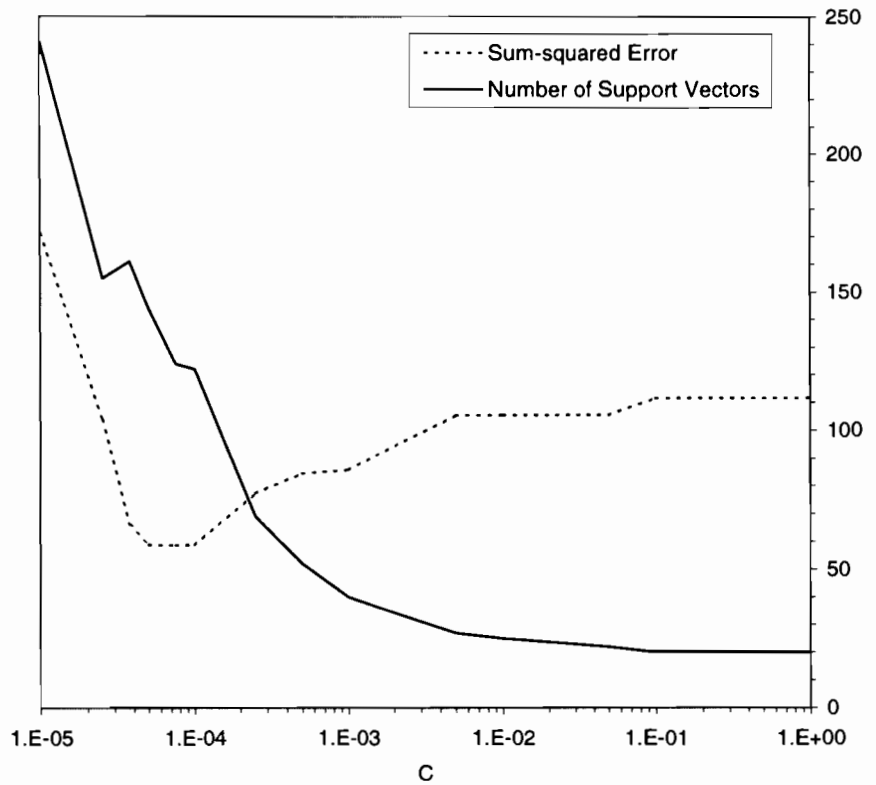

Fig. 6. Change in the number of support vectors and the sum-squared error of mixture estimates with the smoothness constraint C.

training data. In addition, it may be possible to use the knowledge about which data points greatly affect the solution in order to validate or re-design the network. Further, training is now a linear process (constrained quadratic programming problem) with a global minimum and there are fewer tuning parameters to set empirically. It is conjectured that these may be the reasons why the SVM appear to have good empirical generalisation properties.

Non-linear, polynomial transformations of the measured spectral signatures have been proposed within the LSMM framework (Bosdogianni et al., 1997), and the SVM kernels appear to provide a wider range. However, for all non-linear kernels, care must be taken not to overfit the data set and many more data are required to minimise the variance associated with non-linear modelling.

It should be noted that SVMs can also be used for modelling (regression) (Smola and Schölkopf, 1998), and this is particularly relevant for area estimation, as the majority of pixels in any training set will exhibit some degree of sub-pixel class variation. All of the techniques described above assume that only pure pixels are contained in the data set, and that the mixture region is implicitly specified by the boundary pixels. When the target data are continuous variables, SVM use an ε-insensitive region around the prediction in order to select a subset of the data. Points lying within the dead-zone are considered to be well represented and are ignored in the modelling process whereas points lying outside determine the shape of the surface. However, this is not considered further in this paper.

4. Application

In this section, the linear non-separable SVM algorithm is applied to part of a remote sensing

dataset (Lewis et al., 1998). This dataset was generated as part of the EU FLIERS Project. Fig. 4(a) shows a Landsat TM scene of Leicester, UK, from which the land cover within a one square-kilometre tile was identified. The land cover in this tile, Fig. 4(b), consisted of roads, sub-urban housing, garden and parks.

This land cover was grouped into two base classes:

- Developed and other (containing slate, tarmac, concrete, tennis court, etc.).
- Undeveloped and vegetation (containing sand, water, soil, grass, shrubs, etc.).

Hence, the composite classes have a broad core region which proves problematic for conventional LSMM algorithms. Pixels containing a proportion greater than 0.95 of these two classes were considered as pure pixels. This produced a reduced data set of 313 training pairs (76 from the developed and other class and 237 from the undeveloped and vegetation class) compared to the original 1000 data points which describe the mixture between these two classes on the tile (see Fig. 5).

This dataset was used to design three mixture models: a linear, non-separable SVM algorithm, a linear spectral mixture model, and a non-linear, non-separable SVM algorithm. The non-separability of the classes arises from the spectral ambiguity of several pure pixels that lie within the mixing margin. In order to establish the optimal value of the smoothness constraint, C, for this data set, many linear non-separable SVM algorithms were constructed from the 313 pure pixels for different values of C and their performances were calculated using the sum-of-squares error over the remaining $k = 794$ mixed pixels:

$$\mathrm{SSE} = \sum_{t=1}^{k} (y(\mathbf{x}^t) - t^t)^2 \tag{13}$$

A value of $C = 0.0001$ was found to produce a linear, non-separable SVM algorithm that predicted the mixture proportions with the lowest error, SSE = 58.80 (root mean square error =

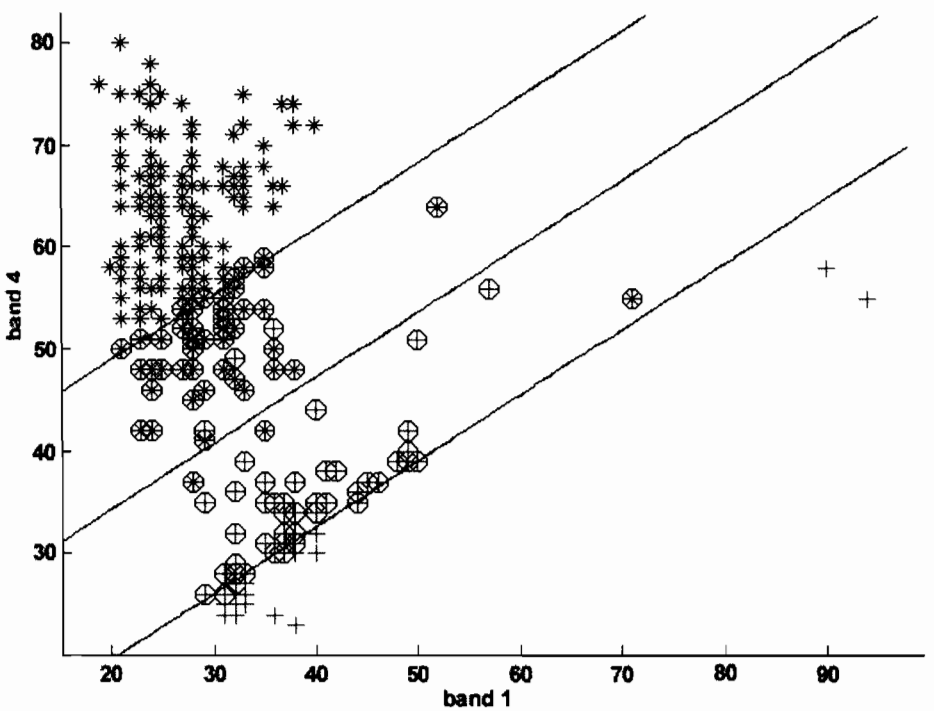

Fig. 7. Optimal linear SVM mixture predictions and selected support vectors for Landsat bands 1 and 4. Crosses are used to represent the developed and other class data and stars represent the undeveloped and vegetation class data. Data selected as support vectors are circled.

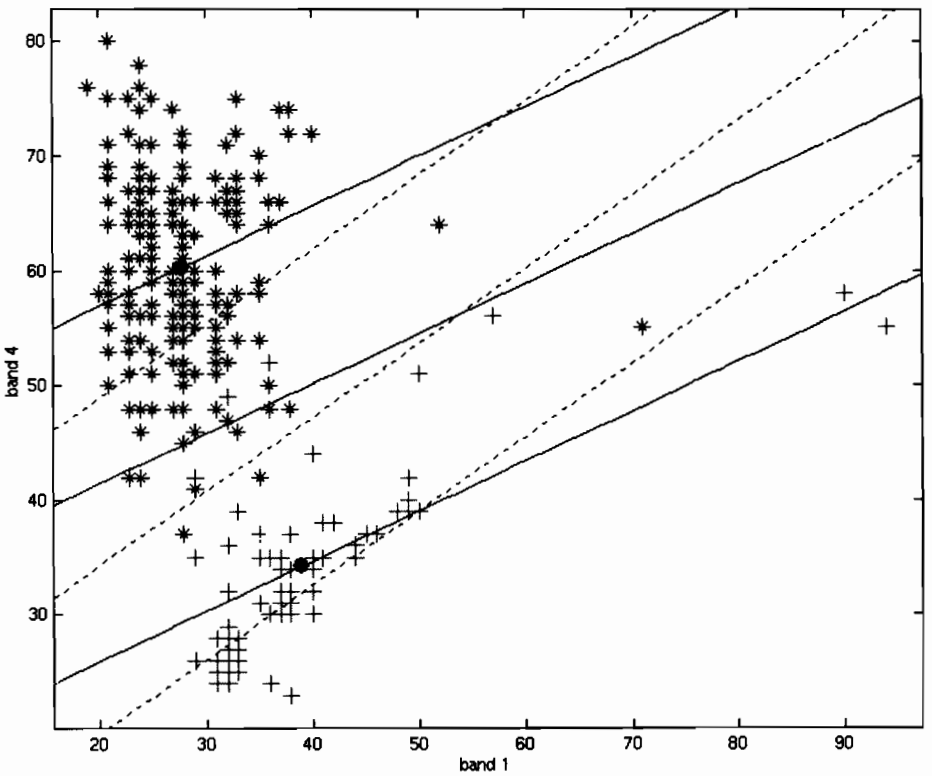

Fig. 8. LSMM mixture margin for Landsat bands 1 and 4. Crosses are used to represent the developed and other class data and stars represent the undeveloped and vegetation class data. The mean of the classes' distributions are shown as black circles and the optimal, linear SVM mixture margin is shown with dotted lines for comparison.

0.272 pixels) using 122 pure pixels as support vectors (39% of the training data). Fig. 6 summarises the optimisation results and Fig. 7 shows the margin boundaries and the support vectors (pure pixels) of this optimal linear, non-separable SVM. It can be seen that the support vectors lie within the mixing margin.

In this case, the classification problem cannot be treated solely as a discrimination problem (i.e. the class conditional distributions overlap); however, the calculated margin, which denotes the region of linear mixing, is sensible. Note that the a priori selection of a set of pure pixels for the classes cores would be extremely problematic. In the SVM example, this simply reduces to estimating a single bound parameter, C, which determines the width of the margin, as illustrated above. In contrast, in the LSMM case the mean of the classes' distributions is generally chosen as the 'pure pixel' spectra. For the data set described here, the LSMM solution predicted the mixture proportions of the mixed pixels with an error, SSE = 87.251 (root mean square error = 0.332 pixels). This LSMM mixing margin is shown in Fig. 8. Since the choice of the classes' means is the 'optimal' solution for the LSMM (assuming a Gaussian distribution of pure pixels), it can be seen from these results that the optimal, linear SVM solution produces better mixture predictions than the optimal LSMM solution on this data set.

For more complex problems, a linear margin may not be appropriate and the non-linear margins that are produced by the kernel-based SVMs could be more appropriate. This is expected to become more significant as the number of spectral bands and classes increases. The result of applying a quadratic (non-linear) kernel to the pure pixel dataset is also illustrated in Fig. 9. As can be seen

from the figure, the margin boundaries are no longer linear or parallel. However, the mixing margin is approximately linear where there exists data to support this. About 30 of the data points have been selected and even though the model appears to fit the data slightly better, there is little evidence in some regions with only a few data points for the extra flexibility introduced. This is a standard manifestation of the bias/variance dilemma, and when more flexible model spaces are used, there should be sufficient evidence in the data to support it.

5. Conclusions

It has been shown that LSMM are related to the basic, linear SVM and under certain circumstances, both algorithms are identical. This is an important observation for the LSMM algorithms, as they appear to possess the 'maximum margin' property. However, the observation is significant in that the SVM performs automatic pure pixel selection, and the associated non-linear techniques mean that non-linear mixture models can be formed from pure pixel, binary target values and continuous mixture data. The type of non-linear mixtures which can be formed have been demonstrated on some real, remote sensing data, although it still remains to quantify the performance of these algorithms.

Acknowledgements

The authors gratefully acknowledge the financial assistance supplied as part of the EU Framework IV FLIERS and the EPSRC GR/K55110 OSIRIS projects.

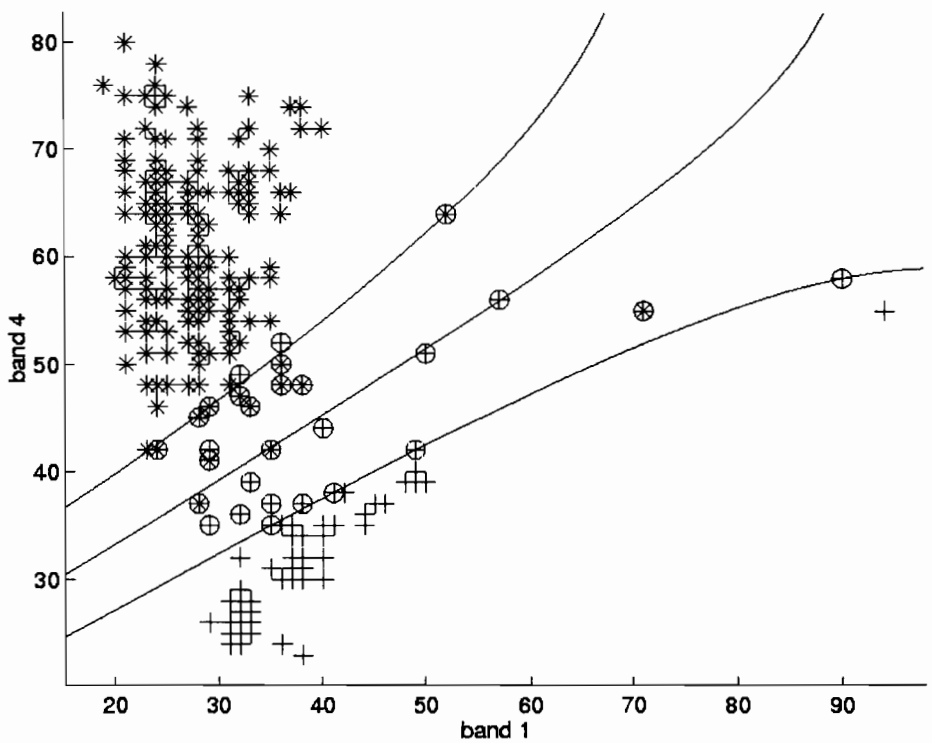

Fig. 9. Quadratic SVM mixture predictions and selected support vectors for Landsat bands 1 and 4 Crosses are used to represent the developed and other class data and stars represent the undeveloped and vegetation class data. Data selected as support vectors are circled

References

Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976 A land use and land cover classification system for use with remotely sensed data. US Geological Survey Professional Paper 964, 28 pp.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, 482 pp.

Bosdogianni, P., Petrou, M., Kittler, J., 1997. Mixture models with higher order moments. IEEE Trans. Geosci. Remote Sens. 35, 341–353

Brown, M., Lewis, H.G., Gunn, S.R., 1999. Linear spectral mixture models and support vector machines for remote sensing. Submitted to IEEE Trans. Geosci. Remote Sens..

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition In: Fayyad, U. (Ed.), Data Mining and Knowledge Discovery. Kluwer Academic Publishers, The Netherlands, pp. 1–43.

Cherkassky, V., Mulier, F., 1998 Learning From Data: Concepts, Theory and Methods. Wiley, New York, 442 pp.

Foody, G.M., Cox, D.P., 1994 Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. Int. J. Remote Sens. 15, 619–631.

Gunn, S.R., 1998. Support vector machines for classification and regression. Technical Report ISIS-1-98, Department of Electronics and Computer Science, University of Southampton (Technical Report), 52 pp.

Haykin, S., 1999 Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, New Jersey, 842 pp.

Hearst, M.A., Scholkopf, B., Dumais, S., Osuna, E., Platt, J., 1998. Trends and controversies–support vector machines IEEE Intell. Syst. 13, 18–28.

Horwitz, H.M., Nalepka, R.F., Hyde, P.D., Morganstern, J.P., 1971 Estimating the proportion of objects within a single resolution element of a multispectral scanner Proceedings of the 7th International Symposium on Remote Sensing of Environment, Environmental Research Institute of Michigan, Michigan, pp. 1307–1320

Lewis, H.G., Brown, M., Tatnall, A.R.L., Nixon, M.S., Manslow, J.F., 1998. Data analysis and empirical classification in FLIERS. FLIERS project report, Department of Electronics and Computer Science, University of Southampton, (Technical Report), 33 pp.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks Cambridge University Press, Cambridge, 403 pp.

Settle, J.J., Drake, N.A., 1993. Linear mixing and the estimation of ground cover proportions. Int. J. Remote Sens. 14, 1159–1177.

Smola, A.J., Schölkopf, B., 1998. A tutorial on support vector regression. Technical Report Neuro COLT TR-1998-030, Royal Holloway College, London, (Technical Report), 73 pp

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 188 pp.

ELSEVIER

Ecological Modelling 120 (1999) 181–197

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Water and carbon fluxes above European coniferous forests modelled with artificial neural networks

M.T. van Wijk *, W. Bouten

Department of Physical Geography and Soil Science, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands

Abstract

Artificial neural networks are used to select a minimal set of input variables to model water vapour and carbon exchange of coniferous forest ecosystems, independently of tree species and without detailed physiological information. Neural networks are used because of their power to fit highly non-linear relations between input and output-variables. Radiation, temperature, vapour pressure deficit and time of the day showed to be the dynamic input variables that determine ecosystem water fluxes. The same variables, together with projected leaf area index are needed for modelling CO_2-fluxes. The results for the individual sites show that the neural networks found mean water and carbon flux responses to the driving variables valid for all sites. The sensitivity analysis of the derived neural networks shows that the LAI-effect of the CO_2-flux model is overfitted because of the low variability of LAI. However, the predictions of CO_2-fluxes of sites not included in the calibration set indicate that the LAI-response of the network is reliable and that results can be used as a first estimate of the net ecosystem carbon exchange of the forest sites. Independent predictions of forest ecosystem vapour fluxes were equally satisfying as empirical models specifically calibrated for the individual sites. The results indicate that both short term water and carbon fluxes of European coniferous forests can be modelled without using detailed physiological and site specific information.

Keywords: Neural networks; Carbon fluxes; Water fluxes; Coniferous forests.

1. Introduction

In recent years, great effort is made in modelling instantaneous carbon and water fluxes at stand scale (Landsberg et al., 1991; Jarvis, 1995). Both top-down and bottom-up approaches are used to model short term forest ecosystem fluxes. Several detailed physiological models use knowledge about photosynthetic and stomatal responses at leaf level and scale these up to canopy level using elaborate radiation interception models (Wang and Jarvis, 1990; Falge et al., 1996). With these detailed models both water and carbon fluxes are predicted. For evapotranspiration only, a widely used approach is the combination of an energy balance, the Penman–Monteith equation, with the Jarvis–Stewart canopy conductance model (Jarvis, 1976; Stewart, 1988). An example of a more simple, purely empirical approach is the

* Corresponding author. Fax: +31-20-5257431.
E-mail address: m.t.wijk@frw.uva.nl (M.T. van Wijk)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00101-5

Makkink model (Makkink, 1957) used in several forest hydrological models (Bouten and Jansson, 1995). More general applicable models for carbon fluxes are in most cases working on higher time and spatial scales (e.g. Williams et al., 1997).

Models for predicting instantaneous water and carbon fluxes in forest ecosystems are usually developed, calibrated and validated for one specific forest. For each application specific parameters of the response functions are optimised or detailed physiological information on species level is used. Extrapolation to other forest sites is in those cases difficult and time-consuming.

Recently the Euroflux project provided measurements of carbon and water vapour fluxes above a large range of forests across Europe (Tenhunen et al., 1998). These measurements give the opportunity to model ecosystem fluxes along a range of biotic and abiotic system inputs, and to evaluate the processes and environmental variables that determine short term forest ecosystem responses.

In this paper a dataset of ecosystem flux measurements (CO_2 and water vapour) of six different coniferous forests in Northwestern Europe is used to explore the possibilities to model these fluxes with a minimal set of explaining variables. The goal is to model water and carbon fluxes independently of tree species and to analyse the model performance over different forest sites without using detailed physiological or site specific information. For this top-down approach artificial neural networks are used because of their power to fit highly non-linear relations (Huntingford and Cox, 1997). Neural networks give the opportunity to have a completely unconstrained optimisation and they estimate input–output responses without a pre-defined mathematical model. The method supplies so called model free estimations (Kosko, 1992). The relations found by the networks are tested by predicting ecosystem fluxes of forest sites with intermediate characteristics not included in the calibration sets. In addition results of the evapotranspiration fluxes are compared to those of the Makkink model.

2. Method

The data were placed at disposal for the Euroflux workshop held in Sesto (Italy) at 26–29th of January 1998. Latent heat (Lh) and CO_2-fluxes of six coniferous forests in North-western Europe were used to model ecosystem water and carbon fluxes. The flux and meteorological data were supplied on a half hourly basis, all made with identical equipment (Tenhunen et al., 1998). Information about the six different sites is given in Table 1. The Vielsalm measurements are described and presented for a longer time period in Aubinet et al. (1999).

Most data are between day number 150–250 of the year 1996 or 1997. All data were supplied with the assumption that there was no soil water stress. Measurements within 24 h after a rain event were skipped from the dataset to model real ecosystem transpiration and not also interception evaporation. The total dataset of the six sites without missing values and after omitting the wet canopy data consisted of 8448 half hourly measurements.

Table 1
Information about the different forest sites

Site	Country	Geographical coordinates	Species	Age (year)	Number of days (dry)	LAI
Flakaliden	Sweden	64°07′ N 19°27′ E	*Picea abies*	34	37 (19)	2.4
Hyytiala	Finland	61°51′ N 24°17′ E	*Pinus sylvestris*	34	98 (52)	3 9
Loobos	Netherlands	52°10′ N 05°44′ E	*Pinus sylvestris*	100	108 (47)	3 0
Tharandt	Germany	50°58′ N 13°38′ E	*Picea abies*	106	24 (9)	5.0
Vielsalm	Belgium	50°18′ N 06°00′ E	*Pseudotsuga menziessu*	60–90	41 (36)	4 2
Weiden Brunnen	Germany	50°09′ N 11°52′ E	*Picea abies*	44	41 (13)	6 5

The night-time fluxes of CO_2 had to be corrected for stable atmosphere effects in combination with storage effects (Baldocchi and Vogel, 1996; Kimball et al., 1997), when measured CO_2 fluxes were often close to zero. The problems associated with measuring night-time CO_2 fluxes are cancelled during windy periods (Lee, 1998). Therefore night-time CO_2 fluxes are skipped from the dataset when wind speed was below 2.5–3.0 m/s which corresponds roughly to the wind criterion used by Black et al., (1996) in screening night-time CO_2 flux data for quantifying the carbon uptake of their forest. This reduced the CO_2 flux dataset to 5776 point measurements.

In order to get a calibration set with high variability of input variables the data of each site were classified into 15 classes per variable, equally distributed over their data range. The input variables that were estimated as most important were global radiation (Rg) (other radiation components like PPFD were not available for all sites), temperature (T) and vapour pressure deficit (VPD). All measurements of each site were incorporated in one of the $15 \times 15 \times 15 = 3375$ classes, so in total there were $3375 \times 6 = 20250$ classes. From each of these classes two data combinations of measurements were randomly selected and put into the calibration set. To prevent artefacts caused by non-equally large calibration subsets of each site, they were made equally large by random draws from the first calibration set. After this there were 248 point measurements for Lh and 238 point measurements for CO_2 for each site in the calibration sets. In total the Lh flux calibration set consisted of 1488 data points and the CO_2 flux calibration set consisted of 1428 data points. The other data points (for CO_2 4348 and for Lh 6960) were placed in the validation set. Data of each site are therefore both in the calibration and in the validation set. The values of the input variables were scaled between zero and one.

A three layer backpropagation neural network was used within Neural Network Toolbox 2.0 of Matlab 4.0, (Demuth and Beale, 1995). The optimisation method applied in the calibration phase was the Levenberg–Marquardt method. The total sum of squared errors (SSE) between measured and modelled values was minimised by tuning the artificial neural network parameters (e.g. scaling factors and inter-neurone connection weights). The number of epochs used in the optimisation was 75 and for each model fifty initialisations were tested. The transfer function for the hidden node layer was the sigmoidal function:

$$\psi(u) = \frac{2}{1 + e^{-2 \times u}} - 1 \quad (1)$$

The other transfer functions available in the software package gave the same (other sigmoidal transformations) or much worse (other linear transformations) results (more details can be found in the Appendix).

The results of different input variable combinations were evaluated using the independent validation set. As measures of the goodness of fit the normalised root mean square error (NRMSE) and the explained variance (R^2) are used. The NRMSE corrects for the size of the data set and for the mean value of the modelled variable (Janssen and Heuberger, 1995).

Besides physical driving variables like Rg, T and VPD, also variables like 'Day of Year' (DoY) and 'Time of Day' (TofD) are tested. TofD is expressed in digital time and corrected for each site so that on cloudless days the maximum global radiation value was at 12.00 h. TofD is used in two ways in the input variable analysis because of the high correlation between the daily pattern of Rg and TofD. First it is used as an input variable together with Rg, T, VPD etc., but it is also used as a variable to analyse the possibilities to improve the results of neural networks using purely physical driving variables. This is done by taking the best simulation results (BS) achieved with physical driving variables together with TofD as input for the network. In this way it is prevented that the neural network will use TofD as the main driving force for the daily pattern of ecosystem fluxes and Rg only as an offset variable to determine the height of this daily pattern. It is important to be careful with just adding input variables because correlation's between the variables can lead to unintended side-effects in the responses that the network finds.

Variables like soil temperature and soil water content, of which it is known that they influence

Table 2
Model misfits (NRMSE) of neural networks using different sets of input variables, explained variance between brackets

Model	Number of hidden nodes	Input-variables	Lh-flux	CO_2-flux
1	3	Rg & T	0 58 (0.79)	0.77 (0.66)
2	3	Rg & T & LAI	0.57 (0.80)	0.73 (0.67)
3a	2	Rg & T & VPD & LAI	0 58 (0 78)	0.76 (0.65)
3b	3	Rg & T & VPD & LAI	0.56 (0.80)	0.71 (0.68)
3c	4	Rg & T & VPD & LAI	0.56 (0.80)	0.70 (0.68)
4	3	Rg & VPD & LAI	0.56 (0.79)	0.73 (0.67)
5a	2	Rg & T & VPD	0.58 (0.78)	0.77 (0.63)
5b	3	Rg & T & VPD	0.56 (0.80)	0.75 (0.64)
5c	3	Rg & T & VPD	0.56 (0.80)	0.75 (0.64)
6	3	Rg & T & VPD & Wind	0.56 (0.80)	0.75 (0.65)
7	3	Rg & T & VPD & TofD	0.56 (0.80)	0.73 (0.67)
8	3	Rg & T & VPD & LAI & TofD	0.52 (0.83)	0.69 (0.70)
9	2	Best Simulation (Rg & T & VPD) & TofdD	0.53 (0.82)	0 73 (0.67)
10	2	Best Simulation (Rg & T & VPD & LAI) & TofD	0.53 (0.82)	0.68 (0.71)
11	3	Rg & T & VPD & LAI & DayNr	–	0.75 (0.65)

soil respiration rates (Freijer et al., 1996) and leaf N-content, which influences leaf maintenance respiration (Barnes et al., 1997), were not available for all sites. To simulate the variability of soil temperature compared to air temperature (more damped and with a short time lag) also the mean air temperature of 2 h preceding the current value were used as input. These calculations were made on a smaller dataset because data could only be used when the air temperature of the 2 h preceding the current measurement were available. N-content of leaves is strongly correlated to leaf area index (Williams et al., 1997), so extra addition of this variable is not expected to improve the performance of the neural networks very much.

The results of neural networks modelling Lh fluxes are compared to the Makkink model, which is a model predicting transpiration when there is no waterstress. For each site the empirical plant factor is calibrated.

$$\mathrm{Lh} = f \times \left[0.65 \times \frac{S}{S+\gamma} \times \mathrm{Rg} \right] \quad (2)$$

In which:

Lh latent heat flux (W/m^2)
f empirical plant factor (−)
S derivative of saturated vapour pressure–temperature curve (hPa/K)
γ psychometric constant (hPa/K)
R_g global radiation (W/m)

Independent predictions were made for sites not included in the calibration sets as a more thorough test of the water and carbon responses which the neural networks had found. This was done by using the Jack Knife method: calibrating the neural networks on only five of the six forest sites and predicting the sixth not included forest site. The carbon fluxes of only four sites could be predicted because of the importance of the variable 'Leaf Area Index' (LAI). The extreme values of LAI of Flakaliden and Weiden Brunnen (lowest and highest value) are not predicted, because these values are clearly outside the interval of LAI values of the other five sites on which the networks are calibrated. The prediction of carbon fluxes of Flakaliden and Weiden Brunnen would therefore be an extrapolation for which the neural network technique is not suited (Huntingford and Cox, 1997).

3. Results and discussion

3.1. Model selection

The validation results of the most important input variable combinations are shown in Table 2.

The optimal construction for most artificial neural network models presented here is using three hidden neurones, except for models 9 and 10 where only two hidden neurones were necessary. Increasing the number of neurones did not improve the model fit of Lh fluxes (see model 5b and 5c), and only led to a minimal improvement in model fit of CO_2 fluxes (see models 3b and 3c). Because of the low variability of LAI (only six different values) it is extremely important to keep the number of hidden neurones as low as possible, otherwise the neural network will use LAI as a kind of individual site index, without using it as a real quantitative variable. With two hidden neurones model performance in simulating Lh fluxes and CO_2 fluxes the model error increased considerably (see models 5a and 3a).

Increasing the number of iterations of the calibration period (the number of epochs) did not improve the model fit. The selection of calibration data over the full experimental validity range used in this article, precludes the problems of overfitting reported by Schaap and Bouten (1996) and Huntingford and Cox (1997), which also use much smaller calibration sets and which have no classification of data.

The fit values when using TofD as an extra input variable (models 7 and 8) and using it as model mismatch analysis factor (models 9 and 10) are not worse (see Table 2). Lh model 9 has even a lower model error than Lh model 7 which uses TofD as an extra input variable. As TofD does not improve modelling results when using it as an extra input variable, the responses of the different physical driving variables are not dependent on the values of TofD; there are no interaction effects between TofD and the other input variables.

Most striking is that the variable LAI does not improve modelling results of the Lh neural network. It seems that there are so many feedback mechanisms working in the process of transpiration (e.g. radiation interception and VPD effects) that LAI has no net influence in this interval of input values (2.4–6.5 m^2/m^2). Forest floor evaporation will also be more important at low LAI values.

The other input variables that were used for modelling CO_2 fluxes did not improve model performance. Soil temperature (at 5 cm depth) was available for two sites (Weiden Brunnen and Vielsalm). Modelling these sites individually adding the variable 'soil temperature' (ST) led to slight decrease in model misfit for the Vielsalm data (NRMSE 0.41 versus 0.40), but no increase in fit was found for the Weiden Brunnen site. The mean value of air temperatures of the 2 h preceding the current value led to a decrease in model error for the Vielsalm site individually (NRMSE 0.38), but when applied to the total dataset adding this variable did not lead to an increase of model fit (both NRMSE 0.70).

The neural network models to be analysed further are Lh flux model 9 with input{BS(Rg, T, VPD) & TofD} and CO_2 flux model 10 with input {BS(Rg, T, VPD, LAI) & TofD}. These models are chosen instead of the models with 'TofD' as an extra input variables to prevent model artefacts due to highly correlated input variables.

3.2. Model performance

The performances for the different sites of the selected neural network models are given for Lh fluxes in Table 3a and for CO_2 fluxes in Table 3b. The results of networks with the same input variables calibrated on the individual sites are also given. Networks calibrated on all data of the individual sites are regarded as the best possible models with these input variables. For Lh fluxes results of the Makkink model are also presented.

Results in Table 3 show that for both Lh and CO_2 fluxes the neural networks calibrated on all sites found a mean ecosystem response to the driving variables. The sequence of fit values is the same for the total model and for the individual fitted models. The NRMSE values for the total model are always slightly higher than the individual fitted networks, which is to be expected because the latter are calibrated on site specific data. These results indicate that the 'total' neural networks did not fit a few sites very well and the other badly, which would be the case if the differ-

ent forest ecosystems did not react to the same extend to the driving variables.

Adding the variable TofD in modelling CO_2 fluxes led to an increase in model misfit of the sites Flakaliden and Weiden Brunnen. This can probably be explained by the very specific TofD response of these sites individually for CO_2 fluxes (see Fig. 6 and below when explaining the response curves). When modelling Lh fluxes adding the variable TofD led to a slight decrease in misfit for all sites.

A typical value for the measurement error of eddy covariance measurements of Lh fluxes is about 5% which results in a standard deviation of about 18 W/m^2 (Bosveld and Bouten, 1992). One can estimate the possibilities of model improvement by calculating the errors of model minus measurements, expressed in variation or standard deviation, and comparing them with the measurement error according to:

$$\text{error (measurements} - \text{model)} \approx \text{error (model)} + \text{error (measurements)} \quad (3)$$

The measurement minus model error of the total dataset expressed in standard deviation is for Lh model number 9 23.3 W/m^2 and the standard deviation of the results of the Vielsalm site, for which the neural networks are performing the best of all sites, is 20.2 W/m^2. According to these results and this estimate of the measurement error the possibilities for improving of the neural network results for modelling Lh fluxes seem small.

The performance results of modelling Lh fluxes and CO_2 fluxes cannot be compared in an easy way. The higher values of NRMSE for CO_2 fluxes are probably partly due to the fact that negative and positive values of this variable lead to a small value of the mean, so that the fit error expressed as a relative value of the mean will be higher than for Lh fluxes.

Table 3
Results for modelling Lh-fluxes for the individual sites (given are NRMSE and between brackets explained variance)

Site	Total neural network		Individual network		Makkink
	Rg & *T* & VPD	BS (Rg & *T* & VPD) & TofD	Rg & *T* & VPD	BS (Rg & *T* & VPD) & TofD	
a					
Flakaliden	0.41 (0 81)	0.39 (0.84)	0.39 (0 83)	0.35 (0 85)	0 48 (0 78)
Hyytiala	0.61 (0.79)	0.59 (0.80)	0.57 (0.80)	0.56 (0.81)	0.64 (0.76)
Loobos	0 58 (0.80)	0.53 (0.83)	0.50 (0.84)	0.47 (0 86)	0.64 (0.78)
Tharandt	0.89 (0.57)	0.83 (0.63)	0.74 (0.66)	0.68 (0.72)	1.13 (0.46)
Vielsalm	0.49 (0.87)	0.47 (0.87)	0.42 (0.90)	0.41 (0.90)	0.56 (0.82)
Weiden Brunnen	0.62 (0.82)	0.61 (0.82)	0.52 (0.86)	0.49 (0.88)	0.69 (0.76)

Results for modelling CO_2 -fluxes for the individual sites (given are NRMSE and between brackets explained variance)

Site	Rg & *T* & VPD & LAI	BS (Rg & *T* & VPD & LAI) & TofD	Rg & *T* & VPD	BS (Rg & *T* & VPD) & TofD
b				
Flakaliden	0.57 (0.76)	0.61 (0.74)	0.50 (0.82)	0.44 (0.86)
Hyytiala	0.71 (0 74)	0.71 (0.75)	0.59 (0.78)	0.56 (0.80)
Loobos	0.85 (0 59)	0 78 (0 63)	0.75 (0.65)	0.73 (0.67)
Tharandt	0.84 (0.50)	0.78 (0.57)	0.76 (0.59)	0.71 (0.64)
Vielsalm	0 49 (0.66)	0 44 (0.72)	0 41 (0.75)	0 36 (0.80)
Weiden Brunnen	1.08 (0.79)	1.13 (0.78)	0.94 (0 84)	0.85 (0.87)

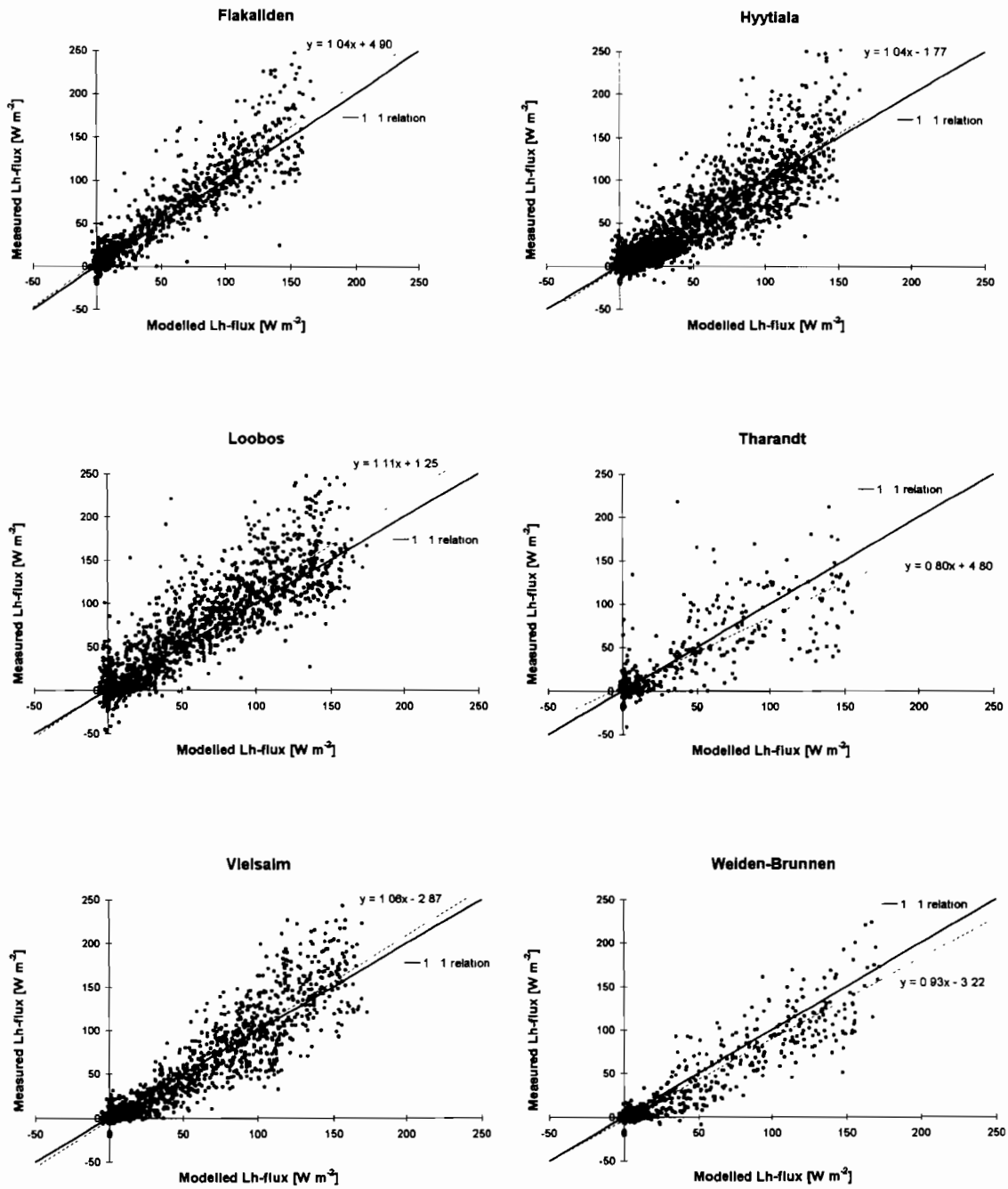

Fig. 1. Measured versus modelled Lh-fluxes for the individual sites.

Graphically the performance of model 9 for Lh fluxes and model 10 for CO_2 fluxes are shown in Fig. 1 and Fig. 2. Regression lines are calculated with the modelled values on the x-axis and measured values on the y-axis (Janssen and Heuberger, 1995). Negative CO_2 flux values are

net carbon uptake by the forests and positive values are net carbon release (respiration). The Tharandt site has both for Lh and CO_2 fluxes the worst performance. The high misfit values of Tharandt cannot be explained by the fact the site has unique properties regarding Rg, T and VPD compared to the other sites, because also performances of individual calibrated net-

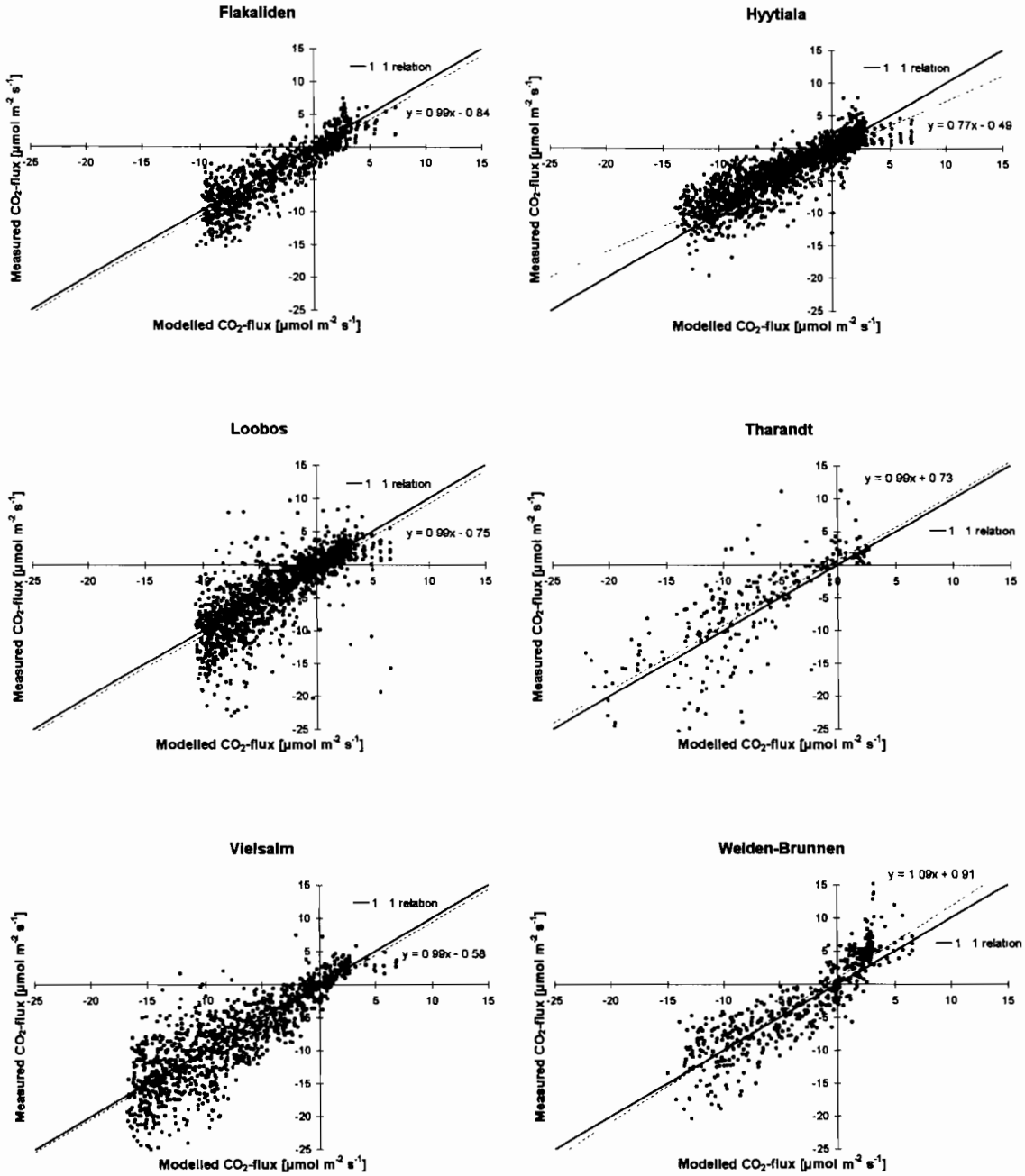

Fig. 2. Measured versus modelled CO_2-fluxes for the individual sites.

works are less than for the other sites (see Table 3a).

For Lh fluxes there is a systematic underestimation of the high values of the measured Lh fluxes. For the plots of CO_2 fluxes an overestimation of the low negative daytime carbon fluxes is calculated whereas the high night-time fluxes are underestimated. Also there can be seen a kind of border, for CO_2 fluxes both at the upper and lower levels of the modelled values whereas for Lh fluxes the border is only visible at the high levels of the modelled values.

This systematic misfit could be due to several causes. Huntingford and Cox (1997) had the same kind of systematic misfit in modelling canopy conductance, and they mention a number of possible explanations like non-constant aerodynamic conductance, missing interannual variability, failure to capture non-linearity in responses and missing input that varies on a timescale longer than one day and shorter than 1 year. Not mentioned are possible artefacts due to the transfer function used in the neural network. The shape of these scatter plots is for the CO_2 fluxes very similar to the shape of the sigmoidal transfer function. If the upper and lower values of the validation data are not defined by unique values of the input variables it could be that in the optimisation procedure followed it is preferable for the network to fit the abundant values close to the mean very well and the extreme and less abundant values less well. The reason that the systematic error is seen at both ends for CO_2 fluxes and only at the higher values for Lh fluxes is probably that for Lh fluxes the lower boundary is well defined (if there is no global radiation there will be almost no evapotranspiration), whereas for the CO_2 fluxes both the upper and lower boundary are not very well defined by any value of an input variable. Another explanation could be the measurement error which is relatively high for the eddy correlation measurement technique compared to variables like temperature, radiation and vapour pressure deficit. Model uncertainty is in such a case small compared to measurement uncertainty. The highest peak measurement values can be caused by measurements errors, they are noise effects, and are not characterised by unique sets of input variables and thereby systematically underestimated by a model.

3.3. Response curves

Responses of artificial neural networks can be evaluated by varying one single input while keeping other inputs at their mean value. These results should be interpreted with care, as the reference value of one variable can influence the response curve of another (Huntingford and Cox, 1997). To evaluate interaction effects in the networks presented here, two variables are varied together while the others are set to their mean value. The most interesting response surfaces are shown in Fig. 3 for Lh fluxes and in Fig. 4 for CO_2 fluxes. All the results presented in Fig. 3 and 4 have to be interpreted with care for extreme values of the graphs are in most cases extrapolations, although the amount of these is kept as low as possible. The values of the inputs that are varied are also constrained by the values of the constant inputs. If one uses for example a constant VPD input of 15 hPa, temperature values below 5°C would be an extrapolation of the neural network responses, as these combinations of input values will not be present in the dataset.

The response surfaces of the Lh model are not very surprising. Transpiration increases with radiation and temperature. Above an optimum value of T (around 18°C) transpiration decreases (Fig. 3A), probably caused by a coupled negative effect of high temperature and vapour pressure deficit values on the stomatal conductance. In Fig. 3D the influence of adding the variable TofD are shown. Best Simulations of Lh fluxes are decreased in the morning and in the afternoon, whereas values at noon are increased.

Figs. 4A and 4B show different day and night-time effects of T and LAI on modelled CO_2 fluxes. As the results of Fig. 4B are calculated with a global radiation of 0 W/m^2 as input, this response surface can be interpreted as the night-time respiration curves of the neural

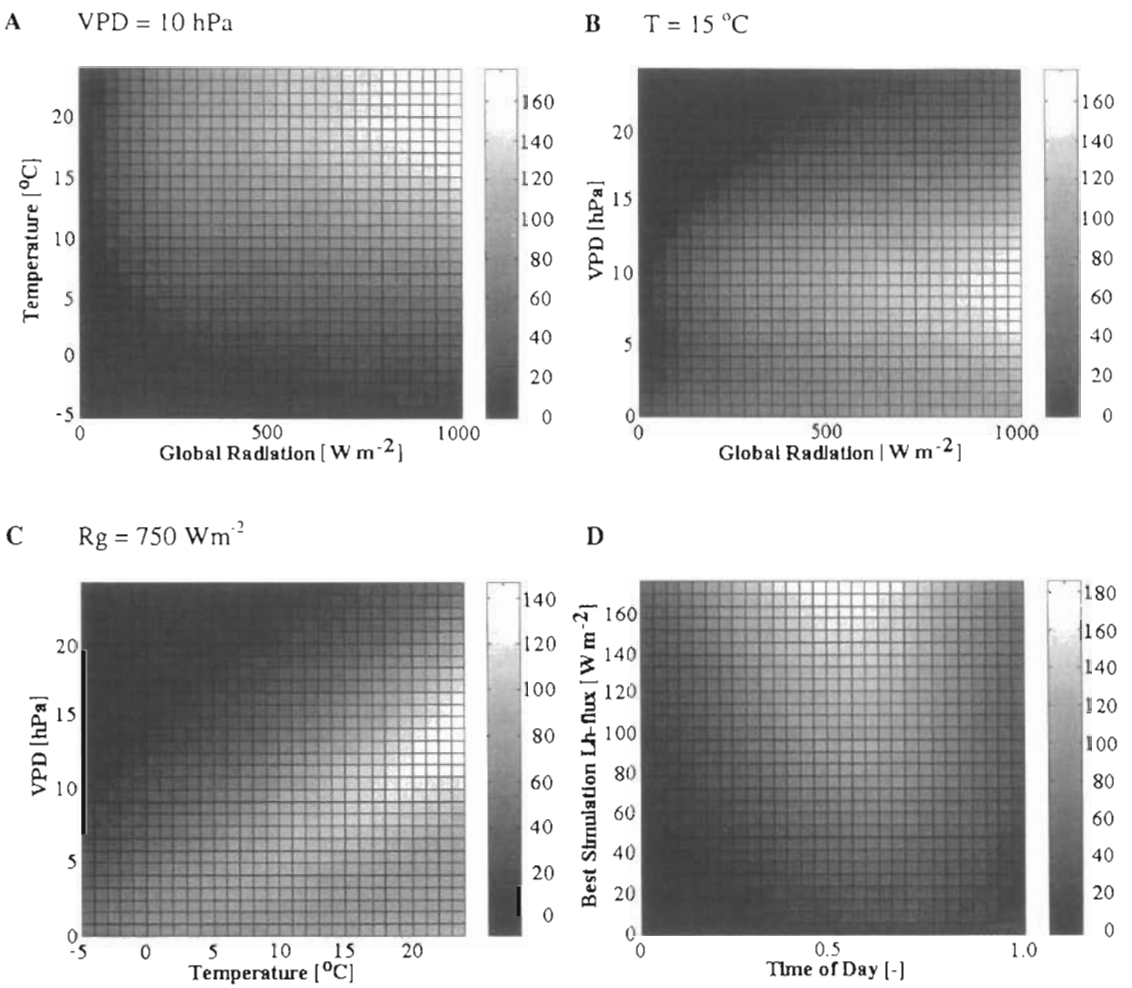

Fig. 3. The response-curves of the neural network modelling Lh-fluxes (grey-tones are Lh-flux values in W/m^2]).

network model. No simple interpretable results follow from these responses. The network seems to find a kind of optimum curve for the net-ecosystem carbon uptake at LAI 5 m^2/m^2, which is the value for Tharandt. This optimum however is dependent of VPD (Fig. 4D). When a value of 20 hPa instead of 10 hPa is used, the highest net ecosystem exchange is at LAI 6.5, which is the value of Weiden Brunnen. However, the value of 20 hPa cannot be used for this response curve fitting, because this value never occurs at the sites of Flakaliden and Hyytiala, and therefore would mean an extrapolation of the network responses for which the method is not fit (Huntingford and Cox, 1997). Different from the daytime carbon fluxes the respiration response for LAI seems to be overfitted: the variability of LAI is too low. The model finds a kind of minimal respiration at LAI 5. To have an accountable CO_2-respiration flux to LAI relation more sites should be included or data of a longer period of the individual sites when the 'LAI' is varying should be used. The LAI response of the neural network will later be tested by predicting independent datasets. In this way the validity, reliability and applicability of this response can be verified.

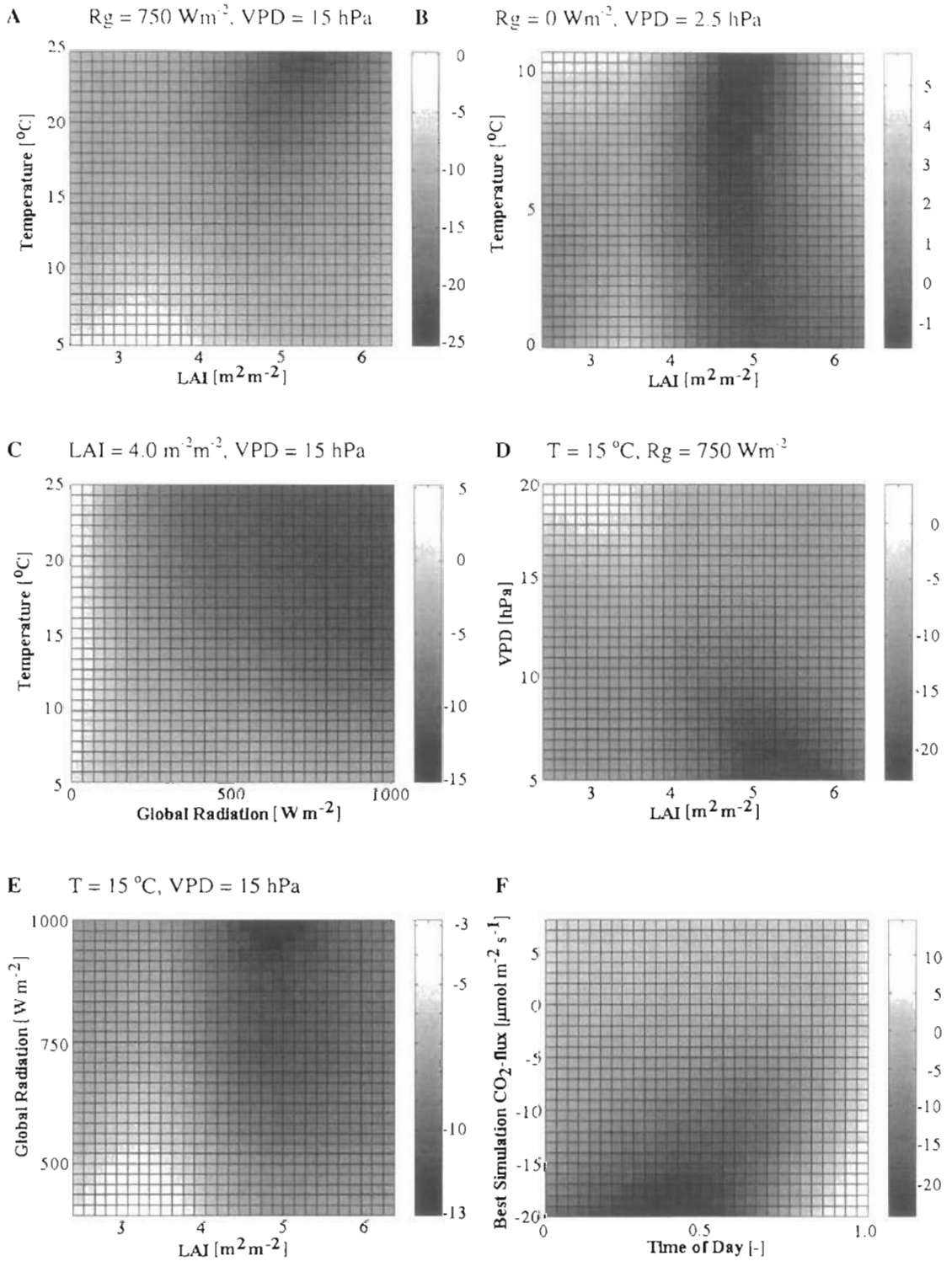

Fig. 4. The response curves of the neural network modelling CO_2-fluxes (grey-tones areCO_2-flux values in µmol/m^2/s^1]).

The response curves of TofD are interesting. Opposite to Huntingford and Cox (1997) we found clear optimum curves both for Lh fluxes as for CO_2 fluxes (see Fig. 3D, Fig. 4F and for the individual sites Figs. 5 and 6). The response curves plotted for the individual sites in Fig. 5 (Lh) and Fig. 6 (CO_2) have a maximum around noon, except for Weiden Brunnen where both the Lh and the CO_2 fluxes show a time shift towards the afternoon. This TofD effect could be due to changes in the fraction of sunlit leafs. At low solar elevation there is more shadowing between the trees, and at high elevations (at noon) radiation can reach much more deeply into the canopy and can also reach the lowest leaf levels (Green and McNaughton, 1997). The effect is not measured in Rg because radiation measurements are done above the forest canopy. This effect can lead to an overestimation of solar radiation effectiveness for the forest at low solar elevations and an underestimation of solar radiation effectiveness at high solar elevations by the neural network. This will be compensated by introducing the variable TofD. The same effect can be achieved by introducing for each site the calculated solar height. The effect will be damped by cloudy days, when solar penetration into the canopy is less dependent on solar height.

The time-shift of the Weiden Brunnen site could be explained by the fact that the forest site of Weiden Brunnen is located on a hill slope (though not very steep, only 10 to 15°) with an exposition towards the south-west. This means that the sun will reach the highest point, from forest viewpoint, in the afternoon.

The differences between the other sites in their reaction to TofD can not be simply explained. You would, for example, expect the TofD effect to be more pronounced in forests with high LAI. This is not clearly visible in the figures.

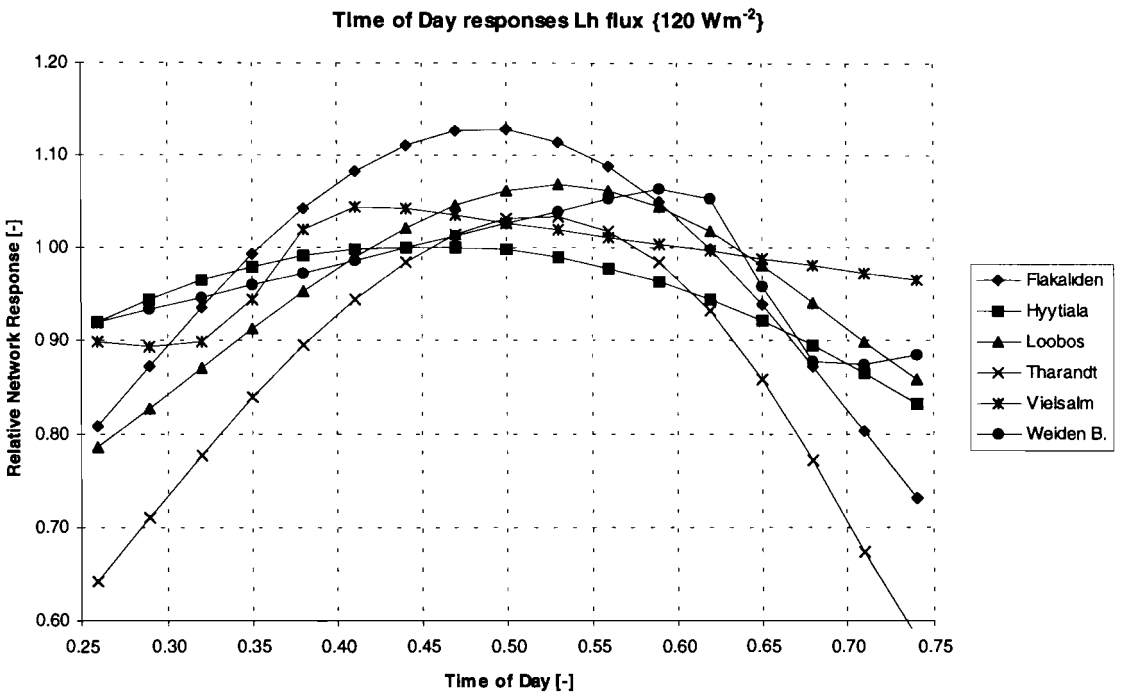

Fig. 5. Response curve of the variable 'TofD' with as input-value of the variable 'Best Simulation' a Lh-flux of 120 W/m^2.

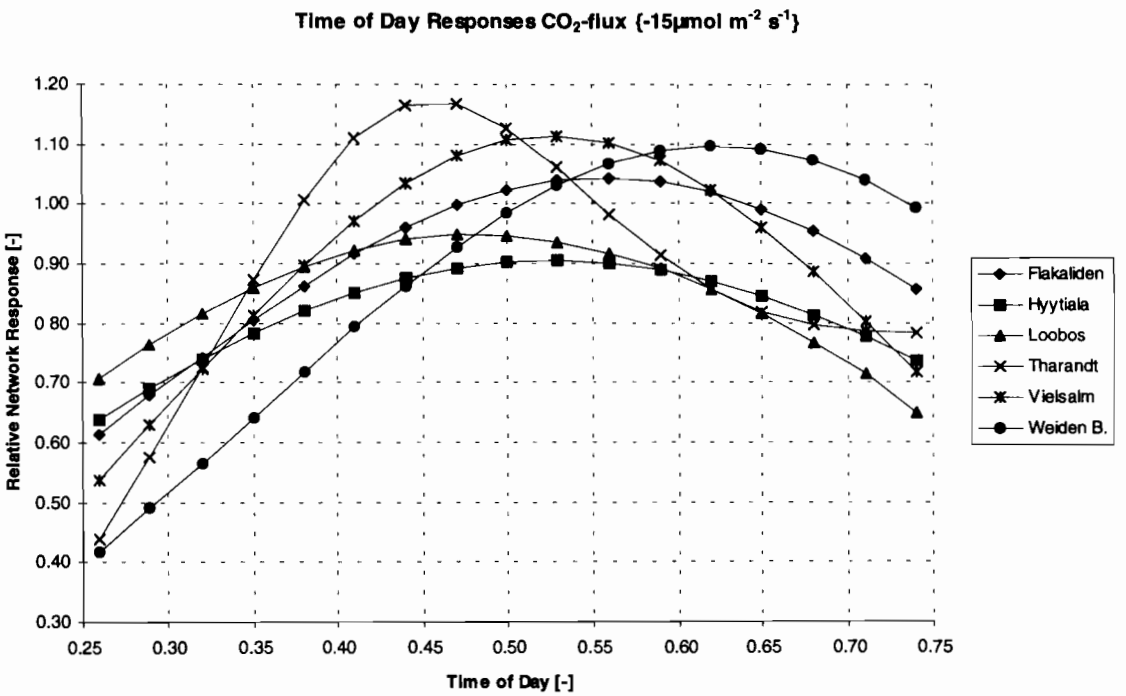

Fig. 6. Response curve of the variable 'TofD' with as input-value of the variable. 'Best Simulation' a CO_2-flux of -15 µmol/m^2/s^1.

3.4. Prediction

Results of independent model predictions are given in Tables 4a and 4b. The predictions of Lh fluxes are very satisfying and in all cases better (expressed in 'NRMSE') than the results of the individually calibrated Makkink model (see Table 3a). The estimates of transpiration sums are also reliable, considering the fact that it is a totally independent prediction. The reliability of the simulated transpiration and carbon exchange sums is an important test for the derived artificial neural networks. Because of the high calculation speed of the networks, once the calibration is completed, the networks can be applied as input for regional models of carbon and water cycling.

The results of the CO_2 fluxes are worse than the results of the Lh fluxes. This is due to the importance of LAI, of which the variability is so small that the networks tend to overfit the relation between CO_2 fluxes and LAI. Otherwise, also these results can be considered satisfying as a first approximation of the carbon balances of the different coniferous forests. Striking is the difference between Hyytiala and the three other sites. Only the Hyytiala sum of CO_2 fluxes is overestimated by the neural network, probably caused by the overfitted LAI response.

3.5. Applicability of neural network models in data analysis

The neural networks derived in this article are black box models with no conceptual mechanisms behind. Therefore they are only communicable by giving the network used and the derived parameters. This is done in the appendix. The neural network models presented are mathematical representations of mean responses of coniferous forests' water and carbon

exchange to different driving variables. As the neural network technique is an empirical method (Kosko, 1992; Huntingford and Cox, 1997) the models given should not be applied to situations outside the maximum values of the variables.

Neural networks are a very powerful tool to extract information from datasets. Here neural networks are used to evaluate the general behaviour of different coniferous forests. Especially the results of independent flux predictions show that satisfying results can be obtained by distracting general behaviour of forests. This method can be applied to all kinds of ecological processes, if enough data are available for this top-down analysis. For these kinds of applications the model free estimations of neural networks are an advantage of the method because there is no pre-defined constraint to the solution which the neural network will find, as in other methods. The disadvantage of a black box method, no clear insight in what the neural network did learn, can be overcome by applying other analysing techniques like fuzzy logic (Kosko, 1992).

4. Conclusions

Both instantaneous water and carbon fluxes can be modelled with artificial neural networks without physiological or site specific information. The variables that are needed for modelling the evapotranspiration are global radiation, temperature, vapour pressure deficit and time of the day. The explained variances of this model for the individual sites are between 0.63 and 0.87 (NRMSE-values are between 0.47 and 0.83). The four input variables of the Lh-flux model together with leaf area index are needed for modelling CO_2 fluxes. The explained variances of the carbon flux model for the individual sites are between 0.57 and 0.78

Table 4
Instantaneous model misfit and misfit in summed values of the independent Lh-flux predictions by neural networks fitted on the other five sites (given are NRMSE and between brackets explained variance)

Site	Rg & T & VPD		BS (Rg & T & VPD) & TofD	
	NRMSE (R^2)	Sum (Measured)/Sum(Modelled)	NRMSE (R^2)	Sum (Measured)/Sum(Modelled)
a				
Flakaliden	0.45 (0.79)	1.17	0.42 (0.83)	1.21
Hyytiala	0.66 (0.75)	1.01	0.61 (0.78)	1.00
Loobos	0 63 (0.75)	1.14	0.58 (0.80)	1.15
Tharandt	1.01 (0.51)	0.92	0.94 (0.57)	0 90
Vielsalm	0.52 (0 85)	0.99	0.51 (0.85)	0.96
Weiden Brunnen	0.69 (0.78)	0.84	0.68 (0.80)	0.82

Instantaneous model misfit and misfit in summed values of the independent CO_2-flux predictions by neural networks fitted on the other five sites (given are NRMSE and between brackets explained variance)

Site	Rg & T & VPD& LAI		BS (Rg & T & VPD& LAI) & TofD	
	NRMSE (R^2)	Sum (Measured)/Sum(Modelled)	NRMSE (R^2)	Sum (Measured)/Sum(Modelled)
b				
Flakaliden	–	–	–	–
Hyytiala	0.82 (0.65)	0.85	0.77 (0.72)	0.85
Loobos	0.86 (0.56)	1.20	0.79 (0.62)	1.09
Tharandt	0 92 (0.45)	1.23	0.89 (0.48)	1.23
Vielsalm	0.53 (0.66)	1.25	0.49 (0.70)	1.25
Weiden Brunnen	–	–	–	–

(NRMSE-values are between 0.44 and 1.13). Independent validations of the individual sites show that the neural networks found mean ecosystem responses valid for all sites. For all sites neural network predictions of water fluxes were better than those of the site specific calibrated one parameter model Makkink. The LAI effect of the neural network describing CO_2 fluxes is probably overfitted because of the low variability of this input. Independent predictions of the CO_2 fluxes, however, showed that the LAI effect is reliable and that the neural network model can be used as first estimate of the net ecosystem carbon exchange.

The variable 'Time of Day' is used in this article as a model mismatch analysis factor by using the best simulations achieved with the physical driving variables 'global radiation', 'temperature', 'vapour pressure deficit' and 'leaf area index' together with 'TofD' as input for the neural networks. The results show a clear optimum curve for 'TofD' with a maximum around noon. This effect is probably due to changes in the fraction of sunlit leaves. At high sun elevations radiation can penetrate deeper into the canopy. This explanation is supported by the time-shift of the maximum of the optimum curve for both Lh and CO_2 fluxes of the Weiden Brunnen site, a site located on a slope with south–west exposition.

Acknowledgements

The funding by the National Research Programme 952232: 'Climate change and Forest Ecosystem Dynamics in Europe' and the University of Amsterdam is gratefully acknowledged. We thank the generous data disposal by the Euroflux-project, and especially the site P.I.'s of which data are used in this article: M. Aubinet, Faculté Universitaire des Sciences Agronomiques de Gembloux; Ch. Bernhofer, University of Technology Dresden; A.J. Dolman, DLO Winand Staring Centre Wageningen; A. Lindroth, SLU Uppsala; J. Tenhunen, E.O Schulze and C. Rebmann, University of Bayreuth; T. Vesala, University of Helsinki. We also thank J.M. Verstraten for a critical evaluation of a previous draft of the manuscript.

Appendix A. Artificial Neural Network model parameters

As neural networks are no conceptual models that are communicable, we give here the derived empirical parameters with a short description of the network architecture.

A.1. Network architecture

First all values of input x are scaled between 0 and 1 using the formula:

$$\text{input}(x) = 1 - \frac{\max(x) - \text{value}(x)}{\max(x) - \min(x)} \quad \text{(A1)}$$

These input values (I) are multiplied with the first connection matrix, $\mathbf{A}_1$, which contains the connections between the input nodes and the second, hidden, node layer (X):

$$\vec{X} = \mathbf{A}_1 \cdot \vec{I} \quad \text{(A2)}$$

The number of input nodes determines the number or columns of the matrix, the number of hidden nodes determines the number of rows of the matrix. A neural network with three input variables and five hidden nodes therefore has a first connection matrix with five rows and three columns, containing 15 connection values.

After multiplication of the input values with the first connection matrix the values of the hidden nodes are known. These values are scaled after adding an offset parameter vector (A_2):

$$\vec{u} = \vec{X} + A_2 \quad \text{(A3)}$$

$$\vec{HN}(\vec{u}) = \frac{2}{1 + e^{-2\vec{u}}} - 1 \quad \text{(A4)}$$

With the above mentioned network with five hidden nodes, this gives another five parameters (for each hidden node one offset parameter).

These scaled values are then multiplied with the connection matrix ($\mathbf{A}_3$), containing the connections between the hidden nodes and the output node.

$$O = \mathbf{A}_3 \cdot \vec{HN} \quad \text{(A5)}$$

In the case of five hidden nodes and one output node this will also give another five parameters. After this multiplication another offset parameter is added to this output node value. This final value is the definitive neural network output value (OV):

$$\mathrm{OV} = O + A_4 \quad \text{(A6)}$$

A.2. Parameter values

The parameter values of four models are given here (see Table 2): for Lh fluxes models 5 and 9, and for CO_2 fluxes models 3 and 10.

max (Rg):	929.0 W/m^2	min (Rg):	−4.0 W/m^2 [input node 1]
max (T):	27.97°C	min (T):	−4.2°C [input node 2]
max (VPD):	23.3 hPa	min (VPD):	0.0 hPa [input node 3]
max (LAI):	6.5 m^2/m^2	min (LAI):	2.41 m^2/m^2 [input node 4]

$$\mathbf{A}_1 = \begin{bmatrix} 1.02 & 2.52 & -3.24 & -3.49 \\ -0.94 & -2.22 & 2.93 & 3.72 \\ -2.88 & 1.12 & -1.02 & 0.11 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} 2.42 \\ -2.56 \\ 3.96 \end{bmatrix}$$

$$A_3 = [81.67 \quad 81.23 \quad -43.43] \qquad A_4 = [34.40]$$

A.6. CO_2 fluxes model 10:

max (BS):	7.7 μmol/m^2/s	min (BS):	−20.6 W/m^2 [input node 1]
max (TofD):	1.0 [−]	min (TofD):	0.0 [−] [input node 2]

A.3. Lh fluxes model 5:

A.3.1. Scaling parameters

max (Rg):	929.0 W/m^2	min (Rg):	−4.0 W/m^2 [input node 1]
max (T):	27.97°C	min (T):	−4.2°C [input node 2]
max (VPD):	23.3 hPa	min (VPD):	0.0 hPa [input node 3]

$$\mathbf{A}_1 = \begin{bmatrix} 0.54 & 2.58 & -2.32 \\ -3.96 & 1.72 & 1.13 \\ -0.59 & 3.83 & -5.38 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} -0.06 \\ 2.75 \\ 2.73 \end{bmatrix}$$

$$A_3 = [-132.0 \quad 62.0 \quad 82.1] \qquad A_4 = [-18.0]$$

A.4. Lh fluxes model 9:

A.4.1. Scaling parameters:

max (BS):	176.9 W/m^2	min (BS):	−7.1 W/m^2 [input node 1]
max (TofD):	1.0 [−]	min (TofD):	0.0 [−] [input node 2]

$$\mathbf{A}_1 = \begin{bmatrix} -0.25 & 2.42 \\ 0.78 & 2.78 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} -0.69 \\ 1.96 \end{bmatrix}$$

$$A_3 = [195.94 \quad -180.34] \qquad A_4 = [-10.5]$$

A.5. CO_2 fluxes model 3:

$$\mathbf{A}_1 = \begin{bmatrix} 0.06 & -2.58 \\ 0.07 & -1.84 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} 2.91 \\ 2.14 \end{bmatrix}$$

$$A_3 = [-73.96 \quad 65.35] \qquad A_4 = [12.39]$$

References

Aubinet, M., Chermanne, B., Vandenhaute, M., Longdoz, B., Yernaux, M., Laitat, E., 1999. Long term measurements of water vapour and carbon dioxide fluxes above a mixed forest in Ardenne's region. Agric. For. Meteor. (in press).

Baldocchi, D.D., Vogel, C.A., 1996. Energy and CO2 flux densities above and below a temperate broad-leaved forest and a boreal pine forest Tree Physiol. 16, 5–16

Barnes, B.V., Zak, D.R., Denton, S.R , Spurr, S.H., 1997. Forest Ecology, 4th edition. Wiley, New York, p. 774.

Black, T.A., den Hartog, G., Neumann, H.H., Blanken, P.D., Yang, P.C , Russell, C., Nesic, Z., Lee, X., Chen, S.G., Staebler, R., Novak, M.D., 1996. Annual cycles of water vapour and carbon dioxide fluxes in and above a boreal aspen forest. Global Change Biol. 2, 219–229.

Bosveld, F.C., Bouten, W., 1992. Comparing transpiration models with eddy-correlation observations of a Douglas fir forest. PhD-thesis W. Bouten: 163–180.

Bouten, W., Jansson, P.-E., 1995. Water balance of the Solling spruce stand as simulated with various forest-soil-atmosphere models. Ecol Model. 83, 245–253.

Demuth, H., Beale, M., 1995. Neural Network Toolbox For Use with Matlab. The MathWorks, Natick.

Falge, E., Graber, W., Siegwolf, R., Tenhunen, J.D., 1996. A model of the gas exchange response of *Picea abies* to habitat conditions. Trees 10, 277–287.

Freijer, J.I., Bouten, W., Jonge, H.D., Verstraten, J.M., 1996. Assessing Mineralization Rates of Hydrocarbons in Soils in Relation to Environmental Factors. Soil Biol. Biochem.

Green, S.R., McNaughton, K.G., 1997. Modelling effective stomatal resistance for calculating transpiration from an apple tree. Agric. For. Meteorol. 83, 1–26

Huntingford, C., Cox, P.M., 1997. Use of statistical and neural network techniques to detect how stomatal conductance responds to changes in the local environment. Ecol. Model. 97, 217–246.

Janssen, P.H.M., Heuberger, P S.C , 1995 Calibration of process-oriented models. Ecol Model 83, 55–66.

Jarvis, 1976 The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field Phil. Trans. R. Soc Lond., Ser. B, 273. 593 - 610.

Jarvis, P G , 1995. Scaling process and problems., Plant. cell and Environ 18, 1079–1089.

Kimball, J S , Thornton, P.E , White, M.A , Running, S.W., 1997. Simulating forest productivity and surface-atmosphere carbon exchange in the BOREAS study region. Tree Physiol. 17, 589–599.

Kosko, B., 1992. Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence. Prentice-Hall, Englewood Cliffs, New Jersey, p. 449.

Landsberg, J.J., Kaufmann, M.R., Binckley, D., Isebrands, J., Jarvis, P G., 1991. Evaluating progress toward closed forest models based on fluxes of carbon, water and nutrients. Tree Physiol 9, 1–15.

Lee, X., 1998. On micrometeorological observations of surface-air exchange over tall vegetation. Agric For Meteorol. 2545, 1–11.

Makkink, G F., 1957. Testing the Penman formula by means of lysimeters. J. Int. Water Eng. 11, 277–288

Schaap, M.G., Bouten, W., 1996. Modelling water retention curves of sandy soils using neural networks. Water Resour Res. 32 (10), 3033–3040.

Stewart, J.B., 1988. Modelling surface conductance of pine forest Agric. For. Meteorol. 43, 19–35.

Tenhunen, J.D , Valentini, R., Kostner, B., Zimmermann, R., Granier, A., 1998 Variation in forest gas exchange at landscape to continental scales. Annal. des Sci. For. 55 (1-2), 1–11.

Wang, Y.P., Jarvis, P.G., 1990. Description and validation of an array model-MAESTRO. Agric. For. Meteorol 51, 257–280.

Williams, M., Rastetter, E.B., Fernandes, D.N., Goulden, M.L., Shaver, G.R., Johnson, L.C., 1997. Predicting gross primary productivity in terrestrial ecosystems. Ecol. Appl. 7 (3), 882–894.

ELSEVIER

Ecological Modelling 120 (1999) 199–211

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artificial neural networks

Rosa M. Barciela [a,*], Emilio García [b], Emilio Fernández [a]

[a] *Departamento de Ecoloxía e Bioloxía Animal, Universidad de Vigo, Facultad de Ciencias, Campus Lagoas-Marcosende s/n, E-36200, Vigo, Spain*

[b] *Departamento de Linguaxes e Sistemas Informáticos, Universidad de Vigo, Facultad de Ciencias, Campus Lagoas-Marcosende s/n, E-36200, Vigo, Spain*

Abstract

Two modelling approaches, dynamic ecological simulation and neural network analysis, were used to describe and predict the main patterns of primary production temporal variability in a coastal embayment affected by upwelling (Ria de Arousa, Western Spanish Coast). A one dimensional, carbon based, size-dependent dynamic simulation model physically forced by solar radiation, temperature, upwelling index and mixed layer depth was developed using object-oriented programming. The model is defined by six biological compartments: nanophytoplankton, microphytoplankton, microzooplankton, mesozooplankton, bacteria and cultured mussels, was tuned with a 3-year data series (1992–1994) from the region and validated using data collected in the same area in 1995 and 1996. The model reproduces both seasonal and interannual patterns and magnitudes of nutrient concentration and phytoplankton biomass. Neural network models were also developed using backpropagation networks with one or two hidden layers and sigmoid and sinusoidal activation functions. The correlation between observed and modelled phytoplankton biomass from 1992 to 1994 were 0.99 and 0.71 for daily and weekly predictions, respectively. Both modelling approaches yield valuable information. The dynamic simulation model contributes to a better understanding of cycling of matter through planktonic food webs but, although reproducing the main patterns of large-scale variability, its predictive potential is low due to the large uncertainty associated with parameter estimation. By contrast, the neural network model, although not providing information on ecosystem functioning, has demonstrated to be a powerful predictive tool for short (daily to weekly) time scales.

Keywords: Ecosystem model; Neural network; Primary production; Temporal variability; Ria de Arousa

1. Introduction

Diverse modelling approaches have been developed in order to gain understanding on the dynamics of planktonic marine ecosystems.

* Corresponding author Present address: James Rennell Division, Southampton Oceanography Centre, European Way, Empress Dock, Southampton, SO14 3ZH, UK. Fax: +44-1703-596400.

E-mail address: rmba@mail soc.soton.ac.uk (R.M. Barciela)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00102-7

Probably, the most intensively adopted approach was based upon a compartmental structure, each compartment representing a trophic level or taxonomic group and the interactions expressed by the different flows occurring among them (Odum, 1971). These models evolved from simple ones, only considering nutrients—phytoplankton–zooplankton interactions (Steele, 1974), to more complex ones, where components such as dissolved organic matter, detritus, bacteria are taken into account and size-based models are implemented (e.g. Fasham et al., 1990; Moloney and Field, 1991; Baretta et al., 1995).

An alternative modelling approach, based on neural network analysis, is currently in progress as a valuable predictive tool in ecological sciences (e.g. Lek et al., 1996; Scardi, 1996). The most important conceptual advantage of neural networks over conventional dynamic ecological models is probably the possibility of collating heterogeneous information in a single computational framework, even though no theoretical guidelines were provided. Artificial neural network systems are known for their capacity to process nonlinear relationships (Hornick et al., 1989; Chen et al., 1990) especially for regressions (Specht, 1991). In this regard, networks with at least one hidden layer can accurately model nonlinear systems even though the underlying casual links were unknown or not fully understood. Neural networks thus represent an approach for predicting mass of compartments from environmental variables, although not providing any insight on its functioning.

The aim of this investigation was to model primary production in an embayment affected by upwelling (Ria de Arousa, NW Spain) in order to predict alterations in carbon incorporation rates by phytoplankton in response to changing environmental conditions by using two different modelling approaches: a dynamic ecosystem model and neural network analysis. The studied region is affected by a wind-driven upwelling where northerly winds prevail from May to October (Blanton et al., 1987), giving rise to the input of dissolved inorganic nutrients into the photic zone and, thereby, to enhanced primary production rates (Àlvarez-Salgado et al., 1996) that are the basis of massive culturing of rafted mussels. As a result, a considerable amount of investigation has been carried out in the region (e.g. Fraga and Margalef, 1979; Tenore et al., 1982; Varela et al., 1984; Hanson et al., 1986; Penas and Varela, 1986; Àlvarez-salgado et al., 1996; Rosón et al., 1997; Zdanowski and Figueiras, 1997, among others). The only biological modelling investigation previously attempted in the Iberian Shelf was developed in the past decade by Penas and Varela (1986). This model, however, was not formulated in quantitative terms and therefore, could not be validated.

2. Dynamic ecosystem model

In this study, a 1D, carbon based, size-dependent compartmental model is presented. The model is physically forced by solar radiation, temperature, vertical advection and mixed layer depth and has two layers: the upper mixed layer and the bottom layer. The upper mixed layer receives nutrients from the bottom layer by physical processes such as vertical advection, mixing and diffusion. In the bottom layer, only the temporal variation in nitrogen concentration was considered. Incoming solar radiation was calculated using equations dependent on declination, latitude, time and atmospheric and oceanic albedo (Peixoto and Oort, 1992). Photosynthetically available radiation (PAR) was calculated as in Baker and Frouin (1987). Vertical extinction coefficients were calculated according to Taylor et al. (1991).

Sea-truth temperature data were used in the model. Upwelling indexes were calculated as in Bakun (1973) and mixed layer depth, defined as the depth where the thermal gradient exceeded 0.2°C m^{-1}, estimated from vertical thermal distributions (Fig. 1).

The abiotic compartments of the model included dissolved inorganic nitrogen (*DIN*) and labile dissolved organic carbon (*LDOC*). Six biological compartments were defined: bacteria (B) (0.2–2 μm); nanophytoplankton (P_N) (<20 μm); microphytoplankton (P_M) (>20 μm); microzooplankton (Z_{Mc}) (<200 μm) and meso-

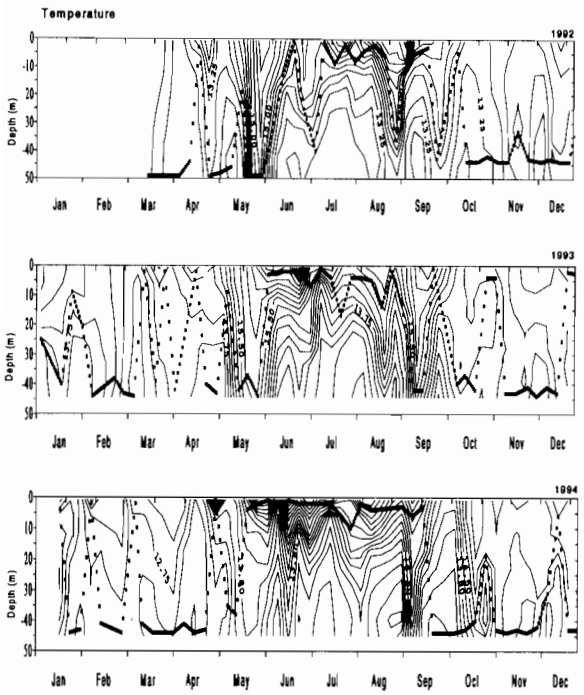

Fig. 1. Temperature (isopleths) and mixed layer depth (dots) variability from 1992 to 1994.

zooplankton (Z_{Me}) (> 200 μm). Cultured mussels (M) were also included in the heterotrophic compartment due to their relevance for nitrogen circulation in the area (Fig. 2).

The model simulates the behavior of the planktonic ecosystem where phytoplankton growth depends on nitrogen, considered as the sum of both nitrate and ammonium; bacterioplankton activity is limited by labile dissolved organic carbon, mesozooplankton consumption is limited by microphytoplankton biomass, microzooplankton by nanophytoplankton and bacteria, and cultured mussels growth is based on particulate organic matter, although they appear to be more efficient filtering phytoplankton than other forms of particulate organic carbon (Cabanas et al., 1979).

2.1. Dissolved inorganic nitrogen

Nitrogen is generally regarded as the limiting nutrient for primary production, and therefore, the ability to model seasonal concentration changes is an essential prerequisite for understanding carbon cycling in the ocean (Fasham et al., 1990). The temporal variation of dissolved inorganic nitrogen (*DIN*) concentration was modelled according to the equation:

$$dDIN/dt = DIN_{adv} + DIN_{dif} + DIN_{reg} - DIN_{phy} \quad (1)$$

where DIN_{adv} represents the increase in nitrogen concentration in the mixed layer due to vertical advection, DIN_{dif} due to turbulent vertical diffu-

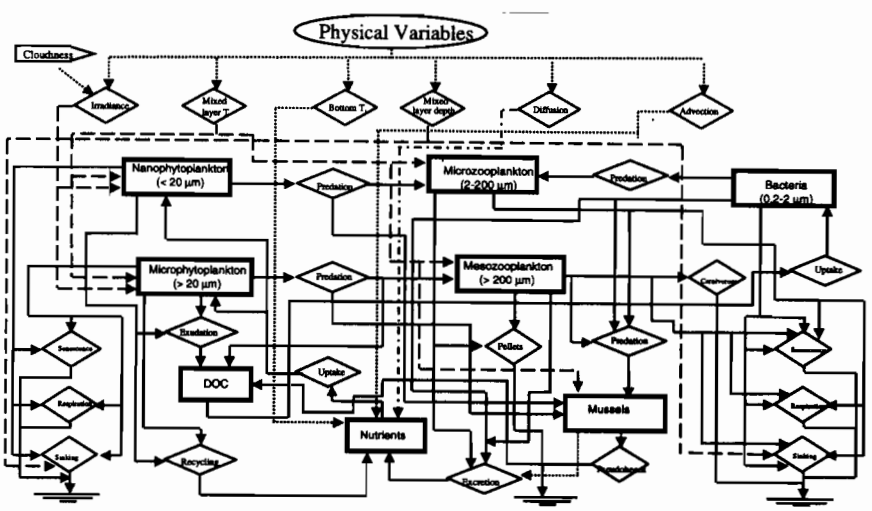

Fig. 2. Diagram of the dynamic ecosystem model.

sion and DIN_{reg} is the biological regeneration term. DIN_{adv} was calculated from upwelling indexes (Bakun, 1973):

$$DIN_{adv} = (1 - V_{up})DIN + V_{up}DIN_f \quad (2)$$

where V_{up} represents the amount of water upwelled in the region and DIN_f nitrogen concentration in the bottom layer. DIN_f was calculated from a function obtained using data collected from a 9 year time series study carried out in the Ria de Vigo. The terms of the function are Fourier terms which depend on total inorganic nitrogen concentration in the bottom layer and bottom temperature (Nogueira, 1998). DIN_{dif} was calculated in several steps. Firstly, vertical mixing (V_m) was calculated as in Taylor et al. (1991) and used to derive vertical mixing coefficients (K_z) according to the expression suggested by Àlvarez-Salgado et al. (1996). Turbulent diffusion rates (M_z) were then calculated from the expression:

$$M_z = [(2K_z area)/(zVol)] \quad (3)$$

where *area* refers to the extension of the studied zone in m^2, z is the average depth of the area and *Vol*, is the volume of the area considered. M_z was used to calculate the amount of nutrients entering the upper mixed layer (DIN_{dif}) as in Àlvarez-Salgado et al. (1996):

$$DIN_{dif} = M_z(DIN - DIN_f) \quad (4)$$

DIN_{reg} was calculated as:

$$DIN_{reg} = DIN_{re} + DIN_z + DIN_B + DIN_M \quad (5)$$

where DIN_{re} refers to the amount of nitrogen remineralized, DIN_z is the amount of nutrients released by zooplankton which is remineralized in the mixed layer, and DIN_g and DIN_M nutrients released by bacteria and mussels, respectively.

DIN_{phy} was calculated as:

$$DIN_{phy} = \mu_{max}\mu_{DIN} \quad (6)$$

Where μ_{max} is the phytoplankton maximum growth rate (see Eq. (11) in the text) and μ_{DIN} represents nitrogen uptake by phytoplankton (see Eq. (12) in the text).

2.2. Autotrophic module

Modelled phytoplankton growth was dependent on temperature, irradiance and nitrogen availability. The losses considered were natural mortality, sinking, respiration, exudation, lysis, zooplankton grazing and mussels filtration rates. The equations for nanophytoplankton (P_N) and microphytoplankton (P_M) growth were:

$$dP_N/dt = P_N(\mu_p - MRP - SRP - RRP - ERP - LRP - GRZ_{MC} - FMR_p) \quad (7)$$

$$dP_M/dt = P_M(\mu_p - MRP - SRP - RRP - ERP - LRP - GRZ_{Me} - FMR_p) \quad (8)$$

where μ_p is the daily phytoplankton specific growth rate expressed as in Taylor et al. (1991):

$$\mu_p = \mu_{IT}\,\mu_{DIN} \quad (9)$$

$$\mu_{IT} = \mu_{max}\,10^{[(T - T_{max})\log(2),10]}(I/(I + I_h)) \quad (10)$$

where μ_{max} was defined as in Aksnes et al. (1995):

$$\mu_{max} = \mu_{max_0}\exp[(\ln Q10/10)T] \quad (11)$$

and

$$\mu_{DIN} = DIN/(DIN + DIN_h) \quad (12)$$

μ_{max_0} is the maximum phytoplankton growth rate at 0°C, $Q10$ the temperature rate constant, T and T_{max} are the daily temperature and annual temperature maximum, I is the daily irradiance, I_h is the irradiance half- saturation constant, DIN is the nitrogen concentration and DIN_h is the nitrogen half-saturation constant. Natural mortality rate, MRP, was modelled as a function dependent on nutrient concentration (Raillard and Ménesguem, 1994):

$$MRP = MR_{min}\mu_{DIN} + MR_{max}(1 - \mu_{DIN}) \quad (13)$$

where MR_{min} and MR_{max} are minimum and maximum natural and daily mortality rates.

Phytoplankton sinking rate, SRP, was calculated as suggested by Taylor et al. (1991), and Gamier et al. (1995):

$$SRP = (SCP/Pc) \quad (14)$$

where SCP represents the sinking coefficient and Pc is the mixed layer depth. Respiration rate,

RRP, was considered as a constant fraction of phytoplankton growth, whereas exudation rate, *ERP*, was modelled as in Fasham et al. (1990):

$$ERP = OMP\mu_p \quad (15)$$

where *OMP* represents the percentage of photosynthetically incorporated carbon released as dissolved organic matter.

It was assumed that lysis rate, *LRP*, is proportional to the difference between maximal and actual growth rate (Baretta et al. 1988; Varela et al. 1995):

$$LRP = LRP_{DIN}\mu_{IT}(1 - \mu_{DIN}) \quad (16)$$

where LRP_{DIN} is the nitrogen dependent lysis rate. Zooplankton consumption on phytoplankton was modelled using a Michaelis–Menten approach (Fasham et al., 1990; Moloney and Field, 1991):

$$GRZ_{MC} = GR_{max}[P_N/(P_N + P_{Nh})] \quad (17)$$

$$GRZ_{Me} = GR_{max}[P_M/(P_M + P_{Mh})] \quad (18)$$

where GRZ_{MC} and GRZ_{Me} are microzooplankton and mesozooplankton grazing rates, respectively and GR_{max} represents the maximum rate of phytoplankton consumption (Kremer and Nixon, 1978):

$$GR_{max} = GR_0 \exp\{[(\ln Q10)/10]T\} \quad (19)$$

GR_0 is the maximum predation rate at 0°C and P_{Mh} and P_{Nh} are microphytoplankton and nanophytoplankton half-saturation constants, respectively.

Filtration rate by mussels, *FRM*p, was modelled as a logarithmic function dependent on temperature (as constants values change according to mixed layer temperature) and mussels size (*L*) as suggested by Peréz Camacho and González (1984).

$$FRMp = a + b \log L \quad (20)$$

The dynamic ecosystem model uses variable phytoplankton chlorophyll:carbon and nitrogen:carbon ratios as suggested by Cloern et al. (1995) and Ietswaart and Flynn (1995), respectively.

2.3. Heterotrophic module

The bacterial component of the model is assumed to be formed by free-living organisms that take up labile *DOC* (*LDOC*) and release ammonium. Other losses considered were natural mortality, sinking, respiration, zooplankton predation and bacterial filtration rates by mussels. The bacterial equation can be written as:

$$dB/dt = B(\mu_B - MRB - RRB - GRZ_{McB} - BF_{Mu} - ERB - SRB \quad (21)$$

where μ_B represents bacterial daily growth rate, *MRB* natural mortality rate, *RRB* respiration rate, GRZ_{MCB} grazing rate by microzooplankton, BF_{Mu} bacterial filtration rates by mussels, *ERB* nitrogen excretion rate and *SRB* sinking rate.

Microzooplankton (Z_{Mc}) and mesozooplankton (Z_{Me}) and growth was dependent on grazing rates on phytoplankton while the losses considered were respiration, predation by carnivorous zooplankton, filtration by mussels, excretion, fecal pellets sedimentation and sinking.

Zooplankton growth was modelled according to the following equations:

$$dZ_{Mc}/dt = Z_{Mc}(GRZ_{Mc} + GRZ_{McB} - RRZ_{Mc} - FRMu_{Mc} - RAE_{Mc} - RFE_{Mc} - SR_{Mc}) \quad (22)$$

$$dZ_{Mc}/dt = ZMe(GRZ_{Me} - RRZ_{Me} - DOMR_{Me} - CMR - FRMu_{Me} - RAE_{Me} - RFE_{Me} - SR_{Me}) \quad (23)$$

where GRZ_{Me}, GRZ_{MC} and GRZ_{McB} are zooplankton grazing rates on microphytoplankton, nanophytoplankton and bacteria respectively, *RRZ* is the respiration rate and *DOMR* represents dissolved organic matter released by sloppy feeding, *CMR* is a constant predation rate by carnivorous zooplankton, *FRMu* is the filtration rate due to mussels, *RAE* is the rate of ammonium excretion, *RFE* is the excretion rate of fecal pellets and *SR* zooplankton sinking rate.

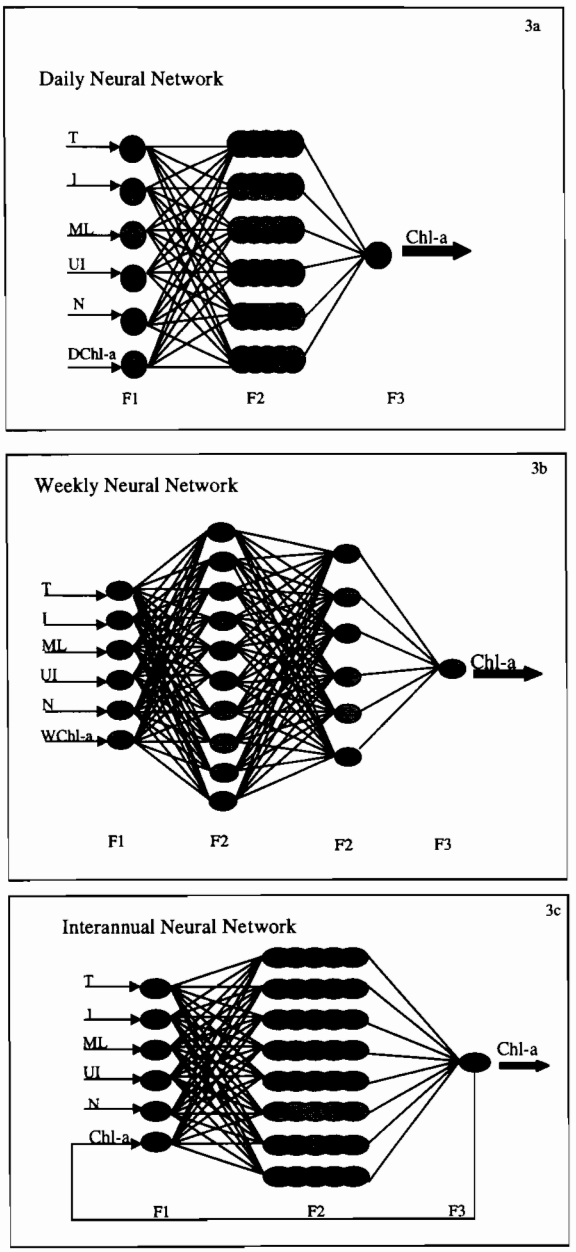

Fig. 3. Structure of the neural network used in this study T, temperature; I, irradiance, ML, mixed layer depth; UI, upwelling index, N, nitrogen; DChl-*a*, daily chlorophyll-*a*; WChl-*a*, weekly chlorophyll-*a*; Chl-*a*, chlorophyll-*a*. F1, input layer of neurons comprising as many neurons as variables at the entry of the system; F2, hidden number of neurons whose number is determined empirically; F3, output layer of neurons with a single neuron corresponding to the single dependent variable.

2.4. Tuning and validation

Tuning and validation of the model were undertaken using data collected weekly at a station located in the Ria de Arousa (42°29′86″N and 08°58′81″W) from May 1992 to December 1996. The different parameters in the equations were chosen from the literature for obtaining the best fit to a set of 3-year empirical data, corresponding to the 1992–1994 period, which was the model tuning. Then, the model simulation was performed by running the model for another 2 years. The data obtained from the simulation were compared with a set of 2-year empirical data corresponding to the 1995–1996 period for the model validation to be carried out.

Upwelling indexes were calculated for Cape Finisterre according to Bakun (1973). Cloudiness was obtained from Vigo airport Meteorological station. Vertical profiles of temperature and salinity were recorded with a SBE-25 CTD. Sigma-t was calculated according to UNESCO (1983). Water samples where collected at 5, 10 and 15 m for determining nutrient concentration (nitrate, nitrite, and ammonium) using an autoanalyzer and chlorophyll-*a* was measured by the spectrophotometric method of Neveux and Panouna (1987).

3. The artificial neural network model

An error back-propagation neural network (Rumelhart et al., 1986) was developed in order to match the same 3-year data series of chlorophyll-*a* used for tuning the dynamic ecosystem model (1992–1994). Three different networks were trained with the aim of obtaining the best fit to the empirical data for different time scales (daily, weekly, seasonal). Five variables were selected for the daily and weekly network: temperature, irradiance, mixed layer depth, upwelling index and nutrients, while phytoplankton feedback was also considered for the annual network.

The number of hidden layers and their neurons were selected by comparing the perfor-

mance of different networks. The daily neural network (Fig. 3a) consisted of six neurons in the input layer (coding the five variables of the environment and phytoplankton biomass from the previous day) and 30 neurons in the hidden layer that had a sigmoid activation function. The weekly neural network (Fig. 3b) was formed by six neurons in the input layer (coding the five variables of the environment and phytoplankton biomass corresponding to the week before) and two hidden layers with ten and six neurons each. The ten-neuron layer and the six-neuron layer had a sinusoidal activation function and a sigmoid activation function, respectively. The annual neural network (Fig. 3c) consisted of six neurons in the input layer (coding the five variables of the environment and phytoplankton feedback) and 40 neurons in the hidden layer that used a sigmoid activation function. All the neural networks had only one neuron in the output layer representing phytoplankton biomass (Lek et al., 1996) and used sigmoid activation functions.

Different learning rates (η) were used. At the beginning, the neural networks used a learning rate (η) of one until no further improvement was observed between neural network results and real data, then the learning rate was reduced to 0.2 in order to gain a better approximation to real data. Momentum terms were not considered. As the neural networks have sigmoid activation functions, data were scaled by dividing them by arbitrary maximum values slightly larger than the maximum observed values.

Training of neural networks was carried out according to the following procedure. Firstly, all summer patterns corresponding to the 1992 data base were selected and randomly introduced into the networks. Winter data were not used due to the noise added by the presence of poleward slope currents in the region (see Section 4 below). After each training cycle or epoch, correlation coefficients between network output and measured data were calculated in order to select the synaptic weights between nodes providing the best fit for each case.

4. Results and discussion

4.1. Dynamic ecosystem model

The annual variation of in situ temperature and mixed layer depth at the sampling station during the 1992–1994 period is shown in Fig. 1. The water column was vertically mixed from September–October to May–June, whereas during the rest of the year, thermal stratification prevailed. In accordance with this pattern, measured dissolved inorganic nitrogen concentration was low during summer and increased progressively as vertical mixing became established (Fig. 4). Modelled nutrients reproduced reasonably well the seasonal trends obtained at sea, being the correlation between both sets of data rather similar for the 3 years studied ($r_{1992}=0.63$, $n=31$; $r_{1993}=0.63$, $n=47$; $r_{1994}=0.57$, $n=45$). The best fits were found in Autumn 1992 and 1993 ($r_{1992}=0.81$, $n=13$; $r_{1993}=0.71$, $n=12$). By contrast, the model significantly overestimated nitrogen concentrations in winter and spring, specially in 1993. This major disagreement is likely to be a consequence of the presence of warm and saline water of subtropical origin on the shelf in these seasons (Frouin et al., 1990; Pingree and Le Cann, 1992), that would eventually enter into the modelled area, a physical process not considered in the formulation of the model. In support of this idea, average winter values of temperature and salinity were higher in 1993 as compared to 1994 (13.4°C and 35.6 in 1993 vs. 12.6°C and 35.2 in 1994). Furthermore, remotely sensed observations of sea surface temperature showed that the pool of warm and salty water flowing poleward was specially intense in 1993 (Pingree, pers. com.). It should be also taken into account that freshwater inflows could be important during this part of the year.

The model reproduces the main patterns of seasonal variability in chlorophyll *a* concentration, as well as the timing of phytoplankton blooms (Fig. 4). The correlation coefficients calculated for 1992, 1993 and 1994 where $r_{1992}=0.30$, $n=31$; $r_{1993}=0.32$, $n=47$; $r_{1994}=0.48$, $n=45$, respectively. The best seasonal fits were observed in summer with correlation coefficients ranging

from $r = 0.64$, $n = 13$ in 1992 to $r = 0.80$, $n = 11$ in 1994. Significant disagreements between real and modelled chlorophyll a concentrations are likely to be explained by the same arguments mentioned for nitrogen, given the strong dependence of phytoplankton growth on nutrient availability. The dynamic ecosystem model yielded daily values of phytoplankton carbon biomass, thus enabling annual rates of primary production to be calculated. These rates ranged from 345 to 606 gC m^{-2} y^{-1}, values of the same magnitude although slightly higher than those previously reported in the literature. Thus, Varela et al. (1984) measured rates of 250 gC m^{-2} y^{-1} in the same area, Fraga (1976) and Prego (1993) estimated carbon incorporation rates of 260 and 350 gC m^{-2} y^{-1} for the Ria de Vigo, and Casas (1995) of 300 gC m^{-2} y^{-1} for the Bay of A Coruña. The relatively high values generated by the model are, however, likely to be typical of the studied area. The lack of primary production data is notorious in the region and, when they exist, do not take into consideration neither interannual nor spatial variability. Moreover, annual primary production data sets available in the literature derive from the integration of monthly observations whereas modelled rates were calculated from hourly values of phytoplankton biomass. Thus, relevant differences are expected to take place as a result of the different time scales involved.

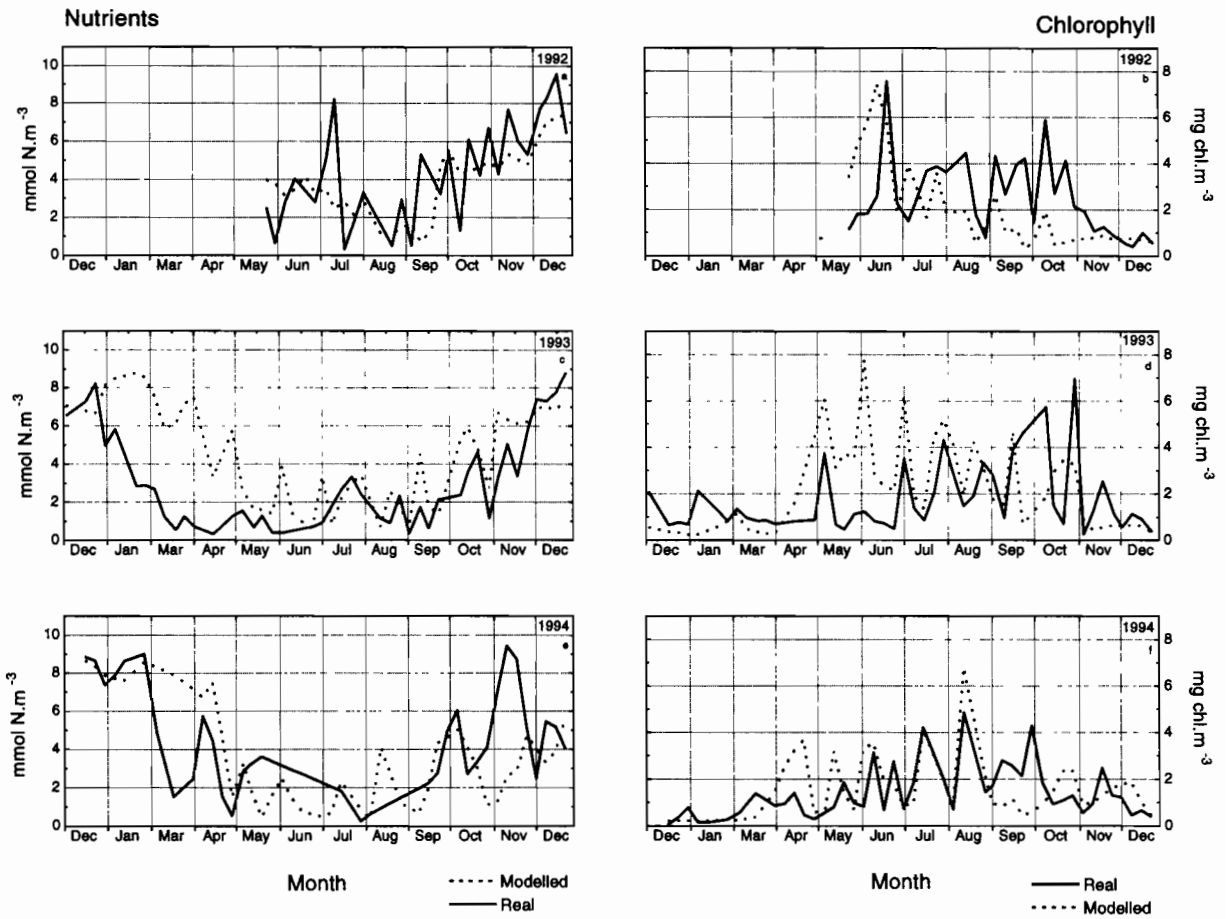

Fig. 4. Results from the dynamic ecosystem model (a, c, e) Real and modelled nutrients in 1992, 1993 and 1994, respectively. (b, d, f) Real and modelled chlorophyll in 1992, 1993 and 1994, respectively.

Table 1
Sensitivity analysis[a]

		std	+5%	−5%	+10	−10%	+20	−20
Nanophytoplankton								
μ_{max_0}	1992	91	25	−24	**b**	−47	**101**	**−56**
	1993	81	42	−21	**179**	−35	**189**	**−50**
	1994	56	11	−7	**405**	−12	**65**	−24
RRD	1992	91	−17	21	−37	**47**	**−51**	**111**
	1993	81	−18	32	−29	**77**	−44	**182**
	1994	56	−7	10	−12	23	−21	**63**
I_h	1992	91	−3	4	−8	10	−14	−19
	1993	81	−3	4	−7	10	−12	19
	1994	56	−1	1	−2	3	−4	5
DIN_h	1992	91	−1	1	−2	2	−4	4
	1993	81	−1	2	−3	4	−6	9
	1994	56	0	0	0	1	−1	2
OMP	1992	91	0	0	0	0	0	0
	1993	81	1	1	1	1	1	1
	1994	56	0	0	0	0	0	0
Microphytoplankton								
μ_{max_0}	1992	515	−5	3	**−72**	7	−21	12
	1993	352	−13	9	**−94**	13	**−62**	18
	1994	289	−3	1	**−98**	1	−9	4
RRD	1992	515	2	−4	5	−7	8	−18
	1993	352	6	−9	12	−24	15	**−50**
	1994	289	0	−3	1	−3	1	−9
I_h	1992	515	0	−1	0	−2	1	−3
	1993	352	1	−2	3	−4	5	−7
	1994	289	−1	−1	0	−1	0	−2
DIN_h	1992	515	0	−1	0	−1	0	−1
	1993	352	1	−1	1	−2	2	−4
	1994	289	−1	−1	−1	−1	−1	−1
OMP	1992	515	0	0	0	0	0	0
	1993	352	0	0	0	0	0	0
	1994	289	−1	−1	−1	−1	−1	−1

[a] Standard values (std) are shown in gC m^{-2} y^{-1}, all other values are presented as a percentage of variation in relation to the standard parameter value. Differences larger than 50% with respect to the standard run appear in boldface.

The dynamic ecosystem model was run iteratively in order to test its sensitivity to changes in parameter values. A total of five parameters were chosen: nitrogen half-saturation constant (DIN_h), irradiance half-saturation constant (I_h), respiration rate (RRP), maximum phytoplankton growth rate at 0°C (μ_{max_0}) and the percentage of photosynthetically incorporated carbon released as dissolved organic matter (OMP). The standard values, chosen from the literature and used for tuning the model, were increased or decreased by 5, 10 and 20% and the effect on annual nano- and microphytoplankton primary production rates evaluated as shown in Table 1. Changes in μ_{max_0} strongly affect model output. Low values of this parameter causes sharp reductions in nanophytoplankton biomass whereas microphytoplankton production displayed a slight increase. Enhanced μ_{max_0} induced drastic increases in nanophytoplankton production of up to 100–400%. Modification of respiration rates also generated significant changes in phytoplankton biomass. Thus, increases in RRP gave rise to reduction or increase of primary production depending on phytoplankton size. Changes in the Michaelis–Menten limitation terms or in OMP induced

hardly noticeable effects on phytoplankton annual production rates.

Validation of the model was performed using a 2-year (1995 and 1996) nutrient and chlorophyll data set collected at the same station mentioned above (Fig. 5). The model reproduced the main seasonal trends in dissolved inorganic nitrogen and chlorophyll concentrations. The seasonal correlation between modelled and sea-truth nitrogen concentrations was rather low during winter ($r_{1995} = 0.10$, $n = 10$; $r_{1996} = 0.27$, $n = 8$), due to model limitations discussed above. The highest correlation coefficients for nutrients were obtained in spring 1995 ($r = 0.84$, $n = 11$) and autumn 1996 ($r = 0.81$, $n = 12$), whereas for chlorophyll, the best fit was observed in winter and autumn 1995 ($r = 0.84$, $n = 11$; $r = 91$, $n = 12$, respectively).

4.2. Neural networks

The results generated by the neural networks, as well as measured chlorophyll values are shown in Fig. 6. The daily neural network yielded excellent results as it was able to reproduce the observed values over the whole temporal period considered, as shown by the very high correlation coefficients presented in Table 2. The weekly neural network also showed good agreement between observed and modelled chlorophyll-*a* values, although not reproduced accurately the increase in chlorophyll measured in summer 1994. The annual network provided the lowest interannual correlation coefficient.

5. Conclusion

The prediction power of the two model types tested, dynamic ecosystem models (DEM) and neural networks (NNM) was determined by calculating interannual and seasonal correlation coefficients between observed and modelled values. The dynamic ecosystem model provided a poor prediction of real primary production whereas the neural networks developed showed a better

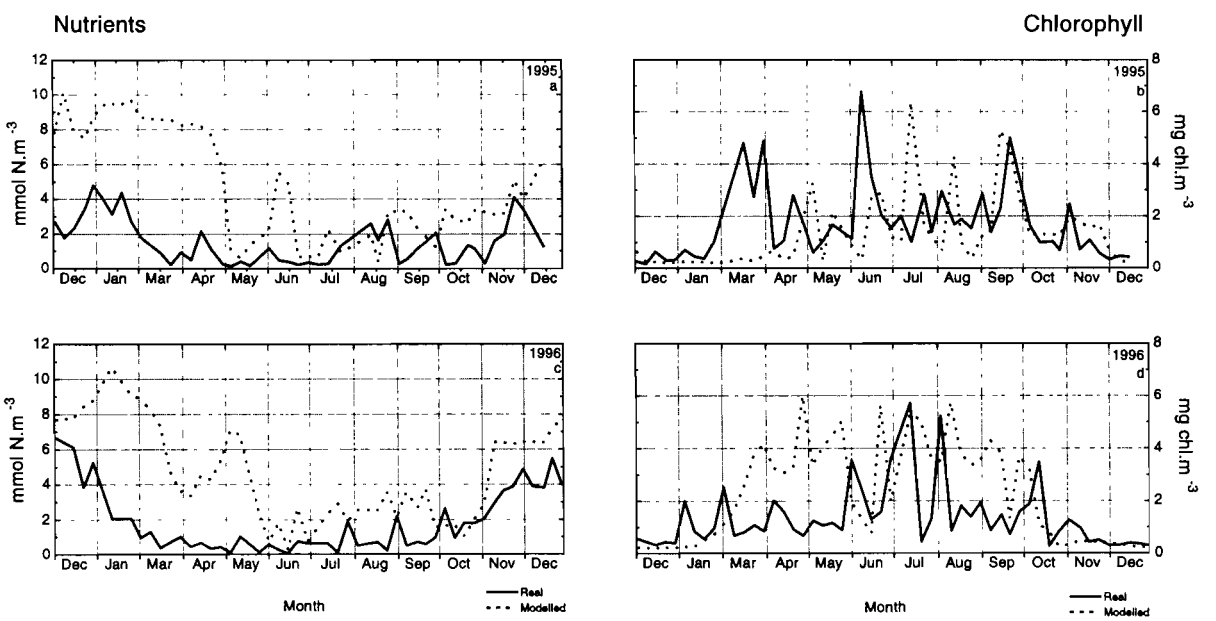

Fig. 5 Dynamic ecosystem model validation (a, c) Real and modelled nutrients in 1995 and 1996, respectively. (b, d) Real and modelled chlorophyll in 1995 and 1996, respectively.

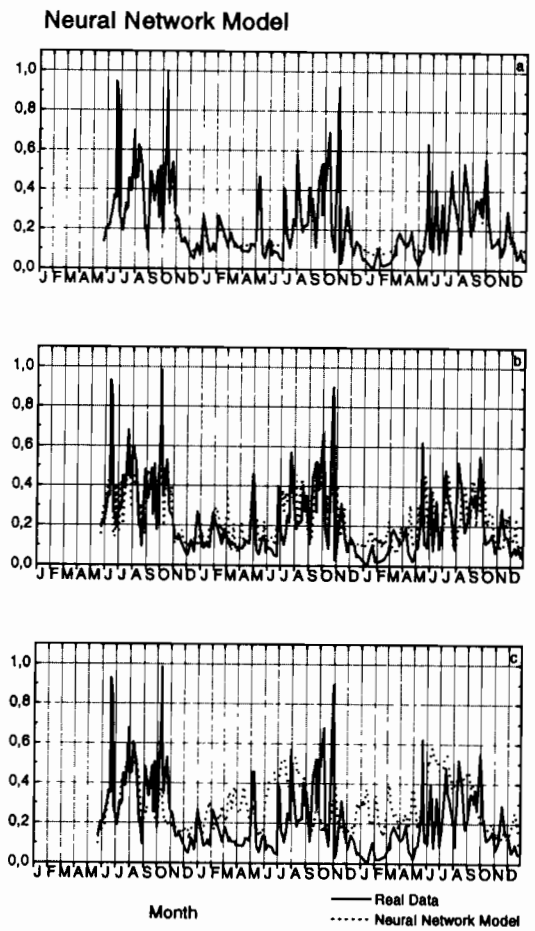

Fig. 6. Neural network model results from 1992 to 1994. (a) Daily neural network; (b) weekly neural network; (c) annual neural network.

performance ($r_{\text{interannual}} = 0.36$ vs. $r_{\text{interannual}} = 0.99$, $r_{\text{interannual}} = 0.71$, $r_{\text{interannual}} = 0.32$ for the daily, weekly and annual neural networks, respectively). In conclusion, the results emerging from this investigation indicate that dynamic simulation models contribute to a better understanding of cycling of matter through planktonic food webs but, although reproducing the main patterns of large-scale variability, its predictive potential is low due to the large uncertainty associated with parameter estimation. This uncertainty is due to the difficulties in estimating parameter values. Some parameters, such as phytoplankton and bacterial growth rates or zooplankton grazing and excretion rates, can be measured experimentally at sea. Others, including phytoplankton natural mortality rate are, at present, almost impossible to measure accurately. So, the approach of most modellers to this problem has been to use experimentally determined parameter values where available or values chosen among a set of determined parameter values that provide a good agreement between the real and the observation sets of data. By contrast, the neural network model, although not providing information on ecosystem functioning, can be successfully employed for modelling very complex and non-linear ecological phenomena and has demonstrated to be a powerful predictive tool for short (daily to weekly) time scales.

Table 2
Correlation coefficients between chlorophyll-*a* concentrations resulting from the different types of neural networks developed and real chlorophyll-*a* values

	Type of neural network		
	Daily neural network	Weekly neural network	Annual neural network
$r_{\text{interannual}}$	0.99	0.71	0.32
r_{summer92}	0.99	0.72	0.76
r_{summer93}	0.99	0.64	0 37
r_{summer94}	0.99	0.05	0.23

References

Aksnes, D.L., Ulvestad, K.B., Balino, B.M., Berntsen, J., Egge, J.K., Svendsen, E., 1995. Ecological modelling in coastal waters: towards a predictive physical-chemical-biological simulation model Ophelia 41, 5–36.

Àlvarez-Salgado, X.A., Rosón, G., Pérez, F.F., Figueiras, F.G., Pazos, Y., 1996. Nitrogen cycling in an estuarine upwelling system, the Ría de Arousa (NW Spain). I. Short-time-scale patterns of hydrodynamic and biogeochemical circulation. Mar. Ecol. Prog. Ser. 135, 259–273.

Baker, K S., Frouin, R., 1987. Relation between photosynthetically available radiation and total insolation at the ocean surface under clear skies. Limmnol. Oceanogr. 32, 1370–1377.

Bakun, A., 1973. Coastal Upwelling Indices, West coast of North America. 1946–1971, NOAA, Technical Report NMSSSRF-671. US Department of Commerce, pp. 103.

Baretta, J.W., Admiraal, W., Colijn, F., Malschaert, J.F.P., Ruardij, P., 1988. The construction of the pelagic submodel In: Baretta, J.W., Ruardij, P. (Eds.), Tidal Flat Estuaries. Simulation and Analysis of the EMS Estuary. In: Ecology Studies, vol. 71. Springer-Verlag, Heidelberg, p. 353.

Baretta, J.W., Ebenhh, W., Ruardij, P., 1995. The European regional seas ecosystem model, a complex marine ecosystem model. Neth. J. Sea. Res. 33 (3/4), 233–246.

Blanton, J.O., Tenore, K.R., Castillejo, F.F., Atkinson, L.P., Schwing, F.B., Lavín, A., 1987. The relationship of upwelling to mussel production in the rías of the western coast of Spain. J. Mar Res 45, 497–511.

Cabanas, J.M., González, J.J., Mariño, J., Pérez, A., Román, G., 1979. Estudio del mejillón y de su epifauna en los cultivos flotantes de la Ría de Arosa III. Observaciones previas sobre la retención de partículas y la biodeposición de una batea. Boll Inst. Esp. Oceanogr. 5, 45–50.

Casas, B., 1995. Composición, biomasa y producción del fitoplancton en la costa de la Coruña: 1989–1992. Tesis Doctoral, Santiago de Compostela, pp 340

Chen, S., Billings, S.A., Grant, P.M., 1990. Non-linear system identification using neural networks. Int. J. Control 51, 1191–1214.

Cloern, J.E., Grenz, C., Vidergar-Lucas, L., 1995. An empirical model of the phytoplankton chlorophyll:carbon ratio—the conversion factor between productivity and growth rate Limnol. Oceanogr.– 40, 1313–1321.

Fasham, M J R., Ducklow, H.W., McKelvie, S.M., 1990. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. J. Mar Res. 48, 591–639.

Fraga, F. and Margalef, R., 1979. Las Rías Gallegas. In: Estudio y explotación del mar en Galicia. Universidad de Santiago de Compostela, pp. 101–121.

Fraga, F., 1976. Fotosíntesis en la Ría de Vigo Inv. Pesq. 40, 151–167.

Frouin, R., Fiúza, A.F.G., Ambar, I., Boyd, T.J., 1990. Observations of a poleward surface current off the coasts of Portugal and Spain during winter. J. Geophys. Res. 95 (C1), 679–691.

Gamier, J., Billen, G., Coste, M., 1995. Seasonal succession of diatoms and *Chlorophyceae* in the drainage network of the Seine River: observations and modeling. Limnol Oceanogr. 40, 750–765.

Hanson, R.B., Àlvarez-Ossorio, M.T., Cal, R., Campos, M.J., Román, M., Santiago, G., Varela, M., Yoder, J.A., 1986. Plankton response following a spring upwelling event in Ría de Arosa, Spain. Mar Ecol. Prog. Ser. 32, 101–113.

Hornick, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Ietswaart, Th., Flynn, K.J., 1995. Modelling interactions between phytoplankton and bacteria under nutrient-regenerating conditions. J. Plank. Res. 17, 729–744.

Kremer, J.N., Nixon, S.W., 1978. A Coastal Marine Ecosystem. Springer-Verlag, Berlin, p. 217.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living. Resour. 9, 23–29.

Moloney, C L., Field, J.G., 1991. The size-based dynamics of plankton food webs. I. A simulation model of carbon and nitrogen flows. J. Plank. Res. 13, 1003–1038.

Neveux, J., Panouna, M., 1987 Spectrofluorometric determination of chlorophylls and pheophytins. Arch. Hydrobiol. 109, 567–581.

Nogueira, E., 1998. Análisis y modelado de la variabilidad temporal de las características hidrográficas en la Ría de Vigo. Tesis Doctoral, Universidad de Vigo, pp. 238.

Odum, E.P., 1971. Fundamentals of Ecology. Saunders, Philadelphia, PA, p. 574.

Peixoto, J.P., Oort, A H., 1992. Physics of Climate. American Institute of Physics, New York, p. 520.

Penas, E., Varela, M., 1986. Submodelo de la producción primaria en la plataforma de Galicia. Boll. Inst. Esp. Oceanogr. 3 (1), 111–130.

Peréz Camacho, A., González, R., 1984. La filtración del mejillón (*Mytilus edulis*) en laboratorio. Cuadernos da área de Ciencias Mariñas. Semin. Estudos Galegos 1, 427–437.

Pingree, R.D., Le Cann, B., 1992. Anticyclonic eddy X91 in the Southern Bay of Biscay, May 1991 to February 1992 J. Geophys. Res. 97 (C9), 14353–14367.

Prego, R., 1993 General aspects of carbon biogeochemistry in the ria de Vigo, north western Spain. Geochim. Cosmochim Acta 57, 2041–2052.

Raillard, O., Ménesguem, A., 1994. An ecosystem box model for estimating the carrying capacity of a macrotidal shellfish system. Mar. Ecol. Prog. Ser. 115, 117–130.

Rosón, G., Àlvarez-Salgado, X.A., Pérez, F.F., 1997. A non-stationery box model to determine residual fluxes in a partially mixed estuary, based on both thermohaline properties: application to the Ría de Arousa (NW Spain). Estuar. Coast Shelf. Sci. 44 (3), 249–262.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production Mar. Ecol Prog. Ser. 139, 289–299.

Specht, D.F., 1991. A general regression neural network IEEE Trans. Neural Networks 2, 568–576.

Steele, J., 1974. Spatial heterogeneity and population stability. Nature 248, 83.

Taylor, A.H., Watson, A.J., Ainsworth, M., Robertson, J E., Turner, D.R., 1991. A modelling investigation of the role of phytoplankton in the balance of carbon at the surface of the North Atlantic. Global Biogeochem. Cyc. 5, 151–171

Tenore, K.R., Boyer, L F., Cal, R.M., Corral, J., García-Fernández, C., González, N., GonzálezGurriaran, E., Hanson, R B , Iglesias, J , Krom, M., López-Jamar, E., McClain, J., Pamatmat, M.M., Pérez, A , Rhoads, D.C., De Santiago, G., Tietjen, J., Westrich, J., Windom, H.L., 1982. Coastal upwelling in the Rías Bajas, NW Spain. contrasting the benthic regimes of the Rías de Arosa and de Muros. J. Mar. Res. 40, 701–772.

UNESCO, 1983. Algorithms for computation of fundamental properties of seawater. UNESCO Technical Papers on Marine Science 44, 1–53.

Varela, R A., Cruzado, A., Gabaldón, J.E , 1995. Modelling primary production in the North Sea using the European regional seas ecosystem model. Neth. J. Sea Res. 33 (3/4), 337–361

Varela, M , Fuentes, J.M., Penas, E., Cabanas, J.M., 1984. Producción primaria de las Rías Baixas de Galicia. Cuadernos da área de Ciencias Mariñas. Seminario de Estudios Galegos 1, 173–182.

Zdanowski, M K., Figueiras, F.G., 1997. Relationships between the abundance of bacteria and other biota and the hydrographic variability in the Ría de Vigo, Spain. Mar. Ecol. Prog. Ser. 147, 257–267.

ELSEVIER

Ecological Modelling 120 (1999) 213–223

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Developing an empirical model of phytoplankton primary production: a neural network case study

Michele Scardi [a,*], Lawrence W. Harding Jr. [b]

[a] *Stazione Zoologica 'A. Dohrn' di Napoli, Villa Comunale, 80121 Napoli, Italy*

[b] *University of Maryland, Horn Point Laboratory (UMCES) and Maryland Sea Grant, Box 775, Cambridge, MA 21613, USA*

Abstract

We describe the development of a neural network model for estimating primary production of phytoplankton. Data from an enriched estuary in the eastern United States, Chesapeake Bay, were used to train, validate and test the model. Two error backpropagation multilayer perceptrons were trained: a simpler one (3-5-1) and a more complex one (12-5-1). Both neural networks outperformed conventional empirical models, even though only the latter, which exploits a larger suite of predictive variables, provided truly accurate outputs. The application of this neural network model is thoroughly discussed and the results of a sensitivity analysis are also presented.

Keywords: Artificial neural networks; Empirical models; Phytoplankton; Primary production; Chesapeake Bay

1. Introduction

Estimates of phytoplankton primary production based on empirical models are increasingly used as an alternative to direct data acquisition that can be both expensive and time consuming. This is particularly true in the era of satellite oceanography because remote measurements of ocean color that provide global coverage of phytoplankton biomass can serve as inputs to models that estimate production. Although empirical models of primary production are usually based on simple linear relationships (e.g. Cole and Cloern, 1987), the estimates they provide are reasonably accurate because primary production is largely regulated by variables that are simple to measure, i.e. downwelling irradiance and phytoplankton biomass.

Despite the usefulness of linear relationships for estimating production, other factors that affect photosynthetic carbon assimilation are related to production in a non-linear manner, such as photosynthetic efficiency of the phytoplankton cells. Therefore, more flexible empirical models that are both simple and capable of reproducing these relationships can theoretically play an important role in improving our ability to estimate production.

Conventional models that attempted to address this problem by means of multiple linear regression (e.g. Eppley et al., 1985), or the use of

* Corresponding author. Fax: +39-81-7641355.
E-mail address: mscardi@mclink.it (M. Scardi)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00103-9

semi-analytic formulations (e.g. Balch et al., 1989), did not perform significantly better than much simpler empirical models. An alternative approach, involving the use of neural networks has recently generated significant improvement in estimating production (Scardi, 1996) or other complex non-linear ecological processes (Lek et al., 1996) where sufficient training data were available. Moreover, neural networks are also able to exploit the heterogeneous information that is provided by other variables that may be correlated to primary production on a regional scale only, and to use this information to achieve refinement of primary production estimates.

The first neural network that was trained as an empirical model of phytoplankton primary production (Scardi, 1996) was essentially a toy model, because of the limited number of training patterns. It was developed with a small data set that was reported in a comprehensive study of phytoplankton photosynthesis in Chesapeake and Delaware Bays (Harding et al., 1986). These data were used in initial efforts because the data on pertinent variables were assembled and readily usable, and comparisons with linear models could be made rather easily. The main purpose of that work was to show that a simple error back-propagation neural network had the potential to outperform conventional empirical models of phytoplankton primary production. Since that initial report, further research has been carried out on primary production and ancillary data for Chesapeake Bay spanning over a decade (Harding et al., in prep.) and on the application of neural networks both to phytoplankton production modelling (Scardi, in prep.) and to other related topics (Recknagel et al., 1996; Recknagel, 1997).

In this paper, we present new results of a case study that focused on developing a reliable modelling tool for Chesapeake Bay. Contemporary studies of trophic dynamics and remotely sensed observations providing synoptic biomass fields in the Bay are components of ongoing research that entail a need for accurate estimates of phytoplankton primary production. Beyond the specific use of neural network analysis to estimate primary production in Chesapeake Bay, however, this approach has general ecological relevance. If successful with data from this very complex estuarine ecosystem in which the principal variables regulating primary production are characterized by variability on a wide range of time and space scales, the likelihood of a broader application to other marine systems is enhanced.

2. Materials and methods

The 1982–1983 data that were used in the initial attempt to develop a neural network model of primary productivity were collected on a series of five cruises in Chesapeake and Delaware Bays. Further development of this model, however, was focused on Chesapeake Bay only. Chesapeake Bay, in Maryland and Virginia, is the largest bay on the Atlantic coast of the US (Fig. 1). It is about 320 km long from north to south and from 5 to 40 km wide. The Susquehanna and the Potomac are the largest of its many tributary rivers and creeks. The bay is a shipping artery, and the bay cities of Norfolk, VA., and Baltimore, MD., are among the nation's leading ports. Waterfowl, fish, oysters, and crabs, long abundant, have been threatened by pollution in recent years. Chesapeake Bay is characterized by strong gradients in salinity, turbidity, dissolved nutrients and chlorophyll as a measure of phytoplankton biomass.

Integral, daily primary production was measured using ^{14}C assimilation in simulated in situ sunlight incubations. Neutral density screens were used to attenuate sunlight and generate a light series, and surface seawater was circulated for cooling. Downwelling irradiance was measured continuously with a LiCor quantum sensor positioned in an unobstructed location on the ship, and vertical profiles were made throughout the day using an underwater LiCor quantum sensor to ascertain the diffuse attenuation coefficient for photosynthetically available radiation (PAR). Further details of the methods are contained in Harding et al. (1986).

Measurements of chlorophyll concentrations were made using standard fluorometric methods (Strickland and Parsons, 1968), nutrient concen-

trations were determined by wet chemistry on a Technicon AutoAnalyzer II, and ancillary data on other properties were collected at the same times and locations as samples were collected for measuring primary productivity. The reference to the original data source (Harding et al., 1986), contains most of the detailed methods and other aspects of the data collection are presented by Fisher et al. (1988).

Measurements of primary production made from 1987–1996 used the same methods as were used in the 1982–1983 cruises. Stations were predominantly located within Chesapeake Bay along the mainstem axis from the limit of salt to the mouth and plume regions nearly 300 km seaward. Approximately ten stations were occupied on each cruise, with the exception of 1995–1996 when more than double this number of stations was occupied per cruise and included sampling lateral to the mainstem axis. Each measurement of primary production was accompanied by collection of a full set of ancillary data.

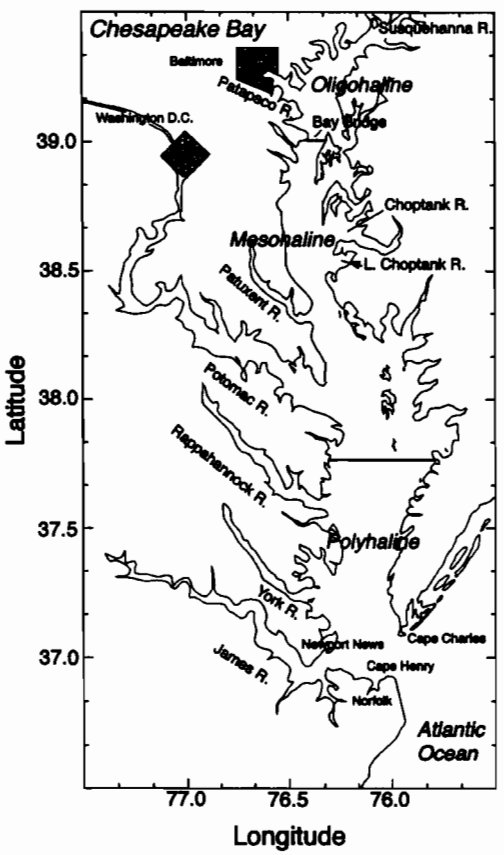

Fig. 1. Chesapeake Bay is the largest bay on the Atlantic coast of the US. It is about 320 km long and from 5 to 40 km wide.

The most recently collected data used in this analysis were from 1995–1997 and were collected on a series of cruises addressing Trophic Interactions in Estuarine Systems (TIES) sponsored by the US National Science Foundation. Of these data, the 1997 measurements were used to test the NN model and not to develop it.

All the neural networks we used as empirical models were multilayer perceptrons with one hidden layer and only one neuron in the output layer (i.e. phytoplankton primary production). This is by far the most common and flexible kind of neural network and it provides good performances in a wide range of applications.

Our applications aimed at training the most generalized neural network, rather than the one that optimally fitted the training test. Therefore, the error backpropagation training algorithm was used in its simplest version, as learning rate, set to a unit value, was not allowed to vary during training and no momentum term was used.

The training procedure was based on a subset of the 1982–1996 data set, which consisted of 326 patterns. In fact, in the case of our final model, only 100 patterns were randomly selected and used as training set, whereas the remaining 226 patterns were used as validation set. Even though a large validation set usually prevents overtraining, other techniques were also applied in order to obtain the most generalized model.

In particular, a small amount of Gaussian noise ($\mu = 0$, $\sigma = 0.01$) was added to the input patterns (Györgyi, 1990) and only a subset ($n = 50$) of the training set was randomly selected for each training epoch. The random selection of the training subset was also needed because a learning per pattern strategy was chosen and therefore it was necessary not to always submit the training patterns in the same order. Moreover, an early stopping strategy was used in the training procedure (i.e. training was stopped as soon as the validation set error started to increase).

The best structure of the neural network models was determined on the basis of empirical tests,

Table 1
Neural network input and output variables[a]

Variable	units	min	max
Input			
$\frac{1}{2}\left[\cos\left(\frac{2\pi \cdot \text{day}}{365}\right)+1\right]$	None	0.0	1.0
$\frac{1}{2}\left[\sin\left(\frac{2\pi \cdot \text{day}}{365}\right)+1\right]$	None	0.0	1.0
Latitude	Degrees	36.8	39 5
Longitude	Degrees	75.6	76.6
Station depth	m	0.0	45.0
Water temperature	°C	0 0	32.0
Salinity	PSU	0.0	32.0
Surface chlorophyll concentration ($\log_{10}$)	mg m^{-3}	−0 8	1.9
Total chlorophyll in the photic zone ($\log_{10}$)	mg m^{-2}	−0.3	2.7
Surface downwelling irradiance	E m^{-2} day^{-1}	0.0	80.0
Light extinction coefficient	m^{-1}	0.0	6.0
Photic zone depth	m	0.0	25 0
Output			
Phytoplankton primary production ($\log_{10}$)	mg C m^{-2} day^{-1}	0.9	3.9

[a] Units and the minimum and maximum values that were used to scale raw data to [0, 1] intervals are also shown. Variable names followed by ($\log_{10}$) indicate that raw values have been log-transformed before scaling them to a [0, 1] interval.

where hidden layers with three to 15 neurons were used. The best performance was obtained with five neurons in the hidden layer both in the case of the simpler model (three inputs) and in the case of the more complex one (12 inputs). However, the differences among neural networks with different structures were not dramatic and only the performance of the 3-x-1 model was perceivably degraded when more than ten hidden neurons were used.

The simpler 3-5-1 model used surface chlorophyll concentration, surface downwelling irradiance and depth of the photic zone as input variables, whereas nine more variables were selected as additional inputs for the more complex neural network. Input and output variables of this neural network are listed in Table 1, where the units and values that were assumed as limits to scale variables into [0, 1] intervals are also given. Inputs for both phytoplankton biomass and primary production were $\log_{10}$-transformed before scaling them to a [0, 1] interval. The log transformation was performed on the basis of both a theoretical assumption and an empirical test. The theoretical assumption was that the mean square error of the neural network output will be biased when raw data are used. This pertains because training patterns containing high values for biomass and primary production, containing proportionately greater sampling and measurement errors, may unduly dominate the output. The empirical test was carried out by comparing the performance of neural networks trained with transformed data to performance with raw data. In the case of the final 12-5-1 neural network, training on log-transformed data outperformed training on raw data, as it allowed the neural network to explain almost 20% more variance (the determination coefficients were $R^2 = 0.546$ and $R^2 = 0.353$, respectively).

The serial number of the day of the year was transformed using sine and cosine functions (see Table 1) that map the date onto a circle. Two inputs—the total chlorophyll in the photic zone and the photic zone depth—were computed on the basis of other input variables. The total chlorophyll in the photic zone (Table 1) was obtained as the product of surface chlorophyll concentration and photic zone depth, assuming that the phytoplankton biomass is homogeneously

distributed in the upper water column. The photic zone depth (Table 1), i.e. the depth where the available downwelling irradiance is the 1% of the surface downwelling irradiance, was obtained as 4.605 (i.e. ln 0.01) divided by the light extinction coefficient (Table 1). If the resulting value was larger than the station depth (Table 1), then the latter was assumed as photic zone depth.

3. Results

The toy model presented by Scardi (1996) performed well using the 1982–1983 data set on which it was trained ($R^2 = 0.940$). When used with a much larger data set spanning 1982–1996 and encompassing a wide range of environmental conditions, this model did not perform nearly as well ($R^2 = 0.156$), as shown in Fig. 2. This relatively simple approach was based on a 3-5-1 neural network that used surface downwelling irradiance, surface chlorophyll concentration, and photic zone depth as inputs. We found that the toy model was unable to reproduce primary production values that were larger than the ones on which it was trained, and that large errors were also obtained even within the range of observations contained in its own training set (0–3 g C m^{-2} day^{-1}).

To ascertain the performance of the toy model compared to other, more conventional approaches, we also used a common model based on linear regression of primary production on a composite variable obtained from the product of the same three variables used as neural network inputs (see Cole and Cloern, 1987). Despite what we term poor performance of the toy model with the larger data set from Chesapeake Bay, the estimates of primary production were significantly better than those obtained using conventional linear models that returned a mean square error almost twice as high.

To overcome the shortcomings of the toy model, a new 3-5-1 neural network was trained on the basis of the entire 1982–1996 data set. Two training procedures were carried out, one on raw data and the other on log-transformed biomass and primary production data, but none produced a synaptic weight set that showed a significant improvement over the toy model.

As in the case of the toy model, these networks were not able to cope with high primary production values, even though they were trained on a quite large data set. This result was not unexpected because primary production in Chesapeake Bay is clearly not regulated by phytoplankton biomass, irradiance and photic zone depth alone; there is a strong landward to seaward gradient in

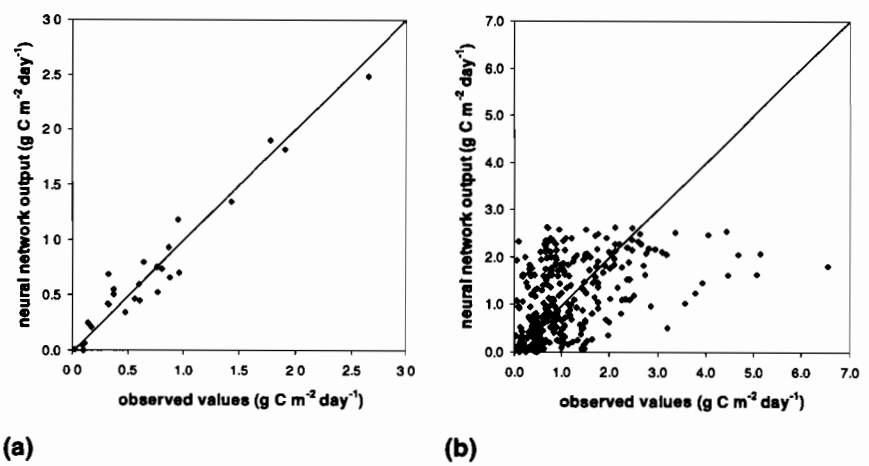

Fig. 2. Scatter plots of neural network outputs versus observed values for the toy model described in Scardi (1996) The 1982–1983 training subset was accurately fitted (a), whereas the whole 1982–1996 data set showed poor generalization (b).

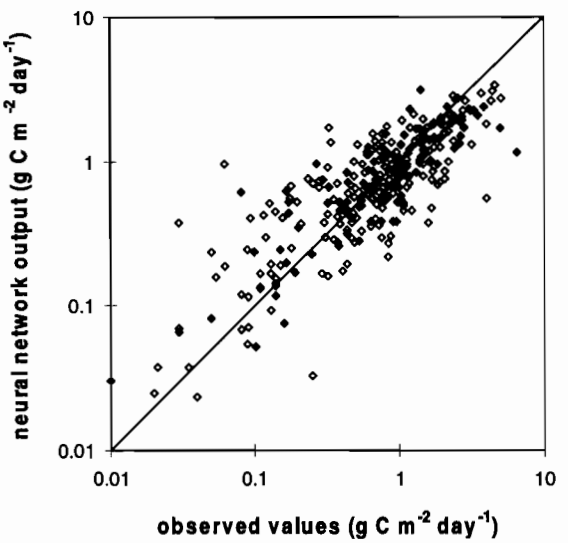

Fig. 3. Scatter plot of the neural network outputs versus observed values for the 12-5-1 model Both training and validation data are shown. The overall agreement between observed and simulated data was satisfactory ($R^2 = 0.546$). The validation set values (white diamonds, $R^2 = 0.614$) were reproduced even better than the training set ones (black diamonds, $R^2 = 0.420$).

dissolved nutrients and much of the Bay is nutrient limited for at least part of the year. Accordingly, we surmised that additional information was needed to improve the model.

The neural network model using a 12-5-1 structure, i.e. a larger suite of predictive variables, shows improved estimates over previous approaches (Fig. 3). This finding pertains both to the training set shown as black diamonds, and to the validation set shown as white diamonds. Primary production values are predicted with greater accuracy ($R^2 = 0.546$) than in the case of linear or other simpler models. The neural network also showed good generalization properties, in that the validation set was fitted even better than the training set ($R^2 = 0.614$ and $R^2 = 0.420$, respectively).

Although data for phytoplankton biomass and primary production were log-transformed before training, the model outputs need to be transformed back to raw data in most applications. Therefore, the error distribution of the model, which was unbiased in log units, was also checked after back-transforming data to raw units and a small bias was detected ($m_{error} = -0.14248$, in the validation set). Obviously, this systematic error depended on the different impact that very large and very small values exerted with or without log transformation.

In order to obtain unbiased primary production estimates, a simple linear correction was defined by least square optimization and applied to the neural network output. The corrected estimates were then computed by multiplying neural network outputs by 1.15575 (this correction could be visualized as a small vertical shifting of all the points in the log–log plot in Fig. 3).

The resulting error distribution was virtually unbiased ($m_{error} = -0.00174$, in the validation set) and almost symmetrical, as shown in Fig. 4. It should be noted that more than 80% of the errors of the primary production estimates in the validation set were within the ± 0.6 g C m^{-2} day^{-1}, i.e. less than 1/10 of the observed data range. Moreover, the accuracy of the model was also slightly improved, as corrected outputs explained 3% more variance ($R^2 = 0.578$) than the uncorrected ones ($R^2 = 0.546$).

The accuracy of the neural network model and its generalization capabilities were also tested on an independent data set ($n = 52$), which was collected during 1997 and therefore was not available during the training phase. In the scatter plot in Fig. 5 predicted versus observed values are shown for both this new testing set (large black circles) and the original training and validation sets (small

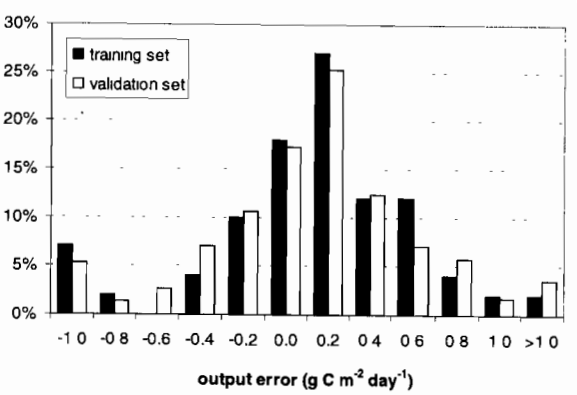

Fig. 4. Error distribution of the corrected neural network outputs. The labels on the error axis indicate the upper limit of each class.

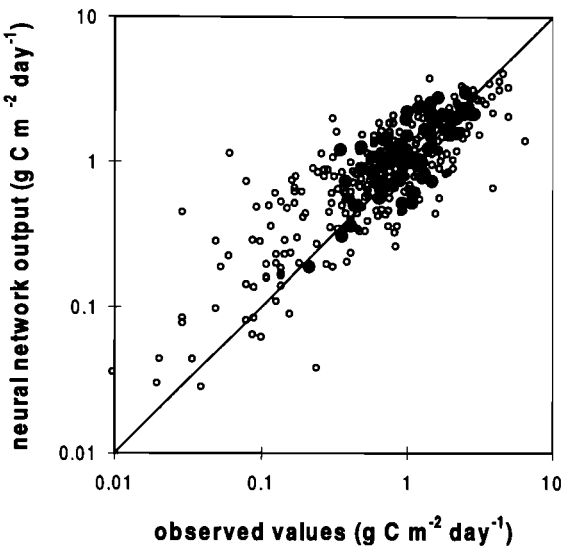

Fig. 5. Scatter plot of the neural network outputs versus observed values. Both the new independent testing set (1997, large black circles) and the original training and validation sets (1982–1996, small white circles) are shown.

white circles). The new primary production values were reproduced by the model with the same accuracy as original data and were almost unbiased, as their mean error was negligible (m_{error} = 0.082).

The error distribution of the new testing set is shown in Fig. 6, where it is compared to the error

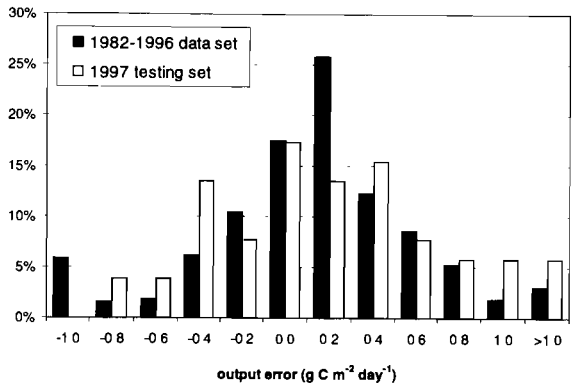

Fig. 6. Error distribution of the corrected neural network outputs. The labels on the error axis indicate the upper limit of each class.

distribution of the original data set. Even though the latter is more regular and symmetrical, the differences between the two distributions are minor and are probably influenced by the smaller number of patterns in the new testing set.

Finally, a sensitivity analysis was carried out using the whole 1982–1996 data set to assess the effect of small changes in each input on the neural network output. The results of this analysis provide a useful insight into the neural network model, but they also help to understand the underlying ecological processes, i.e. the relative importance of the predictive variables to

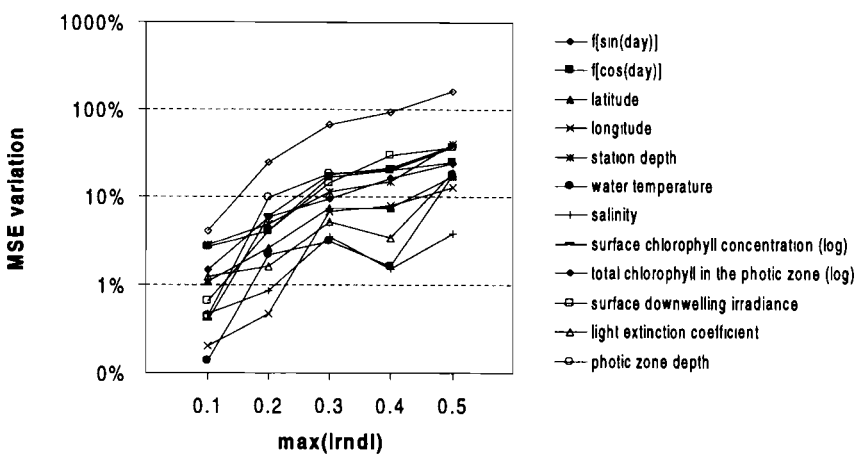

Fig. 7. Percentage variation of the mean square error of the neural network output at increasing levels of input perturbation. White noise ranging from [−0.1, 0.1] to [−0.5, 0.5] was added to each input variable in the whole 1982–1996 data set and the resulting increase in mean square error was expressed as a percentage of the original mean square error.

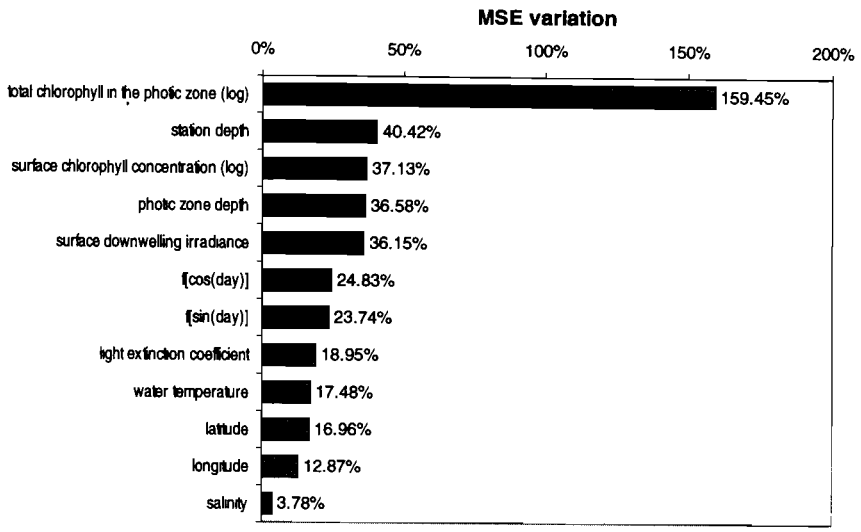

Fig. 8. Percentage variation of the mean square error of the neural network output after the addition of [− 0.5, 0.5] white noise. The input variables are ranked according to their sensitivity.

phytoplankton primary production in Chesapeake Bay.

In the sensitivity analysis, the mean square error of the neural network output is expected to increase as a larger amount of white noise is added to the selected input variable. The mean square error variations that were observed after white noise additions varying from [− 0.1, 0.1] to [− 0.5, 0.5], i.e. from 20 to 100% of the input range, are shown in Fig. 7.

The minimum level of input perturbation was similar in magnitude to the measurement error of the oceanographic data and so were the changes it induced in the mean square error of the neural network output (< 5%). Increasing white noise additions caused increasing mean square errors in the output, even though this relationship was not absolutely monotonic, because less sensitive variables, that did not affect the neural network output very much, showed a few negative increments. However, the relative sensitivity of the input variables did not vary significantly when very large amounts of white noise were added. These results suggest that the primary production model that was embedded in the neural network was probably consistent with the ecological processes as it was not misled by unlikely input patterns.

The most influential variable among the neural network inputs in affecting output was by far the total chlorophyll in the photic zone. It was the only input variable that caused an increase in the mean square error larger than 100% when [− 0.5, 0.5] white noise was added, as it is clearly shown in Fig. 8. As expected for a primary production model, the predictive variables that were related to light availability and phytoplankton biomass had the largest effects on output among the remaining variables. The least influential variable was salinity, despite that it may be viewed as a proxy for freshwater inflow and often covaries with nutrient concentrations.

4. Discussion

The neural network provided accurate and unbiased estimates of phytoplankton primary production for a system that is characterized by high spatial and temporal variability. This is a satisfying result, given the shortcomings of linear models that fail to perform acceptably with the same data and that contain biases that are particularly pronounced at low and high primary productivity rates. In most cases, the error of primary production estimates obtained with the neural network

was within the range of the measurement error. We believe the neural network approach outperformed conventional empirical models because it is inherently much more flexible in dealing with the influences of a number of variables that regulate phytoplankton primary productivity in estuaries.

Our results were obtained using a very conservative approach as far as generalization is concerned, because most of the available data were used for neural network validation and only a restricted subset, i.e. less than one third of the entire data set, was used for neural network training. We also tested our neural network model using an independent data set that was not available during development of the model. The success of this approach implies that the present form of our model is probably less than optimal and that further improvements are still possible. A further consideration is that the training procedure was not optimized (a constant learning rate and no momentum were used), and improvements in this area may refine the model further.

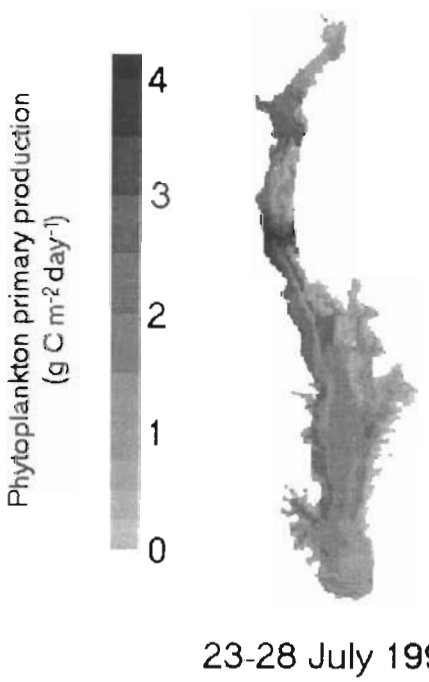

Fig. 9. An example of application of the neural network. The phytoplankton primary production was estimated over the whole Chesapeake Bay mainstem area using interpolated input data. Each pixel in the image corresponds to a 1 km^2 square.

The neural network model of Chesapeake Bay phytoplankton primary production can play an important role in monitoring and research activities, because it may permit reduction of the number of direct primary production measurements that are needed to reconstruct large scale spatial patterns or high frequency time series. An example of such an application is shown in Fig. 9, in which the distribution of phytoplankton primary production in the mainstem area of Chesapeake Bay is presented as a grayscale image. These estimates were based on discrete data collected during a summer cruise (23–28 July 1995) that were interpolated to generate complete input grids. In future applications some of these input grids could be replaced with remotely sensed data. An aircraft remote sensing program (cf. Harding et al., 1992, 1994, 1995) provides high resolution estimates of chlorophyll for the estuary, using sensors designed to replicate band of the satellite ocean color instrument, SeaWiFS, that is now providing global coverage. Data from this source can be substituted for shipboard observations and fill time and space gaps that accompany more routine sampling. Given that the most influential variables on model output are related to the accuracy of biomass inputs, we envision improved performance of the neural network model when remotely sensed data are used.

We believe the principal explanation for the superior performance of the neural network model to that of more conventional approaches is that the complexity of factors that regulate phytoplankton primary production are better captured. From a purely theoretical viewpoint, other empirical models might also obtain this result, provided that their formulation is carefully defined and sufficiently complex to incorporate the data structure. However, in these cases the model formulation has to be explicitly defined by the modeler, who usually opts for an empirical approach when his/her understanding of the processes to be modeled is not complete or he/she thinks that accuracy can be traded for simplicity.

An example of the degree of complexity of the relationships that can be reproduced by a neural

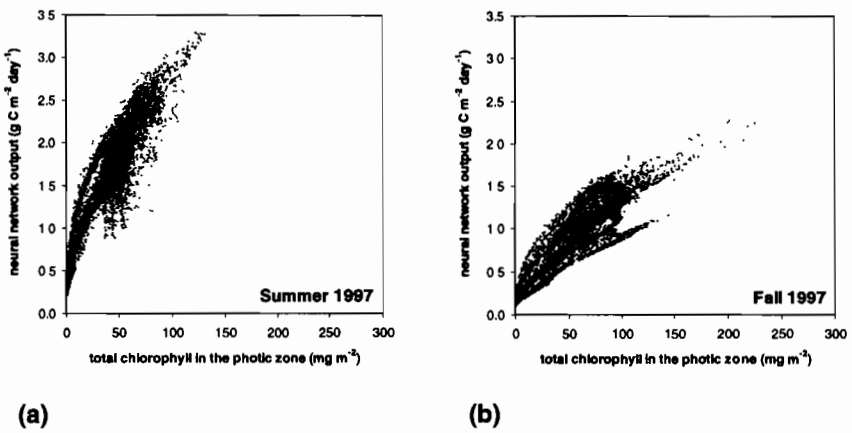

Fig. 10. Neural network output versus total chlorophyll in the photic zone in different seasonal scenarios: summer (a) and fall (b). More than 7000 points, corresponding to pixels in images similar to Fig. 9, are shown in each plot. The position of each point depends on the photosynthetic efficiency of the whole water column and a steeper overall slope implies a higher photosynthetic efficiency, as in the summer 1997 plot (a). It is interesting to notice that the neural network model was able to reproduce a range of different area-specific non-linear relationships.

network model is shown in Fig. 10. The neural network outputs (i.e. estimates of phytoplankton primary production) that were obtained for more than 7000 input patterns were plotted against one of the inputs, the total chlorophyll in the photic zone. Therefore, the position of each point is determined by the photosynthetic efficiency of the whole water column. It is very clear that the two seasonal scenarios that were considered were completely different because the overall photosynthetic efficiency varies in time. However, it is also clear that the biomass/production relationship is also variable within each sub-plot, because all the points are arranged as to form a set of curves, each one having a different slope, that reproduce the spatial variation of the biomass/production relationship in Chesapeake Bay. Even though the sensitivity analysis showed that the total chlorophyll concentration was probably the most relevant input variable, plots obtained with other variables also showed similar patterns.

Finally, it has to be stressed that sensitivity analysis might play an important role in both the optimization of the neural network models and in understanding the processes to be modeled. Sensitivity analysis is not a simple and straightforward task when analytical models are taken into account, but it is even more challenging when neural networks are considered. However, the procedure we used was able to analyze the first-order effects of input perturbation on the neural network output and the results provided a useful insight both into the neural network mechanics and primary production processes. The neural network outputs were almost invariant when small perturbations, similar to those that depend on sampling errors, were introduced. On the other hand, when more noise was added to the inputs, the role of each variable could be defined in terms of relative importance in determining phytoplankton primary production.

The total chlorophyll in the photic zone, i.e. the total biomass that is photosynthetically active, was clearly the most important predictive variable. Other variables, such as salinity, were less sensitive to the addition of white noise and therefore seem to play a less important role. Of course, excluding these variables might help prune the neural network structure. This kind of optimization is not very important from a computational point of view, but could reduce the cost of data acquisition without a significant loss in accuracy of the model.

The neural network model of Chesapeake Bay phytoplankton primary production has been implemented in Java and can be tested at the following URL: http://www.mare-net.com/mscardi/work/nn/cbjavann.htm.

References

Balch, W.M., Eppley, R W , Abbott, M.R., 1989. Remote sensing of primary production-II. A semi-analytical algorithm based on pigments, temperature and light. Deep-Sea Res. 36 (8), 1201–1217

Cole, B.E., Cloern, J.E , 1987 An empirical model for estimating phytoplankton productivity in estuaries. Mar Ecol. Prog. Ser. 36, 299–305

Eppley, R W., Stewart, E., Abbott, M.R , Heyman, U., 1985 Estimating ocean primary production from satellite chlorophyll. Introduction to regional differences and statistics for the Southern California bight. J. Plank. Res. 7, 57–70.

Fisher, T.R., Harding, L.W , Stanley, D.W., Ward, L.G., 1988. Phytoplankton, nutrients and turbidity in the Chesapeake, Delaware and Hudson estuaries. Estuar. Coast. Shelf Sci. 27, 61–93.

Györgyi, G., 1990. Inference of a rule by a neural network with thermal noise Phys. Rev. Lett. 64, 2957–2960.

Harding, L.W. Jr., Meeson, B W., Fisher, T.R. Jr., 1986. Phytoplankton production in two east coast estuaries: photosynthesis-light functions and patterns of carbon assimilation in Chesapeake and Delaware Bays Estuar. Coast. Shelf Sci. 23, 773–806.

Harding, L.W. Jr., Itsweire, E C., Esaias, W.E., 1992. Determination of phytoplankton chlorophyll concentrations in the Chesapeake Bay with aircraft remote sensing Rem Sens. Environ. 40, 79–100

Harding, L.W Jr., Itsweire, E.C , Esaias, W E., 1994. Estimates of phytoplankton biomass in the Chesapeake Bay from aircraft remote sensing of chlorophyll concentrations, 1989 1992 Rem Sens Environ. 49, 41 56.

Harding, L.W Jr., Itsweire, E C., Esaias, W.E , 1995. Algorithm development for recovering chlorophyll concentrations in the Chesapeake Bay using aircraft remote sensing, 1989–1991. Photogr. Eng. Remote Sens. 61, 177–185

Lek, S , Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90 (1), 39–52.

Recknagel, F , 1997. ANNA—artificial neural network model predicting blooms and succession of blue green algae. Hydrobiology 349, 47–57.

Recknagel, F , French, M., Harkonen, P., Yabunaka, K.I., 1996. Artificial neural network approach for modelling and prediction of algal blooms Ecol. Model 96 (1–3), 11–28.

Scardi, M., 1996 Artificial neural networks as empirical models of phytoplankton production Mar Ecol. Prog Scr 139, 289–299.

Strickland, J D H., Parsons, T R , 1968. A practical handbook of seawater analysis. Bull. Fish Res. Bd. Can. 167, 1–311

ELSEVIER

Ecological Modelling 120 (1999) 225–236

ECOLOGICAL MODELLING

www elsevier com/locate/ecomodel

Wedding connectionist and algorithmic modelling towards forecasting *Caulerpa taxifolia* development in the north-western Mediterranean sea

Alex Aussem *, David Hill

Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS), ISIMA-Université Blaise Pascal, Clermont Ferrand II, Campus des Cézeaux-B.P. 125, 63173 Aubiere Cedex, France

Abstract

We discuss the use of supervised neural networks as a metamodelling technique for discrete event stochastic simulation in order to reduce significantly the computational burden involved by discrete simulations. A sophisticated computer model, coupling a geographical information system with a stochastic discrete event simulator, has been developed to anticipate the propagation of the green alga *Caulerpa taxifolia* in the north-western Mediterranean sea. The simulation model provides reliable predictions, a couple of years in advance, of: (i) the local expansion patterns of the alga; (ii) the increase of *C. taxifolia* biomass and (iii) the covered surfaces. However because the algorithmic model accounts for spatial interactions and anthropic dispersion/activities such as eradication, introduction of specific predators etc., simulations are extremely time and memory consuming. Therefore, to reduce the computational burden, a neural network was successfully trained on artificially generated data provided by the simulation runs to provide accurate forecasts 12 years in advance along with associated confidence intervals. The ability of the neural networks to capture the underlying physics of the phenomena is clearly illustrated by several preliminary experiments on a large coastal area. The neural network is able to construct, on this site, estimates of the *Caulerpa taxifolia* expansion 12 years in advance in good agreement with the simulation trajectories.

Keywords: Metamodelling; Neural networks; Discrete event; Simulation; *Caulerpa taxifolia*; Invasive species

1. Introduction

In 1984, the French coast of the Mediterranean sea near Monaco was the initial site of the development of *Caulerpa taxifolia*, a green alga of tropical origin introduced by mistake (Meinesz and Hesse, 1991; Belsher and Meinesz, 1995). After 12 years later, this species had colonized several thousand hectares of the French and Italian coasts and was detected in numerous places of the north-western Mediterranean coast, from Croatia (Adriatic sea) to the Balearic islands (Spain). This surprising development may locally induce an intense and irreversible alteration of the

* Corresponding author. Tel.: +33-4-734050039; fax: +33-4-73405001.

E-mail address: alex@sp.isima.fr (A. Aussem)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00104-0

coastal ecosystems, such as endogenous species distributions (alga, cnidaria, sponges, echinoderms, fishes, etc.) as well as ecosystem functioning (trophic levels relationships). Indeed, the progressive elimination of the benthic flora and fauna is observed at stations heavily occupied by *C. taxifolia*. Though the social and economic repercussions of such damage on the ecosystems are considerable, the precise incidence on fisheries and tourism is difficult to estimate.

With a view to understanding the underlying rules governing the development of *C. taxifolia*, an interdisciplinary joint venture between marine ecologists, biologists and computer scientists was undertaken (Hill et al., 1997, 1998). They concurred on the observation that spatial effects like currents, spatial heterogeneity, fragments spreading etc., are major parameters influencing the colonization process and should therefore be integrated into a sophisticated discrete-event simulation model. The major goals are to provide quantitative results associated with the alga expansion several years in advance, e.g. biomass, production, contaminated surfaces and residual biomass of competing species (especially the protected seagrass *Posidonia oceanica*), given various environmental parameters, e.g. bathymetry, substrates and biocenosis on different sites (Hill et al., 1997). To provide forecasts with sufficient accuracy, the model had to take into account spatial effects as well (heterogeneity of the sites, long and short distances interactions between organisms, etc.).

Unfortunately, one of the major drawbacks of discrete-event simulation is the amount of computational resources required to fully explore numerous possible trajectories of the ecosystem several years ahead according to distinct experimental scenarios, e.g. pattern of currents, anthropic dispersion, etc. Clearly, performing multiple replications for selecting optimal decision variables or policies becomes swiftly inhibitory when dealing with large scale ecosystems.

To supplement the computationally intensive (stochastic) discrete-event simulation software, we developed a so-called metamodel, based on supervised multilayer neural networks, to provide direct forecasts several years in advance in terms of covered surface levels (Coquillard and Hill, 1997). Once trained with artificially generated data provided by the simulation runs, neural networks are shown to be reliable approximations of the underlying biological system that perform satisfactorily and are significantly more computationally efficient than the simulator itself. In addition, such techniques promise insights into the biological mechanisms that discrete simulation alone cannot provide.

This paper first presents the different modelling constraints which were kept in mind while designing the model, as well as the technical choices and the first results. Neural networks are then discussed and applied towards the prediction of the covered surface 12 years in advance.

2. The *Caulerpa taxifolia* simulation model

As discussed above, a sophisticated stochastic discrete-event simulation model was developed few years ago, in collaboration with marine ecologists, biologists and computer scientists, in order to better understand the parameters influencing the colonization of *Caulerpa taxifolia* along the French coast of the Mediterranean sea, and to explore distinct experimental scenarios (Hill et al., 1997, 1998). The major goals were to provide quantitative results associated with the alga expansion several years in advance, given various environmental parameters. For completeness, we briefly detour to present the simulation model before we discuss the metamodelling approach and the experiments.

2.1. The Caulerpa taxifolia settlements

First, a thorough study of *C. taxifolia* settlement and development on precise locations was carried out to facilitate the validation of simulation results. Three settlement sites have been mapped yearly by marine ecologists since the beginning of 1990. All the experiments presented in this paper focus the first zone, namely Villefranche-sur-Mer. This site was mapped at a decametric scale. In spite of the absence of exhaustive mapping, enough data were available in that area to run simulations on a large scale (Hill, 1997).

As discussed above, spatial effects like currents, spatial heterogeneity, fragment spreading, etc. are considered major parameters influencing the colonization process and were therefore integrated in the simulation model. The modelling technique is based on discrete-event simulation used in the last decade for ecological modelling purposes (Hill, 1996). Consequently, the model is no only specified by a mathematical formalism but also by an algorithm describing the system functioning. This required the development of simulation software able to handle any littoral site by interfacing it with a geographical information system (GIS). Part of the simulation model input was initialized with digitized maps created in the GIS software MapGraphix. For each studied site, two maps were used: one for substrates and one for bathymetry. In the same way, the spatial simulation results are provided in GIS format in addition to traditional curves and statistics.

The outputs are composed of means along with confidence intervals obtained by bootstraping techniques (Bradley, 1982). Independence, common mean, variance and normal distribution of the responses were evaluated by a Kolmogorov test applied on results obtained from both 1000 and 10 000 replications. With 1000 replications, the Kolmogorov never rejected the null hypothesis, namely that the observed distribution is normally distributed, for a level of significance of 95%, see (Hill et al., 1997) for more details).

2.2. Modelling elements

The simulation model relies on a set of holistic variables which are estimated and changed with the occurrence of discrete events. The site under study are divided into cells whose size is an important model parameter. The cell size varies from 16 cm^2 up to 370 m^2 depending on the scale of the studied site. Each cell possesses substrate and depth attributes (provided precisely by the GIS). The probability that *C. taxifolia* grows in each cell is linked to a set of evolution rules depending on (i) the depth; (ii) the kind of substrate; and (iii) the number of fragments arriving in the cell. The growth during a simulation session is based on a list of active cells (i.e. those containing *C. taxifolia* individuals). An exploratory approach allowed us to determine the best parameters for each site. Simulations are initialized from experiment files (with 98 parameters). The experiment files contain:

- simulation control parameters (duration, number of replications, GIS map specification etc.)
- model initial values (distribution rate of the fragments, current direction and strength, etc.).
- settlement probabilities as a function of bathymetry, substrate, season, etc.
- monthly parameters for stolon growth, spread of fragments, biomass, and degeneration.

2.3. Model validation

The calibration and the sensitivity analysis have demonstrated that the model is robust enough and the conceptual model is consistent. Numerous replications—up to 100 000—were done for verification purposes (the random number pseudo-period being long enough). The modifications of internal data flow never lead the model to exhibit abnormal behaviour or biased statistical results. Under such conditions the model could be considered as reliable according to the definition in use in the simulation community. Different techniques (Coquillard and Hill, 1997) were used for validation purposes: (i) comparison of results from site to site; (ii) confrontation with empirical knowledge of marine ecologists; and (iii) graphic visualization and animation to make use of the human ability to comprehend spatial relationships.

For this kind of simulation study, the spatial auto-correlation is strong since *C. taxifolia* contaminated zones tend to form aggregated spots. However, from one replicate to the other, peripheral spots distribution is totally different without apparent correlation. Thus, a large number of replicates were carried out and the results combined to perform a discrete spectral analysis. This analysis is useful to point out areas which have a high probability of invasion by *C. taxifolia* although such plots are not always easily interpreted.

2.4. Computational burden

The basic objective of this modelling approach is the quest for a deeper understanding of the *C. taxifolia* expansion phenomena which manifests itself along the Mediterranean coast. Unfortunately, in order to better understand the parameters influencing the colonization process, an exploration of distinct experimental scenarios is required, which in turn translates into a great number of simulation replications. Moreover, many environmental parameters (bathymetry, substrates, biocenosis, etc.) on different sites, have to be varied according to distinct experimental scenarios, e.g. pattern of currents, anthropic dispersion, etc.

Unfortunately, one of the major drawbacks of discrete-event simulation is the amount of computational resources required to fully explore system responses. Under these circumstances, performing multiple replications for the selected values of the decision becomes swiftly inhibitory when dealing with large scale ecosystems, hence the idea of using a so-called metamodel to provide direct forecasts of the relevant variables with sufficient accuracy.

3. The metamodel

To overcome the limitations of discrete-event simulation models, researchers and practitioners in the environmental, management, industrial and production sciences have developed so-called metamodels, which are approximations that perform satisfactorily and are significantly more computationally efficient. Metamodelling techniques can be traced back to the early 1970s, although the term was developed more recently (see Kleijnen, 1987; Kilmer 1994) and references therein. Applications in the literature mainly cover industrial and production applications, such as production planning and control, facilities storage and design and shop floor control (Pierreval and Huntsinger, 1992; Kilmer, 1994). Most of these articles concur in the observation that metamodels are easier to manage and provide more insight than simulation alone.

The main issues in metamodelling are:

- the choice of the underlying functional form,
- the choice of the inputs and their corresponding response variables to be used,
- selection of the appropriate samples from the simulation to construct the metamodel,
- validation of the model.

3.1. Metamodelling principles

To define more explicitly a stochastic simulation metamodel, let X_j, $j=1, \ldots, n$ denote the variables influencing the response, Y, of the physical system. Assume that the system is subject to some additive random fluctuations, then the unknown relationship between Y and the inputs X_j may be written as:

$$Y = f(X_1, \ldots, X_n) + \gamma \tag{1}$$

where γ is the additive noise, with zero mean and is independent of the inputs X_j. It is well known that the minimum mean squared error predictor is then given by

$$\bar{Y} = E[Y|X_1, \ldots, X_n] = f(X_1, \ldots, X_n) \tag{2}$$

A simulation model usually integrates only the input variables which lend themselves to the observation,

$$\hat{Y} = g(X_1, \ldots, X_p) + \delta \tag{3}$$

where $p \leq n$ and δ is some additive noise, with zero mean and is independent of the inputs X_j, and represents the random fluctuation of the simulation model, which was shown, in our case, to be normally distributed.

Now the minimum mean squared error predictor for the simulation model is given by:

$$\bar{\hat{Y}} = g(X_1, \ldots, X_p) \tag{4}$$

Let

$$\bar{\hat{Y}} - \bar{Y} = \varepsilon_s \tag{5}$$

denote the error term accounting for the simulation error, owing to the excluded variables and model miss-specification. Now, a metamodel, h, is a further simplification of the previous model, adjusted to output the optimal simulation response. The metamodel may be written as:

$$\hat{\bar{Y}} = h(X_1, \ldots, X_m) \quad (6)$$

where $m \leq p \leq n$. Since the metamodel is adjusted from the simulation model, let:

$$\bar{\bar{Y}} - \hat{\bar{Y}} = \varepsilon_m \quad (7)$$

be some additional error accounting for the metamodelling error, owing to the further excluded variables and the metamodelling error of fitting the metamodel to the simulation model.

The identification of the optimal metamodel is the identification of the parameters of $h()$. In other words, the metamodel can be viewed as a simplified model of the simulation model acting as a surrogate for the study of the physical system. The metamodelling error is therefore the sum of three terms, namely:

$$Y - \hat{\bar{Y}} = \varepsilon_m + \varepsilon_s + \gamma \quad (8)$$

Now, several metamodelling techniques exist, e.g. kernel-based regression models, neural networks, genetic algorithms, etc. As far as we are concerned, they are distinguished by their trade-offs, e.g. between accuracy and computational expense, between local and global fitting techniques, that must be dealt with when developing a metamodel. The next section aims at introducing the reader to the very basic knowledge of neural networks, making it suitable to readers with interests in non-technical areas.

3.2. Neural network metamodels

It is difficult to present the connectionist network, especially when the audience consists of a mixture of meteorologists, and ecologists, some of whom will know a great deal about connectionist approaches to modelling and prediction, while the knowledge of others is limited in scope. However, we have tried to highlight the specifics in a brief review in which neural networks are analyzed as regression models.

In this paper, we experiment with a multi-layer neural network (also called preceptron or feed-forward network): a simple, well-known and in-depth studied input–output (i.e. static) model (Rumelhart et al., 1986), with non-linear transfer functions, offering universal function approximation capability (Cybenko, 1989). Such a model can accommodate a combination of continuous (usually interval-scaled) and discrete numeric variables, as will be the case in our experiments—even if, in practice, they have severely heterogeneous certainties (Zheng et al., 1997; Murtagh et al., 1998). Additionally, classical neural paradigms are global models, that is, a single neural model is trained to model the entire simulation response surface. This differs from polynomial regression metamodelling, where the regression surface is fitted locally. Finally, the parallel architecture provides robustness to incomplete or erroneous data sets and offers fault tolerant, real-time performance.

It is important to stress that neural networks are deterministic models, meaning that the optimal mean-squared predictor is nothing else than the conditional mean of the desired output, given the input. This intuitive result represents the closed-form optimal solution for a neural network trained using least squares. It should be understood, however, that a neural network of fixed size can only approximate the optimal function thus introducing an effective bias in the solution. Furthermore, training on limited data results in an effective variance over the possible converged solutions. A larger network has smaller bias but requires more samples to train and so has an effectively larger variance. This trade-off is typical in regression theory (Geman et al., 1992). What makes neural networks better at finding solutions than splines for instance is a far more difficult problem to answer and requires thorough statistical understanding of the notions of consistency and generalization. But it is well beyond the scope and purpose of this paper to delve deeper into statistical theory of optimal predictors. Readers interested in more theoretical material are encouraged to consult the literature.

3.3. Measure of fit and confidence intervals

The training set and the test set of a supervised neural network are made up from a number of input–output patterns. The test set is used to assess the performance of the neural network. The target value is the estimate from the simulation of

the conditional mean. We implicitly assume that the average system output is the correct answer, and thus any deviation of the simulation from the actual system is disregarded. On the other hand, for the research presented in this paper, two target outputs were used: the average simulation response and its estimated variance calculated over $n = 10$ simulation replications:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2 \tag{9}$$

Using a single network for simultaneous prediction of both targets can be regarded as detrimental rather than beneficial, because it creates added complexity, and thus increases the risk of overfitting, given the limited number of training patterns. Therefore, two distinct neural networks were used separately in this paper for the prediction of the mean $\hat{\bar{Y}}$ and the variance $\hat{\hat{\sigma}}^2$.

A standard measure of fit (Weigend et al., 1990; Aussem et al., 1995) for the neural network is given by the normalized mean-squared error:

$$NMSE = \frac{1}{\sigma^2 N} \sum_{k=1}^{N} (\varepsilon_m^k)^2 \tag{10}$$

where ε_m^k is the modelling error, i.e. the difference between the average simulation value, at iteration k, and the neural network prediction, and σ^2 is the empirical variance of the simulation values calculated over all the N training patterns. A value of the $NMSE = 1$ thus corresponds to predicting the unconditional mean.

In addition, to gauge the practical relevance of the neural network metamodels, the neural network forecasts were combined to form confidence intervals. Intervals took the form of:

$$\text{Interval} = \hat{\bar{Y}} \pm \frac{\hat{\hat{\sigma}}}{\sqrt{n}} t_{n,\alpha/2} \tag{11}$$

where $\hat{\bar{Y}}$ stands for the neural network forecast of the average simulation response; $t_{n-1,\alpha/2}$ is the Student t-value for a level of significance of $\alpha = 95\%$ and $n = 10$, the number of replications of the simulation (Gordon, 1978). Note again that the normality of the simulation responses was statistically tested on results obtained from more than 1000 replications. Therefore, we implicitly considered that the ten values above were drawn from a normal distribution. Simulation runs were independent so that the necessary statistical assumptions held. The confidence intervals constructed from the neural network predictions were compared to those constructed from the simulation responses, using the same formula.

3.4. Training set-up

A traditional multilayer perceptron trained with a standard back-propagation algorithm was used. Before we turn to the experiments, we detour to review two (very basic) techniques we found to be useful for networks to run optimally. In an attempt to optimize the learning rate η automatically, we allow the η to vary in the course of training since the best values at the outset of training may not be so good later on. Our approach is to check whether the weight updates at each iteration actually decrease the cost function. If not then the process is overshooting and η should be reduced. On the other hand, if several steps in a row have actually decreased the cost function, then η is increased. It appears best in the literature (Hertz et al., 1991; Aussem et al., 1995; Aussem, 1998, 1999) to increase η by a constant and to decrease it geometrically to allow rapid decay when necessary. This gives the overall scheme

$$n = \begin{cases} +\alpha & \text{if } \Delta E \leq 0 \\ -\beta n & \text{if } \Delta E > 0 \end{cases} \tag{12}$$

where α and β are appropriate time constants. In this paper, we took $\alpha = 10^{-3}$ and $\beta = 0.9$.

In order to prevent data overfitting, a topological reduction approach was adopted in these studies. The approach consists in pruning appropriately non-useful connections during training. This is achieved by giving each connection a tendency to decay to zero, such that

$$\omega_{ij}^{\text{new}} = (1 - \varepsilon)\omega_{ij}^{\text{old}} \tag{13}$$

so that the connections disappear unless reinforced (Weigend et al., 1990; Hertz et al., 1991). For small ω_{ij}'s to decay more rapidly than larger ones, we made ε dependent on ω_{ij} by

$$\varepsilon_{ij} = \frac{\gamma\eta}{(1+\omega_{ij}^2)^2} \quad (14)$$

where γ was varied in the course of training. We found out that to it is useful to begin with $\gamma = 10^{-3}$ until performance reaches a minimum and then to decrease the value up to 10^{-4} so the error continues to decrease at a slow rate.

We just stress that more sophisticated and powerful topological reduction approaches exist. They consist in appropriately pruning (or penalizing) non-useful connections as training proceeds, preventing the network from overfitting the data. For further details, the reader is directed to the references.

4. Experiments

This section discusses some preliminary experiments performed with neural network metamodels. Given concerns about the rapid, intense and irreversible alteration of the coastal system, and the computational burden of discrete-event simulation over periods spanning several months, we mainly focused on predictions of *C. taxifolia* contamination on the long-term (12 years in advance). In this regard, our objective was to forecast the *C. taxifolia* expansion in terms of undersea contaminated surface over a 12-year period according to most of the available biological parameters.

4.1. Contaminated surface analysis

Visual inspection of the simulation runs is an important first step towards intuitively understanding how neural networks could help in modelling and forecasting the *C. taxifolia* expansion. One way to assess this effect of the initialization parameters is by visually inspecting different plots obtained after a simulation run. Therefore, we have plotted the most informative curve, namely the contaminated surface versus time, with plausible parameters (Fig. 1). Visual inspection of the same curve but in log-scale (Fig. 2) clearly reveals yearly growing cycles with rapid increases during summer and stagnation during winter seasons. Indeed, the *C. taxifolia* expansion requires light and heat. A second and maybe more subtle point, is that the alga expansion exhibits three distinct regimes: (i) during the first 50 months, a regular growth is observed; (ii) then, a temporary stagnation occurs, corresponding to the bathymetric limit and the coastal edge; and (iii) a new expansion phase takes place, not as strong as the first though, corresponding to the expansion along the coast and in deep water. An easily interpreted plot of *C. taxifolia* expansion shows the variation rate $(x_t - x_{t-1})/x_{t-1}$ versus time (Fig. 3). The three regimes are more apparent. As expected, the

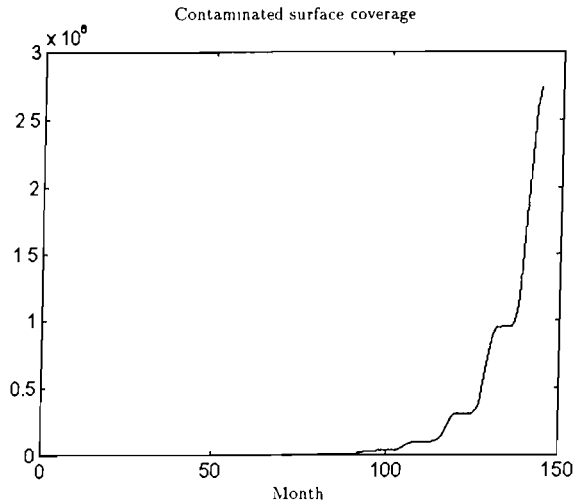

Fig. 1. Contaminated (normalized) surface versus time expressed in months.

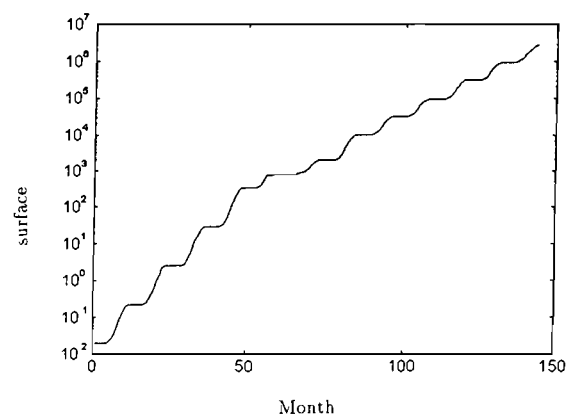

Fig. 2. Contaminated (normalized) surface in log scale versus time expressed in months.

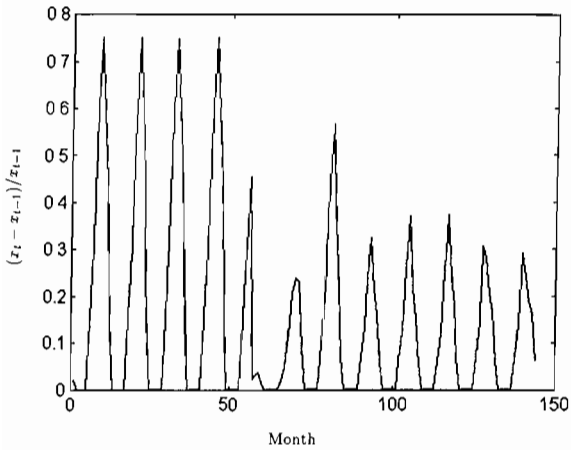

Fig. 3. Contaminated surface growing rate, $(x_t - x_{t-1})/x_{t-1}$, versus time expressed in months.

biomass increase is more or less proportional to the covered surface, therefore, we limit ourselves to the analysis and prediction of the contaminated surface.

To gauge the effects of the *C. taxifolia* microscopic description parameters on the subsequent evolution of the contaminated surface, some trial replications were run. The simulation initialization parameters were all arbitrarily varied within biological plausible ranges. Interestingly, no sensible change on the global shape of the curve, representing the contaminated surface coverage versus time, was observed.

Consequently, instead of trying to reconstruct the overall curve by performing iterated predictions several steps ahead, we decided first to forecast the final value and then fit a prototype curve between the first and the final points.

4.2. Input parameter selection

The objective being set, we now proceed to the selection of the relevant parameters influencing the contamination process. A relevant representation can make useful information explicit and strip away obscuring clutter. Different representations can be equivalent in terms of expressive power but differ dramatically in the efficiency to solve problems (Aussem et al., 1995). Also, the more parameters to be considered, the more parameters are required by the regression model, and thus the higher the risk of overfitting. Therefore, it is best for a small number of input variables to be inserted at the input of the neural network.

According to the experimental knowledge of the ecologists we are collaborating with, a very restricted subset of parameters was selected from the overall simulation initialization parameters for the prediction of the final contaminated surface level, namely:

- the ground projection surface of the cuttings,
- the ground projection surface of the stolons,
- the yearly maximum growing rate of *C. taxifolia*,
- the number of new cuttings per year.

Table 1 shows their respective plausible range of variation. The key parameters being identified, we now proceed to the training set construction.

4.3. 12-year contaminated surface forecasts

Coming up with good data sets addressing parameters of interest is tricky. The practical experience of the ecologists has limitations, which points to the inherent difficulty of selecting the best training set. Therefore, a uniform discretized lattice over the allowable ranges was used for the selection of the neural network training tuples. Afterwards, these values were linearly range-normalized onto the interval [0, 1].

Insufficient time and resources were the major stumbling blocks. For research reported in this paper, the number of replications per training pair was kept constant. Due to the expensive computational effort incurred in conducting each simulation run for obtaining 12-years contaminated

Table 1
Selected parameters and their corresponding range

Parameters/range	Default	Min	Max
Cutting ground projection (cm^2)	10	3	18
Stolon ground projection (cm^2)	0.5	0.2	0.8
Maximum annual growing rate (%)	300	200	500
No. of new cuttings per year	2	1	3

surface values (from a few minutes up to a few hours for a single simulation run on a bi-processor IBM work station), only ten replications per input tuple were carried out. A total of 6000 simulation replications were run.

The contaminated surface 12-year forecasts obtained by simulation were not found to vary drastically over the training tuples. Also instructive is the following observation: although each simulation run-for the same input parameters-yields a somewhat distinct trajectory, they all converged to the an almost identical contaminated surface value (the variance is negligible, below 10^{-2}).

Finally, the data set is made up from 600 pairs consisting of the four selected input parameters and the two target values, namely the average contaminated surface level 12 years in advance, provided by the simulation software, and its variance. Both targets were afterwards range-normalized in the unit interval.

To assess the generalization ability of the metamodels, 5 distinct training + test sets were constructed. The training set (strictly speaking) were made up from 80% randomly selected patterns and the test set made up from the 20% remaining patterns for validation purposes.

4.4. Intermediate results

Preliminary tests have indicated diminishing returns beyond five nonlinear units for both tasks. The two networks used to learn the series have a total of 26 links. The networks were stopped when the minimum of the mean-squared error in the training set was reached and their performance were afterwards tested on the test set.

Predictions of the average surface on the test set are shown in Fig. 4 for a particular test set. Results were far above our expectations. As can be seen, the task was perfectly learned by a relatively small neural network. The five independent trainings performed with the five data sets yielded the same result, in terms of *NMSE* over the test set, here $10^{-2.7}$ The predictions did not degrade from one test set to the other, indicating very good generalization ability. Predictions of the variance (not shown) yielded comparable performance with a $NMSE = 10^{-2.4}$.

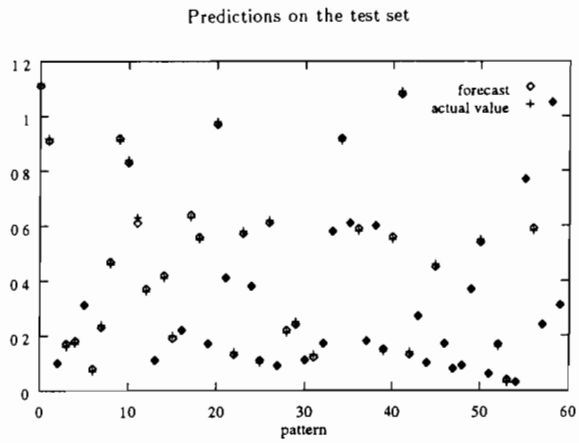

Fig. 4. A 12-year forecasts of the normalized surface contaminated by *C. taxifolia* on the test set.

4.5. Reconstructing the overall curve

With a view towards reconstructing with reasonable assuredness the overall curve—and not only the final value—, all 6000 surface curves obtained by intensive simulation runs were normalized and afterwards averaged to obtain a prototype curve. Unfortunately, our attempts to reconstruct the overall curve by stretching out the prototype curve up to the predicted final surface point, were not as successful. Indeed, as discussed previously, three distinct regimes are observed: existence of these regimes is closely related to the bathymetry.

For the same input parameters, the three regimes occur at the same time. Also, we decided to forecast the month, the magnitude and its variance of the two breakpoints (i.e. the X-coordinate, the Y-coordinate and $Var(Y)$) using the same four input parameters. With this in mind, six training + test sets were constructed from 600 patterns each. Six distinct neural networks were trained separately in order to infer the exact location of these two breakpoints.

The resulting networks had between 20 and 50 links including bias. No more than five nonlinear units were used. Selection of these dimensions were mostly by trial and error, along with various heuristics. In performing these preliminary simulations, we stress that we have made no effort to

obtain an optimal architecture. Instead, we focused on these configurations and tried to obtain the lowest mean-squared error on the training set, and assessed the performance on the test set. Note that in general, selection of dimensions for neural networks remains a difficult problem in need of further research.

Final results are encouraging. The neural networks have learned to infer the month and the average surface with very good accuracy: *NMSE* varied between 10^{-16} and 10^{-3}. Predictions of the surface variance were not as good though: $NMSE = 0.2$ for the first point and 0.3 for the second. We believe that predicting the variance of the surface at these points depends on further parameters that were not taken into account in our approach.

Therefore, we constructed basic confidence intervals for each testing tuple, given the estimate of the average surface at the first point, at the second point, and at the last point 12 years ahead, and their variance. Again, simulations runs were independent and the normality of the responses were verified so that the necessary assumptions held.

To gauge the practical relevance of our metamodels, we ran the neural networks and the simulation model with randomly selected input parameters within their respective range (see Table 1) and compared the curves. Comparisons were done between the 95% confidence intervals provided by the simulator and the neural network metamodel. As seen in Fig. 5, very encouraging results were obtained. The 95% confidence interval of the covered surface reconstructed from the neural network overlaps the confidence interval given by the simulation model. Significant reduction in computational time can thus be achieved without a sacrifice in prediction accuracy.

4.6. Discussion

Our approach may be criticized on two accounts. First, few replications were run for each input tuple: the number of replications (10) was set by balancing the computation cost with the desired accuracy of the estimate. This raises the question of whether our confidence intervals of the covered surface are statistically reliable.

Second, the results presented in this paper only relate to a single site. Though not discussed here, a topic of considerable interest would be to analyze the differences in terms of performance between several metamodels trained at different sites to find out whether the contaminated surface 12 year forecasts is also deterministically related to the same limited set of parameters. This needs further substantiation through more experiments and analysis. We are also interested in assessing the usefulness of the neural network metamodelling approach for other variables and *C. taxifolia* local expansion patterns.

Third, the training and generation of confidence intervals requires several replication of the same process. Therefore, because the actual data about the colonization process only reflect a single trajectory of the process (moreover over a period of time less than 12 years), actual data cannot be used for training purposes. As far as validation is concerned, the simulation model runs were already shown to be in good agreement with the observations (not only in terms of

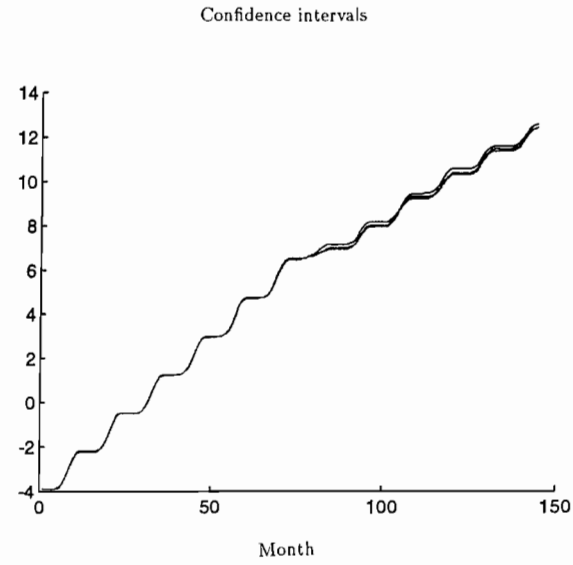

Fig. 5 *C taxifolia* expansion curves in log scale versus time in months; in plain line, the 95% confidence interval output by the simulator, in grey line, the 95% confidence interval reconstructed from the neural network forecasts.

the contaminated surface but also the two-dimensional maps). Actual data collected on different sites will be used when the neural network will take in to account bathymetric and substrate information in order to be able to extrapolate to new sites. Results will be reported in due course.

5. Conclusion

In this article we discussed work carried out with supervised neural networks, as a metamodelling technique for discrete event stochastic simulation, with a view to dramatically reducing the computational burden involved by the simulations. The studies currently being undertaken are aimed to be incorporated later into a prediction package for exploring numerous possible trajectories of *C. taxifolia*, a green alga provoking intense and irreversible alteration of the coastal ecosystems, including endogenous species distribution. The preliminary work presented in this paper addressed the prediction of the covered surface 12 years in advance. A neural network was successfully trained on artificially generated data provided by the simulation runs to provide reliable forecasts. The overall expansion curve was also reconstructed with reasonable accuracy. Future work will address other variables, such as expansion patterns at different sites.

Acknowledgements

We thank Cédrick Laballery, Jérôme Bart, Yannick Tranchier and Damien Azambourg of ISIMA/University of Blaise Pascal, Clermont-Ferrand, France, for their contribution to some ideas reported in this paper. We are grateful to James Paul Hoffman for careful reading of this manuscript. We also thank the anonymous referees for numerous suggestions that greatly improved the clarity of the manuscript. This work was supported by the European LIFE DGXI 95/F/A31/EPT/782 program and is now funded by the French Ministry of Environment (ref. MATE/98154).

References

Aussem, A., 1998 Nonlinear modelling of chaotic processes with dynamical recurrent neural networks In: Proceedings from the Neural Networks and Their Applications NEURAP'98, Marseille, France, pp. 425–433.

Aussem, A., 1999. Dynamical dynamical recurrent neural networks towards prediction and modeling of dynamical processes. Neurocomputing (in press).

Aussem, A., Murtagh, F., Sarazin, M., 1995. Dynamical recurrent neural networks-towards environmental time series prediction. Int. J. Neur. Sys. 6 (2), 145–170.

Belsher, T., Meinesz, A., 1995. Deep-water dispersal of the alga *Caulerpa taxifolia* introduced in the Mediterranean. Aquat. Bot. 51, 163–169.

Bradley, E., 1982. The Jackknife, The Bootstrap and Other Resampling Plans. SIAM, no. 38.

Coquillard, P , Hill, D., 1997. Modélisation et Simulation des Ecosystèmes. Masson, Paris, p. 273.

Cybenko, G , 1989 Continuous value neural networks with two hidden layers are sufficient. Math Control Sign. Sys. 2, 303–314.

Geman, S , Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neur. Comput. 4 (1), 1–58.

Gordon, G., 1978. System Simulation Prentice Hall, Englewood Cliffs, NJ, p. 299.

Hertz, J., Krogh, A , Palmer, R., 1991. An Introduction to the Theory of Neural Computation. Addison-Wesley, Redwood City, CA, p. 124.

Hill, D., 1996. Object-Oriented Analysis and Simulation. Addison-Wesley Longman, UK, p. 291.

Hill, D., 1997. Modélisation de processus d'expansion: Application à *Caulerpa taxifolia*. French Science Academy, Tec and Doc, pp. 219–230.

Hill, D., Coquillard, P , De Vaugelas, J., 1997. Discrete-event simulation of alga expansion. Simulation 68 (5), 269–277.

Hill, D., Coquillard, P., De Vaugelas, J , Meinesz, A., 1998. An algorithmic model for invasive species application to *Caulerpa taxifolia* (Vahl) C. Agardh development in the north-western Mediterranean sea. Ecol Model. 109, 251–265.

Kilmer, R.A , 1994. Artificial Neural Network Metamodels of Stochastic Computer Simulations. Ph D. Dissertation, University of Pittsburgh.

Kleijnen, J.P C , 1987. Statistical Tools for Simulation Practitioners. Marcel Dekker, New York.

Meinesz, A., Hesse, B., 1991. Introduction et invasion de l'algue *Caulerpa taxifolia* en Méditerranée nord occidentale. Oceanol. Acta 14 (4), 415–426.

Murtagh, F., Cambpell, J , Zheng, G , Aussem, A., Ouberdous, M., Demirov, E., Eifler, W., Crepon, M., 1998. Data imputation and nowcasting using clustering and connectionist modelling. In: Proceeding COMPSTAT'98, International Conference on Computational Statistics, Springer-Verlag, Bristol, pp. 401–406.

Pierreval, H., Huntsinger, R.C., 1992 An investigation of neural network capabilities as simulation metamodels. In: Proceedings of the 1992 Summer Simulation Conference, pp. 155–162.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing Explorations in the Microstructure of Cognition, vol. 1, MIT Press, Bradfords Books, Cambridge, MA, pp. 318–362.

Weigend, A.S., Rumelhart, D.E., Huberman, B.A., 1990. Predicting the future: a connectionist approach. Int. J. Neur. Sys. 1, 195–220.

Zheng, G., Rouxel, S., Aussem, A., Campbell, J., Murtagh, F., Ouberdous, M., Demirov, E., Eifler, W., Crepon, M., 1997. Forecasting of ocean state using satellite-sensor data. In: Murtagh F., Campbell J.G., McKevitt P. (Eds.), Eighth Ireland Conference on Artificial Intelligence (AI–97), vol. 1, University of Ulster, pp. 234–240.

ELSEVIER

Ecological Modelling 120 (1999) 237–246

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Applying artificial neural network methodology to ocean color remote sensing

Lidwine Gross [a,*], Sylvie Thiria [a], Robert Frouin [b]

[a] *LODYC, UPMC (Paris 6), 4 place Jussieu, 75252 Pariscedex 05, France*
[b] *SIO, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0221, USA*

Abstract

Artificial neural networks (ANN) are widely used as continuous models to fit non-linear transfer functions. In this study we used ANN to retrieve chlorophyll pigments in the near-surface of oceans from Ocean Color measurements. This bio-optical inversion is established by analyzing concomitant sun-light spectral reflectances over the ocean surface and pigment concentration. The relationships are complex, non-linear, and their biological nature implies a significant variability. Moreover, the sun-light reflectances are usually measured by satellite radiometers flying at 800 km over the ocean surface, which affect the data by adding radiometric noise and atmospheric correction errors. By comparison with the polynomial fit usually employed to treat this problem, we show the advantages of neural function approximation like the association of non-linear complexity and noise filtering.

Keywords: Ocean color; Inversion; Noise filtering; Artificial neural networks

1. Introduction

Quantitative assessment of oceanic primary production and its role in the global carbon cycle is a critical environmental and scientific issue (JGOFS, 1987; Abbot et al., 1994; Falkowski, 1994). Knowledge of primary production is necessary to determine the biomass variability of the ocean, derive the effect of biological processes on the partial pressure of carbon dioxide (CO_2), and therefore, better understand how phytoplankton carbon fixation affects the net CO_2 flux across the air-sea interface (Behrenfeld et al., 1998). Primary production depends on light availability and other environmental factors (temperature, nutrients), and on the amount of phytoplankton present for photosynthesis (Morel, 1991). The amount of phytoplankton and their optical properties (absorption, scattering) affect the spectral diffuse reflectance of the ocean, defined as the ratio of upwelling to downwelling irradiance at a given depth. Since phytoplankton pigments generally absorb more in the blue than in the green, the greener the water, the more phytoplankton (Clarke et al., 1970). Thus by measuring ocean color, i.e. the spectral reflectance at zero depth, $R_w(\lambda)$, one can obtain estimates of phytoplankton pigment concentration.

* Corresponding author. Fax: +33-1-44277159.
E-mail address: lgr@lodyc.jussieu.fr (L. Gross)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00105-2

A variety of optical transfer functions (bio-optical models) have been proposed to quantify the influence of chlorophyll pigments on spectral reflectance. The bio-optical relationships are generally established by analyzing concomitant reflectances and pigment data (concentration, inherent optical properties). They are complex and non-linear, making inversion difficult for phytoplankton content retrieval. In remote sensing, the most popular algorithms to estimate phytoplankton pigment concentration utilize simple ratios of reflectances in the blue and green or combinations of ratios (Aiken et al., 1995; O'Reilly et al., 1998). Standard algorithms are based on the ratio of reflectances at 443 and 555 nm, $R_w(443)/R_w(555)$, or 490 and 555 nm, $R_w(490)/R_w(555)$ (denoted hereafter RR443 and RR490). The logarithm of the pigment concentration, C, is computed from the logarithm of the reflectance ratio using a third order polynomial fit (Andre and Morel, 1991; O'Reilly et al., 1998).

Blue–green ratios, when applied to satellite-derived marine reflectances, are affected by atmospheric correction errors. Atmospheric correction is difficult to be done accurately, because typically 80% of the signal measured at satellite altitude originates from the atmosphere (Viollier et al., 1980). Typical errors of 5–10% on the reflectance in the blue are expected with current atmospheric correction schemes, but they may be much larger in the presence of dust or pollution-type aerosols (Gordon and Wang, 1994; Gordon, 1997). Even though atmospheric correction errors are correlated spectrally, they may not cancel in a ratio, yielding significant, even unacceptable errors on phytoplankton pigment retrievals. Estimates obtained with the Coastal Zone Color Scanner (CZCS) had errors of about 40–50% at low pigment concentrations (Gordon et al., 1980, 1983), but part of the errors might be due to phytoplankton type variability.

As shown in Thiria et al. (1993), artificial neural networks (ANN) are good candidates for modeling geophysical transfer functions for they can approximate a wide range of non-linear continuous functions (Bishop, 1995). This property and the ability of ANN to model noise can be exploited, in certain conditions, to filter measurement noise during a model calibration, which is interesting when dealing with real data. They have been used in a number of geophysical applications, but it is only recently that attempts have been made to retrieve ocean color variables with the help of ANN (Schiller and Doerffer, 1999; Keiner and Brown, 1999). In the present study, we propose a multi-layered perceptron (MLP), to compute phytoplankton pigment concentration (chlorophyll-*a* plus phaeophytin) from satellite-derived marine reflectances. We examine the differences between this function approximator and the cubic polynom that is usually employed to do this task, and the contribution of additional spectral bands in the case of the use of MLP. We perform a formal analysis of the capability of ANN to filter noise, thus we focus mainly on methodology. To calibrate our models, we use simulated datasets which take into account radiometric noise and residual atmospheric correction errors. This procedure allows us to verify the importance of the presence and nature of simulated noise when calibrating the ANN, since we show that taking into account noise is necessary to extract the inherent information of the geophysical signal and obtain an operational function. We choose a particular ocean color radiometer, the sea-viewing wide field-of-view sensor (SeaWiFS) onboard the SeaStar satellite, which measures reflected sunlight in five spectral bands centered at 412, 443, 490, 510, and 555 nm (Hooker et al., 1992), but the same type of analysis could be performed for any other ocean color radiometer.

2. Datasets used for ANN calibration

Since it is difficult to gather the necessary amount of data to educate ANN properly, we used simulated datasets (pairs of marine reflectances and pigment concentration). By contrasting results obtained using clean (i.e. non-noisy) and noisy data, we expected to gain information on the applicability of the theoretical model to real observations. We only considered Case I waters, i.e. waters for which optical properties depend mostly on phytoplankton pigments, since these

waters constitute more than 90% of the world ocean (Morel, 1988).

We calculated spectral marine reflectance, R_w, as a function of phytoplankton pigment concentration, C, using the bio-optical model of Morel (1988). This model incorporates average bio-optical parameters determined by regression analysis on in-situ measurements. Those parameters like absorption and scattering coefficients, on the other hand, vary with the type of phytoplankton population (natural assemblages) and biological cycles (Bricaud et al., 1995; Garver and Siegel, 1997). In the simulations, however, we did not take into account variability due to phytoplankton type. The modeled marine reflectances, therefore, depend only on C. We varied C from 0.03 to 30 mg m^{-3}, the domain of validity of the model. The simulations were made for the SeaWiFS spectral bands centered at wavelengths of 412, 443, 490, 510, and 555 nm (λ_i, with $i=1, 2,\ldots, 5$, respectively). We used new values for the absorption coefficient of optically pure sea water reported by Pope (1993). Consequently, we adjusted the Morel (1988) diffuse attenuation coefficient for phytoplankton, since this coefficient was computed by subtracting the contribution of pure oceanic waters based on previous estimates.

Two types of data were generated, they are summarized in Table 1. The first type of data, hereafter referred to as Type 1 data, was obtained using the Morel (1988) model, modified as indicated above. No noise was added to the modeled reflectances, nor to the pigment concentrations. These non-noisy data will be used to demonstrate the ability of the MPL network to inverse a complex bio-optical function; they will also serve as a reference in the study of the effects of noise on the performance of the MLP. The second type of data, hereafter referred to as Type 2 data, includes simulated SeaWiFS-derived reflectances. These reflectances were obtained by adding a three-component noise, Δ_{atm}, to the Type 1 reflectances. The noise is due to (1) radiometric performance; (2) imperfect atmospheric correction; and (3) passage from bi-directional reflectance just above the surface (the SeaWiFS product after atmospheric correction) to irradiance reflectance just below the surface (the variable related to C in Morel's model). When computing Δ_{atm}, we assumed that the atmospheric correction is performed according to Gordon and Wang (1994), that is by obtaining aerosol information in the near-infrared where the ocean is 'black' and extrapolating the information to the visible.

Table 1
Statistical ensembles for models calibration

Ensemble	Description
Type 1	$R_{w1}(\lambda_i) = g(C, \text{mg m}^{-3})$ from Morel (1988)
Type 2	$R_{w2}(\lambda_i) = R_{w1}(\lambda_i) + \Delta_{atm}(\lambda_i)$

Table 2
Notation of the different inverse models

Ensemble used for calibration	Artificial neural networks	Polynomial fits
Type 1	ANN-1	RR443-1 and RR490-1
Type 2	ANN-2	RR443-2 and RR490-2

In order to have statistically significant results, we dealed with a large amount of data. The inverse models (ANN, RR443 and RR490) were calibrated using the same 5000 vectors of simulated pairs of $\{C^k, R^k_{wn}(\lambda_i), i=1\ldots5\}$ ($n=1$ or 2, $k=1, 2,\ldots5000$) and tested with independent test sets of 10 000 simulated vectors with same characteristics called Test-1 for $n=1$ or Test-2 for $n=2$. We obtained consequently two different neural inverse models and four different classical inverse models denoted as described in Table 2. We estimated models performance using different index, RMS error and relative RMS error which allows to free from absolute values:

$$\text{RMS} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} [C^k - F(\vec{R}^k_w, W)]^2} \quad (1)$$

$$\text{rel.RMS} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \left[\frac{C^k - F(\vec{R}^k_w, W)}{C^k}\right]^2} \quad (2)$$

where N_{test} is the number of patterns in the test set and $F(\vec{R}_w^k, W)$ refers to the inverse model.

3. Artificial neural network methodology

In this section we provide a basic description of the multi-layered perceptrons (MLP), and of the properties they offer for non-linear regression. Then we discuss the numerical methodology employed to optimize the parameters of such non-linear models.

A neuron is an elementary transfer function which calculates an output s when an input A is applied:

$$s = f(A) \tag{3}$$

where f is called the transition function and is usually non-linear. An artificial neural network is formed by interconnected neurons, each neuron receiving and sending signals only to the neurons to which it is connected. Multi layered perceptron (MLP) is a particular class of artificial neural networks in which neurons are organized in several layers. The state s_j of a neuron j is computed by $s_j = f(A_j)$, where A_j is the total information received from the other neurons s_h computed as a weighted sum $A_j = \Sigma_h w_{jh} s_h$. The transition function f can be linear, i.e. $f(u) = u$ for the exit layer, or a sigmoid with

$$f(u) = a\frac{\exp(\alpha u) - 1}{\exp(\alpha u) + 1} \tag{4}$$

We dealt with $a = 1.7159$ and $\alpha = 1.3333$ so that f was quasi-linear in the range $[-1, 1]$, $f(-1) = -1$ and $f(1) = 1$. The w_{jh} are the connection weights from h to j; they are real numbers parameterizing the influence of the connected neurons. The weight matrix $W = [w_{jh}]$ defines the MLP specificity.

Theoretical considerations show that MLP's are universal function approximators (Bishop, 1995). Given the flexibility of ANN, we chose to take into account all the available information, so we related the five SeaWiFS spectral reflectances to the pigment concentration. In order to avoid numerical saturation, we used a logarithmic coding for C and then we normalized both the inputs and the output. Our problem was then to determine the architecture of the MLP, i.e. to decide the number of neurons and the way they are connected, which represents choosing a function's family in which we seek the best multidimensional real function allowing us to approximate the transfer function between the vector $\{R_w(\lambda_i),\ i = 1...5\}$ and the scalar C. We determined the best architecture using a constructing methodology which makes intensive use of cross validation (Bishop, 1995). For inverting the Morel (1988) model (Type 1 data), we set a completely connected MLP with five inputs (the five spectral in-waters reflectances), two hidden layers of six and four sigmoid neurons, and one linear output which gives the concentration C. This network, denoted ANN-1 hereafter, has 69 parameters to be adjusted (see its structure Fig. 1). We used the same architecture to invert the noisy reflectances (Type 2 data) and denoted ANN-2 the MLP obtained when trained on those data (see Table 2).

We then solved a regression problem since we had to optimize the network's weights to obtain the best estimated model. The w_{jh} values were computed by a calibration process (called learning phase) in which the inputs and output of the MLP were well-defined data sets, and the w_{jh} values were the control variables. This learn-

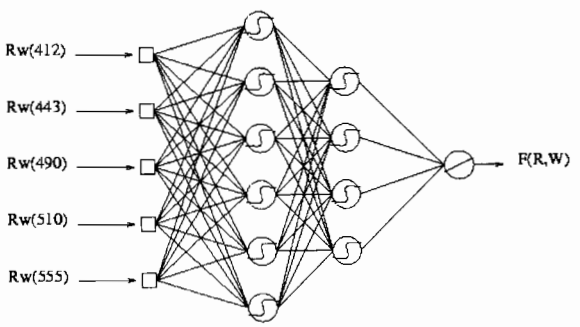

Fig 1 Architecture of the MLP which models the inverse problem of ocean color for the case of SeaWiFS data, the $R_w(\lambda_i)$ are the five spectral reflectances and $F(\vec{R}_w^k, W)$ is the output of the network giving the estimated pigment concentration C.

ing process was based on a minimization where the cost function is:

$$J(W) = \sum_{k=1}^{N} [C^k - F(\vec{R}^k, W)]^2 \quad (5)$$

where N is the number of observations in the learning ensemble, C^k is the desired concentration of the observation k, and $F(\vec{R}^k, W)$ the corresponding concentration computed by the neural model, which is a function of the reflectance vector and internal parameters set by the weight matrix W. A necessary condition to minimize J is to find the neural weight matrix W^* so that:

$$\nabla J(W)|_{W=W^*} = 0 \quad (6)$$

To approach the minimum of this multi-dimensional cost function, we used a classical gradient descent technique which is an iterative optimization method, adapted to MLP by the mean of the gradient backpropagation (Bishop, 1995). The weights of the MLP were first randomly initialized between -1 and 1 according to a uniform probability distribution (matrix W_0). Then, each step of the algorithm modified the whole weight matrix by the equation:

$$W_{t+1} = W_t - \varepsilon \cdot \nabla J(W_t) \quad (7)$$

where ε, the learning rate, was set in our case to the same value for the whole MLP ($\varepsilon = 0.01$). Series of cross validation tests allowed us to control the quality of the minimum estimation and of the generalization. The theory shows that, if the architecture of the MLP is well-chosen and the learning phase is well achieved, the MLP gives an approximation of the mean field of the variable C, more precisely the conditional average of the concentration C for each point $\{R_w(\lambda_i),\ i = 1 \ldots 5\}$. When the calibration is done, the MLP inverse model does algebraic operations only, leading to fast computation. The three order polynoms RR443 and RR490 were also calibrated using a least square method, so that the performance of classical approaches and ANN were comparable.

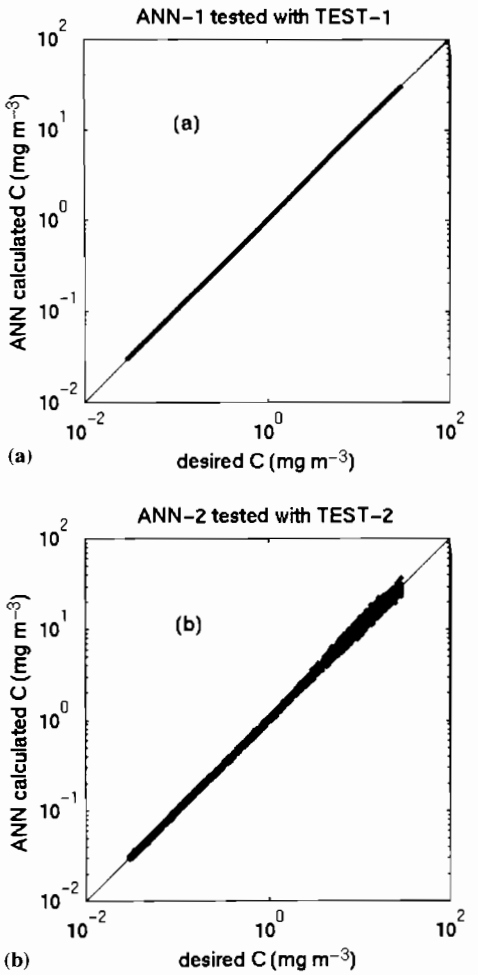

Fig. 2. (a): Test of ANN-1 on Test-1 ensemble. (b): Test of ANN-2 on Test-2 ensemble.

4. Performance of the ANN-1 and ANN-2 models

This section is a formal study using simulated data. We explore the ability of ANN to take into account the problem of residual atmospheric correction. We discuss the performance of ANN-1 and ANN-2 models after calibration on two independent test sets related to the two different data type. In the following, the restitution of pigment concentration by ANN inverse models is controlled by mean of scatter plots. Fig. 2a shows the good neural inversion ANN-1 on Type 1 data, and so the ability of ANN to invert complex mathematical functions. Fig. 2b

shows the performance of ANN-2, the neural inverse model calibrated with atmospherically noisy data (Type 2). The curve is scattered by the simulated uncertainty of the measurements but there is no bias. Further investigations using cross tests allow to understand the properties of the two ANN inverse models. We first test ANN-2 on Type 1 data (i.e. data with no noise). The scatter plot of Fig. 3a proves that ANN-2 is a generalization of ANN-1. On the contrary, ANN-1 cannot deal with noisy data: the scatter plot of ANN-1 when testing with noisy data shows a degradation of the performance (Fig. 3b). Table 3 gives the different quality index obtained for the four different experiments. Clearly, the performance of 5.48% reached by ANN-2 when dealing with Test-2 indicates a good fit of the data. ANN-1 and ANN-2 exhibit a similar behavior showing the ability of ANN-methodology to take into account noise effect. We give in the following an extensive study of this result.

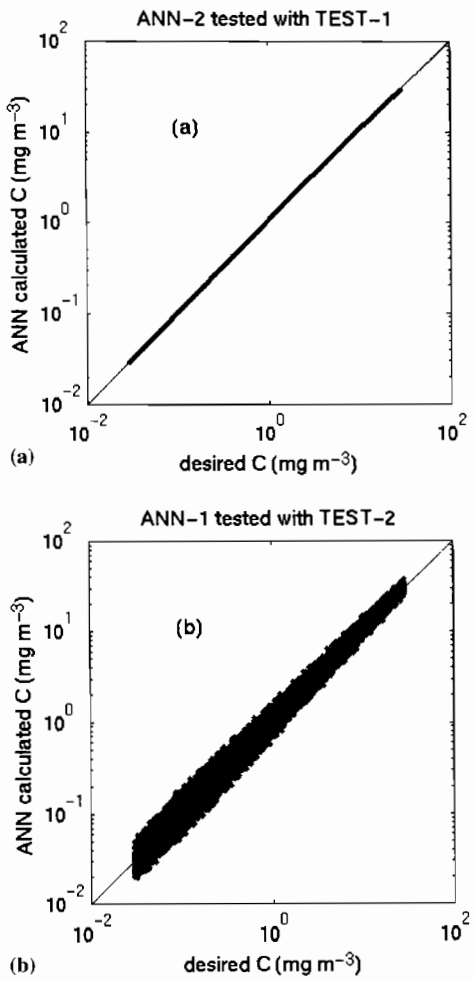

Fig. 3 (a): Test of ANN-2 on Test-1 ensemble. (b): Test of ANN-1 on Test-2 ensemble.

Table 3
Performance of ANN-1 and ANN-2 tested on Test-1 and Test-2 ensembles

Statistical parameter	Test-1		Test-2	
	ANN-1	ANN-2	ANN-1	ANN-2
RMS error (mg m^{-3})	0.024	0.135	0.926	0.730
rel. RMS (%)	0.14	2.85	21.90	5.48

5. Comparison with band ratios

We can enlighten the results we get by comparing the neural models performances to those of the classical polynomial inverse fits of ocean color. We calculated the performances of the polynomial fits based on band ratio RR443-1 and RR490-1 calibrated on Type 1 data and of RR443-2 and RR490-2 calibrated on Type 2 data (see Table 2). The inversion of Type 1 data is correct for both ANN and polynomial fits (Figs. 4a and 4b). While the ANN-1 inversion is quasi-perfect from small to large values of concentration, the accuracy given by the polynomial fits decreases (in a oscillating way) with the augmentation of concentration value (from ± 2 to $\pm 10\%$ for RR490-1 and ± 5 to $\pm 14\%$ for RR443-1, but those performances stay reasonable. The scatter plots for the two methods RR443-2 and RR490-2 given in Figs. 5a and 5b are to be compared with Fig. 2b. The fit is less pigment concentration dependant for ANN-2 inverse model than for RR443-2 and RR490-2 which present difficulties in recovering small and

high pigment concentration. All the performances are summarized in Table 4.

Fig. 6 shows the relative RMS error of the three inverse models, ANN-2, RR443-2 and RR490-2, plotted for different ranges of C. When dealing with atmospheric correction error, the polynomial fits have great difficulties to give a sufficient accuracy. For small values of concentration, the most usual values, relative errors arise between ± 30 and $\pm 50\%$. Those results correspond to reality when satellite data are treated with band ratios. On the other side, ANN-2 gives an accuracy around $\pm 3\%$, when staying under 10 mg m^{-3}.

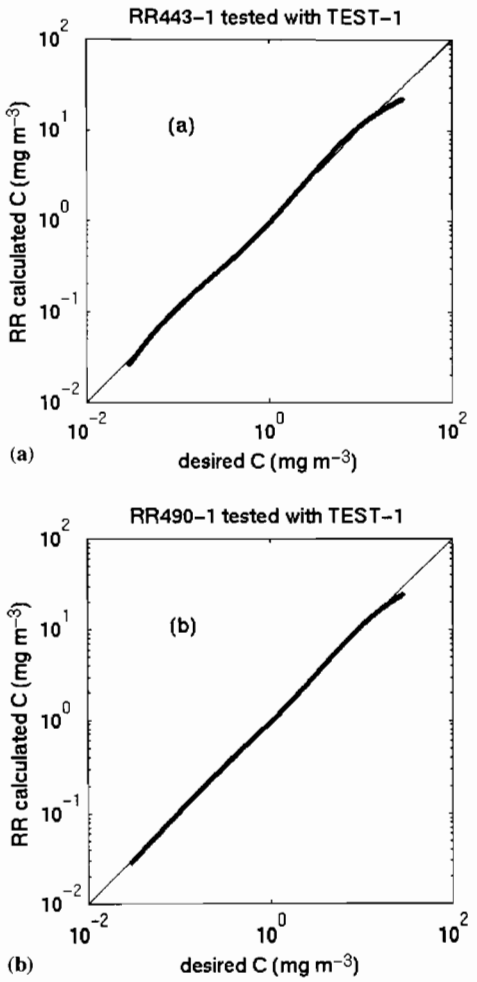

Fig. 4. (a): Test of RR443-1 on Test-1 ensemble. (b): Test of RR490-1 on Test-1 ensemble.

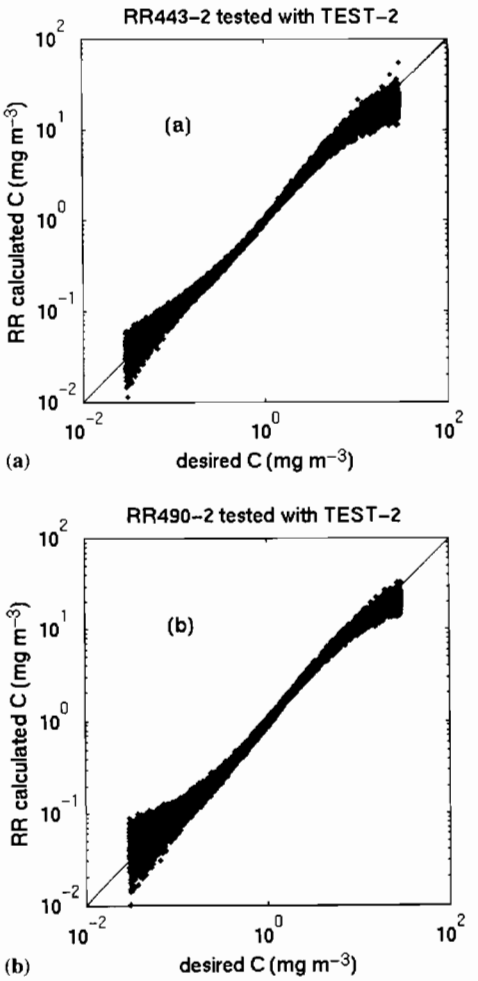

Fig. 5. (a): Test of RR443-2 on Test-2 ensemble (b): Test of RR490-2 on Test-2 ensemble

6. Discussion and conclusion

In the present paper, we propose an artificial neural network methodology to solve Ocean Color inverse problem, e.g. to retrieve the ocean chlorophyll pigment concentration from satellite derived in-water reflectances. We calibrated several models using two different type of simulated data. The first artificial neural network

Table 4
Performance of RR443-1, RR490-1, RR443-2 and RR490-2 tested on Test-1 and Test-2 ensembles

Statistical parameter	Test-1		Test-2	
	RR443-1	RR490-1	RR443-2	RR490-2
RMS error (mg m^{-3})	1.245	0 809	2 105	1.693
rel. RMS (%)	8.59	5.14	18.59	23.29

(ANN-1) inverts a mathematical model which is the bio-optic model of Morel (1988) while the second (ANN-2) was designed to invert satellite derived reflectances. To simulate the data corresponding to ANN-2, we added to the in-water reflectances calculated by the bio-optic model an estimation of the radiometric noise and the atmospheric correction errors resulting from atmospheric correction algorithms (we took the case of SeaWiFS). The performance of each neural model were compared to those of the classical three order polynomial fits based on band ratio usually employed to retrieve pigment concentration. These experiences allow us to understand when ANN improve the restitution of chlorophyll and to show the importance of simulating the appropriate noise when the data sets used to calibrate the models are simulated.

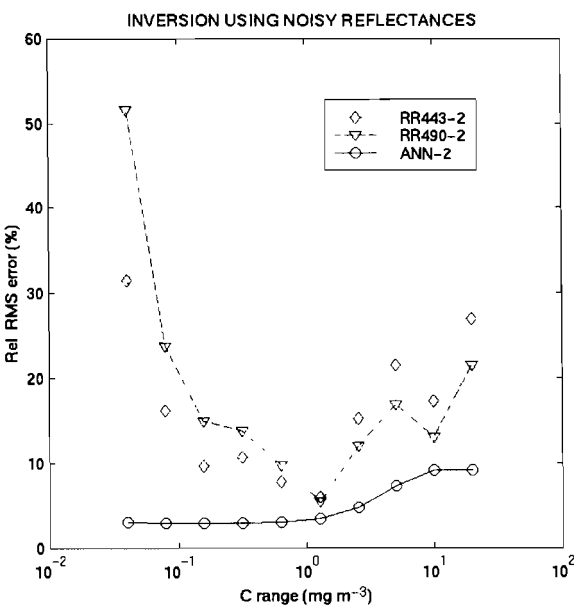

Fig. 6 Compared relative RMS error between ANN-2, RR443-2 and RR490-2 all tested on Test-2 ensemble

The ANN-1 model gives a quasi-perfect inversion of the model of Morel ($\pm 0.14\%$ of relative RMS error), and the polynomial fits gives also reasonable results for biological studies (less than $\pm 15\%$ of error). We although denote the instability of the accuracy given by the polynomial fits on the whole range of concentration C, which is probably due to the intrinsic oscillations of the polynomial family. The ANN is by definition a very soft function approximator, which is here an obvious quality.

The performance of the ANN-2 model are much better than the polynomial fits (ANN-2 gives a $\pm 3\%$ accuracy and the polynomial fits are between ± 30 and $\pm 50\%$ for small concentration values). This shows that ANN-2 is able to filter atmospheric correction errors using the information given by the five channels of SeaWiFS.

ANN are complex and stable function approximators able to deal with noise measurement and very adaptive systems. If the neural models are calibrated with simulated data as in the present experiments, the quality of the simulation of the measurement noise is fundamental. Using a model like Morel (1988) to simulate in-waters reflectances and adding a simple estimation of atmospheric correction noise allowed us to accomplish an accurate inverse transfer function between satellite-derived marine reflectances and chlorophyll concentration. This formal experiment made only on simulated data allows us to envisage ANN to deal with satellite reflectances. A more realistic data simulation including the biologic variability of phytoplankton type should lead to real application. We can

also envisage the use of real measurements in the calibration process to improve the ANN performance.

Acknowledgements

This work was supported by the Centre National de la Recherche Scientifique, the Centre National d'Etudes Spatiales, by Aerospatiale, by the European Community NEUROSAT Programme (ENV4-CT96-0314), by the National Aeronautics and Space Administration under grants NAG5-6202 (to R. Frouin) and the National Space Development Agency of Japan under contract G-0035 (to R. Frouin). We thank M. Crépon and C. Mejia of the Laboratoire d'Océanographie Dynamique et de Climatologie, Université Pierre et Marie Curie for stimulating discussions.

References

Abbot, M.R., Brown, O.B., Evans, R.H., Gordon, H.R., Carder, K.L, Muller-Karger, F.E., Esaias, W.E., 1994. In: Hooker, S.B., Firestone, E.R. (Eds.), Ocean color in the 21th century. a strategy for a 20 year time series NASA tech, Memo 104566(29), SeaWiFS technical report series, p. 34.

Aiken, J., Moore, G.F., Trees, C.C., Hooker, S.B., Clark, D.K., 1995 In: Hooker, S.B., Firestone, E.R. (Eds.), The SeaWiFS CZCS-type pigment algorithm. NASA tech, Memo 104566(29), SeaWiFS technical report series, p. 34.

Andre, J.M., Morel, A., 1991. Atmospheric corrections and interpretation of marine radiances in CZCS imagery, revisited. Oceanological Acta 14, 3–22.

Behrenfeld, M.J., Falkowski, P.G., Esaias, W.E., Balch, W., Campbell, J.W., Iverson, R.L., Kiefer, D.A., Morel, A., Yoder, J.A., 1998. In: Hooker, S.B., Firestone, E.R. (Eds.), Toward a consensus productivity algorithm for SeaWiFS. NASA tech, Memo 104566(42), 18-30, SeaWiFS technical report series, p. 36

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford, p. 482.

Bricaud, A., Babin, M., Morel, A., Claustre, H., 1995. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: analysis and parameterization. J. Geophys Res. 100 (C7), 13321–13332.

Clarke, G.L., Ewing, G.C., Lorenzen, C.J., 1970 Spectral of backscattered light from the sea obtained from aircraft as a measurement of chlorophyll concentration Science 167, 1119–1121

Falkowski, P.G., 1994 The role of phytoplankton photosynthesis in global biogeochemical cycles. Photosyn. Res. 39, 235–258

Garver, S.A., Siegel, D.A., 1997. Inherent optical property inversion of ocean color spectra and its biogeochemical interpretation. 1. Time series from the Sargasso Sea J. Geophys. Res 102, 18607–18625

Gordon, H.R., Wang, M., 1994. Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm. Appl. Optics 33 (3), 443–452

Gordon, H.R., 1997. Atmospheric correction of ocean color imagery in the Earth Observing System Era J Geophys Res 102, 17081–17106

Gordon, H.R., Clark, D.K., Mueller, J.L., Hovis, W.A., 1980. Phytoplankton pigments derived from Nimbus 7 CZCS: Initial comparisons with surface measurements. Science 210, 63–66

Gordon, H.R., Clark, D.K., Brown, J.W., Brown, O.B., Evans, R.H., Broenkow, W.W., 1983. Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparisons between ship determinations and coastal zone color scanner estimates Appl. Optics 22, 20–36

Hooker, S.B., Esaias, W.E., Feldman, G.C., Gregg, W.W., McClain, C.R., 1992. An overview of SeaWiFS and Ocean Color, NASA tech. Memo 104566 (1), SeaWiFS Technical report Series, Greenbelt, MD

JGOFS, The Joint Global Ocean Flux Study: background, goals, organization, and next steps, 1987. SCOR, ICSU, Paris.

Keiner, L.E., Brown, C.W., 1999. Estimating oceanic chlorophyll concentrations with neural networks. Int. J. Rem. Sen. 20 (1), 189–194.

Morel, A., 1988 Optical modeling of the upper ocean in relation to its biogenous matter content (case 1 waters). J. Geophys. Res., 93 (C9), 10749–10768.

Morel, A., 1991. Light and marine photosynthesis: a spectral model with geochemical and climatological implications. Prog. Oceanogr. 26, 263–306.

O'Reilly, J., Maritorena, S., Mitchell, B.G., Siegel, D.A., Carder, K.D., Garver, S.A., Kahru, M., McClain, C., 1998. Ocean color algorithms for SeaWiFS. J. Geophys. Res 103, 24937–24953

Pope, R.M., 1993. Optical absorption of pure water and sea water using the integrating cavity absorption meter, PhD thesis, Texas A&M University College park, Texas, p. 243

Schiller, H., Doerffer, R., 1998 Neural network for emulation of an inverse model-operational derivation of Case

II properties from MERIS data. Int. J. Rem. Sen. 20 (9), 1735–1746.

Thiria, S., Mejia, C., Badran, F., Crepon, M., 1993. A neural network approach for modeling nonlinear transfer functions: application for wind retrieval from spaceborne scatterometer data. J. Geophys. Res. 98 (C12), 22827–22841.

Viollier, M., Tanre, D., Deschamps, P.Y., 1980. An algorithm for remote sensing of water color from space. Boundary Layer Meteorology 18, 247–267.

ELSEVIER

Ecological Modelling 120 (1999) 247–260

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Predictive models of collembolan diversity and abundance in a riparian habitat

Sithan Lek-Ang [a,*], Louis Deharveng [a], Sovan Lek [b]

[a] *LET-UMR 5552, CNRS-University Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France*
[b] *CESAC-UMR 5576, CNRS-University Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France*

Abstract

The artificial neural network (ANN) was used in this work for modelling the abundance and diversity of hydrophilous *Collembola* on the microhabitat scale. The procedure was applied to a Collembolan assemblage of the northern Pyrenees. Six variables were retained to describe its structure: abundance of the three dominant species, species richness, overall abundance of Collembola, and Shannon index. Seven environmental variables were selected as explanatory variables: distance to water, soil temperature, water content, and proportion of mineral soil, moss, litter and rotten wood in the substrate. Correlations between observed values and values estimated by ANN models of the six dependent variables were all highly significant. The ANN models were developed from 83 samples chosen at random and were validated on the 21 remaining samples. The role of each variable was evaluated by inputting fictitious configurations of independent variables and by checking the response of the model. The resulting habitat profiles depict the complex influence of each environmental variable on the biological parameters of the assemblage, and the non-linear relationships between dependent and independent variables. The main results and the ANN potential to predict biodiversity and structural characteristics of species assemblages are discussed.

Keywords: Biodiversity, Species richness; Community structure; Artificial neural network models; Multiple linear regression; Wet habitats

1. Introduction

Biodiversity conservation is a growing concern in western environmental policies. While species and habitats are disappearing at an alarming rate, we are however unable to evaluate, even roughly, the extent of this biodiversity loss, not to mention predicting it. In fact, estimating biodiversity is a tedious task when thousands of species may inhabit the same patch of forest, so taxonomist training would be advantageously coupled here with the development of forecasting techniques based on habitat characteristics, or on a subset of the overall biodiversity. Surprisingly, attempts to predict biodiversity on such grounds are scarce in the literature, except with a few animal groups such as birds (McArthur et al., 1966). Conversely, a wealth of works deal with abundance and biomass prediction (Verner et al., 1986), obviously

* Corresponding author. Fax: +33-5-61556196
E-mail address: ang@cict.fr (S. Lek-Ang)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00106-4

in relation to their more direct socio-economic importance. There are a-priori no specific mathematical tools for predicting biodiversity, so the techniques used for predicting abundance also should work for biodiversity or any other measurable biological variable.

A lot of theoretical models have been proposed in this respect (McArthur et al., 1966; Fretwell, 1972; Tilman, 1982; Schoener, 1983) using a wide range of multivariate techniques, including several methods of ordination, canonical analysis, univariate and multivariate linear, curvilinear, and logistic regressions. A thorough and critical review by James and McCulloch (1990) shows that these conventional models, usually based on multiple regression, assume smooth, continuous, and either linear or simple polynomial relationships between variables. They are capable of solving many problems, but also have serious shortcomings since the main processes that determine the level of biodiversity or species abundance are often non-linear, whereas the methods are based on linear principles. Such models are for example not able to adequately reproduce the behaviour of real systems when very low or high values of the variables are considered (Lek et al., 1996b). Non-linear transformation of variables (logarithmic, power or exponential functions) may improve the results only to a limited extent. The artificial neural network (ANN) approach as proposed here emerges as a different and original methodology which is not constrained by assumptions about the type of relation between the studied variables (Rumelhart et al., 1986). The number of papers using ANN methodology published in ecological sciences has grown rapidly in recent years, e.g. modelling of greenhouse climate (Seginer et al., 1994), identification of the major goals of underwater acoustics (Casselman et al., 1994), prediction of density and biomass of brown trout redds (Lek et al., 1996a), prediction of density and biomass of trout (Baran et al., 1996; Lek et al., 1996b), prediction of the penetration of wild boar into cultivated fields (Spitz et al., 1996), prediction of phytoplankton production (Scardi, 1996), prediction of production/biomass (P/B) ratio of animal populations (Brey et al., 1996), and prediction of fish species richness on a global scale (Guégan et al., 1998), etc.

In the field of soil ecology, multiple linear regression (MLR)-based models relating environmental variables to community structure have been proposed by some authors (Boudjema et al., 1991) sometimes using non-linear transformations of independent or/and dependent variables to improve results (Vegter et al., 1988; Cancela Da Fonseca, 1991). Even so, the results have often remained insufficient, with a low percentage of variance explained. On the other hand, it has been shown that ANN can efficiently model non-linear systems in ecology (Lek et al., 1996b; Scardi, 1996). In the present study, we apply this method to relate the structure and diversity of an assemblage of hydrophilous *Collembola* to microhabitat characteristics. Hydrophilous *Collembola* often constitute the most abundant and diversified arthropods in a large range of wet habitats (Deharveng and Lek, 1995). As such, and because their specific richness is relatively constant along the year as long as water is present, they may provide an interesting raw material to evaluate predictive methods in population ecology. Though it does present sound data on the structure of hydrophilous assemblages of *Collembola*, this case study should be seen first as an attempt to develop predictive tools that are urgently needed for the study and the monitoring of biodiversity.

2. Material and sampling methods

2.1. Study sites and sampling

The studies were undertaken at the site of Ruau, located in the Northern Pyrenees (Arbon, Haute Garonne, France) at an altitude of 784 m. A small permanent spring used for watering livestock flows at the foot of a steep 3-m high slope covered with small trees. Above are large meadows on deep soils. We selected four transects perpendicular to the streamlet, each with four sample points at increasing distance from the water. The distance was 1.50 m between transects and 0.20–0.40 m between sample points on a transect, with the starting points 0–5 cm from the streamlet. Sampling was carried out every 2

months at 12–16 points from December 1993 to December 1994, for a total of 104 samples. The lowest points were sampled at all sampling periods, but the distal row was only occasionally sampled. Each sample was a substrate core of 125 cm^3. Extraction by Berlese technique lasted 2 weeks, until complete drying of the substrate. The animals were preserved in alcohol and sorted under the stereo-microscope. The Collembolan specimens not directly identifiable were mounted in Marc-André II after clearing in lactic acid, and examined with a Nachet 300 microscope under interferential contrast. After identification, the adult and juvenile individuals of each species were counted. After completion of the faunistic analysis, six variables describing the community structure were retained for each sample: abundance of *Collembola* (total number of specimens), species richness (total number of species), relative abundance of the three dominant species (*Isotomurus cassagnaui* [Icas], *I. prasinus* [Ipra] and *Brachystomella parvula* [Brp]), and Shannon diversity index (Table 1). Two of these species are strictly hydrophilous, while the third one, Brp, has both hydrophilous populations (Deharveng and Lek, 1995) and merely open habitat populations (Ponge, 1993).

Seven environmental variables were selected to describe the studied habitats (Table 1), on the basis of their known or supposed biological importance. Temperature and water content, which have a strong impact on most insects, including soil species (Boudjema et al., 1991; Argyropoulou et al., 1993; Deharveng and Bedos, 1993), both showed large fluctuations during the year at Ruau, with patterns varying with the spatial location of the sample points. The relative importance of mineral soil, litter, moss and rotten wood in the substrate has rarely been investigated so far (Deharveng and Lek, 1995), although it is a long established fact that specialized assemblages occupy each of these four substrates (Linnaniemi, 1907; Ponge, 1980; Weiner, 1981).

Distance to water and soil temperature were recorded in situ at the sampling points. Water content (= fresh weight − dry weight of the sample) was measured in the laboratory. The proportion of the different elements of the substratum (mineral soil, moss, litter and rotten wood) was visually estimated and assigned to five ordinal classes defined by their upper limits: absent (0), present up to 25% in volume (1), from 25 to 50% in volume (2), from 50 to 75% in volume (3), more than 75% in volume (4). Volumes were preferred to weights in this estimation because of

Table 1
Independent (i) and dependent (d) studied variables with methods of measurement

Variable	Type	Abbreviated	Methods of measurement
Distance to water	i	WAT	In situ, with ribbon centimetre
Temperature	i	TEM	In situ, with digital thermometer
Water content	i	HUM	Fresh weight-dry weight of substratum
Miner	i	MIN	Proportion of mineral soil in the substratum (visual estimated)
Moss	i	MOS	Proportion of moss in the substratum (visual estimated)
Decaying leaves	i	LIT	Proportion of dead leaves in the substratum (visual estimated)
Decaying wood	i	WOO	Proportion of rotten wood in the substratum (visual estimated)
Isotomurus cassagnaui	d	Icas	Number of *Isotomurus cassagnaui* in the sample (counted under binocular loupe)
Isotomurus prasinus	d	Ipra	Number of *Isotomurus prasinus* in the sample (counted under stereomicroscope)
Brachystomella parvula	d	Brp	Number of *Brachystomella parvula* in the sample (by counting in stereomicroscope)
Total abundance of *Collembola*	d	Nind	Count by stereomicroscope
Species richness	d	SR	Identification by stereomicroscope
Shannon index	d	SI	$SI = -\pi^*\log(\pi)$

the very large differences in density and spatial structure (i.e. spaces available for animals) of the substrates.

2.2. Technique of modelling

We analyzed our data set with: (i) the traditional method of multiple linear regression (MLR), to obtain a predictive model of reference; (ii) optimal non-linear transformation using the SAS Transreg procedure (SAS Institute, 1988; this procedure seeks an optimal transformation of variables, using a method of alternating last squares, a B-spline transformation); (iii) an artificial neural network (ANN) method, to evaluate the performance of this recent method in non-linear modelling. To compare these three methods the whole set of available data was used. To justify the predictive capacity of ANN and MLR methods, modelling was carried out in two steps. First, to fit the models, the matrix (104 records × 7 environmental variables) was used to perform the MLR, the alternating last squares and the ANN methods. The correlation coefficient between observed and predicted values was used to quantify the capability of models to produce the right answer through the training procedure. Second, to test the ANN models, we selected at random a training set (80% of the records, i.e. 83) and a validation set (20% of the records, i.e. 21). This operation was repeated three times giving rise to test 1, test 2 and test 3 which we studied by ANN and MLR. For each of the three sets, the model was determined with the training set and then validated with the test set. The quality of the model was judged through the correlation between observed and predicted values in the validation set.

For classical statistical analysis, univariate, bivariate and multivariate analyses were performed by the SPSS Software release 6.0 (Norusis, 1993). The univariate analyses estimated the mean, standard deviation, coefficient of variation, minimum, maximum, median and quartiles. In bivariate analyses we studied the correlation between variables using Pearson correlation coefficients (values and probabilities of significance at 5 and 1% of confidence intervals). In multivariate analyses, MLR procedures were applied. Examination of studentized residuals for normality, independence and homogeneity was used to test the validity of the models.

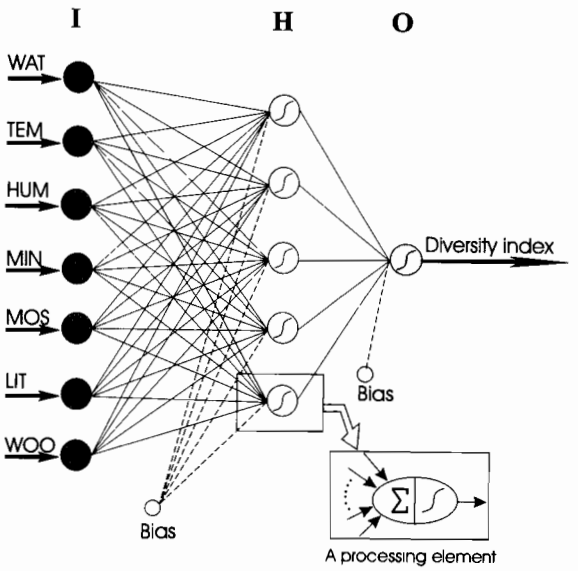

Fig. 1. Representation of the structure of the neural network used. Seven input nodes (I), five hidden layer nodes (H) and one output node (O) are shown. WAT, distance to water; TEM, soil temperature; HUM, water content in the substratum; MIN, proportion of mineral soil in the substratum; MOS, proportion of moss in the substratum; LIT, proportion of litter in the substratum; WOO, proportion of wood in the substratum.

For ANN modelling, the classic multilayer feed-forward neural network was used throughout the analyses. The processing elements in the network, called neurons are arranged in a layered structure (a typical three-layer network is shown in Fig. 1). The first layer, called the input layer, connects with the input variables. In our case, it comprises seven neurons corresponding to the seven habitat variables. The last layer, called the output layer, connects to the output variables. It comprises a single neuron which gives the value of the dependent variable to be predicted. The layers between the input and output layers are called the hidden layers. There can be one or more hidden layers and the number of neurons in each layer is an important parameter of the network. The network configuration is determined empirically by

testing various possibilities and selecting the one that provides the best compromise between bias and variance (Geman et al., 1992; Kohavi, 1995). In our study, a network with one hidden layer of five neurons was selected for each of the six dependent variables studied.

Each neuron is connected to all neurons of adjacent layers (neurons within a layer and in non-adjacent layers are not connected). Neurons receive and send signals through these connections. In feed-forward networks, signals are transmitted only in one direction: from input layer to output layer through hidden layers (no feed-back connections are permitted). Connections are given a weight which modulates the intensity of the signal they transmit.

Training the network consists in using a training data set to adjust the connection weights in order to obtain the best fit between expected and observed values. This training was performed according to the back-propagation algorithm (Rumelhart et al., 1986). The connection weights, initially taken at random in the range $[-0.3, 0.3]$, are iteratively adjusted by a method of gradient descent based on the difference between the observed and expected outgoing signals. Many iterations are necessary to guarantee the convergence of estimated values toward their expectations, without obtaining an overfit, i.e. incapability of the model to generalize (Smith, 1994). The computational program was realized in Matlab environment and computed with an Intel Pentium processor.

Input data have orders of magnitude that differ greatly according to the variables. So as to standardize the measurement scales, inputs were converted into standardized variables. The dependent variable was also scaled in the range [0...1] to adapt it to the demands of the transfer function used (sigmoid function).

2.3. Sensitivity of independent variables

A disadvantage of ANN in comparison with MLR models is their lack of explanatory power. MLR analysis can identify the contribution of each individual input in determining the output and can also give some measures of confidence for the estimated coefficients. On the other hand, there is currently no theoretical or practical way of accurately interpreting the weights in ANN. For example, weights cannot be interpreted as regression coefficients nor easily used to compute causal impacts or elasticities. Therefore, ANN are generally suited for forecasting or prediction rather than for explanatory analysis. But in ecology it is necessary to be able to explain the impact of the variables. To illustrate the importance of explanatory variables inside the ANN, Garson (1991) and Goh (1995) proposed a procedure for the partitioning of the neural network connection weights in order to determine the relative importance of the various input variables. Lek et al. (1995, 1996a,b) have built an algorithm allowing the visualization of the profiles of explanatory variables. In this work, an experimental approach has been used to determine the response of the model to each input variable separately by applying the technique described by Lek et al. (1996a,b).

3. Results

The 104 samples contained a total of 11 637 specimens of *Collembola* of which 11 312 were identified at species level, representing 55 species. Hydrophilous species were dominant in number with Icas, 2658 specimens i.e. 22.8% of the total, Ipra, 1272 specimens i.e. 10.9% and Brp, 1170 specimens i.e. 10%. However, Brp was present in a higher proportion of samples (66.35% occurrence) than the two other species (about 30% occurrence). Large variations in abundance of these three species were observed between samples (Table 2), with a high coefficient of variation (188, 237 and 309% for Brp, Icas and Ipra, respectively).

All samples contained *Collembola*. Mean species richness was 10.64 ($SD = 3.75$, $N = 104$). This is a low diversity compared to forest litter habitats in the same area where over 15 species are recorded on average for samples of the same size, but similar to values obtained in wet habitats at another Pyrenean site, the Arize mountain (9.2 with $SD = 4.50$ and $N = 60$, Deharveng and Lek,

1995). However, as the sample volume was only 125 cm^3 at Ruau, but 250 cm^3 in Arize, the former site is significantly richer, probably in relation to its lower elevation. Species richness and Shannon index were relatively stable with coefficients of variation below 37%. The abundance of *Collembola*, with a coefficient of variation of 84%, reflects fairly large seasonal and spatial fluctuations, but remains well under the variation level of the hydrophilous species of the assemblage.

Among environmental variables, much of the variation was due to the seasonal cycle (particularly temperature: 9.48°C (SD = 3.8), with a minimum of 3.9°C in February, and a maximum of 17.8°C in August). The largest variations were observed for litter and rotten wood content of the substrate (CV = 122 and 175%, respectively), independently of the season.

3.1. Correlation between assemblage characteristics and environmental variables

Among the environmental variables (Table 3), correlation coefficients are significant or highly significant in most cases but with relatively low values: only three correlations above $|0.5|$ ($P < 0.001$) were observed, involving MIN, MOS, HUM and WAT. Some correlations between independent and dependent variables are highly significant: Icas with HUM, WAT and MOS; Ipra with HUM and WAT; Nind with HUM and WAT; SI with WAT. Water content and distance to water therefore appear as major determinants of assemblage characteristics. Other correlations were significant at a lower level (Brp and TEM, Nind and MIN, SR and TEM, SI and TEM) and most (30) were not significant. In particular, species richness was very poorly related to environmental variables.

Among the dependent variables, a high correlation was found between Nind and the abundance of each of the two most abundant species ($r = 0.71$ for Icas, $r = 0.70$ for Ipra) indicating the numerical importance of these species in the community. The correlation was even higher between SI, the Shannon index, and SR, one of the measures on which it is built ($r = 0.76$, $P < 0.001$). Correlation between Icas and Ipra was relatively high ($r = 0.53$, $P < 0.001$) reflecting their strong dependence on water. Brp was conversely poorly related to Icas ($r = -0.10$, $P = 0.293$) or Ipra ($r = -0.10$, $P = 0.327$). Other highly significant correlations were between SR and Brp, between SI and Icas and between SR and Icas. Correlations were weaker with the other variables; the low value of correlation between species richness and the total number of Collembolan specimens is particularly noticeable.

Table 2
Summary statistics[a]

	Minimum	Q1	Median	Q3	Maximum	Mean	SD	CV%
TEM	3.9	6.5	8.45	11.95	17.8	9.48	3.8	40.08
HUM	3.6	25.4	34.5	46.55	89.4	36	18.86	52.39
WAT	0	2.5	20	50	100	35.77	31.89	89.15
MIN	0	2	2	2	4	1.9	0.78	41.05
MOS	0	1	1	2	3	1.21	0 89	73.55
LIT	0	0	0	1	3	0.6	0.73	121.67
WOO	0	0	0	1	2	0 28	0.49	175.00
Brp	0	0	2	12	113	11.25	21.18	188.27
Icas	0	0	0	4	296	25.56	60 69	237.44
Ipra	0	0	0	1	275	12.23	37.79	308.99
SR	1	8	11	13	21	10.64	3.75	35.24
Nind	1	53	74	149.5	561	111 89	94.23	84.22
SI	0	1.94	2.45	2.92	3.51	2.26	0.83	36.73

[a] SD, standard deviation; CV%, coefficient of variation in percentage; Q1, Q3, first and third quartile.

Table 3
Pearson correlation coefficient matrix between studied variables

	TEM	HUM	WAT	MIN	MOS	LIT	WOO	Brp	Icas	Ipra	SR	Nind
HUM	−0.41**											
WAT	0.01	−0.55**										
MIN	0.23*	−0.25*	0.36**									
MOS	−0.21*	0.47**	−0.50**	−0.63**								
LIT	−0.16	−0.18	0.16	−0.29**	−0.38**							
WOO	0.25*	−0.21*	0.11	0.02	−0.27**	−0.36**						
Brp	−0.25*	−0.10	0.08	−0.14	0.12	0.03	−0.03					
Icas	−0.06	0.46**	−0.45**	−0.15	0.26**	−0.09	−0.11	−0.10				
Ipra	−0.11	0.30**	−0.30**	−0.06	0.02	0.02	−0.11	−0.10	0.53**			
SR	−0.24*	0.00	0.14	−0.17	0.02	0.18	−0.05	0.35**	−0.28**	−0.04		
Nind	−0.15	0.36**	−0.32**	−0.24*	0.11	0.18	−0.12	0.22*	0.71**	0.70**	0.15	
SI	−0.22**	−0.14	0.30**	0.08	−0.17	0.15	−0.05	0.11	−0.52**	−0.14	0.76**	−0.16

* Significant ($P<0.05$).
** Highly significant ($P<0.01$).

Table 4
Multiple linear regression between parameters of Collembolan assemblages and environmental variables[a]x

	Brp	Icas	Ipra	SR	Nind	SI
TEM	−0.279*	0.277**	−0.083	−0.212	−0.001	−0.276*
HUM	−0.157	0.470**	0.187	−0.027	0.307*	−0.123
WAT	0.080	−0.196	−0.296*	0.186	−0.212	0.213
MIN	0.078	0.305	−2.016**	−0.917	−0.805	−0.410
MOS	0.300	0 375	−2.547**	−0.866	−0.942	−0.544
LIT	0.122	0.365	−1.903**	−0.652	−0.477	−0.379
WOO	0.120	0.174	−1.251**	−0.475	−0.437	−0.306

[a] The models shown the standard coefficients of seven independent variables with their significant level
* Significant at 0.05.
** Significant at 0 01

3.2. Multiple linear regression (MLR) analysis

For the 104 samples, the MLR procedure using the 7 independent variables gives the following coefficients of multiple correlation: with Brp, $R^2 = 0.10$ ($F_{7,96} = 1.55$, $P = 0.17$); with Icas, $R^2 = 0.32$ ($F_{7,96} = 6.53$, $P < 0.001$); with Ipra, $R^2 = 0.22$ ($F_{7,96} = 3.91$, $P < 0.001$); with SR, $R^2 = 0.13$ ($F_{7,96} = 2$, $P = 0.07$); with Nind, $R^2 = 0.24$ ($F_{7,96} = 4.22$, $P < 0.001$); with SI, $R^2 = 0.16$ ($F_{7,96} = 2.65$, $P = 0.02$). Low correlation coefficients reflect the low percentages of explained variance (less than 33% for all studied variables). With $\log(x+1)$ transformation of variables, we obtained R^2 equal to, respectively 0.29, 0.64, 0.28, 0.31, 0.40 and 0.32 for Brp, Icas, Ipra, SR, Nind and SI. All models were highly significant ($P < 0.001$). Values of determination coefficients indicate a clear improvement of MLR models after non-linear transformation of variables. As this operation improves their linearity, we can conclude that non-linear relationships exist between the dependent and independent variables. Thus, a method based on alternating last squares was used to try to linearise the variables. With the Transreg procedure in SAS Software after maximum transformation of variables using the B-spline function, we obtained a squared multiple correlation equal to 0.41, 0.67, 0.47, 0.55, 0.49 and 0.48 for Brp, Icas, Ipra, SR, Nind and SI, respectively, i.e. a significant improvement of model quality.

Returning to the results of the MLR analysis, we give in Table 4 the standard coefficients of seven independent variables for the six dependent variables characterizing the Collembolan assemblage. Except in the SR model where none of the variables were significant, other models had at least one significant variable. The maximum was recorded for Ipra with five significant variables (Table 4).

3.3. Artificial neural network (ANN)

We used an ANN of one hidden layer of five neurons with seven independent variables, i.e. a 7-5-1 neural network (46 parameters in total: $7 \times 5 + 5 + 6$). Results after 500 iterations of the training procedure are presented in Fig. 2. The correlation coefficient (r) between observed and estimated values was close to 1 for Icas, Ipra, Brp and Nind ($r = 0.996$, $r = 0.965$, $r = 0.944$ and $r = 0.914$, respectively, $P < 0.001$). The lowest correlation coefficients were observed for SR and SI ($r = 0.847$ and $r = 0.872$, respectively, $P < 0.001$). The ANN therefore gave satisfactory results practically over the whole range of values of the dependent variables (Fig. 2). For the variables which represent species abundances (Icas, Ipra, Brp and Nind) most points were well aligned on the perfect fit diagonal (coordinates 1:1). Although poorly represented, the strong values of the output variable are clustered around this same perfect line. Only a few points lie far off, with some weak values slightly underestimated (Ipra

and Brp). For the remaining dependent variables SR and SI, which measure assemblage diversity, fitting is acceptable in spite of poorer results.

The sensitivity of the seven independent habitat variables on the six dependent variables obtained from ANN modelling is illustrated in Fig. 3. The 12 points cover the range of variation of each of the variables tested, with a class interval which was modified according to the variables. As illustrated in Fig. 3, we can distinguished seven sensitivity types:

- Exponential contribution: the independent variables contribute only at their low values. This is the case of WAT and MIN for Icas, and MOS, WAT and TEM for Ipra.

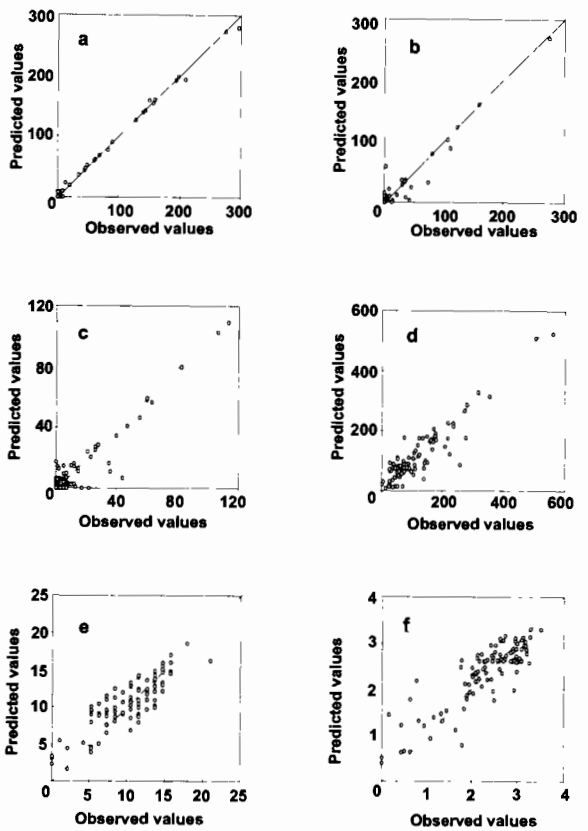

Fig. 2. Correlation graph between observed values and values estimated by the model. The solid line indicates the perfect fit line (coordinates 1:1). (a) Icas (*Isotomurus cassagnaui*); (b) Ipra (*Isotomurus prasinus*), (c) Brp (*Brachystomella parvula*); (d) Nind (total abundance of *Collembola*); (e) SR (species richness), (f) SI (Shannon index).

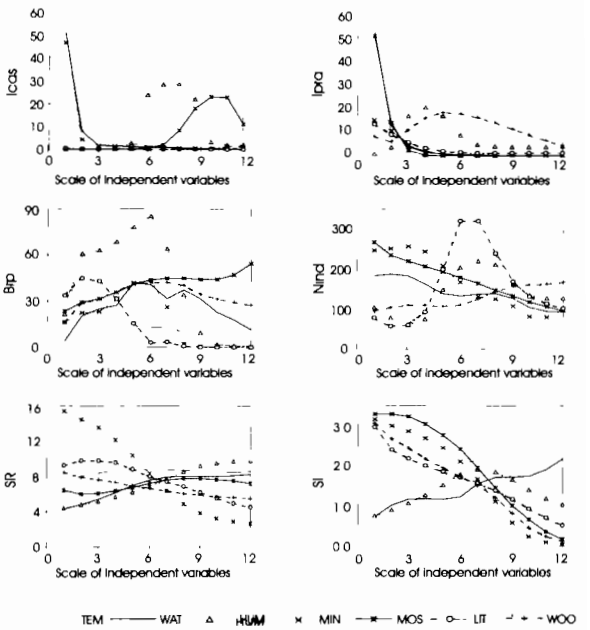

Fig. 3. Contribution profile for each independent variable to the determination of assemblage characteristics of *Collembola* fauna by ANN. Icas (*Isotomurus cassagnaui*), Ipra (*Isotomurus prasinus*), Brp (*Brachystomella parvula*), Nind (Total abundance of *Collembola*), SR (species richness), SI (Shannon index). The abscissa represents the 12 variation intervals of the independent variables between their minimum and their maximum.

- Gaussian contribution: the independent variable affects the dependent variable mostly around its average value, and has little influence at extreme values. This is the case of HUM for Icas, HUM and WOO for Ipra, WAT and WOO for Brp, LIT, TEM and HUM for Nind, and TEM and HUM for SI.
- Increasing contribution: dependent variable is low for low values of the independent variable and increases to a maximum at high values. This was observed for MOS for Brp abundance, HUM and WAT for SR, WOO for Nind, and WAT for SI.
- Decreasing contribution: dependent variable is relatively high at low values of the independent variable and decreases gradually afterwards. This was the case of TEM for Icas, MIN and LIT for SR, and MOS, MIN, WOO and LIT for SI.

- Skewed-to-the-left curve: the dependent variable is high only for high values of the independent variable. This sensitivity type was observed only for MOS to explain the abundance of Icas.
- Skewed-to-the-right curve: the dependent variable is high for low values of these independent variables; it decreases more or less rapidly afterwards to become virtually null thereafter. This contribution is present only for Brp abundance for four environmental parameters (TEM, HUM, LIT and MIN).
- Weak contribution: the contribution of the independent variable is very low, and not altered over its range, with a profile represented by a quasi-horizontal line. This is the case of WOO and LIT for Icas, and TEM, MOS and WOO for SR.

3.4. Test of the models

To test the variability, the prediction power of the different models determined from three training fractions was tested on three independent test fractions (Table 5). The lowest correlation between observed and predicted values was obtained for SR ($r = 0.66$–0.82, $P < 0.001$) and SI ($r = 0.79$, $P < 0.001$). Correlations for Nind and Brp ($r = 0.80$–0.87, $P < 0.001$ and $r = 0.84$–0.90, $P < 0.001$) were higher. The best results were obtained with Icas ($r = 0.95$–0.99, $P < 0.001$) and Ipra ($r = 0.88$–0.98, $P < 0.001$), like in the models based on the complete set of 104 samples. The same tests realized with MLR-models (Table 6) give clearly inferior results (maximum correlation coefficient equal to 0.57 for Icas in the second test set).

Table 5
Correlation coefficient between predicted and observed values by ANN models for three independent testing sets for the six studied parameters of Collembolan assemblages

	Training set			Testing set		
Set no	1	2	3	1	2	3
Icas	0 986	0.990	0 990	0.990	0.950	0 956
Ipra	0.985	0.990	0.990	0.979	0.877	0.889
Brp	0.883	0.938	0.935	0 904	0.843	0.854
SR	0.864	0.851	0.834	0.700	0.656	0.823
Nind	0 940	0.946	0.906	0.865	0.820	0 797
SI	0.865	0.901	0.879	0.796	0 798	0.789

Table 6
Correlation coefficient between observed and predicted values by MLR-models for three independent testing sets for the six studied parameters of Collembolan assemblages

	Training set			Testing set		
Set no.	1	2	3	1	2	3
Brp	0.461	0.413	0 464	0.347	0 551	0.427
Icas	0.559	0 531	0.574	0 275	0.566	0.242
Ipra	0.630	0.582	0.610	0.161	0.515	0.287
SR	0.402	0.342	0.331	0.081	0.349	0.473
Nind	0 574	0.527	0 565	0.301	0.556	0 194
SI	0.494	0 460	0.426	0.053	0.266	0.511

On the whole, the coefficients in the training set were nearly identical to those of the models based on 104 samples. These results indicate a great stability (small standard deviations) of the prediction performance of the ANN models for different testing sets. The small decrease in performance in the test set compared to the training set can be related to the small size of the data set combined with the fact that each sample is likely to have some kind of unique information that is relevant to the model. The correlation coefficients were clearly not as low when the data were analyzed by MLR, in particular for Shannon index and specific richness.

4. Discussion

Two kinds of results emerge from this study: those related to the artificial neural network methodology and its ability to predict the characteristics of a species assemblage; and those related to the ecology of hydrophilous *Collembola*, which are of interest for Collembologists and wet habitat ecologists. MLR, spline regression and backpropagation of the ANN were applied on the same dataset with the aim to develop stochastic models of biodiversity prediction, using Collembolan assemblages and habitat features on a microhabitat scale. The backpropagation procedure of the

ANN gave much higher correlation coefficients than other methods. This may point to the predominantly non-linear relationships between the studied variables on the one hand, and on the other hand the ability of ANN to directly take into account any non-linear relationships between the dependent variables and each independent variable (Lek et al., 1996b). These results are in agreement with literature data, where performances of ANN have been repeatedly reported to overpass those of more traditional method such as MLR (Ehrman et al., 1996; Lek et al., 1996b; Scardi, 1996). However, the comparison between the predictive power of MLR and that of ANN is not quite fair, in particular as the number of parameters is different. In any case, ANN constitutes a new and powerful alternative in predictive ecological modelling, where poor fitting of biological characteristics to conventional models (mostly MLR) is often the rule.

Collembola are often the dominant group of Arthropods in wet habitats, yet literature references related to the ecology of hydrophilous species are scarce. On these grounds, it is hardly surprising that a large amount of novel information has been generated by the present study. Most characteristics of the Collembolan assemblages studied have been satisfyingly fitted to measured environmental parameters through ANN analysis. Variations in abundance among dominant species (Icas, Ipra, Brp) are in particular strongly connected to a set of environmental variables: temperature, distance to water, structure of the substratum and type of organic matter. A second important finding is the complexity of the response of Collembolan assemblages to changes in environmental parameters, so far largely overlooked in the relevant literature (e.g. Van Straalen, 1994). On the whole, emerging patterns of species abundance response to environmental parameter fluctuations appear both mostly non-linear and very heterogeneous, in spite of the high ecological similarity of the studied species. Some factors are clearly predominant, but they are not the same for the different measured biological variables. Conversely, sensitivity of species richness and Shannon index often follow similar patterns for different environmental variables, making it difficult to detect which one(s) is (are) the driving factors (Fig. 3). Nevertheless, there are some positive outcomes of this study, that are summarized below.

1. An unexpected result is that distance to free water has more impact than water content of the substrate for hydrophilous species abundance. Distance to water is weakly correlated to water content on the scale of our study because of its independence from season, local microtopography and superficial water circulation. *Isotomurus* species in particular experience an abrupt numerical decrease as soon as their distance to water increases. Their abundance peaks for medium-range water content. In contrast, *B. parvula* abundance reaches its maximum value at medium distance to water and medium water content, in agreement with empirical observations suggesting that its stenohygry is lower than that of *Isotomurus* (Deharveng and Lek, 1995). The overall abundance of Collembola follows yet another pattern which is likely to be explained by the strong impact of non-hydrophilous species, not documented in detail in this paper.
2. Distance to water has a slight but positive impact on biodiversity indices on our study scale, reflecting a more general trend of increasing species richness from water edge to mesophilous litter (Deharveng and Lek, 1995). The decreasing saturation of the mineral part of the substrate on this gradient gradually gives more micro-voids and new microhabitats for colonization by terrestrial mesofauna, and may contribute to the observed patterns.
3. Water content of the substrate is known to have an overwhelming importance for Collembolan populations (Vannier and Verhoef, 1978; Verhoef and Witteveen, 1980) but studies are lacking at the community level. According to Vegter et al. (1988), moisture heterogeneity has a strong influence on the abundance of epigeomorphic *Collembola*, but not on the assemblage structure. In our study, both abundance and assemblage structure were clearly affected by variations in water content of the substrate.

4. Surprisingly large differences were observed among species in response to variations of the studied variables (Fig. 3). The impact of temperature, the most documented environmental variable (Hopkin, 1997) is different on different ecological categories of *Collembola* as already stated in the literature (Van Straalen and Joosse, 1985; Van Straalen, 1994). In the present study, it further appears that, even among the same ecological category, species response may strongly vary between the most strictly hydrophilous species (Icas and Ipra) and those less so (Brp). The relationship of temperature to overall abundance of the hydrophilous *Collembola* assemblages still follows another pattern which is not that of these dominant species. The same comments also apply to other variables, particularly water content and mineral soil content of the substrate for which even the profiles of the two most hydrophilous species strongly diverge. A direct implication of these results is that extrapolation of ecological information from single, even dominant, species to communities may be strongly misleading: communities may be highly heterogeneous assemblages even at a relatively narrow functional level.
5. The relationships between environmental variables and biological parameters characterizing living communities have rarely been evaluated in the literature related to soil science. Boudjema et al. (1991) expresses these relations as a polynomial function, with pH and temperature as driving variables. Van Straalen (1994) reported a correlation between egg development and a measure of enzyme activity linked to temperature. But correlations, when measured, remained fairly low in all documented cases. The ecological profiles obtained from ANN models (Fig. 3) clearly exhibit the complexity and non-linearity of interacting processes, which may account for the difficulty in predicting species and community responses using traditional methods.
6. Intuitively, soil ecologists are aware of the prime importance of ligneous material for soil living assemblages, but this variable is rarely if ever taken into account in the literature. The same could be said for decaying leaves or moss content of the substrate. The classical measures of organic matter do not give any information on the relative proportion of these three elements, though it is likely to be of higher biological significance than overall amount of organic matter itself. By introducing these variables in our analysis, we expected to obtain some sound information about their influence on species abundance and assemblage structure. The results were, in fact, difficult to interpret. No influence was detected on the profiles of Icas, the most water-dependent species of the assemblage, and only a limited one on Ipra. The less strictly hydrophilous species Brp appeared more sensitive to these variables. Unexpectedly, increase in leaf litter and wood content of the substrate were associated with decreasing biodiversity, in apparent contrast to the usual (but again poorly documented) trend of increasing biodiversity from open to forested habitats. An appealing hypothesis is that leaf and wood litter is less important in wet habitat than in mesophilous habitats, because decomposition processes are less active in water or water-saturated substrate, providing a lower diversity of fungal species on which most *Collembola* feed.

Is it finally possible to predict the level of biodiversity of a living group from environmental variables? Because they largely control the presence and abundance of individual species, environmental variables necessarily contribute to the control of community structure, hence of biodiversity. Hard data, are however, lacking to support this commonplace statement in soil ecosystems and previous attempts to detect simple and linear relationships between edaphic factors and diversity have failed to give clear-cut results (for instance in tropical Collembolan assemblages, Deharveng and Bedos, 1993). Three reasons (or a combination of these) may explain this failure: (i) the pertinent variables have not been identified; (ii) interactions between species play a major role; and (iii) relationships between abiotic factors and biodiversity are non-linear. This last hypothesis was considered here. The results obtained indicate that species interaction, or consideration of addi-

tional variables, is not needed to satisfactorily predict the characteristics of the observed assemblage patterns. Several parameters relevant to biodiversity were efficiently predicted by the ANN-based models in the studied community. Additional data sets, experimental manipulations and repeated mathematical analyses would be necessary to assess this first result more firmly, but the ANN has demonstrated here a promising potential in the field of community ecology, as a tool to evaluate, understand, predict and manage biodiversity.

Acknowledgements

This work was supported by a grant from the European Community (contract DGXII niPL93-1917: High Endemism in Areas, endemic biota and the conservation of biodiversity in Western Europe).

References

Argyropoulou, M.D., Asikidis, M.D., Iatrou, G.D., Stamou, G.P., 1993. Colonization patterns of decomposing litter in a maquis ecosystem. Eur. J. Soil Biol. 29, 183–191.

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996 Stochastic models that predict trout population densities or biomass on macrohabitat scale. Hydrobiologia 337, 1–9

Boudjema, G , Julien, J.M., Sarkar, S., Cancela Da Fonseca, J.P., 1991. Etude par analyse statistique multilinéaire de l'impact des facteurs physico-chimiques sur l'abondance des Microarthropodes édaphiques d'une forêt de mousson en Inde orientale. Rev Ecol. Biol. Sol. 28, 303–322.

Brey, T., Jarre-Teichmann, A., Borlich, O , 1996 Artificial neural network versus multiple linear regression: predicting P/B ratios from empirical data. Mar. Ecol. Progr. Ser. 140, 251–256.

Cancela Da. Fonseca, J.P., 1991. Ecological diversity and ecological systems complexity: local or global approach? Rev. Ecol. Biol. Sol. 28, 51–66.

Casselman, F.L., Freeman, D.F., Kerrigan, D.A., Lane, S.C., Magley, D.M., Millstrom, N.H., Roy, C R., 1994. A neural network-based underwater acoustic application. In: Proceedings of the IEEE International Conference on Neural Networks IEEE, Orlando. pp. 3409–3414.

Deharveng, L., Bedos, A., 1993. Factors influencing diversity of soil *Collembola* in a tropical mountain forest (Doi Inthanon, Northern Thailand). In: Paoletti, M.G., Foissner, W., Coleman, D. (Eds), Soil Biota, Nutrient Cycling and Farming Systems. Lewis, pp. 91–111.

Deharveng, L., Lek, S., 1995. High diversity and community permeability the riparian *Collembola* (Insecta) of Pyrenean massif Hydrobiologia 312, 59–74

Ehrman, J M., Clair, T.A., Bouchard, A., 1996. Using neural networks to predict pH changes in acidified Eastern Canadian lakes. Artif. Intell. Appl. 10, 1–8.

Fretwell, S.D., 1972 Populations in a Seasonal Environment. Monography. Population Biology, vol. 5. Princeton University, Princeton NJ, p. 217.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artif. Intell Expert 6, 47–51

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemna Neural Comput. 4, 1–58.

Goh, A.T.C., 1995. Back-propagation neural networks for modelling complex systems Artif. Intell. Eng. 9, 143–151.

Guégan, J.F , Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384

Hopkin, S.P., 1997. Biology of the Springtails (Insecta: *Collembola*). Oxford University, Oxford, p. 330.

James, F.C , McCulloch, C E., 1990. Multivariate analysis in ecology and systematics panacea or Pandora's box? Ann. Rev. Ecol. Syst. 21, 129–166.

Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection In· Proceeding of the 14th International Joint Conference on Artificial Intelligence Morgan Kaufmann, Montreal. pp 1137–1143.

Lek, S., Belaud, A., Lauga, J., Dimopoulos, I., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar. Fresh Res 46, 1229–1236

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Resour. 9, 23–29.

Lek, S., Delacoste, M , Baran, P., Dimopoulos, I., Lauga, J., Aulanier, S , 1996b. Application of neural networks to modelling non-linear relationships in ecology. Ecol. Mod. 90, 39–52.

Linnaniemi, W.M., 1907. Der Apterygotenfauna Finlands. I. Allgemeiner Teil. Acta Soc. Sci. Fen. 34, 1–134.

McArthur, R H , Reicher, H., Cody, M.L., 1966. On the relation between habitat selection and bird species diversity. Am. Nat. 100, 319–332.

Norusis, M.J., 1993. SPSS for Windows. Base system user's guide release 6.0. SPSS Inc, pp. 828.

Ponge, J.F., 1980. Les biocénoses des Collemboles de la for_t de Sénart. In: Pesson (Ed.), Actualités d'Ecologie Foresti_re P. Gauthier Villard, Paris, pp. 151–176.

Ponge, J.F., 1993. Biocenose of *Collembola* in Atlantic temperate grass-woodland ecosystems. Pedobiologia 37, 223–244.

Rumelhart, D.E., Hinton, G.E , Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

SAS Institute, 1988 SAS Technical report P-179. Additional SAS/STAT procedure, release 6.03. SAS Institute, Cary, NC, pp 255.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Progr. Ser. 139, 289–299.

Schoener, T.W., 1983. Field experiments on interspecific competition. Am. Nat. 122, 240–285.

Seginer, I., Boulard, T., Bailey, B.J., 1994 Neural network models of the greenhouse climate. J Agr. Eng. Res. 59, 203–216.

Smith, M., 1994. Neural networks for statistical modelling. Van Nostrand Reinhold, New York.

Spitz, F., Lek, S., Dimopoulos, I., 1996. Neural network models to predict penetration of wild boar into cultivated fields. J. Biol. Syst. 4, 433–444.

Tilman, D., 1982. Resource competition and community structure. In: Monography Population Biology, vol. 17. Princeton University, Princeton NJ, p. 296.

Van Straalen, N.M., 1994. Adaptive significant of temperature responses in *Collembola*. Acta Zool. Fenn. 195, 135–142.

Van Straalen, N M., Joosse, E N.G , 1985. Temperature responses of egg production and egg development in two species of *Collembola*. Pedobiologia 28, 265–273.

Vannier, G., Verhoef, H.A., 1978. Effect of starvation on transpiration and water content in the populations of two coexisting *Collembola* species Comp Biochem. Physiol.(A) 60, 483–489.

Vegter, J.J., Joosse, E.N.G., Ernsting, G., 1988. Community structure, distribution and population dynamics of *Entomobryidae* (*Collembola*) J. Anim. Ecol. 57, 971–981.

Verhoef, H.A., Witteveen, J., 1980. Water balance in *Collembola* and its relation to habitat selection; cuticular water loss and water uptake. J. Insect Physiol. 26, 201–208.

Verner, J , Morrison, M.L., Ralph, C.J., 1986. Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates. Wisconsin University, Madison WI, p. 478.

Weiner, W.M., 1981. Collembola of the Pieniny national park in Poland. Acta Zool. Cracoviense 25, 417–500

ELSEVIER

Ecological Modelling 120 (1999) 261–270

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Prediction of response of zooplankton biomass to climatic and oceanic changes

Ichiro Aoki [a,*], Teruhisa Komatsu [b], Kangseok Hwang [c]

[a] *Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, University of Tokyo, Yayoi, Bunkyo, Tokyo 113-8657, Japan*

[b] *Ocean Research Institute, University of Tokyo, Minamidai, Nakano, Tokyo 164-8639, Japan*

[c] *National Fisheries Research and Development Agency, Shirang-ri, Kijang, Pusan, South Korea*

Abstract

This paper examines the long-term variation in zooplankton biomass in response to climatic and oceanic changes, using a neural network as a nonlinear multivariate analysis method. Zooplankton data collected from 1951 to 1990 off the shore of northeastern Japan were analyzed. We considered patterns of the Kuroshio and the Oyashio, sea surface temperature, and meteorological parameters as environmental factors that affect zooplankton biomass. Back propagation neural networks were trained to generate mapping functions between environmental variables and zooplankton biomass. The performance of the network models was tested by varying the numbers of input and hidden units. Changes in zooplankton biomass could be predicted from environmental conditions. The neural network yielded predictions with smaller errors than those of predictions determined by linear multiple regression. The sensitivity analysis of networks was used to extract predictive knowledge. The air pressure, sea surface temperature, and some indices of atmospheric circulation were the primary factors for predictions. The patterns of the Kuroshio and the Oyashio demonstrated different effects among sea areas.

Keywords: Zooplankton, Neural networks; Biomass prediction; Kuroshio–Oyashio, climatic change

1. Introduction

The northeastern sea area of Japan is well known as a highly productive fishing ground under the influences of the Kuroshio (western boundary current of the subtropical gyre) and the Oyashio (western boundary current of the subarctic gyre) dynamics in the North Pacific. Pelagic fishes, such as sardine, mackerel, and skipjack tuna, migrate into this sea area during the summer to feed and store nutrition. In addition, larvae and juveniles of small pelagic fish hatched in winter and spring in the sea area near the Kuroshio south of Japan are transported and dispersed widely in the northeastern sea area of Japan in summer. Zooplankton biomass in this area is suggested to be an important factor that affects recruits of the Japanese sardine (Aoki and Komatsu, 1997). One hypothesis is that the abun-

* Corresponding author. Fax: +81-3-5841-8165.
E-mail address. mail@hongo.ecc.u-tokyo.ac.jp (I. Aoki)

0304-3800/99/$ - see front matter
PII. S0304-3800(99)00107-6

dance and distribution of zooplankton are controlled by oceanic conditions including the dynamics of the Kuroshio and Oyashio system. Meteorological conditions also directly and indirectly affect biological processes in the ocean.

Odate (1994) reported on the long-term variations in zooplankton biomass in the Oyashio, the Kuroshio, and their transition regions in the northeastern sea area of Japan during 1951–1990. Based on these data, Tomosada and Odate (1995) analyzed the interrelationship between zooplankton biomass and water temperature and meteorological parameters by the use of the correlation coefficient.

A number of environmental factors are assumed to be associated with the change in zooplankton biomass. Moreover, these associations may involve nonlinear relationships. The neural network has the advantage of being able to be applied to a nonlinear multivariate analysis method without loss of accuracy in these situations. Thus, this paper examined a predictive model relating the long-term variation in zooplankton biomass to climatic and oceanic changes using neural networks.

2. Methods

Zooplankton data used in this study were taken from Odate (1994), which reference provided a time line series of monthly mean zooplankton densities (wet weight/m^2) between 1951 and 1990 in the northeastern sea area of Japan from 33 to 46°N and from the east coast of Japan to 160°E (Fig. 1). Zooplankton samples were collected most intensively in the area 34–41°N and west of 150°E. The samples were taken by vertical hauls from 150 m to the surface using Marutoku net (45 cm diameter, 0.33 mm mesh). Three regions were defined based on the 100m depth temperature: < 5°C, Oyashio region; 5–15°C, transition region; and > 15°, Kuroshio region. The original data of zooplankton were averaged over the 12 months of each year and smoothed further by a 5-year running mean (Fig. 2), because the aim of this study was to examine long-term changes.

We considered the Kuroshio, the Oyashio, sea surface temperature, and meteorological parameters as environmental factors affecting zooplankton biomass. These variables were used in analysis as follows (Table 1).

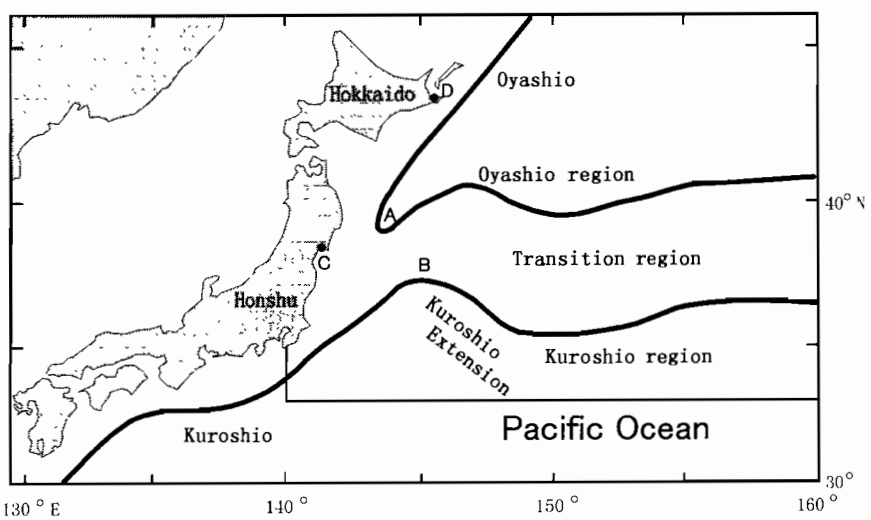

Fig 1. Study area. A, Southern limit of the First Intrusion of the Oyashio current; B, Northern limit of the Kuroshio Extension; C, Ishinomaki; D, Nemuro

1. Southernmost latitude of the First Intrusion of the Oyashio (OY) and northernmost latitude of the Kuroshio Extension axis (KE): The data from 1951 to 1988 given in Kawai (1989) were used. Coordinates from 1989 to 1990 were determined from monthly 100 m depth temperature charts compiled by the Japan Meteorological Agency (JMA), according to the indicative temperature of the Oyashio Front defined by Kawai (1972), and that of the Kuroshio Extension axis defined by Murakami (1993).
2. Sea surface temperature in the sea area of 35–45°N and 140–150°E (SST): we used ten-day mean sea surface temperature anomaly data by JMA.
3. Mean air temperature (AT) and mean sunshine time (SN) in northern Japan: annual anomalies at four locations in Hokkaido and northeastern Honshu districts published by JMA (1994) were used.
4. Air pressures at Nemuro (NM) and Ishinomaki (IS): monthly mean air pressures data by JMA were used.
5. Atmospheric circulation indices: Far East Polar Vortex Index (PV), Far East Zonal Index (ZI), East Sea Index (ES), Subtropical Index (STI), and Southern Oscillation Index (SOI) compiled by JMA were used. These indices were postulated to affect oceanic circulation and sea temperature. PV, ZI, ES, and STI are defined by monthly mean anomaly of 500 hPa height in the Far Eastern area, and represent the pressure field and the amplitude of the westerlies. SOI indexes the amplitude of the trade wind relating to the El Niño in the equatorial Pacific Ocean.

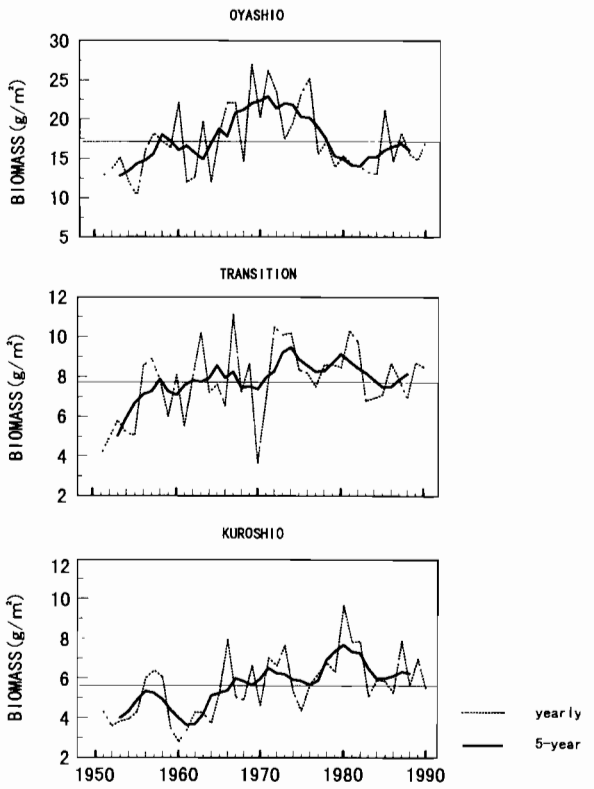

Fig. 2. Changes in zooplankton biomasses during 1951–1990 in the northeastern sea area of Japan. Dotted lines, yearly average; solid lines, 5-year running mean. Thin horizontal lines indicate total averages of each sea region.

Table 1
List of input variables used for prediction of zooplankton biomass [a]

	Variables (Abbreviation)
Hydrographic	Southern limit of the Oyashio current (OY)
	Northern limit of the Kuroshio Extension (KE)
	Sea surface temperature in the northeastern sea area of Japan (SST)
Meteorological elements	Air temperature in the northern Japan (AT)
	Sunshine time in the northern Japan (SN)
	Air pressure at Nemuro (NM)
	Air pressure at Ishinomaki (IS)
Atmospheric circulation	Far East Polar Vortex Index (PV)
	Far East Zonal Index (ZI)
	East Sea Index (ES)
	Subtropic Index (STI)
	Southern Oscillation Index (SOI)

[a] See text for full explanation.

These environmental original data, in the same manner as zooplankton biomass data, were processed to 5-year running means. Back propagation neural networks, which have the

followingthree types of layers, were used: an input layer, one or more hidden layers, and an output layer. Neural networks were constructed for each of the three sea regions. Units in the input layer corresponded to the 12 variables shown in Table 1. One unit was assigned to the output layer to represent the zooplankton biomass in a given sea region. Here, we compared the performance of neural network models by varying the number of input variables or units. The object of neural network learning is to generalize from a training set to the systems as a whole. There is a danger that too many input variables lead to an overfit model and subsequently lower the ability to yield accurate generalization. Therefore, we first trained the network (one hidden layer and four hidden units) using all 12 input variables and data (1953–1988), and determined the mean connection weight of each input variable. The mean connection weight was defined as $\Sigma_{j=1}^{n} w_{ij} w_{jo}/n$, where w_{ij} is the weight between the input unit i and the hidden unit j; w_{jo} the weight between the hidden unit j and the output unit; and n the number of hidden units. Then we considered two additional models with different input variables selected according to the absolute mean connection weight (Table 2). Further, different numbers of layers and units in the hidden layers were also tested.

The effective data set consisted of 36 pairs of input and output vectors for each year ranging from 1953 to 1988. The data set was divided into four subsets of 9 years: (1) 1953–1961; (2) 1962–1970; (3) 1971–1979; and (4) 1980–1988. Then, we trained the network, leaving out one of the subsets, and after training, ran the network using the omitted subset as a test set. We define Case-*i* in which the subset (i) was left out from training and used to generate a test prediction. Therefore, this training and test operation was repeated four times. The resulting estimate of generalization error on test sets was used for evaluating different network models. The generalization error was estimated by using the mean absolute error of predicted values for test sets.

We used a commercial neural network simulator, RHINE (CRC Inc.) with a personal computer. In computations, input values for each variable were linearly normalized with max = 1 and min = 0, and initial values for the weights of the connection were set at random in the range of ± 0.3. The number of learning cycles was 5000. The learning error decreased as the number of learning cycles increased, and showed scarcely any changes after 5000 cycles.

Table 2
Input variables included in each model [c]

Variables	Sea region		
	Oyashio	Transition	Kuroshio
OY	◎[a]	○[b]	○
KE	◎		◎
SST	○	◎	◎
AT	○	◎	
SN	◎	○	
NM		◎	○
IS	◎	◎	◎
PV		◎	◎
ZI	○		○
ES		○	
STI			◎
SOI	◎	◎	

[a] Absolute value of the mean connection weight > 1.0.
[b] Absolute value of the mean connection weight > 0.5.
[c] See Table 1 for abbreviation of variables. Model 1: all 12 variables, Model 2: ◎ + ○, Model 3: ◎.

Table 3
Comparison of errors in prediction test among different sets of input variables and layer structures

Network model	No. of units per layer			Mean error	
	Input	Hidden-1	Hidden-2	Absolute (g/m^2)	Relative (%)
Oyashio region					
Model 1-a	12	4	–	1.72	9.86
Model 1-b	12	8	–	1.80	10 31
Model 1-c	12	4	4	1.75	10.03
Model 2-a	8	3	–	1.34	7.69
Model 2-b	8	6	–	1 72	9.85
Model 2-c	8	3	3	2.05	11.75
Model 3-a	5	3	–	1.44	8.28
Model 3-b	5	5	–	1 41	8.10
Model 3-c	5	3	2	1.65	9.43
Transition region					
Model 1	12	4	–	0.621	7.89
Model 2	9	3	–	0.591	7.51
Model 3	6	3	–	0.546	6.94
Kuroshio region					
Model 1	12	4	–	0.397	6 96
Model 2	8	3	–	0.387	6.80
Model 3	5	3	–	0.589	10.3

3. Results

For the Oyashio region, networks were trained and tested varying the hidden layer structure. As a result, the simple structure in the hidden layers minimized the error for each model (Table 3). Among them, Model 2-a best minimized error. Predicted changes in zooplankton biomass by test sets are illustrated in Fig. 3 for the three models each of which had the simplest configuration in the hidden layer. Model 2-a provided good prediction of zooplankton biomass in the 1970s and 1980s, though it reproduced less successfully the peak and trough in the late 1950s and mid 1960s, respectively.

Because conditions of one hidden layer and few hidden units proved satisfactory, the following networks for the transition and the Kuroshio regions were trained with one hidden layer and 3 or 4 hidden units. For the transition region, Model 3 with 6 input variables resulted in the smallest error during prediction tests (Table 3). Additional input variables to these six variables did not improve the performance of the networks. The neural net model reproduced the fluctuation in zooplankton biomass in the transition region with satisfactory accuracy except in the case of a

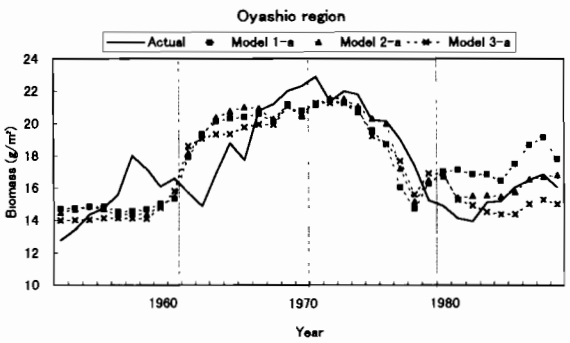

Fig. 3 The actual zooplankton biomass in the Oyashio region and the outputs of test predictions by the neural networks. The test predictions were based on four trained networks by different subsets of data. Vertical lines indicate periods of subsets of data.

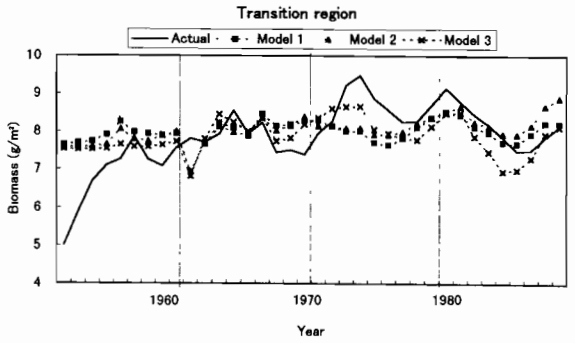

Fig. 4. The actual zooplankton biomass in the transition region and the outputs of test predictions by the neural networks The test predictions were based on four trained networks by different subsets of data. Vertical lines indicate periods of subsets of data.

low level that occurred in the mid-1950s (Fig. 4). For the Kuroshio region, the predictive performance of the network lowered when the number of input variables was reduced to 5 (Table 3). Predictions generated by Models 1 and 2 were very close to the actual biomass (Fig. 5).

A way to study the influence of each input variable on the output is to vary input values and see the result on the output. In this sensitivity analysis, we used four trained networks (Cases 1 to 4), selecting the best model for each sea region. Then we varied the values of a given input variable among nine levels with eight equal intervals over the variable range, fixing the values of all

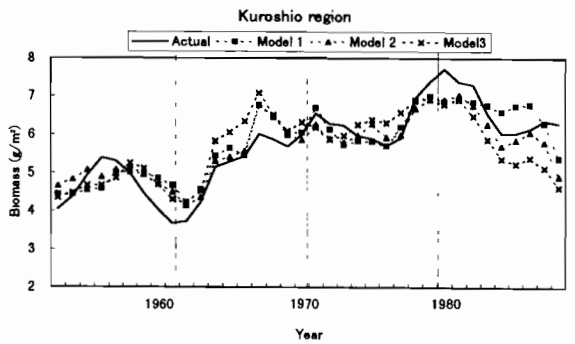

Fig. 5. The actual zooplankton biomass in the Kuroshio region and the outputs of test predictions by the neural networks. The test predictions were based on four trained networks by different subsets of data. Vertical lines indicate periods of subsets of data.

other input variables at mean values. For every input variable, output values increased or decreased monotonously with input values. Fig. 6 shows the ratios of the range of output values to the range of the actual changes in zooplankton biomass.

Four input variables, OY, SOI, KE, and IS, heavily impacted on zooplankton biomass in the Oyashio region (Fig. 6a). The first two variables demonstrated positive relations and the latter two inverse relations to biomass. The other four variables contributed little to the prediction, based on the negligible difference in errors for Model 2a and Models 3a and b (Table 3). In the transition region, SST and AT were most effective, and IS and SOI were also important factors (Fig. 6b), which findings were similar to those in the Oyashio region. In the Kuroshio region, KE, SST, STI, and IS showed greater impact on zooplankton biomass (Fig. 6c). The variable STI which was the most significant constituted a characteristic factor in this region. The effects of SST and IS were common to the other two regions, while KE demonstrated a positive relation to zooplankton biomass only in the Kuroshio region.

The prediction errors were compared with linear multiple regression models using the same input variables as those used in the neural net models (Table 4). In all network models but model 3 for the Kuroshio region, generalization errors were smaller than in corresponding multiple regression models. In linear multiple regression, the error increased when many input variables were used. On the other hand, the neural net maintained stability against the larger number of input variables. When the 12 input variables were used in the network, errors were smaller than those that occurred with linear multiple regression models using fewer variables.

4. Discussion

This study showed that the long-term variation in zooplankton biomass can be predicted by generating mapping functions between environmental variables and zooplankton biomass. In the predic-

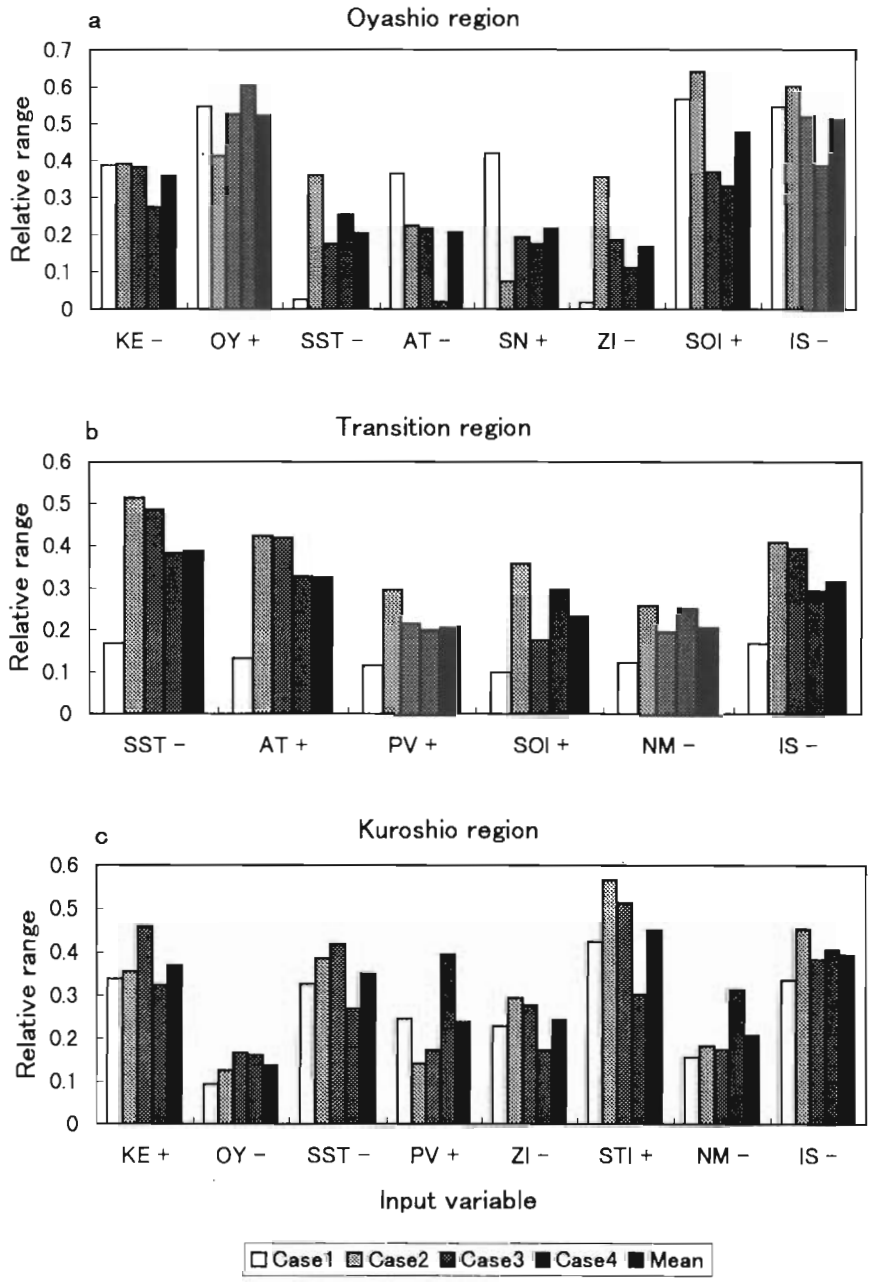

Fig. 6. The influence of each input variable on the output. Relative range is ratio of the range of the output value to the range of the actual change. The results were derived from Model 2a, Model 3, and Model 2 for the Oyashio, the transition, and the Kuroshio regions, respectively. The plus and minus signs indicate positive and inverse relation to the output, respectively.

tion tests, the neural network yielded predictions with smaller errors than those generated by the linear multiple regression.

Changing the number of hidden units controls the complexity of the function represented by a neural network (Ripley, 1996). To retain the net-

work's ability to generalize, it may be not an optimal strategy to train it to perfection (Haykin, 1994). The neural net risks overfitting data, and one way to avoid this overfitting is limiting the number of hidden units (Smith, 1996). For our data, one hidden layer and 3 or 4 hidden units were sufficient to achieve good results. The simple structure of the hidden layer proved better for generalization.

The number of input units is also important to prevention of overfitting. In cases of the Oyashio and the transition regions, selection of input variables slightly improved the predictive performance. On the other hand, in the case of the Kuroshio region, a limit of five variables in Model 3 reduced the capacity to predict. This is probably because some primary factors were removed. Model 1 with 12 variables provided a prediction close to the actual values, as did Model 2 with eight variables. In the linear multiple regression, too many input variables resulted in a poor prediction. The neural network, however, maintained better stability than the linear multiple regression when all 12 variables were included. Input values are transformed to compound variables in hidden units, and a selection of input variables is made in the form of the interconnecting weights in the network. This process can be viewed as a joint use of non-linear principal component analysis and multiple regression analysis (Hirafuji et al., 1988).

Table 4
Comparison of mean errors in prediction test between neural networks and multiple linear regression [a]

Model	Neural network (g/m^2 (%))	Multiple regression (g/m^2 (%))
Oyashio region		
Model 1	1.72 (9.86)	2.83 (16.22)
Model 2	1.34 (7.69)	1.75 (10.02)
Model 3	1.44 (8.28)	1.71 (9.81)
Transition region		
Model 1	0.621 (7.89)	0.716 (9.10)
Model 2	0.591 (7.51)	0.846 (10.76)
Model 3	0.546 (6.94)	0.670 (8.52)
Kuroshio region		
Model 1	0.397 (6.96)	0.815 (14.29)
Model 2	0.387 (6.80)	0.423 (7.42)
Model 3	0.589 (10.3)	0.467 (8.19)

[a] Input variables included in each model are shown in Table 2. The results of the neural networks for Oyashio region are by Model-a (Table 3). Numerals in parenthesis are relative errors in%.

The neural net predictions were not fully satisfactory, for several of the years surveyed, for regions of the transition and the Oyashio. Since the trained network deals with new data by interpolation, predictions derived from the new data become inaccurate in the case of extrapolation. Zooplankton biomass in the transition region was at a low level in the 1950s. The actual values in the mid-1950s were below the range of training data (1962–1988). Consequently, the estimates stayed within the range of training data. It is difficult to predict an unknown event that has not occurred in training data. In other words, the values of training data should cover as wide a range as possible.

For the Oyashio region, the neural net produced less successful predictions for the mid and late 1950s. The actual values of zooplankton biomass at the peak (1958) and trough (1963) were within the range of training data (Cases1 and 2). The poor prediction is likely due to the exclusion some factors from this study; factors that contribute to the change of zooplankton abundance. One possible factor is the wind on the sea surface. The wind can affect biological processes through turbulence and vertical mixing. Brodeur and Ware (1992) reported a positive correlation between the intensity of winter winds and summer zooplankton biomass in the subarctic Pacific Ocean. Although air pressure and atmospheric circulation indices represented by 500 hPa height were included in the neural net models, a direct measure of the intensity of the wind may be in order.

Correlation analysis of zooplankton biomass among three regions showed that there was a significant correlation between the Oyashio and the transition regions ($R = 0.411$, $P < 0.05$), and between the transition and the Kuroshio regions ($R = 0.625$, $P < 0.01$), though not between the Oyashio and the Kuroshio regions ($R = 0.163$, $P > 0.05$). The transition region mediates between the Oyashio and Kuroshio waters. The contribution of environmental factors in the transition region were common with the Oyashio and/or the

Kuroshio regions: SST, PV, and NM with the Kuroshio region; SOI with the Oyashio region; and IS with both regions. Environmental factors may affect zooplankton production in a given sea area directly and/or through that in other regions.

Among environmental parameters, the air pressure at Ishinomaki (IS) consistently had a strong negative correlation with zooplankton biomass in the three regions; that is, the lower air pressure led to a higher zooplankton biomass. This result agrees with the findings of Tomosada and Odate (1995). It is probable that the disturbance of sea surface layers by low air pressure causes vertical mixing and subsequent nutrient enrichment.

Kotani and Odate (1992) reported a significant negative correlation between zooplankton biomass and sea surface temperature off the shore of northeastern Japan. Our results also showed that zooplankton abundance in the transition and the Kuroshio regions increased as SST lowered. This is probably due to the southward advection of the Oyashio water, which is rich in plankton and nutrients.

The behaviour of the Kuroshio and the Oyashio currents had different respective effect on zooplankton abundance in these regions. Zooplankton in the Oyashio region becomes abundant when the Oyashio is confined north and the Kuroshio Extension south. On the contrary, zooplankton in the Kuroshio region increases when the Kuroshio Extension shifts northward. Plankton and nutrients are rich in the Oyashio region, poor in the Kuroshio region, and intermediate in the transition region. It seems possible that the Oyashio water is subject to a negative effect when it shifts southward and that the Kuroshio water is subject to a positive effect when it extends northward.

It has been suggested that the southward intrusion of the Oyashio is associated with the ENSO (El Niño-Southern Oscillation) phenomenon, during which the value of SOI becomes negative (Nitta and Yamada, 1989; Hanawa, 1991; Sekine, 1993). The contributions of OY and SOI were consistent in the Oyashio region, and SOI influences in the transition region similarly.

It is difficult to explain the positive correlation between zooplankton and air temperature (AT) for the transition region, since sea surface temperature (SST) had an inverse correlation. The variable AT may represent a mechanism other than the effect of temperature. An interpretation of the connection between the Subtropical Index (STI) and zooplankton biomass in the Kuroshio region as strong is speculative. STI indicates the intensity of the westerlies at low latitudes (Nomoto and Chiba, 1986). The strong westerlies may be presumed to promote zooplankton production through a change of behaviour and sea surface disturbance of the Kuroshio, which is a subtropical gyre. The associations of zooplankton abundance with environmental factors were modelled in the form of the trained neural network, and the sensitivity analysis of the nets demonstrated an ability to extract predictive knowledge.

In this study, original data were smoothed by using 5-year running averages. This process was likely to eliminate complex relationships from the data, which may be why improvement in prediction by using neural networks was slight compared with the results of linear multiple regression. The advantage of neural networks may become more evident in the case of predicting annual changes of zooplankton biomass using original data. In such an analysis, we would need to consider the time lag between changes in zooplankton biomass and driving environmental variables. The time lag probably differs among variables, and a number of combinations of input variables with different time lags are assumed. Based on this study, further analysis will promote a clearer understanding of zooplankton dynamics in relation to oceanic changes.

References

Aoki, I., Komatsu, T., 1997. Analysis and prediction of the fluctuation of sardine abundance using a neural network. Oceanol. Acta 20, 81–88.

Brodeur, R D., Ware, D.M., 1992. Long-term variability in zooplankton biomass in the subarctic Pacific Ocean. Fish Oceanogr 1, 32–38.

Hanawa, K., 1991 Long-term variations of the atmospheric circulation over the North Pacific and the Oyashio. Bull. Hokkaido Natl Fish. Res. Inst. 55, 125–139.

Hirafuji, M., Ono, Y., Kobayashi, K., 1988. Nonlinear multivariate analysis with a neural network and a method of developing neural expert-systems. Proc.5th Meeting of Japan. Soft. Sci., pp 113–116.

Haykin, S., 1994. Neural networks. a comprehensive foundation Macmillan, NY, p. 696.

Japan Meteorological Agency, 1994. Report of climatic change '94. P. 444.

Kawai, H., 1972. In: Masuzawa, J. (Ed.), Hydrography of the Kuroshio and the Oyashio. In Physical Oceanography II. Tokai Univ. Press, Tokyo, pp. 129–321.

Kawai, H., 1989. Long-term fluctuations in the north limit of the Kuroshio E xtension axis and in the south limit of the Oyashio water near the east coas t of Japan Bull. Japan. Soc. Fish. Oceanogr. 53, 353–363.

Kotani, Y., Odate, K., 1992. Variations of zooplankton biomass in the waters off Tohoku province. Bull. Japan. Soc. Fish. Oceanogr. 56, 182–185.

Murakami, M., 1993. On the 100 meter depth temperature indicative of the Kuroshio Extension axis in Tohoku area. Umi no Kenkyu 2, 343–349.

Nitta, T., Yamada, S., 1989. Recent warming of tropical sea surface temperature and its relationship to the northern hemisphere circulation. J. Meteorol. Soc. Japan 67, 375–383.

Nomoto, S., Chiba, M., 1986 Relation between the monthly mean temperature, monthly total precipitation and the 500 mb circular indices in Japan. Tenki 33, 31–39.

Odate, K., 1994. Zooplankton biomass and its long-term variation in the western North Pacific Ocean, Tohoku Sea Area, Japan Bull. Tohoku Natl. Fish. Res. Inst. 56, 115–173.

Ripley, B.D., 1996. Pattern recognition and neural networks. Cambridge University Press, NY, p 403.

Sekine, Y., 1993. Variation in subarctic circulation in the North Pacific coupled with the variation in atmospheric circulation Uni to Sora 69, 73–80

Smith, M., 1996. Neural networks for statistical modeling. International Thompson Computer Press, Boston, p. 235.

Tomosada, A., Odate, K., 1995. Long-term variability in zooplankto biomass and environment. Umi to Sora 71, 1–7.

ELSEVIER

Ecological Modelling 120 (1999) 271–286

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks

Ingrid M. Schleiter [a,*], Dietrich Borchardt [a], Rüdiger Wagner [c], Thomas Dapper [c], Klaus-Dieter Schmidt [c], Hans-Heinrich Schmidt [b], Heinrich Werner [c]

[a] *Department of Sanitary and Environmental Engineering, University of Kassel, Kurt-Wolters-Strasse 3, D-34125 Kassel, Germany*
[b] *Limnologische Flußstation, Max-Planck-Gesellschaft, Schlitz, Germany*
[c] *Department of Mathematics/Computer Science, University of Kassel, Kassel, Germany*

Abstract

The assessment of properties and processes of running waters is a major issue in aquatic environmental management. Because system analysis and prediction with deterministic and stochastic models is often limited by the complexity and dynamic nature of these ecosystems, supplementary or alternative methods have to be developed. We tested the suitability of various types of artificial neural networks for system analysis and impact assessment in different fields: (1) temporal dynamics of water quality based on weather, urban storm-water run-off and waste-water effluents; (2) bioindication of chemical and hydromorphological properties using benthic macroinvertebrates; and (3) long-term population dynamics of aquatic insects. Specific pre-processing methods and neural models were developed to assess relations among complex variables with high levels of significance. For example, the diurnal variation of oxygen concentration (modelled from precipitation and oxygen of the preceding day; $R^2 = 0.79$), population dynamics of emerging aquatic insects (modelled from discharge, water temperature and abundance of the parental generation; $R^2 = 0.93$), and water quality and habitat characteristics as indicated by selected sensitive benthic organisms (e.g. $R^2 = 0.83$ for pH and $R^2 = 0.82$ for diversity of substrate, using five out of 248 species). Our results demonstrate that neural networks and modelling techniques can conveniently be applied to the above mentioned fields because of their specific features compared with classical methods. Particularly, they can be used to reduce the complexity of data sets by identifying important (functional) inter-relationships and key variables. Thus, complex systems can be reasonably simplified in clear models with low measuring and computing effort. This allows new insights about functional relationships of ecosystems with the potential to improve the assessment of complex impact factors and ecological predictions.

Keywords: Artificial neural networks; Stream invertebrates, Population dynamics; Impact assessment; Bioindication; Time-series

* Corresponding author. Fax: +49-561-8043642.
E-mail address: schleit@hrz.uni-kassel.de (I.M. Schleiter)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00108-8

1. Introduction

The physical and chemical properties of running waters and their effects on the community are driven by numerous environmental variables such as climatic conditions, production–respiration ratio, urban storm-water run-off and waste-water effluents. The underlying interactions and dependencies are only partially understood. Furthermore, data for the calibration of theoretical models often are qualitatively or quantitatively insufficient. Because the knowledge of species–habitat interrelations remains insufficient, an integrative and, in consequence, prognostic assessment of ecosystem properties is not presently available (e.g. Vannote et al., 1980; Statzner et al., 1988; Townsend, 1989; Statzner et al., 1994; Townsend and Hildrew, 1994; Bayerisches Landesamt für Wasserwirtschaft, 1998).

Ecosystem analysis and prediction with empirical statistical and analytical methods are often limited by the spatially complex and temporally dynamics of ecological processes. This is one reason for the typically non-linear interrelations of variables and species with data being not normally distributed. Therefore, alternative mathematical methods have to be developed. Artificial neural networks (ANNs) provide an attractive alternative tool for analysing ecological data and for modelling due to their specific features such as non-linearity, adaptivity (i.e. learning from examples), generalisation and model independence (no a-priori model needed).

ANNs have been applied to various fields of aquatic sciences and engineering, such as modelling water quality (e.g. Daniell and Wundke, 1993; Maier and Dandy, 1993, 1994, 1996a,b; Lachtermacher and Fuller, 1994; Schizas et al., 1994; Maier, 1995a; Winkler and Voigtländer, 1995; Kaluli et al., 1998; Wen and Lee, 1998) and relating community characteristics with environmental variables (e.g. Chon et al., 1996; Lek et al., 1996; Recknagel, 1997; Recknagel et al., 1997, 1998; Guégan et al., 1998; Lee et al., 1998; Maier et al., 1998). Additional articles are found in this issue and in a review by Maier (1995b), focusing on the prediction of environmental, hydrological and water resources data.

In this paper we present results of the applicability of ANNs in the following fields: (1) dynamics of water quality as influenced by meteorology, urban storm-water run-off and waste-water effluents; (2) bioindication of chemical and hydromorphological habitat characteristics with benthic macroinvertebrates; and (3) prediction of population dynamics of aquatic insects.

The general objectives of this paper are to demonstrate the potential and limitations of ANNs and other modelling techniques for data analysis, impact assessment and ecological prediction in running waters, and to specify the general conditions for applications of ANNs, such as selection of relevant input variables, training conditions, network type and forecasting period.

2. Material and methods

We used multi-layer-perceptrons based on the Backpropagation (BP) algorithm (Rumelhart et al., 1986) and two-dimensional, motoric feature maps (FM; Ritter et al., 1994). A special variant of the BP-network type, the so-called senso-net (Dapper, 1998), was also used to determine the most important input variables (sensitivity analysis). Senso-nets include an additional weight for each input neuron representing the relevance (sensitivity) of the corresponding input parameter for the neural model. The sensitivities are adapted during the training process of the network. Appropriate subsets of potential input variables can be selected according to these sensitivities. In contrast to most statistical methods, the dimension-reducing techniques based on neural networks have the ability to map non-linear coherences (in this application, between species abundance and environmental variables). The neural networks and modelling techniques used for our experiments are described in Werner et al. (1999). The generalisation performance, E, of the networks was calculated as:

$$E = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{c} (y_{ij} - \hat{y}_{ij})^2 \quad (1)$$

where n is the number of samples in the test set; c is the number of output neurons; y_{ij} is the measured values; and $\hat{y}_{ij}$ is the modelled values.

For each variable and network type, we trained a number of networks and calculated E (minima, maxima, mean and median). We used the determination coefficient R^2 between measured and modelled data as a second standard measure of model quality. Unless otherwise specified, the results indicate E_{min} and R^2_{max}.

Prediction of water quality variables at the river Lahn (topic 1) was based on both on data from studies on a small running water body (Kuhbach (Germany); Borchardt et al., 1997) and a larger stream (Lahn at Limburg (Germany)), using different ANNs. Basic data were daily meteorological variables (precipitation, radiation), discharge and water quality data (oxygen, conductivity, pH, water temperature) obtained above and below the waste water treatment plant of Limburg (main characteristics summarised in Table 1). Modelling daily maxima and minima was based on routine government data in 1995 (365 day-sets), whereas for modelling the diurnal fluctuation was based on data from a scientific study (150 day-sets) from April to October 1996. The data were used without pre-processing either as compact time intervals or as alternating days distributed into a training and a test data set at a ratio of 1:1, 2:1 and 3:1, respectively. Different network types were trained with several combinations of input variables and histories of varying lengths (see Table 2).

Table 1
Main characteristics of the River Lahn at Limburg[a]

Catchment	Annual precipitation (mm)	675
	Area (ha)	42998
River Lahn	Mean annual discharge (l/s)	47000
	Mean low discharge (l/s)	10000
Limburg city	Number of inhabitants	44051
	Channelized catchment (ha)	1522.4
	Impervious area (ha)	765.0
	Specific water consumption (l/(Inh.*d))	151.1
	Number of storage tanks (–)	40
	Specific storage volume (m^3/ha)	23.7
	Discharge WWTP (l/s)	304.0

[a] From Mang et al. (1998), modified.

Functional relationships between water quality, habitat characteristics and colonisation patterns of benthic macroinvertebrates (topic 2), were based on measurements of ten chemical and seven hydromorphological variables (Table 3), and the abundance of 248 species from eight small streams in Central Hesse (Germany). The experiments used the daily maxima of each variable over a one year period, except for oxygen where the minimum was used. The data were analysed with Pearson correlation and stepwise forward regression analysis (SPSS, 1997), and sensitivity analysis on Senso-nets. The pre-processing was used to identify those species which have significant interrelations with the variables in question, and thus may be used as bioindicators. The chemical and morphological variables were modelled with BPN, FM and senso-nets, respectively, using the abundance of the five most relevant species which were identified by stepwise forward regression analysis, here called predictors. The use of the best five predictors (input variables) provided the best models (lowest generalisation error) when testing different procedures for the selection of input variables (correlation, regression, factor analysis, sensitivity analysis on senso-nets, reduction of trained BPN, bottleneck nets) modelling four chemical variables: oxygen, conductivity, BOD_5, NH_4–N (Dapper, 1998; Schleiter et al., in prep.). The R^2 values of these neural models, calculated on unknown test-data, were nearly as high as those of the regression models with the best five predictors as input, calculated on the whole data: $R^2_{(BPN)} = 0.89$, $R^2_{(regression)} = 0.90$. For the neural modelling, all variables were standardised linearly into the interval [0, 1] and in circulation divided into a training-set and a test-set in a ratio of 2:1.

Long-term population dynamics (topic 3) are based on monthly data of aquatic insect emergence and environmental variables from 1969 to 1994 for the small stream Breitenbach (Central Germany; Wagner and Schmidt, 1999). The data were collected by the Limnologische Flußstation, Schlitz. We compared the accuracy of the abundance prediction of *Apatania fimbriata* (Pictet,

Table 3
Correlation matrix of water quality (ten parameters), habitat structure (seven parameters) and benthic macro-invertebrates[a]

	NH_4-N	COD	BOD_5	Conductivity	NO_2-N	P_{tot}	Oxygen	NH_3-N	NO_3-N	pH	Discharge regime	Diversity of substrate	Fine sediment	Width/depth ratio	Diversity of habitat features	Extension of riparian zone	Structure of river bank	Average score 5 bed-parameters	Average score 2 riparian parameters	Average score 7 morphological parameters
Chironomus thummi-Gr	0 35	0 58	0 52	0 54	0 63	0 55	0 52	0 48			0 42	0 50	0 61	0 49	0 51			0 57		0 51
Gammarus pulex	0 35			0 35		0 43					0 64	0 42	0 50		0 67			0 54		0 56
Baetis rhodani		0 42	0 36	0 42	0 48	0 40								0 43	0 55			0 40		0 52
Elmis sp	0 36	0 36	0 35	0 38					0 36			0 40						0 40		
Hydrius fuliginosus (I)		0 48		0 66		0 65	0 46					0 43	0 42					0 40		
Tubifex sp		0 39	0 44									0 55	0 48					0 46		
Eristalinae indet				0 58		0 41	0 42					0 40	0 46							
Simulium ornatum-Gr				0 56		0 38						0 52	0 40							
Simulium aureum-Gr				0 59								0 41								
Tanytarsini indet				0 43										0 41						
Ecdyonurus venosus-Gr									0 36		0 46							0 41		
Amphinemura sp				0 39					0 37										0 49	
Siphlonurus lacustris	0 47	0 64	0 71		0 51			0 81												
Hydrobius fuscipes	0 48	0 55	0 55				0 41	0 53												
Heteroceras sp	0 36	0 45	0 56		0 44			0 69												
Chironomus plumosus-Gr		0 42	0 57		0 52			0 53												
Anacaena limbata (I)	0 42	0 37				0 43														
Limnephilus binotatus	0 36									0 42										
Limnodrilus sp	0 41		0 35																	
Haliplus laminatus (I)					0 36	0 35														
Sigara lateralis	0 48																			
Diamesinae indet				0 48																
Copelatus haemorrhoidalis							0 48													
Siphlonurus sp										0 40										
Culex sp					0 39															
Daphnia pulex					0 39															
Limnephilus nigriceps	0 38																			
Dolichopodidae indet		0 37																		
Erpobdella octoculata	0 37																			
Helobdella stagnalis	0 36																			
Sigara sp		0 36																		
Limnephilidae indet			0 36																	
Bezzia sp					0 36															
Cloeon dipterum	0 35																			
Chironomus sp	0 35																			
Anacaena globulus	0 34																			
Ironoquia dubia			0 34																	
Nepa rubra							0 34													
Potamonectes depressus										0 34										
Oreodytes sanmarkii (I)									0 35		0 41				0 51	0 51				0 57
Rhyacophila fasciata											0 63	0 53	0 48	0 62	0 59	0 67	0 60	0 62	0 65	0 70
Hydropsyche sp											0 53	0 39	0 50		0 57	0 58		0 49	0 54	0 60
Gammarus fossarum											0 41	0 39	0 41				0 57	0 43	0 53	0 49
Radix ovata												0 45	0 44			0 53	0 51	0 42	0 53	0 61
Sericostoma sp														0 51	0 50	0 66	0 65	0 41	0 67	0 66
Chaetopteryx villosa											0 42				0 58	0 49	0 53		0 53	0 54
Limnius perisi (I)											0 49		0 42			0 50		0 45		0 49
Potamophylax latipennis											0 42		0 39		0 55			0 42		0 52
Elmis aenea (I)															0 74	0 58	0 60		0 61	0 70
Leuctra nigra											0 49	0 48	0 40					0 48		
Pisidium casertanum											0 41	0 43		0 44				0 45		
Tanypodinae indet												0 58	0 50	0 55				0 50		

Table 3 (Continued)

	NH_4–N	COD	BOD_5	Conductivity	NO_2–N	P_{tot}	Oxygen	NH_3–N	NO_3–N	pH	Discharge regime	Diversity of substrate	Fine sediment	Width:depth ratio	Diversity of habitat features	Extension of riparian zone	Structure of river bank	Average score 5 bed-parameters	Average score 2 riparian parameters	Average score 7 morphological parameters
Rhithrogena venicolorata-Gr												0 46	0 43	0 54				0 50			
Lasiocephala basalis															0 51	0 57			0 53	0 53	
Orectochilus villosus															0 57	0 52			0 50	0 58	
Drusus annulatus											0 42			0 40				0 42			
Sienophylax permistus											0 40			0 42				0 41			
Annitella obscurata											0 41			0 43				0 42			
Halesus radiatus											0 42			0 43				0 42			
Dugesia gonocephala													0 44	0 49				0 40			
Hydropsyche pellucidula											0 42									0 52	
Ancylus fluviatilis											0 41			0 48							
Potamophylax rotundipennis									0 51							0 43					
Potamophylax nigricornis											0 40							0 40			
Rhyacophila sp											0 49							0 47			
Sialis fuliginosa											0 44									0 49	
Limnephilus sp												0 41									
Bathyomphalus contortus												0 40									
Glyphotaelius pellucidus											0 40										
Baetis niger											0 42										
Limnius sp											0 42										
Leuctra digitata											0 42										
Protonemura meyeri											0 42										
Potamophylax cingulatus														0 49							
Ephemerella ignita															0 50						

[a] Sampling sites from eight rivers, sample size, $n = 3$ for each sampling

1843), (Insecta, *Trichoptera*) based on canonical correspondence analysis (CCA, ter Braak, 1988, 1990) and ANNs. *A. fimbriata* is a dominant species in the Breitenbach. Environmental variables included into the models were maximum monthly water temperature (T) and discharge (D), both measured at the Breitenbach, and monthly precipitation (P) determined close to the catchment. Further variables had marginal or no influence on species abundance (individuals/m^2) and were thus omitted from the models (Borchardt et al., 1997). Environmental variables and populations were related with correlation, regression (SPSS, 1997) and ordination (Wagner and Schmidt, 1999). We calculated and tested the significance of abundance differences between groups of years with different discharge patterns. ANNs used all available data of P, D, T, and abundance (A) of preceding periods to predict species abundance in the target month (training set: test set ratio was 4:1, $n = 300$). Modelling with the entire database was compared with methods of a preceding reduction of vector dimensions by correlation, regression or sensitivity analysis (see below), to reduce computing time. Reduction of dimension in this case means the deletion of variables with evidently low or no influence on the target variable, and not a loss of information due to the computation of a mean and a variability measure (Dapper, 1998).

3. Results

3.1. Temporal variability of water quality

The daily maxima and minima of all target variables could be modelled successfully using the data of the previous day as network input (Table 2). The generalisation performance of BPN was higher than those of FM. As can be seen from the generalisation error in Table 2, the performance of both network types could be increased clearly, when specifying the nets to one output variable (maximum or minimum). Furthermore, an improvement of the generalisation performance was reached by increasing training effort. The accuracy of the network predictions decreased with

Table 2
Summary of generalisation errors for different combinations of target-oxygen concentrations, conductivity and pH and input-parameters[a]

Target	Input-parameters	Generalisation error BPN	FM
O_2-min/max ($t+1$)	O_2-min/max (t)	0.00463	0.01427
	O_2-min/max (t)+discharge (t)	0.00571	0 01460
	O_2- diurnal variation (t)	0.00270	0.00410
O_2-min ($t+1$)	O_2-min (t), 200 days in the training set	0 00227	0.00313
	O_2-min (t), 212 days in the training set	0.00185	0.00283
	O_2-min (t), 320 days in the training set	0.00199	0.00159
O_2-min ($t+1$)	O_2-min (t), 300 days in the training set	0.00225	0.00185
O_2 min ($t+2$)	O_2-min (t), 300 days in the training set	0.00303	0.00443
O_2 min ($t+3$)	O_2-min (t), 300 days in the training set	0.00515	0 00733
O_2 min ($t+4$)	O_2-min (t), 300 days in the training set	0.00677	0 00881
O_2 min ($t+5$)	O_2-min (t), 300 days in the training set	0.01064	0.01049
O_2-min ($t_{31}-t_{60}$)	O_2-min (t_1-t_{30}), 30 days in the training set	0.00035	0 00035
O_2-min ($t+1$)	O_2-min (t)+discharge (t)+water temperature (t)+rainfall (t)	0.00193	0.00338
	O_2-min (t)+rainfall (t)+rainfall ($t+1$)	0.00235	0.00270
	O_2-min (t)+discharge (t)+discharge ($t+1$)	0.00234	0.00329
	Water temperature (t)	0.00861	0.00948
	Water temperature (t)+discharge (t)	0.00793	0.00863
	Water temperature (t)+discharge (t)+rainfall (t)	0 00754	0.00570
	Water temperature (t)+discharge (t)+rainfall (t)+pH-value (t)+conductivity (t)	0.00726	0.01009
conductivity-min/max ($t+1$)	Conductivity-min/max (t)	0.00599	0.00431
Conductivity-max ($t+1$)	Conductivity-max (t)	0.00143	0.00369
	Conductivity-max (t), 320 days in the training set	0.00341	0.00379
	Conductivity diurnal variation (t)	0.00495	0.00620
	Conductivity(t)+discharge (t)+water temperrature (t)+rainfall (t)	0.00331	0.00627
	Conductivity (t)+pH-value (t)+discharge (t)+water temperature (t)+rainfall (t)	0.00346	0.00520
pH-min/max ($t+1$)	pH-min/max (it)	0.01661	0.01853
	pH-min/max (t), 212 days in the training set	0.00198	0.00199
pH-max ($t+1$)	pH-max (t), 212 days in the training set	0.00067	0.00065
	pH-max (t), 320 days in the training set	0.00020	0.00020
	pH-values diurnal variation (t)	0.00180	0.00280
	pH-max (t)+discharge (t)	0.00090	0 00092
	pH-max (t)+rainfall (t)	0.00075	0.00064
	pH-max (t)+discharge (t)+water temperature (t)+rainfall (t)	0.00091	0.00137
	pH-max (t)+conductivity (t)+discharge (t)+water temperature (t)+rainfall (t)	0 00113	0.00182

[a] Unless otherwise specified, 200 training-sets were utilized and 60-min-average of 5-min-measurements, except for discharge (daily maximum), were used; $t=$, days.

increasing forecast period. The daily minima of oxygen for the following month could be predicted with low error by both network types using the oxygen minima of the previous month.

The water quality at the time t proved to be the most important input variable, predicting water quality at the time $t+1$ with different input modalities (Table 2). The inclusion of other or supplementary input variables caused no improvement of the generalisation performance of the networks.

While the daily maxima and minima of oxygen and other water variables could be predicted with relative low error, the forecast of the diurnal

variation appeared to be more difficult. This is shown by predictions of the diurnal variation of oxygen with a simple model using the sum of precipitation and the oxygen values of the previous day (Fig. 1). The generalisation power of the 25-8-24 BPN was slightly higher than those of a 15×15 FM ($E_{BPN} = 0.05307$, $R^2_{BPN} = 0.79$ versus $E_{FM} = 0.071021$, $R^2_{FM} = 0.75$).

A series of experiments on network training with variations of data length, histories and different measurement modes (60- and 30-min measurements of oxygen, daily sum and 5-min-values of precipitation) showed no general trend and could therefore be considered to be of minor importance for model performance (Borchardt et al., 1997).

3.2. Colonisation patterns of benthic macro-invertebrates

A correlation analysis provided high significance levels ($\alpha < 0.01$) for 40 out of 248 species with at least one of the chemical variables, and for 47 species with at least one of the morphological variables (Table 3). The number of highly significant species for chemical variables varied between 3 (pH) and 16 (NH_4–N), whereas those for hydromorphological variables varied between 9 (extension of riparian zone) and 27 (discharge regime) (compare Table 3). Most species showed highly significant relationships with only one type of variables, either chemical or hydromorphological.

The stepwise regression analysis provided more or less complex models for the different chemical variables: the number of predictors were 10 for BOD_5, 11 for COD, 16 for oxygen and total phosphorus, 17 for NH_4–N, 21 for conductivity and NO_2–N, 23 for NH_3–N and NO_3–N, and 25 for pH-value.

Using the abundance of only the five best predictors for each variable, the 10 chemical factors could be modelled with good agreement between measured and modelled values ($E < 0.01$, $R^2 > 0.8$; Table 4). This indicates functional relationships between the chemical variables and the selected species groups. The generalisation performance of BPN was higher than those of FM (average $E_{BPN} = 0.00558$, $E_{FM} = 0.01225$; Table 4), except for NH_3–N ($E_{BPN} = 0.01634$, $E_{FM} = 0.00970$, $R^2 < 0.35$). This is because modelling NH_3–N on sampling site 8, provided a high generalisation error, particularly with BPN ($E_{BPN} = 0.13268$, $E_{FM} = 0.11006$).

The generalisation performance of the reduced networks was clearly higher than those based on all potential input variables (average $E_{BPN(5)} = 0.00316$ via $E_{BPN(248)} = 0.02088$ comparing four models: oxygen, conductivity, BOD_5, NH_4–N). The calculation effort decreased to 2% for FM and 0.9% for BPN compared with those based on all 248 input variables.

Even for the seven morphological variables simple neural models could be generated (Table 5). The average performance of senso-nets based on the abundance of the five best predictors was $E = 0.0074$, $R^2 = 0.85$. The best generalisation was reached for the average score of the seven hydromorphological variables—in Germany, the assessment used for the morphological structure of streams is called Gewässerstruktur-Güteklasse (Fig. 2).

3.3. Population dynamics of aquatic insects

The results of CCA-ordination indicated a strong dependence of the population density of *A. fimbriata* on the discharge pattern (Wagner and Schmidt, 1999). Abundance was highest at high discharge with low flow variability (D), was lower at winter and spring floods (E), and lowest during periods of low flow (F) or after seasonally unpredictable discharge events (B in Fig. 3). Based on monthly data, no significant dependence of $D-P$ was detected. Abundance between patterns was significantly different. However, predictions could only be made with an error of hundreds of specimens per year.

The precision of the ANN model with the original data was quite high ($R^2 = 0.63$). All months with any abundance were predicted correctly. The abundance magnitude differed between prediction and actual data (Fig. 4a). Pre-selection of five variables ($abundance_0$, $abundance_{11}$, $temperature_1$, $temperature_6$, $temperature_7$) with correlation analysis increased

Table 4
Generalisation error of 5-3-1-BPN and 5 × 5-FM for predictions of chemical parameters from the abundance of five macroinvertebrates at each case identified with regression analysis

Target parameter	Oxygen		Conductivity		BOD_5		NH_4–N		pH	
Network type	BPN	FM	BPN	FM	BPN	FM	BPN	FM	BPN	FM
Error										
Minimum	**0.00278**	0.00836	**0.00465**	0.00631	**0.00290**[a]	0 01029	**0.00229**	0.01227	**b**	0.01851
Maximum	0.01232	0 01006	0.01645	0.01089	0.00971	0.01726	0 01678	0.03063	0.02241	0.02117
Mean	0.00611	0.00887	0.00834	0.00943	0.00682	0.01247	0.00710	0.01770	0.01479	0.01915
Median	0 00562	0.00872	0.00745	0.01013	0 00742	0.01159	0.00493	0.01524	0.01374	0.01851
Standard deviation	3.49E-06	3.76E-07	8.48E-06	2.95E-06	5.64E-06	6.00E-06	1.86E-05	4.33E-05	1.71E-05	1.07E-06
Species	*Chironomus thummi*-Gr		*Limnius volckmari*		*Agabus* sp.		*Sigara lateralis*		*Glossiphonia complanata*	
	Dolichopodidae indet.		*Elmis sp*		*Dolichopodidae indet*		*Dolichopodidae indet*		*Radix peregra*	
	Goera pilosa		*Goera pilos*		*Limnodrilus* sp.		*Limnophila* sp.		*Sericostomatidae indet*	
	Hydroporus sp.		*Oreodytes sanmarkii*		*Calopteryx splendens*		*Tanytarsini indet.*		*Sigara fossarum*	
	Anacaena limbata		*Chironomus plumosus*-Gr		*Nemoura avicularis*		*Chironomus plumosus*-Gr		*Daphnia pulex*	

Target parameter	COD		NH_3–N		NO_2–N		NO_3–N		P_{tot}	
Networktype	BPN	FM	BPN	FM	BPN	FM	BPN	FM	BPN	FM
Error										
Minimum	**0.00292**	0.02624	0.01634	**0.00970**	**0.00731**	0.00899	**0.00369**	0.01212	**0.00552**	0.00970
Maximum	0.05646	0.03338	0.02264	0.01037	0.01566	0.02873	0.01830	0 02657	0 01412	0.02017
Mean	0.02708	0.03054	0.01930	0 01016	0.01103	0.01518	0.00793	0 02312	0.00890	0.01564
Standard deviation	1.24E-04	6.70E-06	2.21E-06	5.58E-08	3.70E-06	4.89E-05	1.57E-05	3.11E-05	4.94E-06	1.54E-05
Species	*Ilybius fuliginosus*		*Ilybius fuliginosus*		*Culex* sp		*Chironomidae indet.*		*Gammarus pulex*	
	Enallagma cyathigerum		*Bathyomphalus contortus*		*Chironomus* sp.		*Goera pilosa*		*Limnephilus rhombicus*	
	Dolichopodidae indet		*Limnophila* sp		*Rhyacophila* sp.		*Radix auricularia*		*Erpobdella octoculata*	
	Pyrrhosoma nymphula		*Sigara* sp.		*Limnephilus rhombicus*		*Tanytarsini indet.*		*Helobdella stagnalis*	
	Bathyomphalus contortus		*Culex* sp		*Tubifex* sp.		*Hydropsyche angustipennis*		*Chaetopteryx villosa*	

[a] Bold, lowest error

the accuracy of the model to $R^2 = 0.86$ (Fig. 4b). Cross correlation indicated almost no influence of precipitation on *A. fimbriata* abundance. Pre-selection by regression analysis found other variables relevant ($abundance_0$, $abundance_{10}$, $abundance_{11}$, $temperature_1$, $precipitation_{12}$) and increased the accuracy of the model to $R^2 = 0.86$ (Fig. 4c). The best of three different sensitivity analyses selected the variables $abundance_0$, $abundance_{11}$, $temperature_0$, $temperature_6$, $discharge_6$, and had an accuracy of $R^2 = 0.93$ (Fig. 4d). An overview of these experiments indicates the best models were computed with a pre-selection of the best five variables by sensitivity analysis or regression. The models with variables pre-selected by correlation or without any pre-selection resulted in lower measure of accuracy (Table 6). ANN models generally explained much more variability (20–30%) than linear regression models.

4. Discussion

The corresponding chemical data of the previous day proved to be the most important network input, when modelling water quality of the river Lahn. Using other or supplementary input variables, we achieved no significant improvement of the generalisation performance of the networks. This result is attributed to the strong autocorrelation of the values at the time t and $t+1$.

In our experience, the most important basis for successful neural modelling is a sound and representative data base. For example, from our data it was not possible to predict the temporal variation

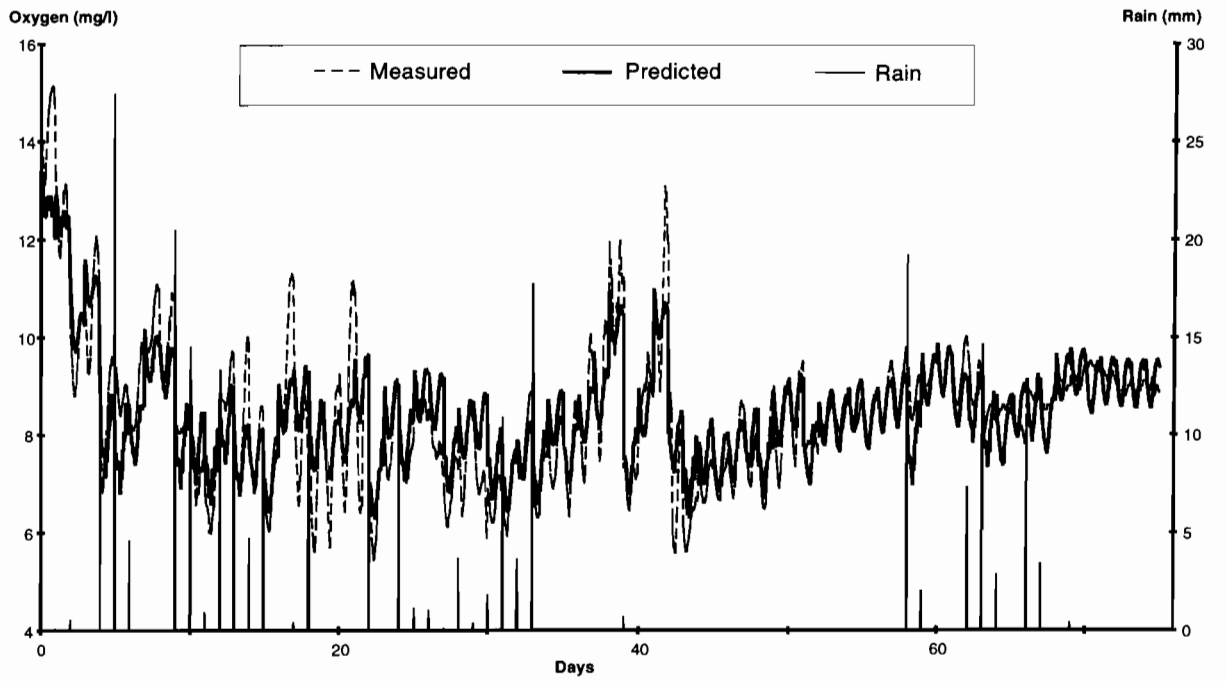

Fig. 1. Comparison of measured and modelled temporal patterns of oxygen based on daily sum of precipitation and oxygen (24 60-min-values) of the previous day with 25-8-24-BPN (151 daily data sets for training and test).

Table 5
Modelling morphological habitat characteristics with 5-3-1-senso-nets from the abundance of five macroinvertebrates at each case identified with regression analysis

Target	$Error_{min}$	R^2	Species
Discharge regime	0.0067	0.87	*Gammarus pulex* *Gammarus roeseli* *Nemoura cinerea* *Dolichopodidae indet.* *Agabus guttatus* (I)
Width/depth ratio	0.009	0.83	*Rhyacophila fasciata* *Dugesia gonocephala* *Tanypodinae indet.* *Elmis aenea* (I) *Agabus guttatus* (I)
Diversity of substrate	0.0132	0.82	*Tanypodinae indet* *Gammarus pulex* *Agabus didymus* *Tubifex* sp. *Pilaria* sp.
Fine sediments	0.0137	0.64	*Chironomus thummi*-Gr *Gammarus pulex* *Tanypodinae indet.* *Pilaria* sp. *Agabus didymus*
Diversity of habitat features	0.0096	0.83	*Gammarus pulex* *Pilaria* sp. *Gyraulus albus* *Gyrinus* sp. *Hydropsyche* sp.
Extension of riparian zone	0.0048	0.90	*Rhyacophila fasciata* *Electrogena* sp. *Ephemerella ignita* *Limnodrilus* sp. *Sigara* sp.
Structure of river bank	0.0048	0.95	*Sericostomatidae indet.* *Pilaria* sp *Hydropsyche angustipennis* *Plectrocnemia conspersa* *Dytiscidae indet.*

of water quality as a function of meteorological data. For this purpose, precise data gathering or a spatial-temporal allocation of the input (radiation, precipitation) and target variables (e.g. oxygen) are necessary on compatible time scales.

Generally, whether the temporal dependence of output data is derived on a specific time scale or integrated over an indefinite period of time is decisive. Accordingly, different network approaches, either (non-linear) auto-regression or (partial) recurrent networks (e.g. Jordan-nets; Pham and Oh, 1992), are more suitable for successful modelling. The time dependent integration of previous states and events/processes is a major problem when modelling time series. We expect better modelling with specific time-dependent networks with feedback onto specific neurons storing internal network states.

Due to their specific features, particularly the ability to handle non-linearities, ANNs combined with specific procedures for the selection of input variables provide an attractive tool for modelling species/species traits and habitat relations. A series of chemical and hydromorphological properties could be modelled with low error from the abundance of only a few specific macroinvertebrates identified with regression analysis. This dimension-reducing, pre-processing caused an increase of the generalisation performance of the networks and a considerable reduction of the calculation effort. The results clearly indicate functional relationships between colonisation patterns of benthic macroinvertebrates and chemical and hydromorphological habitat characteristics within lotic ecosystems. Furthermore, a hierarchy of factors determining the community structure of invertebrates may be identified from theoretically numerous impact variables.

The species groups selected for each chemical and morphological model showed no or little congruence (see Tables 4 and 5). Even for related variables, different species groups were detected. Some species were selected by several models (e.g. *Gammarus pulex*: discharge regime, diversity of habitat features, P_{tot}, diversity of substrate, fine sediments), whereas others appeared in only one model. Because of restrictions in the basic data set due to the narrow geographical region and limited

abiotic gradients, the selected species groups in our examples may not be generalized. The results obtained here need further analysis with ecological information and validation based on additional data. Thereby more approaches have to be tested as genetic algorithms (Goldberg, 1989) to detect relevant predictors for non-linear models based on general regression neural networks (Specht, 1991), equation synthesis (e.g. Roadknight et al., 1997), weight analysis (Balls et al. 1996,) and correlated activity pruning (Wiersma et al., 1995). Based on more comprehensive data, we would expect it to be possible to verify if key species or species assemblages for definite abiotic environmental states can be identified independently and reproduced for different sites. This may also be possible using the species traits hypothesis (Resh et al., 1994).

Discharge and water temperature are two main abiotic factors controlling the structure and dynamics of stream invertebrate populations (Ward and Stanford, 1979, 1982), as well as the variability of habitats and the reproductive success of lotic species through metabolic processes (e.g. Feminella and Resh, 1990; Céréghino and Lavandier, 1998). However, many interdependencies between the environment and the species remain less well known. The long-term population dynamics of aquatic insects could be meaningfully described with ANNs in addition to classical statistics. Regression and correlation models have repeatedly been used to explain patterns in communities and they provided useful insights on environmental control of ecosystems, but their predictive power is low (ter Braak and Verdonschot, 1995; Paruelo and Tomasel, 1997; Walley and Fontama, 1998). With classical statistical methods and ordination (CCA; ter Braak, 1988, 1990), the variability between year abundance of individual species was attributed mainly to the discharge pattern during larval development (Wagner and Schmidt, 1999). Due to the necessity to recognise patterns and not single discharge events, ANNs are an alternative method to model species abundance (Colasanti, 1991; Lek et al., 1996).

Larvae of *A. fimbriata* are grazers that avoid sandy substratum, they undergo a dormancy from November to the next February underneath larger stones (Aurich, 1992). Concerning the life history traits, pattern D (Fig. 3) provided low discharge

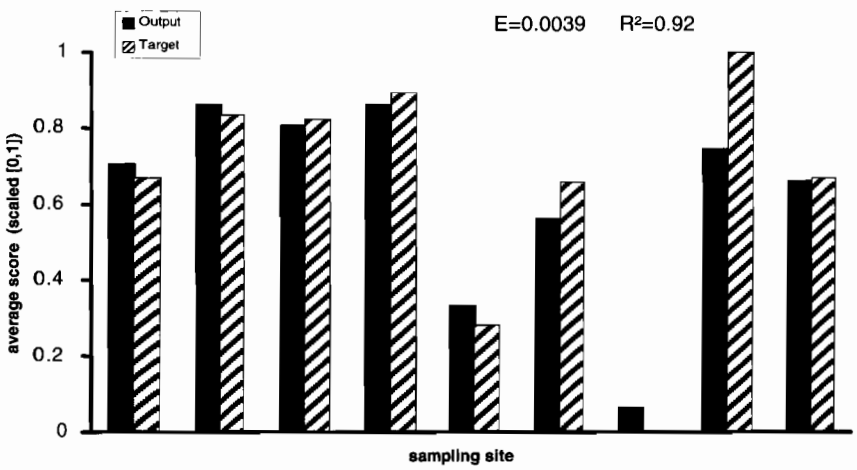

Fig. 2. Modelling average score of seven hydromorphological parameters with 5-3-1-senso-nets from the abundance of five macroinvertebrates at each case identified with regression analysis.

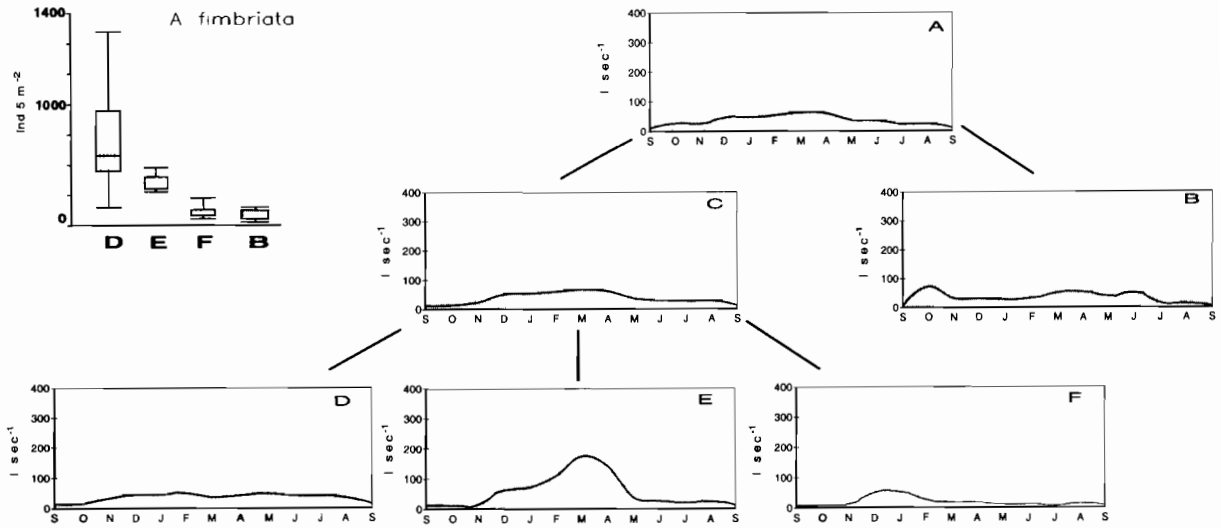

Fig. 3. Discharge patterns of the Breitenbach discriminated with CCA: (A) 25-years mean of monthly maximum discharge; (B) years with non-seasonal events; (C) the seasonal pattern; (D) permanent, good discharge, (E) winter and spring floods; (F) long-term low flow [mean of within group monthly maximum (line) $\pm$ 1 SD (raster)], and the density (ind/5m^2) of *A. fimbriata* at patterns D, E, F and B (top left).

variability for almost an entire year at high mean flow. Strongly increased discharge in winter and spring disturb larval populations in dormancy, and low flow conditions over almost the entire year are disadvantageous for the development of eggs in summer, larvae in dormancy, and the pupae in early summer, due to increased deposition of organic or inorganic material. Lowest success at the non-seasonal pattern is interpreted as the interaction of the magnitude and the duration of floods. Most other aquatic insects, with the exception of the mayfly *Baetis vernus* (Curtis), have their lowest abundance at these discharge pattern (Wagner and Schmidt, 1999).

In ANN models the best pre-selection method was a sensitivity analysis, whereas other methods were less accurate in the prediction of *A. fimbriata* abundance. *B. vernus* was also best modelled by sensitivity pre-selection, but in *B. rhodani* pre-selection by correlation was optimal (Wagner et al., 1999). Variables selected for the best models of all three species were abundance of the parent generation and temperature during the emergence or oviposition period of the parents. In addition, in *A. fimbriata* temperature and discharge 6 months before emergence are among the most relevant predictors. During this period larvae are in winter dormancy, and higher temperature or discharge may have disturbed the larvae or their habitat. This demonstrates the potential of precise abundance predictions some months before emergence of the adults, and of the preselection methods, in particular sensitivity analysis, that detected sensitive conditions or periods in the life cycle of *A. fimbriata*.

5. Conclusion and perspectives

The results show that ANNs can successfully and meaningfully be applied in the analysis of effect-relations (e.g. species/species traits with habitat) including the identification and assessment of complex impact factors and for the prediction of system behavior (e.g. critical water states with an early-warning-system and long-term population dynamics depending on environ-

mental variables) having specific features compared with conventional methods (see Werner et al., 1999). Particularly, they have advantages if the relationships are unknown, very complex or non-linear. Combined with specific procedures for the selection of the most important impact variables, they can be used to reduce the input dimension and therefore the complexity in a reasonable way. This causes an increase of the generalisation performance and a simplification of the model

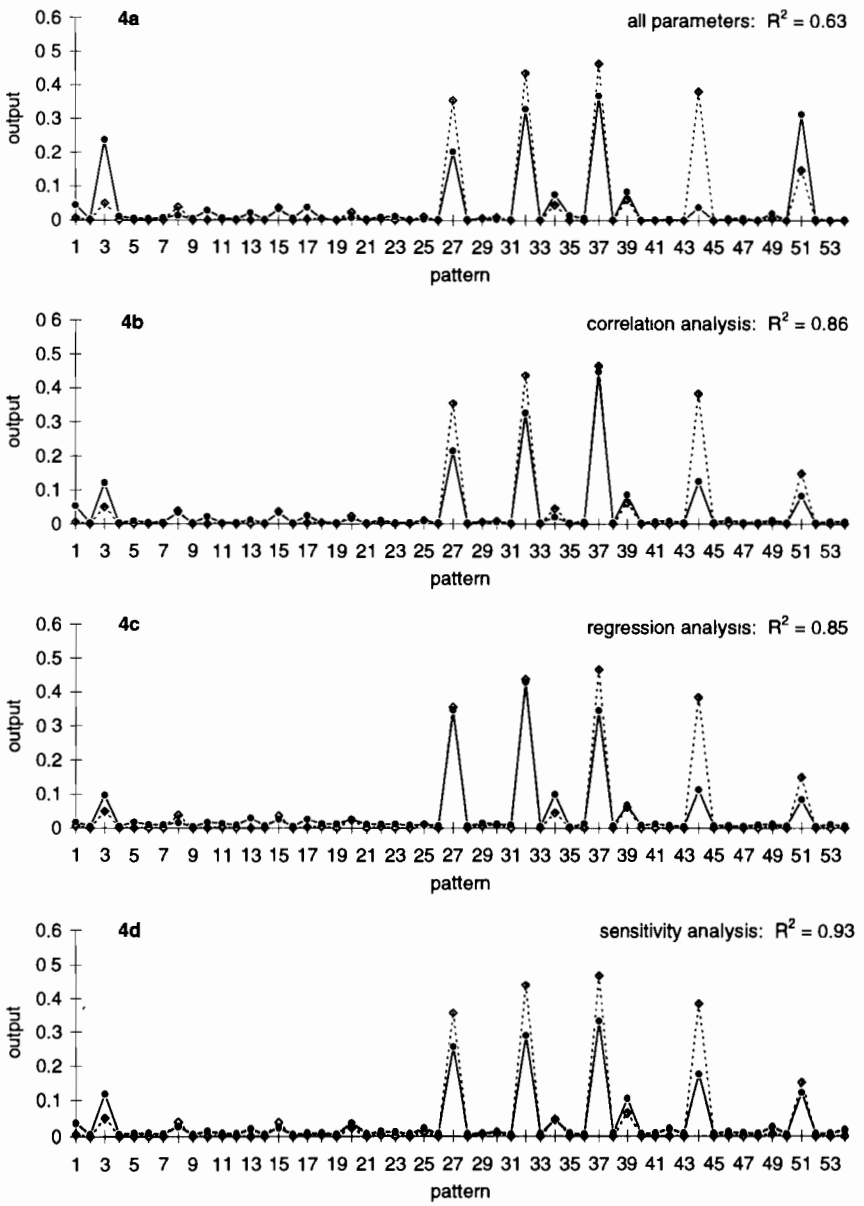

Fig. 4. Abundance prediction of *A. fimbriata*. (4a) Model, all input-variables, best generalisation (lowest error): 51-10-1-senso-net; (4b–d) Model, pre-selection by different methods, best generalisation: 5-3-1-senso-net. (Model, full line, diamond; observed data, dotted line, square).

Table 6
Overview of some ANN models with different network input modes to predict the abundance of *A. fimbriata*

Pre-processing	Sensonets, generalisation performance		
	$Error_{min}$	R^2	$Error_{mean}$
All variables	0.0022	0.63	0.0040
Correlation	0.0010	0.86	0.0019
Regression	0.0009	0.86	0.0017
Sensitivity	0.0010	0.93	0.0020

and allows a better understanding of the underlying relations.

Because the networks learn from examples, the quality of the neural models heavily depends on the quality of the data base, in particular whether it is representative for the given problem, the given site or the given study period. Therefore, representative and compatible data are the main requirement for neural models.

We expect that with the aid of neural models and specific dimension-reducing, pre-processing methods, bioindication, ecological prediction and the analysis of cause-effect-relations can be improved substantially. Generally, modelling complex non-linear relationships can be handled considerably better with ANNs and supplementary modelling techniques than with classical methods. The best results may be obtained with specific combinations of linear and non-linear techniques.

Acknowledgements

This study was supported by a grant through the German Research Foundation (Deutsche Forschungsgemeinschaft) to D. Borchardt (BO 1012). Numerous assistants and students have contributed for many years in the collection and determination of the insects and the preparation of the data, particularly E. Döring, H. Quast-Fiebig, G. Stüber, E. Turba, B. Landvogt-Piesche. M. Obach assisted in the revision and prepared Fig. 4. T. Horvath provided linguistic advice. All this help is gratefully acknowledged.

References

Aurich, M., 1992 The life-cycle of *Apatania fimbriata* Pictet in the Breitenbach. Hydrobiologia 239, 65–78.

Balls, G.R., Palmer-Brown, D., Sanders, G E., 1996. Investigating microclimatic influences on ozone injury in clover (*Trifolium subterraneum*) using artificial neural networks. N. Phyto. 132, 271–280.

Bayerisches Landesamt für Wasserwirtschaft (Hrsg), 1998. Integrierte ökologische Gewässerbewertung—Inhalte und Möglichkeiten. Münch. Beitr. Abwasser Fisch. Flußbiol. 51, 683.

Borchardt, D., Dapper, T., Schleiter, I., Schmidt, K.-D., Werner, H., Wagner, R , 1997 Modellierung von Wirkungszusammenhangen in Fließgewässern mit Hilfe Neuronaler Netzwerke. Abschlußbericht zum DFG-Forschungsvorhaben We 959/5-2, Universität-Gh Kassel, pp. 141.

Céréghino, R., Lavandier, P., 1998 Influence of hypolimnetic hydropeaking on the distribution and population dynamics of Ephemeroptera in a mountain stream. Freshw Biol. 40, 385–399.

Chon, T.-S., Park, Y S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecol. Model. 90, 69–78.

Colasanti, R.L., 1991. Discussions of the possible use of neural network algorithms in ecological modelling. Comput. Microbiol. 3, 13–15.

Daniell, T.M., Wundke, A.D , 1993. Neural Networks—Assisting in Water Quality Modelling. Watercomp, Melbourne, Australia, March 30–April 1, Internal report, pp 51–57.

Dapper, T., 1998 Dimensionsreduzierende Vorverarbeitungen fur Neuronale Netze mit Anwendungen in der Gewässerökologie. Dissertation im Fachbereich Mathematik/Informatik der Universität Gh Kassel, Berichte aus der Informatik D 34, Shaker Verlag, Aachen, pp. 243

Feminella, J W., Resh, V.H., 1990. Hydrologic influences, disturbance, and intraspecific competition in a stream caddisfly population. Ecology 71, 2083–2094.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. korr. Nachdruck d. 1. Auflage, Addison-Wesley, Reading MA, pp. 412.

Guégan, J.-F , Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity global riverine fish diversity. Nature 391, 382–384.

Kaluli, J.W., Madramootoo, C.A., Djebbar, Y., 1998 Modelling nitrate leaching using neural networks. Water Sci Technol. 38 (7), 127–134.

Lachtermacher, G , Fuller, J D., 1994. Backpropagation in hydrological time series forcasting, in stochastic and statistical methods. In: Hipel, K.W., McLeod, A.I , Panu, U.S., Singh, V.P. (Eds.), Hydrology and Environmental Engineering. Kluwer Academy, Norwell, MA, pp. 229–242.

Lee, H.-L., DeAngelis, D., Koh, H.-L., 1998. Modelling spatial distribution of the unionid mussels and the core-satellite hypothesis. Water Sci. Technol. 38 (7), 73–79.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some environmental variables in trout abundance models using neural networks Aquat. Living Resour. 9, 23–29.

Maier, H.R., 1995a. Use of artificial neural networks for modelling multivariate water quality time series. PhD Thesis. Department of Civil and Environmental Engineering, The University of Adelaide, pp. 464.

Maier, H.R., 1995b. A review of artificial neural networks Research Report No. R131. Department of Civil and Environmental Engineering, The University of Adelaide, Adelaide, pp. 94

Maier, H.R., Dandy, G.C., 1993. Use of Artificial neural networks for forecasting water quality. Reprints, Stochastic and Statistical Methods in Hydrology and Environmental Engineering An International Conference in Honour of Professor T.E. Unny, University of Waterloo, Ont., Canada, June, 21–25, pp. 509–511.

Maier, H R., Dandy, G.C., 1994. Forecasting salinity using neural networks and multivariate time series models. Reprints, Water Down Under 94, Adelaide, South Australia, November 21–25, pp 297–302.

Maier, H.R., Dandy, G C , 1996a. Neural network models for forecasting univariate time series. Neural Netw World 5 (96), 747–771.

Maier, H.R , Dandy, G.C., 1996b. The use of artificial neural networks for the prediction of water quality parameters. Water Resour. Res. 32 (4), 1013–1022

Maier, H.R , Dandy, G.C., Burch, M D., 1998. Use of artificial neural networks for modelling cyanobacteria *Anabena* spp in the River Murray, South Australia. Ecol. Model. 105, 257–272.

Mang, J , Geffers, K., Borchardt, D., 1998. Fallbeispiel Lahn bei Limburg (Hessen)—ein staureguliertes Fließgewässer 2. Ordnung. gwf-Wasser Abfall 139 (7), 408–417.

Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems—a comparison of artificial neural networks and regression models. Ecol. Model. 98, 73–186.

Pham, D.T., Oh, S.J., 1992. A recurrent backpropagation neural network for dynamic system identification. J. Syst. Eng. 2, 213–223.

Recknagel, F , 1997. ANNA—artificial neural network model for predicting species abundance and succession of blue–green algae. Hydrobiologia 349, 47–57.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Recknagel, F., Fukushima, T , Hanazato, T., Takamura, N., Wilson, H., 1998. Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumigaura by artificial neural networks. Lakes Reserv. Res. Manag. 3, 123–133.

Resh, V.H., Hildrew, A.G., Statzner, B., Townsed, C.R., 1994. Theoretical habitat templets, species traits, and species richness: a sysnthesis of long-term ecological research on the Upper Rhône River in the context of concurrently developed ecological theory. Freshw. Biol. 31, 539–554.

Ritter, H., Martinez, T., Schulten, K., 1994. Neuronale Netze: eine Einfuehrung in die Neuroinformatik selbstorganisierender Netzwerke (2 Aufl). Addison-Wesley, Bonn, pp. 325.

Roadknight, C M., Balls, G R., Mills, G.E., Palmer-Brown, D , 1997. Modelling complex environmental data. IEEE Trans Neural Netw. 8, 852–862.

Rumelhart, D.E., Hinton, G.E., Williams, R J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McChelland, J.L. (Eds.), Parallel Distributed Processing, vol. 1. MIT, Cambridge, MA, p. 318.

Schizas, C.N., Pattiches, C.S, Michaelides, S.C., 1994. Forecasting minimum temperature with short time-length data using artificial neural networks Neural Netw. World 4 (2), 219–230.

Specht, D.F., 1991. A general regression network. IEEE Trans Neural Netw. 6, 568–576.

SPSS Inc., 1997. SPSS for Windows, Base system user's guide, Release 7.5 1., Chicago, USA, pp. 628.

Statzner, B , Gore, J.A , Resh, V.C., 1988. Hydraulic stream ecology: observed patterns and potential applications. J N Am. Benthol. Soc 7, 307–360.

Statzner, B , Resh, V.H., Dolédec, S. (Eds.), 1994. Ecology of the upper Rhône River: a test of habitat templet theories. Special issue. Freshw. Biol. 31 (3), 556.

ter Braak, C.J.F., Verdonschot, P F.M., 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat. Sci 57, 255–289.

ter Braak, C.J.F., 1988. CANOCO—a Fortran program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (version 2.1). Agricultural Mathematics Group, Wageningen, The Netherlands, pp. 95

ter Braak, C.J.F., 1990. Update notes, CANOCO version 3.10. Agricultural Mathematics Group, Wageningen, The Netherlands, pp. 35.

Townsend, C.R., 1989. The patch dynamics concept of stream community ecology. J. North Am. Benthol. Soc. 8, 36–50.

Townsend, C.R., Hildrew, A G., 1994. Species traits in relation to a habitat templet for river systems. Freshw. Biol. 31, 265–275.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. The river continuum concept. Can. J. Fish. Aquat. Sci. 37, 130–137.

Wagner, R., Dapper, T., Schmidt, H.-H., 1999. The influence of environmental variables on the abundance of aquatic insects—a comparison of ordination and artificial neural networks Hydrobiologia (in press).

Wagner, R., Schmidt, H.-H Relationship of aquatic insects emergence (*Ephemeroptera, Plecoptera, Trichoptera*) to environmental variables: a 25 years analysis of the breitenbach Freshw. Biol. (in preparation).

Walley, W.J , Fontama, V.N., 1998 Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. Water Res. 32, 613–622.

Ward, J.V., Stanford, J.A., 1979. Ecological factors controlling stream zoobenthos with emphasis on thermal modifications of regulated streams. In: Ward, J.V., Stanford, J.A. (Eds.), The ecology of regulated streams. Plenum, New York, pp. 35–55.

Ward, J.V., Stanford, J.A., 1982. Thermal responses in the evolutionary ecology of aquatic insects. Ann. Rev. Entomol. 27, 97–117.

Wen, C.-G., Lee, C.-S., 1998 A neural network approach to multiobjective optimization for water quality management in a river basin. Water Resour. Res. 34, 427–436

Werner, H., Dapper, T., Schmidt, K.-D., Borchardt, D., Schleiter, I.M., Wagner, R., Schmidt, H.-H., 1999. In: Neural network tools for the analysis of ecological data aquatic systems (In preparation)

Wiersma, F.R., Poel, M., Oudshoff, A M., 1995. The BB neural network rule extraction method. In: Kappen, B., Gielen, S. (Eds.), Proceedings of the 3rd Annual SNN Symposium on Neural Networks. Springer-Verlag, pp. 69–73.

Winkler, U., Voigtlander, G., 1995. Anwendung neuronaler Netze für die Simulation von Prozeßabläufen auf vorhandenen Kläranlagen. Korrespond. Abwasser 10, 1784–1792.

ELSEVIER

Ecological Modelling 120 (1999) 287–297

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Individual-based modelling of fishermen search behaviour with neural networks and reinforcement learning

Michel Jules Dreyfus-León [a,b,*]

[a] *Facultad de Ciencias Marinas, U.A.B.C., Mexico*

[b] *Instituto Nacional de la Pesca, México,Programa Nacional de Aprovechamiento del Atún y Protección del Delfín, Apartado Postal 453, Ensenada, B.C. Mexico*

Abstract

A model to mimic the search behaviour of fishermen is built with two neural networks to cope with two separate decision-making processes in fishing activities. One neural network deals with decisions to stay or move to new fishing grounds and the other is constructed for the purpose of finding prey within the fishing areas. Some similarities with the behaviour of real fishermen are found: concentrated local search once a prey has been located to increase the probability of remaining near a prey patch and the straightforward movement to other fishing grounds. The artificial fisherman prefers areas near the port when conditions in different fishing grounds are similar or when there is high uncertainty in its world. In the latter case a reluctance to navigate to other areas is observed. The artificial fisherman selects areas with higher concentration of prey, even if they are far from the port of departure, unless a high uncertainty is related to the fishing ground. Connected areas are preferred and followed in orderly fashion if a higher catch is expected. The observed behaviour of the artificial fisherman in uncertain scenarios can be described as a risk-averse attitude. The approach seems appropriate for an individual-based modelling of fishery systems, focusing on the learning and adaptive characteristics of fishermen and on interactions that take place at a fine scale.

Keywords: Fishermen; Fleet dynamics; Neural networks; Q learning; Search behaviour; Modelling; Reinforcement learning

1. Introduction

Understanding fleet dynamics helps us to describe and analyze fisheries, as well as the implications that management regulations will have on a fishery due to changes in the behaviour of the fleet or individual vessels adjusting to the new rules. Consequently, it is an important factor to account for in decision making. The spatial allocation of fishing effort is a variable driven by human behaviour, related to the spatial distribution of fish. It is an important component of the dynamic behaviour of fishing fleets which should be a major part of fisheries research (Hilborn, 1985). The knowledge of how fishermen allocate their fishing effort in space is essential to understand how a fishery develops and the relationship between catch rate and abundance (Hilborn and Walters, 1987), and in the formulation of management policy.

* Corresponding author. Fax: +52-561745637.

E-mail address: dreyfus@cicese.mx (M. J. Dreyfus-León)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00109-X

Although many technological advances aid in the finding and catching of fish, open ocean fishermen are fundamentally hunters. The search for fish may occupy a considerable amount of time (Mangel, 1981). Therefore, the search behaviour of fishermen should be incorporated in spatial models of fishery systems in order to understand the implications of management regulations, which are to be applied in the future with biological, social and economic objectives and consequences. Knowledge or understanding of search activities should improve representations of fisheries, since the search is a basic and substantial element of fishing effort. Simulation models, which are necessary because of our scarce knowledge and the uncertainty related to all fishery systems, can help us understand the relationship between fish and fishermen. They give us the opportunity to experiment and assess our knowledge and understanding of the real system and hopefully guide in the never-ending quest to assemble new improved models of this class of complex systems.

There has been some theoretical and applied work on spatial distribution of effort, varying from qualitative anthropological observations to simulation modelling and quantitative studies (Gillis et al., 1993). Those representations tend to model fishing effort in an aggregate manner from the effort or the space perspective (Mangel and Clark, 1983; Mangel and Beder, 1985; Allen and McGlade, 1986; Mangel and Clark, 1986; Hilborn and Walters, 1987; Anganuzzi, 1996). Furthermore those representations establish a priori movement rules for effort distribution between areas, and few works relying on observations of movement dynamics in specific fisheries have yielded some assumptions on behavioural patterns of movement (Hilborn and Ledbetter, 1979; Hilborn and Walters, 1987; Gillis et al., 1993). However, analysis on a fine scale of searching effort by individual vessels has seldom been done (Kleiber and Edwards, 1988; Polacheck, 1988)

In this work the approach is focused at an individual level and at a fine scale of spatial search. The purpose is to assign information to an artificial fisherman, called from now on fishermat, in relation to the term animat that is frequently used in the behaviour-based artificial intelligence field (Maes, 1993). The fishermat is assembled with the sole purpose of learning a spatial search strategy to find fish; It is constructed with the tools to improve its search methods in a simulated world through learning mechanisms. Artificial neural networks inspired by the neuronal structure of the brain have been widely used since the 1980s (Thagard, 1996), and form the basic information-processing system of the fishermat. Many neuronal representations and learning mechanisms exist, although their neurological plausibility is doubtful. They are extremely simple compared to real brain neurons, nevertheless, they have proven to be useful for many purposes. The neural networks' learning abilities let us use a bottom-up approach, which expects the emergence of new behaviours rather than assuming movement patterns.

Learning in neural networks is frequently achieved by supervised learning, from examples provided by a knowledgeable external supervisor (Sutton and Barto, 1998). On the other hand, reinforcement learning appears to be a good strategy to mimic human behaviour when combined with the rewarding actions of neural networks that promote higher fitness (Bonarini, 1997). Reinforcement learning is tantalizing because learning occurs through trial and error experimentation within the environment. Feedback is a scalar payoff, hence no explicit teacher is required, and little or no prior knowledge is needed (Whitehead and Lin, 1995). Reinforcement learning allows the acquisition of adapted behaviours by interacting with the environment. A positive or negative feedback can be given to the fishermat in relation to costs and benefits that drive fisheries activities. Besides, learning is the essential adapting tool for humans, and in complex and variable environments learning allows the possibility of prompt adjustment.

2. The model

2.1. World description and experiment

The world is a 200 × 200-square space, divided into 16 toroidal areas of 50 × 50 pixels (Fig. 1).

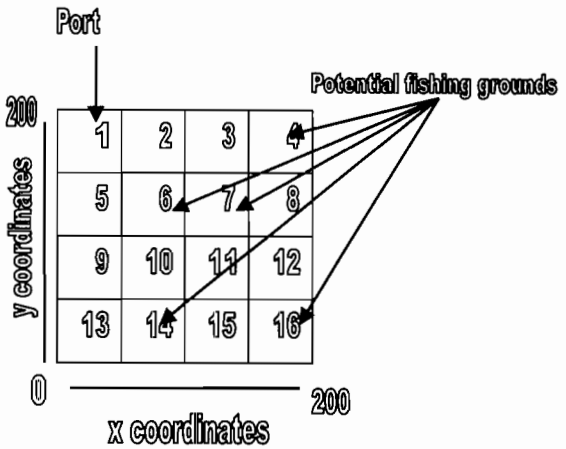

Fig. 1. Fishermat's world.

The reason for the areas to be of toroidal nature is for the fishermat to stay within a particular area while searching, until he decides to change to another area. Five of those toroidal areas are considered potential fishing grounds (areas 4, 6, 7, 14 and 16) in four of the scenarios considered, and the number is reduced to four fishing grounds in the remaining scenarios (Table 1), where the fishermat interacts with the environment seeking for a fishing strategy. The port of departure is in area 1, area 6 is neighbouring it and area 7 is close to the former as well as to area 4. These three areas are considered close to port or at least connected to port. Areas 14 and 16 are at a longer distance from port and are not close to other fishing grounds. Scenarios differ from homogeneous to very uncertain environments in relation to prey presence, taking into account fishing ground locations.

Fish are allocated in patches at random in all the fishing grounds. Up to 2000 fish are allocated in total to each epoch and the prey remain motionless. An epoch consists of 160 movement decisions between areas by the fishermat. When the fishermat is located in the same position as a fish, the prey disappears (catch has occurred) until the next epoch when the initial abundance of fish is redistributed. The fish unrealistically remain motionless. However, this is not a problem for the focus of the study. The fishermat performs 40 search motion steps within an area before deciding to stay or change to another sector to continue its exploration. When the decision to change to a different area is made, movement is directed to the centre of the new chosen zone.

Table 1
Scenarios considered as learning environments for adaptive searching skills by the fishermat

Scenarios	Characteristics
S1	Five similar fishing grounds (same initial amount of prey), in areas 4, 6, 7, 14 and 16
S2	Areas near the port of departure (area 1) have the same initial prey concentration. Areas 14 and 16 have a higher density
S3	In some epochs fish are distributed only in areas close to port and in other epochs they are exclusively present in areas 14 and 16, far from port
S4	In every epoch fish may be distributed or not in any of the five potential fishing grounds
SR20	Area 14 is no longer a fishing ground. Fish are always present in lower density in areas 4, 6 and 7, close to port. Area 16 has a 0.2 probability of being a desert environment in a particular epoch, but has a higher density of prey otherwise
SR40	Area 14 is no longer a fishing ground. Fish are always present in lower density in areas 4, 6 and 7, close to port. Area 16 has a 0.4 probability of being a desert environment in a particular epoch, but has a higher density of prey otherwise
SR60	Area 14 is no longer a fishing ground. Fish are always present in lower density in areas 4, 6 and 7, close to port. Area 16 has a 0.6 probability of being a desert environment in a particular epoch, but has a higher density of prey otherwise
SR80	Area 14 is no longer a fishing ground. Fish are always present in lower density in areas 4, 6 and 7. Area 16 has a 0.8 probability of being a fishing ground in a particular epoch, but has a higher density of prey otherwise

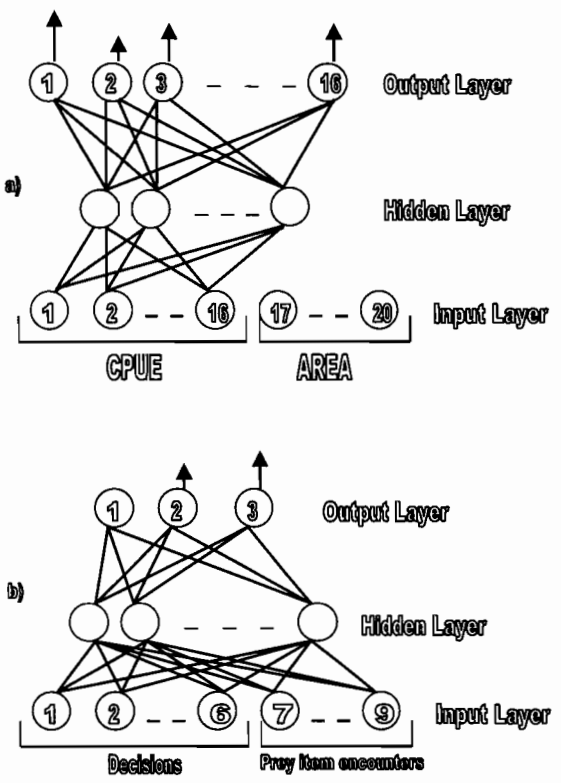

Fig. 2. Fishermat's neural networks structure. (a) NN1; (b) NN2.

2.2. Fishermat description

Two neural networks represent the fishermat, each with a three-layer structure. One neural network (NN1, Fig. 2a) is designed to accomplish the task of learning a strategy of decisions to move or stay in a particular area under the circumstances in which the fishermat is involved. This neural network keeps track of the fishermat's area-location in binary code representation in four neurons, and with 16 other input neurons keeps a knowledge of the relative value of each area, with a measure of catch per unit effort over its past experience. The hidden layer consist of 40 neurons and the output layer has 16 neurons, each representing an area. The highest output defines the next area where the fishermat will search for prey.

The second neural network (NN2, Fig. 2b) has the aim of learning search patterns, confined within an area. NN2 has a short memory of the last three movement decisions triggered by this neural network and each is represented with two bits (01, 10 or 11) that constitute part of the input to NN2 in six neurons. Three more neurons keep a performance success record of prey encounter (0 or 1) in the last three time steps within the local search. The hidden layer consists of ten neurons. In the output layer three neurons represent different types of movement: (a) no change of direction; (b) slight relative change of direction to the right or left from current track; and (c) a more abrupt change in direction. In each case turning right or left is kept as a random event.

Both neural networks represent the whole decision system of the fishermat, receiving information through interaction with its world and modifying the environment with each action (Fig. 3). The neural networks are used alternately since they are not required at the same time, although both networks influence each other's performance. Modelling the fishermat's brain with two separate processing networks was chosen to harmonize with the plausible idea that real fishermen

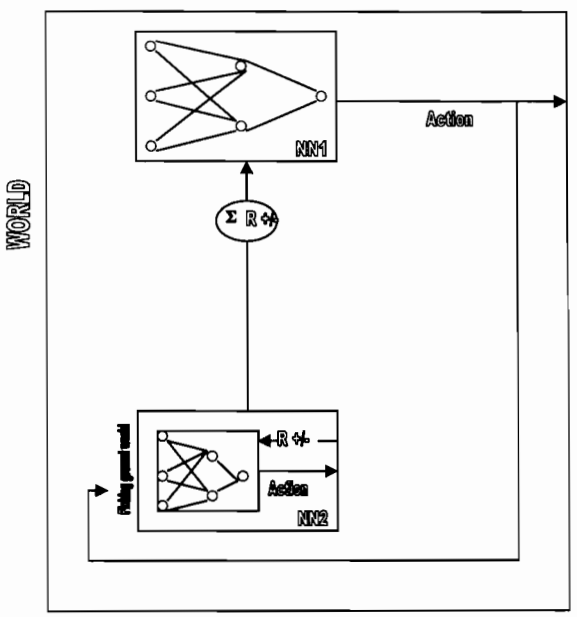

Fig. 3. Fishermat-World model.

get to learn where potential fishing grounds are found, even though they vary in extension. Nevertheless, once in the fishing ground, exact position of prey is unknown. The predator's precise position might not be relevant when a search is done in this space scale, with scarce or no clues to prey whereabouts. Different information and skills are needed to perform at both levels, and two neural networks might improve performance without interfering with each other, and avoid sending confusing and deceiving signals to the networks.

2.3. Learning

Reinforcement learning algorithms solve goal-directed problems by determining a behaviour for the agent that maximizes its total positive and negative rewards. Reinforcement learning algorithms specify such behaviours as state-action rules, called policies. Q-learning is used as a reinforcement learning technique, which seems appropriate because the agent experiences different actions in different circumstances and in the process of learning, predicts the reward or penalty of future actions in a different state of affairs (Watkins and Dayan, 1992). It can also learn from raw experience and without a model of the environment's dynamics (Sutton and Barto, 1998). The aim is to maximize the discounted cumulative reward (Whitehead, 1991). The objective of learning is to predict which action maximizes the agent's performance (Lin, 1991). Output neurons should signal an estimated utility function Q(s, a), over states s and actions a. Each output neuron, representing a particular action or policy to perform, signals the expected Q value or utility for that action-state pair. The utility is a numeric value, which is the predicted future reward that will be achieved for that action-state pair (Maclin and Shavlik, 1996). The neuron with the highest expected reward triggers the specific behaviour that it represents. This means that each output neuron predicts the reward to be obtained by performing the action it represents under the world status. To avoid premature convergence (Mahadevan, 1994), random exploration of the environment is allowed 10% of the time for the first 40 epochs, instead of the action with maximum Q. This diminishes the probability of finding a local maximum. After each action the corresponding utility estimate is adjusted by calculating the difference between the output from the winning neuron and the real utility of the action in a particular state with the following algorithm, from (Sutton and Barto, 1998):Initialize $Q(s, a)$ arbitrarily by randomly choosing the initial connection weightsRepeat for each epoch:

- Initialize the state (condition) s of the world
- Choose action a for state s using policy derived from Q, but allow some random exploration of actions.
- Repeat for each step of the epoch:

  (1) Take action a, observe r (reward), s' (next state)

  (2) $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max a' Q(s', a') - Q(s, a)]$;

  (where α is the learning rate and γ is the discount rate)

  (3) s $\leftarrow s'$;

- Until the final epoch

The difference between the real and predicted utility of the action taken is used with standard backpropagation as a measure of error. Standard backpropagation training is a gradient descent method to minimize the total square error of the output computed by the net (Fausett, 1994) and during training distributes the error between prediction and performance to all the connections between neurons involved in that decision.

Positive or negative rewards for NN2 appear at each time step it performs, with or without prey item encounter. Reward for NN1 is related to the cumulative reward achieved by NN2 after the last NN1 decision (Fig. 3) and decisions to navigate to a different area include costs proportional to distance travelled. Neuron output is scaled between -1 and $+1$, using the binary sigmoid function, and costs and benefits to be gathered from movement decisions are adjusted so as not to pass these values. All the neural networks and simulations are programmed in C++.

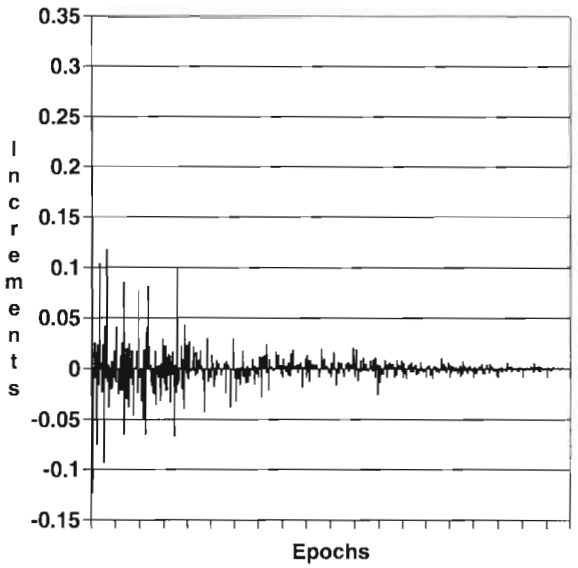

Fig. 4. Variations in weight connections during the learning process.

3. Results

After setting initial values to the neural networks connections, the learning process begins. Differences between old and new weights are higher at the beginning of the learning process and exponentially decrease over the experiment, as shown (Fig. 4) by several connections between the hidden layer and output neurons in NN1. The strategies to cope with this artificial fishing space vary with the scenarios. In relation to area preference (Fig. 5), in S1 the fishermat prefers to search for prey in areas 4, 6 and 7 in orderly fashion, and avoids any exploration in zones 14 and 16. In S2, denser prey areas 14 and 16 are visited, with a preference for the latter, while areas 4 and 7 are neglected. When good quality habitat for fishing shifts back and forward from areas 4, 6 and 7 to areas 14 and 16, preference is for the foremost, but there is exploration and exploitation of the distant areas. When all the fishing grounds randomly change between a desert-type environment and an inhabited one, the fishermat explores all the five zones, with preference for those closer to port, i.e. area 6 and area 7 neighbouring it (Fig. 5a). In the SR20 scenario, the distant fishing ground (area 16) has a high concentration of prey and a 0.2 probability of being uninhabited and is preferred for fishing activities. When this probability increases area 16 ceases to be an option for the fishermat (Fig. 5b).

With respect to the conditions that trigger a decision to change to another fishing ground, negative rewards are highly associated with exploration when the environment is more uncertain (Fig. 6), as shown by scenarios S1 to S4 and SR20 to SR80. In the yellow fin tuna fishery in the eastern Pacific Ocean, the decision to either continue searching in a particular area or navigate after having explored is made in relation to the catch per day. In purse seine vessels of 1000 tons of carrying capacity of the Mexican fleet in 1997, decisions to continue exploiting a particular area

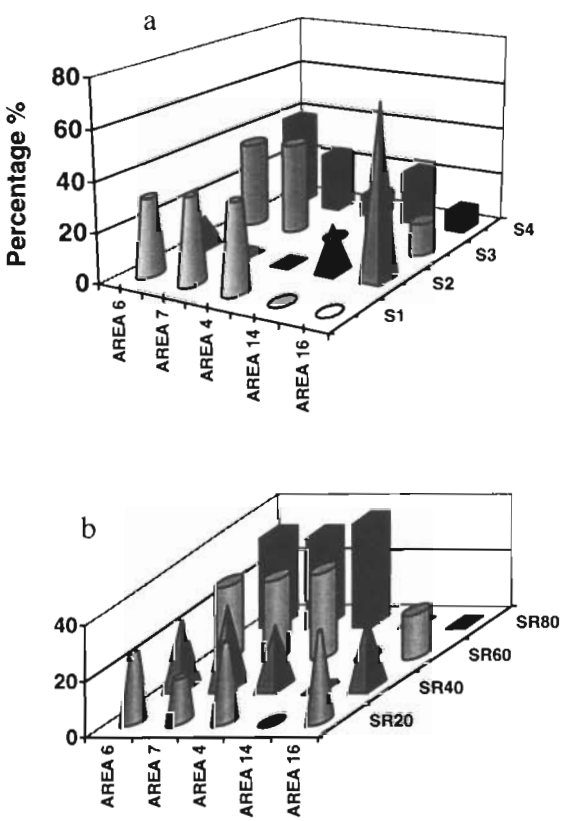

Fig. 5. (a) Search time per fishing ground in four scenarios with five fishing grounds, S1, S2, S3 and S4. (b) search time per fishing ground in the scenarios with gradual increase in uncertainty, SR20, SR40, SR60, SR80.

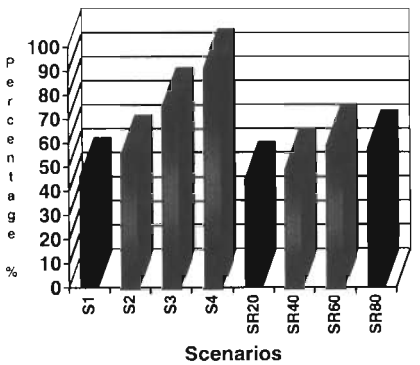

Fig. 6. Percent of decisions to change to another fishing ground in relation to negative rewards (r).

came after having higher catches (a mean of 33 tons per day), whereas decisions to navigate after a day of searching appeared after lower success (with a mean of 14 tons per day). Since the frequency distribution of catch per day for each policy is skewed toward low values (Fig. 7), the median is more representative and the contrast is large (from 25 tons to 1 ton, respectively). Figs. 8 and 9 show the frequency distribution of decisions to navigate in relation to the relative level of catch per unit time that the fishermat senses and the frequency distribution for searching decisions in all scenarios, respectively. All frequency distributions for shifting area determinations are skewed toward low catch per day values. Judgement to keep searching in the same area is biased in particular towards high values of catch per unit time, in especially for S2, S3, SR20, SR40. Skewness toward low values is also part of the distributions, particularly for S4 and SR80.

Part of the NN2 strategy can be seen comparing the degree of movement direction change after finding a prey item (Fig. 10). When no catch is the most recent event, maintaining the same track is preferred around 80% of the time. When catch is the most recent experience, both low and high degree of shift in direction are experienced more frequently, and in the same proportion. One way to holistically understand the search patterns of the fishermat is with a graphic description of its course track. Those tracks can be compared with real vessel paths, in this case from the tuna fishery (Fig. 11). Coordinates and the continental contour are eliminated to keep the information private and make a better comparison between them.

4. Discussion

The fishermat seems to cope with the world, avoiding risky or uncertain decisions. In the ho-

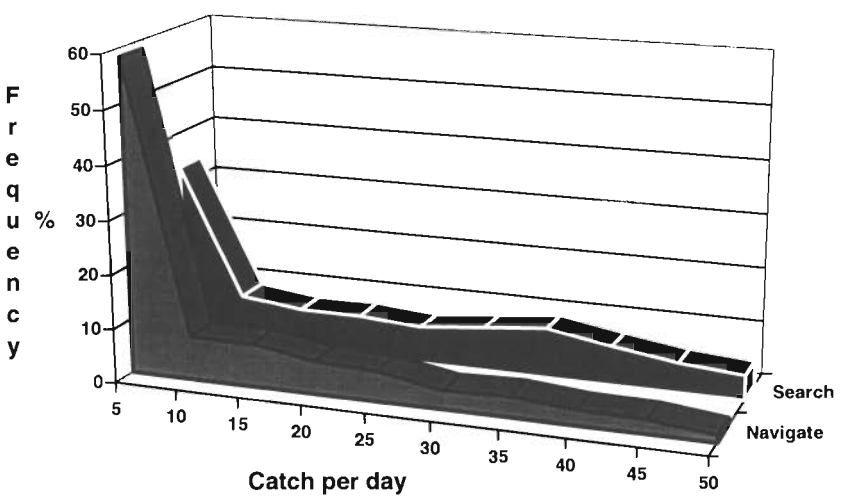

Fig. 7. Decisions to keep searching or navigate to a different area in the tuna fishery, in relation to catch after a day of searching activity.

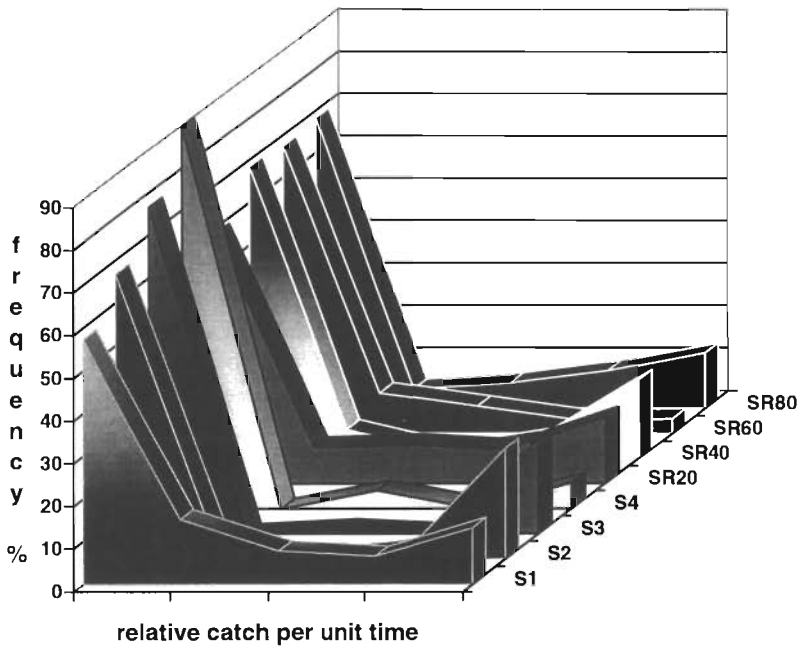

Fig. 8. Decisions made by the Fishermat to navigate in all scenarios.

mogeneous S1 scenario, near-port fishing grounds were chosen instead of the distant areas, which were excluded from further exploration. In S2, basically area 6 was adopted as an intermediate fishing ground towards the high concentration areas where it resolved to stay longer time intervals. In both scenarios, as expected, the fishermat chose the most profitable areas, thus eliminating the cost of long-distance travel. When some uncertainty was incorporated into the S3, S4, SR20, SR40, SR60 and SR80 scenarios, a risk-averse attitude can be interpreted from the results. Under risky situations the fishermat sometimes preferred areas close to port and avoided as much as possible movement between areas. The behaviours described are linked to the particular structure of the scenarios as well as to costs and benefits.

The fishermat was built with some information about its surrounding world and its performance within it. The type of information that was explicitly incorporated for neural network processing appears to be logical information that would be analyzed by real fishermen. The relative value of the fishing grounds was incorporated as part of the knowledge of the fishermat. This has been discussed in the analysis of the spatial distribution of effort in fisheries (Hilborn and Ledbetter, 1979; Abrahams and Healey, 1990; Gillis et al., 1993) and of foragers getting information of potential forage patches (Clark and Mangel, 1984). The artificial fisherman was banned from having complete information of the world, similar to real life circumstances that fishermen deal with. Nevertheless, the neural networks incorporated from their experience some knowledge and adopted behaviours related to situations that they did not explicitly know. Neural networks go through a classification process, which is incorporated into their structure. Experience is locked within the connection weights, through experience in the scenarios. This can be seen when decisions do not seem to relate to the particular input that they use as inquiry for the next determination, i.e. expecting higher payoffs after a bad experience in an area, the fishermat sometimes decides to stay.

Similar outcomes showed up between real and artificial fishery decisions. Frequency distributions of decisions to navigate in all scenarios are shifted

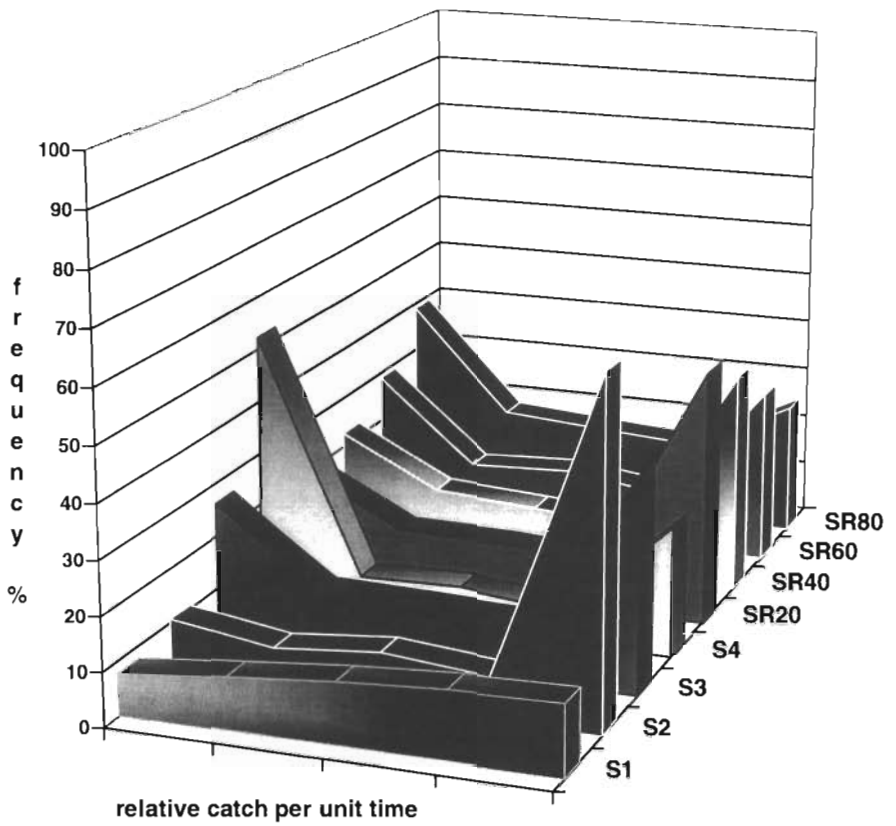

Fig. 9. Decisions made by the Fishermat to keep searching in all scenarios.

to the relative low catch per time values compared to those found in searching decisions. Scenario S4, with high uncertainty incorporated into all the fishing grounds, has frequency distributions for navigation and searching resolutions that match the ones found in the tuna fishery in particular. In both cases, reliable decisions do not always result in high rewards. This is caused by the uncertainty related to the dynamic system and the fact that fish location remains unknown until detected by the crew or technological equipment. Fishermen sometimes have clues with respect to prey whereabouts but they remain as tools only for enhancing the probability of finding fish.

The local search strategy that emerged in NN2 is one that has been widely described with predators in patchy environments. A higher sinuosity of movement once a prey item has been found might improve the possibility of finding another one (Pyke, 1978). Since a predator or a fisherman does not know the extension of a patch or their relative position within it, this strategy improves the chances of remaining in or near the patch. In the same manner moving straightforward in higher proportion will certainly drive a predator toward

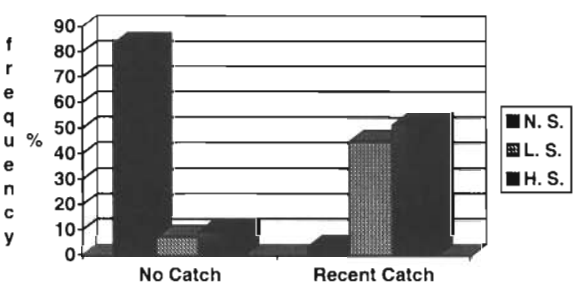

Fig. 10. Degree of movement sinuosity in relation to prey item encounter success in the last time step.

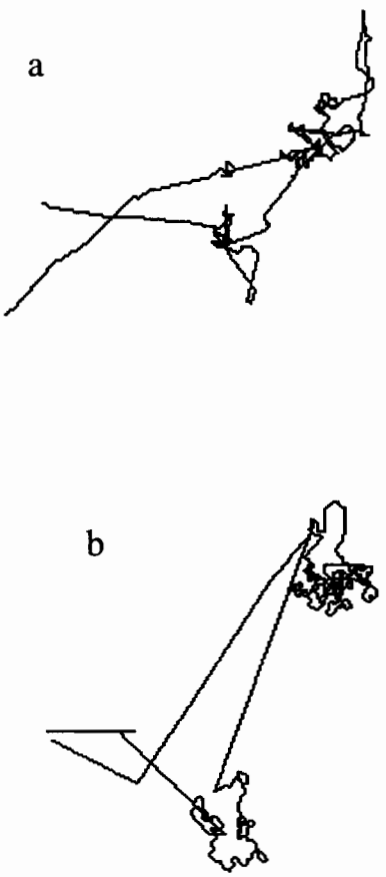

Fig. 11. (a) Real track of a fishing vessel; (b) track of a fishing trip made by the fishermat.

new areas. The fishermat's local position is not sensed by NN2 because it does not seem relevant due to the above. Lack of knowledge of areas of high prey concentration and of vessel's position within them seems to be the pattern in pelagic fisheries, causing search to be described as a random walk (Clark and Mangel, 1979; Mangel and Clark, 1986). Also in NN2 a short memory of failure or success in movement decisions was incorporated which can be viewed as a success variable comparable to a perception of fish density (Kleiber and Edwards, 1988).

The whole movement path left by the fishermat resembles real vessel tracks and could be used as a Turing test to determine whether the fishermat's behaviour is intelligent if a human observer can not differentiate between the real and the artificial behaviours. Apart from that test, similarities exist between the tuna vessel and the fishermat's track, i.e. more or less straightforward paths followed by some extensive local search paths. The result of the latter is that a greater distance is searched within clusters of prey than outside them, as described in tuna purse seine cruises (Polacheck, 1988). From another perspective of the modelling framework, the use of two or more neural networks to cope with different but related processes seems a reasonable scheme to follow, and in this way to avoid sending misleading signals to the neural networks.

Furthermore, since discovery and exploitation are key elements in hunting and fishing and involve aspects of adaptiveness, creativity and learning (Allen and McGlade, 1986), there is a potential benefit of applying neural network methodology as well as reinforcement learning techniques. They permit a more flexible and dynamic modelling due to the emphasis on learning and adaptation instead of the fixed behaviours proposed in mathematical representations. Incorporating decision making as an internal process into the model results in very different and more complex model behaviour (Smith et al., 1982). To use this framework in a specific fishery, rewards should be proportional to economic costs and benefits, which are the main driving forces of the activity.

The approach used in this model is intended to focus on individual components of the system as well as in a local spatial scale, neglected in traditional fishery models, instead of on higher hierarchical levels. Traditional models assume that all individuals are the same and interact homogeneously within the system (Kawata and Toquenaga, 1994). This model can be extended to represent not only an individual but also fleets, in an individual-based manner, with the emergence of new behaviours to cope with the interaction among vessels. In this manner complex spatio-temporal patterns can appear. This is a shift in the way modelling is done in ecology (Judson, 1994) and should incorporate more dynamics into the system being represented.

Acknowledgements

The author is grateful to the Consejo Nacional de Ciencia y Tecnología of México and the Centro de Investigación Científica y de Educación Superior de Ensenada for their support. Thanks to Dr Pierre Kleiber and the reviewers for their valuable comments and to Dr Kim Murphy and Christine Harris for their help editing this document.

References

Abrahams, M.V., Healey, M.C., 1990. Variation in the competitive abilities of fishermen and its influence on the spatial distribution of the British Columbia salmon troll fleet. Can. J. Fish. Aquat. Sci. 47, 1116–1121.

Allen, P.M., McGlade, J.M., 1986. Dynamics of discovery and exploitation: the case of the scotian shelf groundfish fisheries. Can. J. Fish. Aquat. Sci. 43, 1187–1200.

Anganuzzi, A., 1996. An aggregate model of effort distribution. In: Status of Interaction of Pacific Tuna Fisheries in 1995. Proceedings of the second FAO Expert Consultation on Interaction of Pacific Tuna Fisheries. Shimizu, Japan, 23–31 January 1995. FAO Fisheries Technical Paper 365, Rome, FAO, 1996, pp. 612.

Bonarini, A., 1997. Anytime algorithms. Adapt. Behav. 5 (3/4), 281–315.

Clark, C.W., Mangel, M., 1979. Aggregation and fishery dynamics: a theoretical study of schooling and the purse seine tuna fisheries. Fish. Bull. 77 (2), 317–337.

Clark, C.W., Mangel, M., 1984. Foraging and flocking strategies: information in an uncertain environment. Am. Nat. 123, 626–641.

Fausett, L., 1994. Fundamentals of Neural Networks. Prentice Hall, Englewood Cliffs, NJ, p. 460.

Gillis, D.M., Peterman, R.M., Tyler, A.V., 1993. Movement dynamics in a fishery: application of the ideal free distribution to spatial allocation of effort. Can J. Fish. Aquat. Sci. 50, 323–333.

Hilborn, R., Walters, C.J., 1987. A general model for simulation of stock and fleet dynamics in spatially heterogeneous fisheries. Can. J. Fish. Aquat. Sci. 44, 1366–1369.

Hilborn, R., Ledbetter, M., 1979. Analysis of the British Columbia salmon purse-seine fleet: dynamics of movement. J. Fish. Res. Board Can. 36, 384–391.

Hilborn, R., 1985. Fleet dynamics and individual variation: why some people catch more fish than others. Can. J. Fish. Aquat. Sci. 42, 2–13.

Judson, O.P., 1994. The rise of the individual-based model in ecology. TREE 9 (1), 9–14.

Kawata, M., Toquenaga, Y., 1994. From artificial individuals to global patterns. TREE 9 (11), 417–421

Kleiber, P., Edwards, E.F., 1988 A model of tuna vessel and dolphin school movement in the eastern tropical Pacific. technical description of the model US NMFS, SWFC Admin. Report LJ-88-28, pp. 18

Lin, L J., 1991. Self-improving reactive agents: case studies of reinforcement learning frameworks. In: Meyer J.A., Wilson, S.W. (Eds.), From Animals to Animats, Proceedings of the First International Conference on Simulation of Adaptive Behavior, pp. 297–305.

Maclin, R., Shavlik, J.W., 1996. Creating advice-taking reinforcement learners. Mach Learn. 22, 251–281.

Maes, P., 1993. Behavior-based artificial intelligence. In: From Animals to Animats 2. Proceedings of the Second International Conference on Simulation of Adaptive Behaviour, pp. 2–10.

Mahadevan, S., 1994. To discount or not to discount in reinforcement learning: a case study comparing R learning and Q learning. In: Cohen W W., Hirsh H. (Eds.), Machine Learning. Proceedings of the Eleventh International Conference, pp. 164–172.

Mangel, M., Clark, C.W., 1986. Search theory in natural resource modelling. Nat. Res. Mod. 1 (1), 3–54.

Mangel, M., Clark, C.W., 1983. Uncertainty, search, and information in fisheries. J. Cons. Int. Explor. Mer. 41, 93–103.

Mangel, M., Beder, J.H., 1985. Search and stock depletion. Can. J Fish. Aquat. Sci. 42, 150–163.

Mangel, M., 1981. Search effort and catch rates in fisheries. Eur. J. Operat. Res. 11, 361–366.

Polacheck, T., 1988. Analyses of the relationship between the distribution of searching effort, tuna catches, and dolphin sightings within individual purse seine cruises. Fish. Bull. 86 (2), 351–366.

Pyke, G.H., 1978. Are animals efficient harvesters? Anim. Behav. 26, 241–250.

Smith, C.L., Stander, J.M., Tyler, A.V., 1982. Human behaviour incorporation into ecological computer simulations. Environ. Manag. 6 (3), 251–260.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. An introduction. MIT, Cambridge, MA, p. 322.

Thagard, P., 1996. Mind. Introduction to Cognitive Science. MIT, Cambridge, MA, p. 213.

Watkins, C.J.C.H., Dayan, P., 1992. Technical note Q-learning. Mach. Learn. 8, 279–292.

Whitehead, S.D., Lin, L.J., 1995. Reinforcement learning of non-Markov decision processes. Artif. Intell. 73, 271–306.

Whitehead, S.D., 1991. Complexity and cooperation in Q-Learning. In: Proceedings of the Eighth International Workshop on Machine Learning. San Mateo, CA, pp. 363–367.

ELSEVIER

Ecological Modelling 120 (1999) 299–311

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake

Sébastien Brosse [a,*], Jean-François Guegan [b], Jean-Nöel Tourenq [a], Sovan Lek [a]

[a] *CNRS, UMR 5576 CESAC, Université Paul Sabatier, 118 Route de Narbonne 31062 Toulouse cedex, France*
[b] *Centre de recherche IRD de Montpellier, CEPM/UMR CNRS-IRD 9926, BP 5045, 34032 Montpellier cedex 1, France*

Abstract

The present work describes a comparison of the ability of multiple linear regression (MLR) and artificial neural networks (ANN) to predict fish spatial occupancy and abundance in a mesotrophic reservoir. Models were run and tested with 306 observations obtained by the sampling point abundance method using electrofishing. For each of the 306 samples, the relationships between physical parameters and the abundance and spatial occupancy of various fish species were studied. For the 15 fish species occurring in the lake, six main fish populations were retained to perform comparisons between ANN and MLR models. Each of the six MLR and ANN models had eight independent environmental variables (i.e. depth, distance from the bank, slope of the bottom, flooded vegetation cover, percentage of boulders, percentage of pebbles, percentage of gravel and percentage of mud) and one dependent variable (fish density for the considered population). To determine the population assemblage, principal component analysis (PCA) was performed on the partial coefficients of the MLR and on the relative contribution of each independent variable of ANN models (determined using Garson's algorithm). The results stress that ANN are more suitable for predicting fish abundance at the population scale than MLR. In the same way, a higher level of ecological complexity, i.e. community scale, was reliably obtained by ANN whereas MLR presented serious shortcomings. These results show that ANN are an appropriate tool for predicting population assemblage in ecology.

Keywords: Artificial neural networks; Multiple linear regression; Principal component analysis; Population assemblage; Fish ecology; Lake

1. Introduction

Interactions between organisms and their biotic and abiotic environmental characteristics strongly influence the habitat use, the spatial occupancy of species, the proportion of each species within the community and, thus, the community composition and structure (Schoener, 1989; Eklöv, 1997). Modelling and simulation are useful tools to roughly mimic the ecosystem structuration and functioning but their ability to model individual

* Correponding author. Fax: +33-5-61556096.
E-mail address: brosse@cict.fr (S. Brosse)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00110-6

distribution, populations and ecosystems depends on the available modelling techniques and computing power (Giske et al., 1998). For example, Ricker (1975) used correlation analysis to assess the influence of the environment on recruitment using abundance data. Canonical correspondence analysis (ter Braak and Verdonschot, 1995) and multiple least-square regression (Binns and Eiserman, 1979) have frequently been used as qualitative methods to explore the relationships between biological assemblages of species and their habitat preferences. The MLR method is now a statistical tool which is used in routine in ecology, but it suffers from some drawbacks in that the relationships between variables in environmental sciences are often non-linear (James and McCulloch, 1990), while the method used is based on linear principles. Transformation of non-linear variables by logarithmic, power or exponential functions can appreciably improve the results, but have often failed to fit data (Lek et al., 1996b). The artificial neural network (ANN), with the error back-propagation procedure, is at the origin of an interesting approach comparable with regression analysis, but particularly efficient for non-linear data (Rumelhart et al., 1986). Up to now, ANN have been used in ecology for modelling phytoplankton production (Scardi, 1996), fish species richness prediction (Guegan et al., 1998), and prediction of density and biomass of various fish populations (Baran et al., 1996; Lek et al., 1996a,b; Mastrorillo et al., 1997). Nevertheless, ANN have scarcely been applied at the community scale, and the work of Tan and Smeins (1996) is probably the only study at this scale which used ANN performance to predict grassland community changes. Moreover, their work only predicted the density of each species taken one by one, and did not deal with the existence of interactions between species.

The aim of the present study is to model the spatial distribution and abundance of six fish populations according to measurable environmental characteristics. Here, we use two distinct modelling methods and we compare their respective capacities to fit observed patterns: (1) multiple linear regression (MLR); (2) artificial neural networks (ANN). Then we quantified the influence of the eight environmental variables on the spatial distribution and habitat use of each population, leading to an approach of the spatial assemblage of the six fish populations studied.

2. Materials and methods

2.1. Study site and sampling

Lake Pareloup is located in the southwest of France, near the city of Rodez. It covers a total surface area of 1350 ha for a volume of about 168 10^6 m^3. The maximum depth is 37 m and the average depth is 12.5 m. It is a warm monomictic lake, which therefore undergoes a summer thermal stratification, with a low oxygen content below the thermocline (located at about 10 m depth from early June to mid-September) preventing the fish from colonising deep water during this period. Fish sampling was performed weekly from late June to late August in a restricted littoral zone of the lake providing a wide range of topographical characteristics. Point abundance sampling by electrofishing (Nelva et al., 1979) modified for young fish (Copp, 1989) was employed to evaluate the microhabitat of the main fish populations. Each week, 30–40 sampling points were investigated in the same area of the lake. For each of the resulting 306 sampling points, nine habitat variables were taken into account: distance from the bank (DIS) in metres, depth (DEP) in metres, local slope of the bottom at each sampling point (SLO) expressed in four classes from zero (nil slope) to three (sheer slope), percentage of flooded vegetation cover (VEG) and percentages of five substrata: boulders (BOU), pebbles (PEB), gravel (GRA), sand (SAN) and mud (MUD). Fishes collected were preserved in 4% formaldehyde solution. In the laboratory, 0+ roach (*Rutilus rutilus*, L. 1758), 0+ perch (*Perca fluviatilis*, L. 1758), 0+ rudd (*Scardinius erythrophthalamus*, L. 1758), 0+ gudgeon (*Gobio gobio*, L. 1758), 0+ pike (*Esox lucius*, L. 1758) and adult perch were identified and numbered for each sampling point.

2.2. Modelling techniques

Modelling was carried out after $\log_{10}(x+1)$ transformation of the dependent variables. This transformation was applied to avoid an undue influence of outliers on the models (ter Braak and Looman, 1995). The Pearson correlation matrix showed a strong correlation between SAN and MUD ($r = -0.98$) and therefore, the variable SAN was removed from the data matrix in order to deal with colinearity. MLR and ANN models were set up using the same dataset (i.e. 306 samples × (eight environmental variables + six fish populations)) with the aim of comparing the two methods.

For MLR, models were set up using all the variables simultaneously. Calculations were done using SPSS software (Norusis, 1993). For 0+ pike, which is a top-predator fish with low density, we considered its absence (coded 0) and presence (coded 1). To process these categorical variables, a logistic regression was used to model 0+ pike distribution. For each of the six models, final values of the partial standardised regression coefficients of MLR were retained to define the influence of environmental factors at the population scale. Then, they were used to perform principal component analysis (PCA) in order to assess the spatial occupancy of fish populations within the entire fish assemblage.

For ANN modelling, a multilayer feed-forward neural network was used. The processing elements in the network, called neurons, are arranged in a layered structure. The first layer, called the input layer, connects with the input variables. In our case, it comprises eight input neurons corresponding to the eight environmental variables, respectively. The last layer, called the output layer, comprises a single neuron which corresponds to the dependent variable to be predicted (fish density for the population considered) (Fig. 1). The layer between input and output layers is called the hidden layer. We could have used a single neural network with six output neurons (one for each of the six fish populations), but we preferred to use six networks with the same architecture, each one predicting the abundance of one fish population, as to easily extract from the models the influence of the eight environmental variables on each fish population. The network configuration is approached empirically by testing various possibilities and selecting the solution that provides the best compromise between bias and variance (Geman et al., 1992; Kohavi, 1995). Training the network consists of using a training data set to adjust the connection weights in order to minimise the error between observed and predicted values. This training was performed according to the back-propagation algorithm (Rumelhart et al., 1986). The computational program was written in a Matlab® environment and computed with an Intel Pentium® processor.

The modelling was carried out in two steps: first, model training was performed using the whole data matrix. This step was used to estimate the performance of the ANN to learn data. Second, we used the 'leave-one-out' bootstrap cross-validation test (Efron, 1983; Efron and Tibshirani, 1995), where each sample is left out of the model formulation in turn and predicted once,

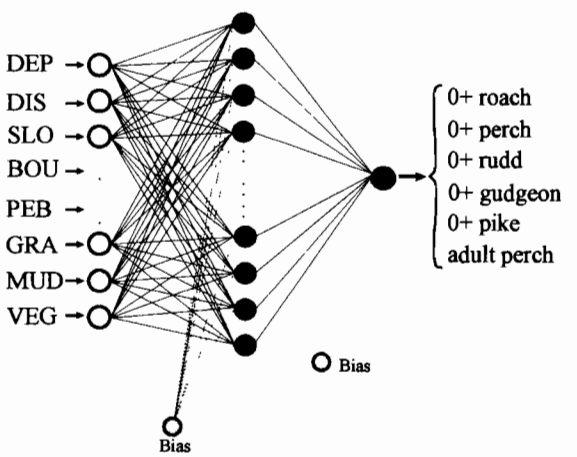

Fig. 1. Typical three-layered feed-forward artificial neural network. Eight input neurons corresponding to eight independent environmental variables (DEP = depth, SLO = slope, DIS = distance from the bank, BOU = boulders, PEB = pebbles, GRA = gravel, MUD = mud, VEG = flooded vegetation), ten hidden layer neurons and one output neuron for estimating one fish population density. Each of the six fish populations was predicted in turn. Connections between neurons are shown by solid lines: they are associated to synaptic weights that are adjusted during the training procedure. The bias neurons are also shown; their input value is one.

to validate the models. This procedure is appropriate when the amount of data is quite small and/or when each sample is likely to have 'unique information' (Efron and Tibshirani, 1995; Kohavi, 1995). This step allows the prediction capabilities of the network to be assessed.

One disadvantage of ANN is their lack of explanatory power. Classical analyses, like MLR, can identify the contribution each independent variable (i.e. input) has on the dependent variable (i.e. output) and can also give some measures of confidence about the estimated coefficients. On the other hand, currently, there is no theoretical or practical way of accurately interpreting the weights attributed in ANN. For example, weights cannot be interpreted as regression coefficients. Therefore, ANN are generally better suited for forecasting or prediction than for explanatory analysis. Some authors have proposed methods for interpreting neural network connection weights to illustrate the importance of explanatory variables in the ANN (Garson, 1991; Dimopoulos et al., 1995; Goh, 1995; Lek et al., 1996a,b). These studies have demonstrated the potential of ANN approaches to explain non-linear interactions between variables in complex systems, and have proposed a procedure for partitioning the connection weights to determine the relative importance of the various input variables. In the present work, Garson's algorithm (Garson, 1991), modified by Goh (1995), was used to determine the influence of the environmental variables. Ten models were set up for each of the six fish populations studied. Then the influence of environmental variables was defined for the ten models and used to assess the spatial distribution of the six populations within the entire community using PCA. In this case, each model was considered as a statistical unit. Thus, PCA was performed on a data matrix containing 60 units (ten units per population × six populations) and the eight environmental variables. Finally, to separate fish population spatial occupancy within the community, cluster analysis was performed on the PCA results using the coordinates of the 60 units on the first two PCA axes.

3. Results and discussion

3.1. Performance of the models

3.1.1. Multiple linear regression models

Examination of Fig. 2 shows some pitfalls which may exist when developing MLR models. Two of the six models were not significant to fit the relationships between fish density and the eight environmental variables: 0+ pike ($r = 0.15$, $P = 0.54$) and adult perch ($r = 0.19$, $P = 0.22$). In both these models, the predicted values showed only nil or close-to-nil values (except one point for adult perch) (see Fig. 2). Overall, we obtained 94% of correct performance estimated using a performance index (PI), based on the proportion of responses within plus or minus 10% of the actual value, but samples with fish were never well-predicted. For the four significant models, correlation coefficients were quite low. Only two models gave a correlation coefficient higher than 0.5 (0.59 for 0+ rudd and 0.70 for 0+ gudgeon), furthermore, the best of these two coefficients was biased as this high value was due to only one non-nil sample well-predicted. Moreover, for the six models, most of the high values of fish abundance were always underestimated and some low predicted values were aberrant, i.e. negative fish densities. The points were not well-distributed along the line of perfect prediction (coordinates 1:1). The residuals tended to increase with estimated values, and their distribution was far from normal. To determine the optimal predictive capacity of traditional methods, we used a non-parametric regression technique: generalized additive models (GAM) (Hastie and Tibshirani, 1990), using the locally-weighted smoother of Cleveland (1979) currently called 'lowess', were set up for the six populations. With this method, the six models were significant ($P < 0.01$) and we obtained a clear improvement of the correlation coefficients: $r = 0.54$ for 0+ roach, $r = 0.38$ for 0+ perch, $r = 0.74$ for 0+ rudd, $r = 0.74$ for 0+ gudgeon, $r = 0.27$ for 0+ pike and $r = 0.37$ for adult perch. These improvements of the quality of the model's predictions testifies to the non-linear behaviour of the relationships between dependent (i.e. fish populations) and independent (i.e. envi-

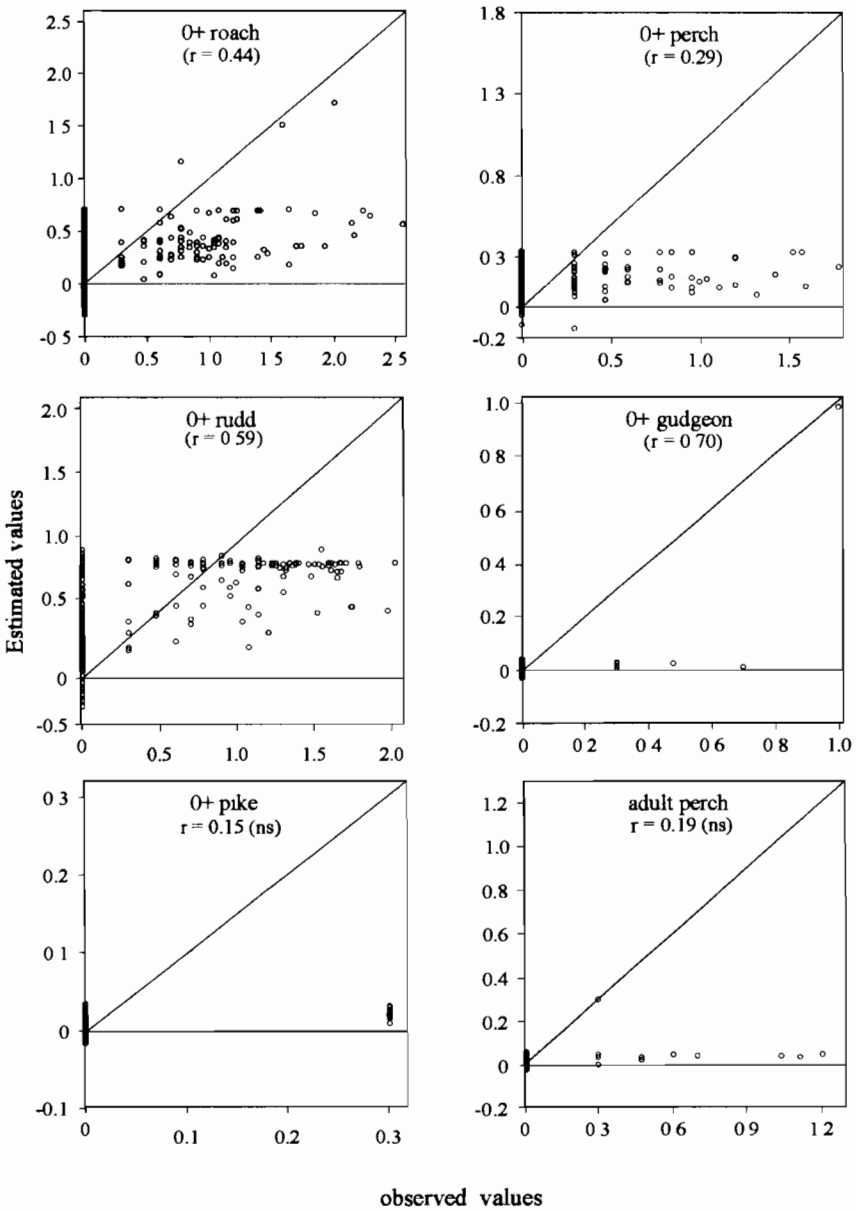

Fig. 2. Recognition performance of the MLR models for the six fish populations. Scatter plots of predicted values vs observed values. The solid line indicates the perfect fit line of prediction (coordinates 1:1)

ronmental variables) variables. In addition, it justifies the use of ANN, which are known to be able to deal with non-linear relationships between dependent and independent variables when compared with classical MLR methods.

3.1.2. Artificial neural network models

The ANN structure used was a three-layered ($8 \rightarrow 10 \rightarrow 1$) feed-forward network with bias (Fig. 1). There were eight input neurons to code the eight different independent variables. The hidden

layer had ten neurons, determined as the optimal configuration giving the lowest error in the training and testing sets of data with minimal computing time (Geman et al., 1992; Lek et al., 1996b,c). The output neuron computed the value of the dependent variable (fish density). We thus had a total of 101 parameters: (eight input neurons × ten hidden neurons) + (ten hidden neurons × one output neuron) + 11 bias parameters.

The ANN with back-propagation gave much higher correlation coefficients between observed and predicted values (Fig. 3) than MLR. Fig. 3 shows that both low and high values of fish densities were well-predicted even for scarce pop-

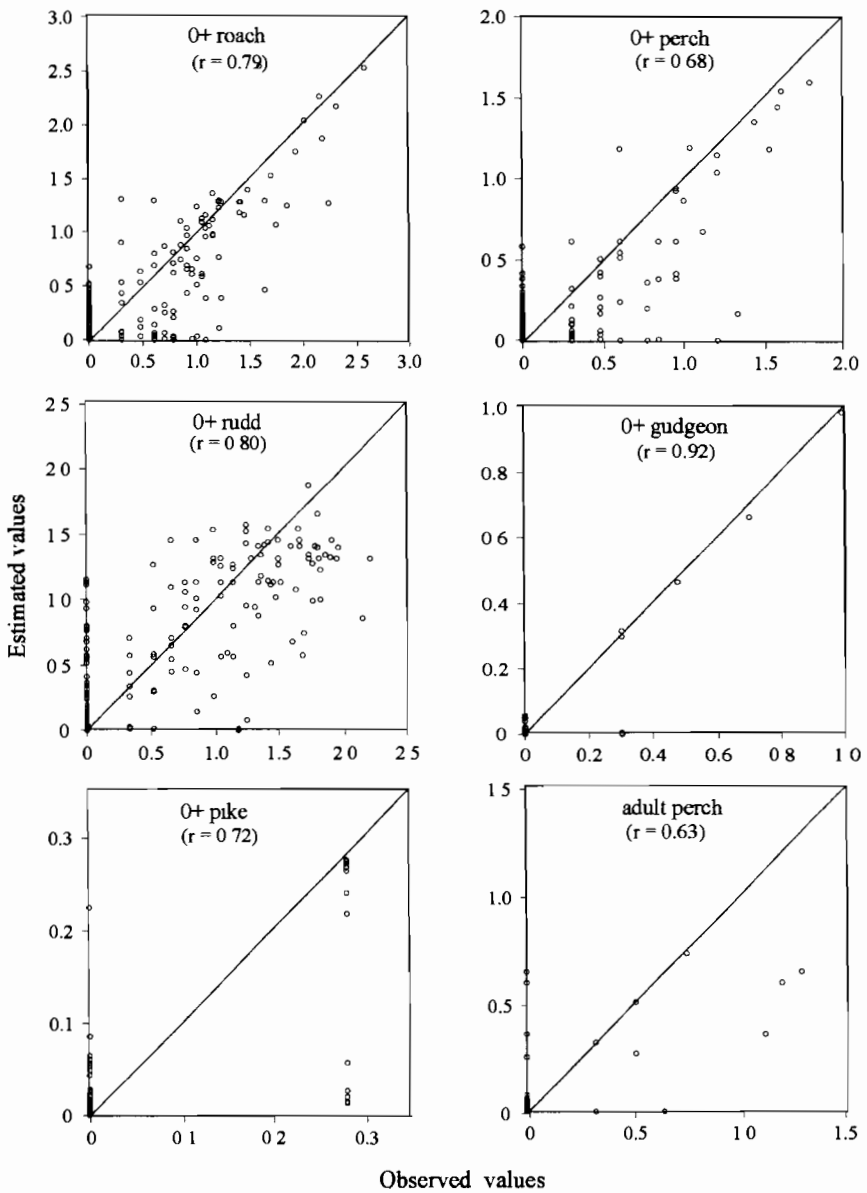

Fig. 3. Recognition performance of the ANN models for the six fish populations. Scatter plots of predicted values vs. observed values in the training procedure. The solid line indicates the perfect fit line of prediction (coordinates 1:1).

Table 1
Performance index (PI) and sum of squared errors (SSE) in ANN training and testing and in MLR training for the six populations[a]

	ANN				MLR	
	Training		Testing		Training	
	PI	SSE	PI	SSE	PI	SSE
0+ Roach	66	7.32	63	11 98	50	70.74
0+ Perch	72	6.23	69	14.02	65	29.46
0+ Rudd	69	9.26	61	13.46	46	65.27
0+ Gudgeon	97	0.66	96	2.65	98	1.04
0+ Pike	90	2.26	91	1.80	90	5.50
Adult Perch	94	7.74	91	8.25	95	1 18

[a] PI is the percentage of well-predicted values with an error rate lower than 10%.

ulations such as for 0+ gudgeon, for 0+ pike and for adult perch. For these three fish populations, non-nil values were rarely predicted as nil values by the network (only two samples for 0+ gudgeon and for adult perch) and a large proportion of the high values were well or perfectly predicted. For 0+ roach, 0+ rudd and 0+ gudgeon, points were well-distributed along the diagonal of best fit. 0+ perch, adult perch and 0+ pike abundances were underestimated, but the results remained clearly better than those obtained using MLR. Moreover the distribution of residuals was close to normal with a mean value of 0.007 (S.D. = ±0.152) for 0+ roach, 0.017 (S.D. = ±0.128) for 0+ perch, 0.006 (S.D. = ±0.183) for 0+ rudd, −0.004 (S.D. = ±0.056) for 0+ gudgeon, −0.001 (S.D. = ±0.172) for 0+ pike and 0.001 (S.D. = ±0.010) for adult perch.

A cross-validation testing procedure (i.e. leave-one-out bootstrap) was performed to validate the ANN models. Models could have been evaluated using the determination coefficients (r^2) or correlation coefficients (r), but because of the scarcity of high values of fish densities (especially for 0+ gudgeon, 0+ pike and adult perch), we preferred to use performance index (PI) and sum of squared errors (SSE) to assess model prediction performance. The PI was based on the proportion of responses within plus or minus 10% of the actual value.

The PIs obtained after the testing procedure were very close to those obtained after training for each of the six species (Table 1). SSE of the test were low and close to those obtained during the training procedure. MLR gave high PIs due to the abundance of nil values; however the SSE values were clearly higher than for ANN, except for 0+ gudgeon and 0+ perch due to the scarcity of non-nil values. Thus, compared with MLR, ANN gave better results both in training and testing procedures.

3.2. Importance of the environmental variables in population abundance

In MLR, the influence of each variable can be roughly assessed by checking the final values of the partial standardized regression coefficients. Each coefficient of a linear model is the partial derivative of the response of the model with respect to the variable of that coefficient. The standardized coefficients of MLR therefore generally give a way to compare the relative influence of each independent variable on the dependent variable, when all other independent parameters have been kept constant in the models. Table 2 shows the MLR standardised partial coefficients of the eight variables for each population. Few among these coefficients were significant (5 for 0+ roach, 3 for 0+ perch and 0+ rudd, 2 for 0+ gudgeon, 1 for 0+ pike and adult perch). Moreover, three of the eight variables are usually considered as essential for 0+ fish microhabitat choice: distance from the bank (DIS), depth

(DEP), and flooded vegetation (VEG), but MLR considered only distance from the bank (DIS) as significant (except for adult perch). MLR shows that 0+ fish abundance was significantly correlated to low values of distance from the bank (DIS) (i.e. negative coefficients), this is in accordance with ecological studies (Haberlehner, 1988; Copp, 1992). Nevertheless, according to MLR models, 0+ roach and 0+ gudgeon abundance increase with depth (DEP), which seems illogical, as deep littoral areas are usually avoided by 0+ fish. Finally, the flooded vegetation (VEG) was never considered as a significant variable, whereas it is logically one of the most important variables for 0+ fish (Persson and Eklöv, 1995; Eklöv, 1997).

For ANN, the results of Garson's algorithm stress the importance of environmental variables in the model (Fig. 4). Standard errors calculated for each variable after ten training procedures were very low, showing the stability of the network models. The contribution of each environmental variable to the model for the six populations was in accordance with previous ecological studies (Holland and Huston, 1984; Haberlehner, 1988; Copp, 1992; Mastrorillo et al., 1996): 0+ roach, 0+ perch, 0+ rudd and 0+ pike are closely linked to the flooded vegetation (VEG) and the distance from the bank (DIS) whereas 0+ gudgeon is indifferent to the flooded vegetation (VEG) but strongly influenced by the distance from the bank (DIS). Finally, adult perch habitat is known to be largely governed by the depth (DEP) and the distance from the bank (DIS) (Persson, 1983; Persson and Eklöv, 1995). Moreover, fish microhabitat is defined by several variables showing that microhabitat results from a complex combination of different habitat characteristics (only 0+ gudgeon show a quite simple diagram, with only one important variable, the distance from the bank (DIS), which contributes more than 50%). The main processes that determine fish habitat and distribution can be approximated by linear functions only to a limited extent. Even when simple (e.g. logarithmic) transformations of variables to linearize their distribution are used, the results remain unsatisfactory. The use of complex transformations of the variables (e.g. GAM) improves the results, but they remain lower than those obtained by ANN. On the other hand, ANN with only one hidden layer can model non-linear systems in ecology without complex transformations of the data (Goh, 1995; Lek et al., 1996b; Scardi, 1996). The microhabitat of the six fish populations studied here was reliably fitted by ANN to the measured environmental characteristics of the points sampled in the lake. The ANN models clearly show the influence of each variable on the microhabitat of each population whereas MLR gives aberrant values which are irrelevant from an ecological point of view. Thus, MLR models are unable to represent ecological reality due to non-linear relationships such as those which probably exist between the densities

Table 2
MLR partial standardised coefficients for the six fish populations studied[a]

	0+ Roach	0+ Perch	0+ Rudd	0+ Gudgeon	0+ Pike	Adult perch
DEP	0.117*	−0.050	−0.011	0.689**	−0.016	0.003
DIS	−0.418**	−0 246**	−0.167**	−0 213**	−0.126*	0.008
SLO	0.054	0.180**	−0.294**	0.033	−0.033	0.075
BOU	−0.033	−0 019	−0.073	−0.013	−0.014	0.123*
PEB	−0.112*	−0.061	−0.031	−0.019	−0.018	−0.021
GRA	0.184**	−0 047	−0 049	−0.014	−0.013	−0.013
MUD	0.263**	0.144*	−0.256**	−0.073	0.094	0.067
VEG	−0.098	−0.087	0 044	−0.066	0.025	−0.037

[a] Environmental variables were lettered as follows: DEP = depth, SLO = slope, DIS = distance from the bank, BOU = boulders, PEB = pebbles, GRA = gravel, MUD = mud, VEG = flooded vegetation.

* Significant coefficient ($P<0.05$).

** Highly significant coefficient ($P<0.01$).

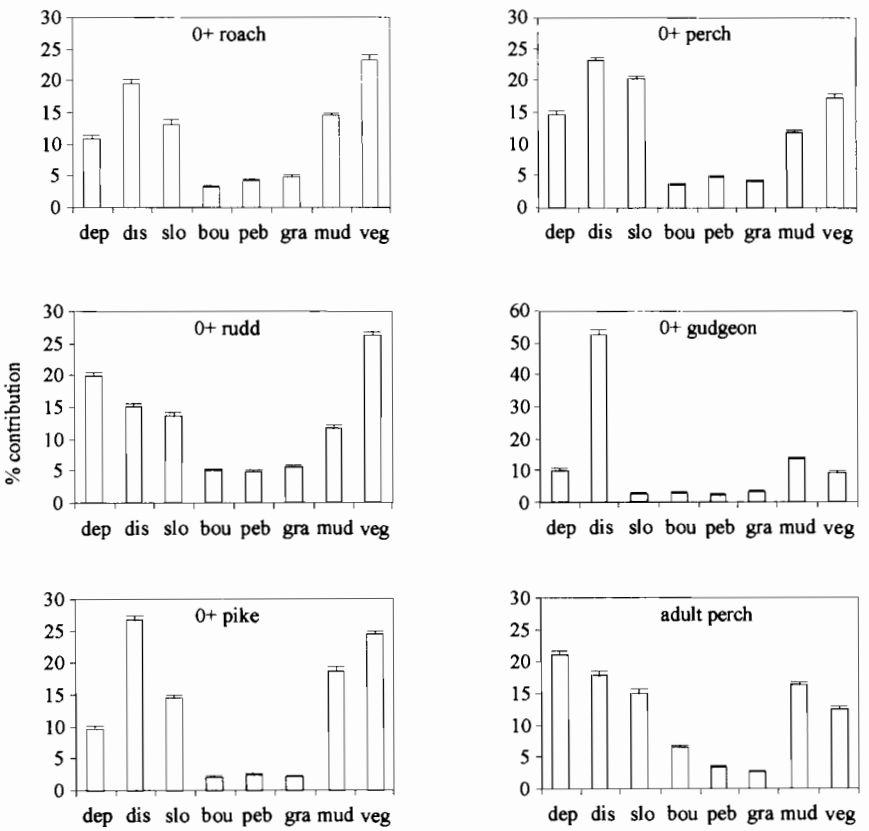

Fig. 4. Percentage contribution of each of the eight independent variables to the prediction of fish densities, obtained by Garson's algorithm (variables lettered as in Fig. 1). Bars indicate the mean value of the results of the ten models for each fish population, horizontal lines represent standard errors of the mean.

of the six fish populations considered and some environmental variables.

3.3. Population assemblage studies

To visualise the spatial distribution of the six fish populations studied within their environment (i.e. population assemblage), on the basis of the information provided by the models, PCA techniques were used.

On its first and second axes which accounted for 49.2 and 29.5% of the total information, respectively (Fig. 5a), the PCA performed on the partial coefficients of the MLR revealed a significant correlation ($P < 0.01$) between distance from the bank (DIS), pebbles (PEB), flooded vegetation (VEG) and 0+ pike and 0+ roach; gravel (GRA), mud (MUD) and 0+ perch and 0+ roach; slope of the bottom (SLO), distance from the bank (DIS), boulders (BOU) and adult perch. We can notice, on the first axis, an opposition between (0+ pike, 0+ rudd) and (0+ roach, 0+ perch). The second axis shows an opposition between adult perch, and 0+ rudd (Fig. 5b). These results based on MLR models conflict with general agreement on habitat use by both 0+ roach and 0+ perch individuals since, during the larval and juvenile periods, they are generally located close to shelters such as flooded vegetation (Haberlehner, 1988; Persson and Eklöv, 1995).

Concerning ANN, the PCA performed on the contribution factors (Goh's algorithm results) allowed the microhabitat of the six fish populations

to be taken into account simultaneously to better define their spatial occupancy and thus to approach the population assemblage. On its first and second axes, which accounted for 43.1 and 20.6% of the total information, respectively (Fig. 6a), the PCA revealed a significant correlation ($p < 0.01$) between flooded vegetation (VEG) and 0+ roach, 0+ rudd and 0+ pike; between depth (DEP) and adult perch; between distance from the bank (DIS) and 0+ gudgeon. We can see, on the first axis an opposition between 0+ gudgeon individuals and the other fish species individuals except for 0+ pike. The second axis shows an opposition between adult perch and the group 0+ roach, 0+ rudd, 0+ pike and 0+ perch (Fig. 6b). The representation of the ten statistical units for each population reveals the range of microhabitat variation for each fish population. Moreover, the cluster analysis distinguishes several groups and enables an approach to be made to the spatial range of microhabitat characteristics for each population (Fig. 6c). The separation of some fish populations such as 0+ gudgeon or for top-predators (i.e. 0+ pike and adult perch) has already been observed in natural environments, and the spatial occurrence of 0+ roach, 0+ rudd and 0+ perch, as illustrated by the cluster analysis, is well-known by ichthyologists. The fish assemblage visualised in the PCA was in accordance with various ecological studies concerning the microhabitat of these species (Persson, 1983; Haberlehner, 1988; Copp, 1992; Hosn and Dowing, 1994; Persson and Eklöv, 1995; Mastrorillo et al., 1996). As a consequence, the fish assemblage was reliably predicted using ANN. This predicted spatial occupancy can be easily visualised on a PCA plane. Thus, ANN are more suitable than MLR to reproduce the operation of real complex multispecies systems (i.e. population assemblage) on the basis of the ecological variables introduced in the model.

4. Conclusion

The back-propagation of ANN constitutes a more efficient tool than MLR to predict fish abundance and spatial occupancy from the environmental characteristics of the littoral area of a lake. The selection of input variables introduced into the modelling procedures, their ecological significance and the constitution of testing sets of

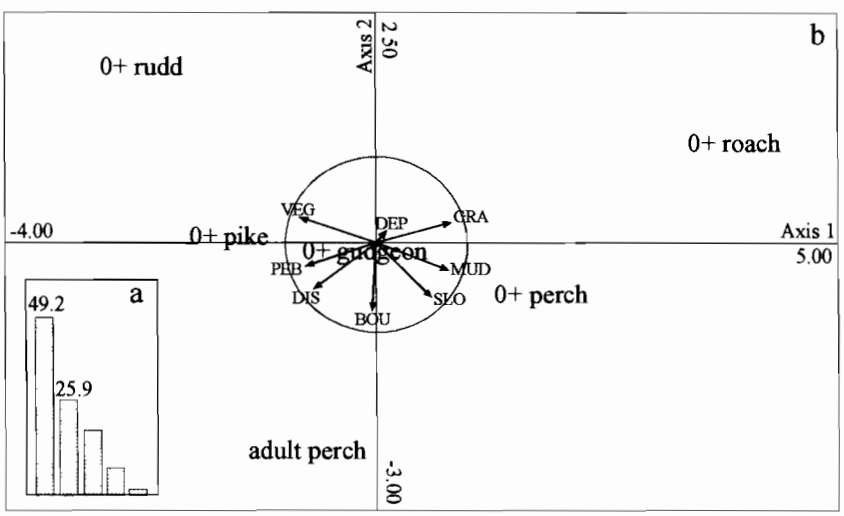

Fig. 5. Principal component analysis (PCA) performed on MLR results using the standardised partial regression coefficients for the six fish populations. (a) Histogram of eigenvalues; (b) distribution of the six samples (i.e. populations) and the eight environmental variables (DEP = depth, SLO = slope, DIS = distance from the bank, BOU = boulders, PEB = pebbles, GRA = gravel, MUD = mud, VEG = flooded vegetation) on the F1 × F2 plane.

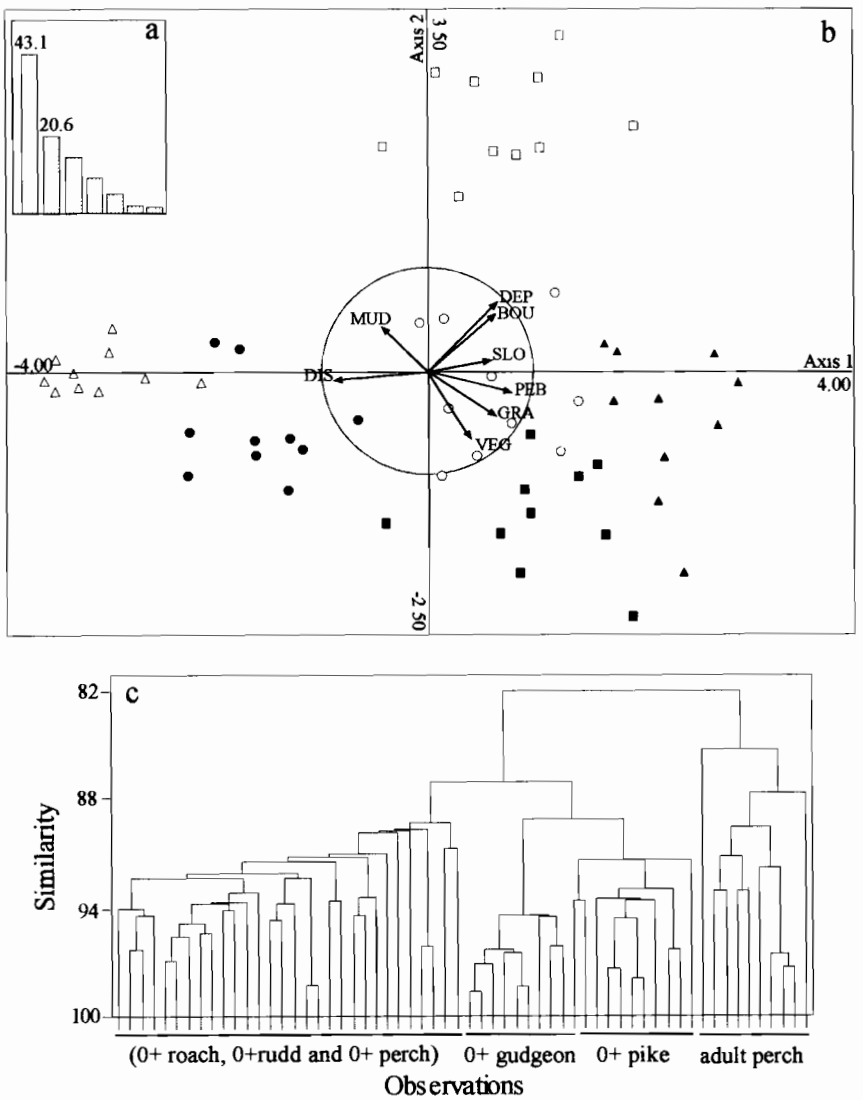

Fig. 6. Principal Component Analysis (PCA) performed on ANN results using Garson's algorithm for the six fish populations. For each population, the statistical units (samples) were the results of the ten ANN models (a) Histogram of eigenvalues; (b) distribution of the 60 samples and the eight environmental variables ((DEP = depth, SLO = slope, DIS = distance from the bank, BOU = boulders, PEB = pebbles, GRA = gravel, MUD = mud, VEG = flooded vegetation) on the F1 × F2 plane. (□) adult perch; (■) 0+ roach; (o) 0+ perch; (●) 0+ pike; (△) 0+ gudgeon; (▲) 0+ rudd; (c) cluster analysis of the first two coordinates of PCA showing a separation between adult perch, 0+ pike and 0+ gudgeon, the three other populations are dispersed across the similarity gradient.

data to assess the performance of the model are important elements for this type of approach (Faush et al., 1988). The ANN modelling approach used here is a fast and flexible way to incorporate multiple input parameters into a single model. In addition to the predictive value of the model, the combination of ANN and multivariate analysis simultaneously visualise the results provided by several ANN models with the same data matrix at the input. It is this ability to

deal with multiple information sources that provides the power of this approach, resulting in a significant improvement in ANN modelling over conventional techniques. These results on the use of ANN for population assemblage analyses are promising and open new fields for their applications to ecology.

Acknowledgements

The authors are grateful to S. Beker for correcting the English version. The authors thank William Silvert and an anonymous reviewer for helpful discussions on the subject leading to improve the manuscript. This research was supported in part by a doctoral grant (S. Brosse) provided by the French electricity agency (E.D.F.).

References

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996. Stochastic models that predict trouts population densities or biomass on microhabitat scale. Hydrobiologia 337, 1–9.

Binns, N.A., Eiserman, J.P., 1979. Quantification of fluvial trout habitat in Wyoming. Trans. Am. Fish. Soc. 198, 215–228.

Cleveland, W.S., 1979. Robust locally-weighted regression and scatterplot smoothing. J. Am. Stat. Assoc 74, 829–836.

Copp, G.H., 1989. Electrofishing for fish larvae and juveniles: equipment modifications for increased efficiency with short fishes. Aquacult. Fish. Manage. 20, 453–462

Copp, G.H., 1992. Comparative microhabitat use of cyprinid larvae and juveniles in a lotic floodplain channel. Environ. Biol. Fishes 33, 181–193

Dimopoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity critera for choosing networks with good generalization ability. Neural Process. Lett 2, 1–4.

Efron, B., 1983. Estimating the error rate of a prediction prediction rule: some improvements on cross-validation. J Am. Stat. Assoc. 78, 316–331.

Efron, B., Tibshirani, R., 1995. Cross-validation and the Bootstrap: estimating the error rate of a prediction rule. Tech. Rep. 176. Department of statistics, Stanford Univ., 27 pp ftp://utstat.toronto.edu/pub/tibs/cvboot.ps

Eklöv, P , 1997. Effects of habitat complexity and prey abundance on the spatial and temporal distributions of perch (*Perca fluviatilis*) and pike (*Esox lucius*). Can J Fish. Aquat. Sci. 54, 1520–1531.

Faush, K.D., Hawkes, C.L., Parsons, M.G., 1988. Models that predict the standing crop of stream fish from habitat variables: 1950–85. Gen. Tech. Rep. PNW-GTR-213. U.S. Department of agriculture, Forest service, Pacific north reaserch station, Portland, OR, 52 pp.

Garson, G.D., 1991. Interpreting neural network connection weights Artif. Intel. Expert. 6, 47–51.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilema. Neural Comput. 4, 1–58

Giske, J., Huse, G., Fiksen, O., 1998. Modelling spatial dynamics of fish Rev. Fish. Biol. Fish. 8, 57–91.

Goh, A T.C., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intel. Eng. 9, 143–151.

Guegan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384

Haberlehner, E., 1988. Comparative analysis of feeding and schooling behaviour of the Cyprinidae *Alburnus alburnus* (L., 1758), *Rutilus rutilus* (L., 1758), and *Scardinius erythrophthalamus* (L. 1758) in a backwater of the Danube near Vienna Int. Rev. Hydrobiol. 73, 537–546.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models Chapman and Hall, London, p. 333.

Holland, L.E., Huston, M.L., 1984. Relationship of Young-of-the-Year northern pike to aquatic vegetation types in backwaters of the upper Mississippi river. North Am. J. Fish. Manage. 4, 514–522.

Hosn, W.A., Dowing, J.A., 1994. Influence of cover on the spatial distribution of littoral-zone fishes. Can. J. Fish Aquat. Sci. 51, 1832–1838.

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Ann. Rev. Ecol. Syst. 21, 129–166.

Kohavi, R., 1995. A study of the cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), Montreal, pp. 1137–1143

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Res. 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagner, S., 1996b. Application of neural networks to modeling non-linear relationships in ecology. Ecol. Model. 90, 39–52.

Lek, S., Dimopoulos, I., Fabre, A., 1996c. Predicting phosphorus concentration and phosphorus load from watershed characteristics using back-propagation neural networks. Acta Oecol. 17, 43–53.

Mastrorillo, S., Dauba, F., Belaud, A., 1996. Utilisation des microhabitats par le vairon, le goujon et la loche franche dans trois rivières du sud-ouest de la France Ann. Limnol. 32, 185–195.

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biol. 38, 237–246.

Nelva, A., Persat, H., Chessel, D., 1979. Une nouvelle méthode d'étude des peuplements ichtyologiques dans les grands cours d'eau par échantillonnage ponctuel d'abondance. C. R. Acad. Sci. Paris Serie III 289, 1295–1298.

Norusis, M.J., 1993. SPSS for Windows. Base system user's guide release 6.0, SPSS Inc., 828 pp.

Persson, L., 1983. Food consumption and competition between age classes in a perch *Perca fluviatilis* population in in a shallow eutrophic lake. Oikos 40, 197–207.

Persson, L., Eklov, P., 1995. Prey refuges affecting interactions between piscivorous perch and juvenile perch and roach Ecology 76, 70–81.

Ricker, W.E., 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can. 191, 1–382

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Prog. Ser 139, 289–299.

Schoener, T.W., 1989. Food webs from the small to the large. Ecology 70, 1559–1589.

Tan, S.S., Smeins, F E., 1996. Predicting grassland community changes with an artificial neural network model. Ecol Model 84, 91–97.

ter Braak, C.J.F., Looman, C W.N., 1995. Regression. In: Jongman, R.G.H., ter Braak, C.J F., Van Tongeren, O.F.R. (Eds), Data analysis in community and landscape ecology. Cambrige University Press, pp 29–77.

ter Braak, C.J.F., Verdonschot, F.M , 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat. Sci. 57, 254–289.

ELSEVIER

Ecological Modelling 120 (1999) 313–324

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Microsatellites and artificial neural networks: tools for the discrimination between natural and hatchery brown trout (*Salmo trutta*, L.) in Atlantic populations

Didier Aurelle [a,*], Sovan Lek [b], Jean-Luc Giraudel [c], Patrick Berrebi [a]

[a] *Laboratoire Génome et Populations, CNRS UPR 9060, Cc063, Université Montpellier II, Place Eugène Bataillon, 34095, Montpellier Cedex 05, France*

[b] *CESAC, UMR 5576, Bat 4R3, CNRS-Univ. Paul Sabatier, 118 route de Narbonne, 31062, Toulouse Cedex, France*

[c] *IUT Périgueux Bordeaux IV, Département Génie biologique, 39 rue Paul Mazy, 24019, Périgueux Cedex, France*

Abstract

Artificial Neural Networks (ANN) were applied to microsatellite data (highly variable genetic markers) to separate genetically differentiated forms of brown trout (*Salmo trutta*) in south-western France. A classic *feed-forward* network with one hidden layer was used. Training was performed using a back-propagation algorithm and reference samples representing the different genetic types. The hold-out and the leave-one-out procedures were used to test the validity of the network. They were chosen according to the populations and the questions analysed. The informative content of the different variables used for the distinction (the alleles of the different loci) was also evaluated using the Garson–Goh algorithm. The results of learning gave high percentages of well-classified individuals (up to 95% for the test with the hold-out analysis). This confirms that ANNs are suitable for such genetic analyses of populations. From a biological point of view, the study enabled evaluation of the genetic composition and differentiation of different river populations and of the impact of stocking.

Keywords: Artificial Neural Network; Classification; Microsatellites; Stocking; Brown trout

1. Introduction

Salmonids are extensively studied fishes both from a practical point of view (fisheries management) and for some more theoretical aspects (ecology and evolution). The brown trout (*Salmo trutta* L.) displays some interesting biological characteristics for the study of genetic intraspecific differentiation: brown trout lives in the upper part of the rivers and is philopatric. Genetic studies have shown that the species *S. trutta* includes several genetic entities. For example, in the western part of the French Pyrenees, two wild forms are present naturally: ancestral Atlantic and modern Atlantic (the first one was called ancestral according to Hamilton *et al.*, 1989). Moreover, stocking practices led to the introduction there (and more generally in most French rivers) of a third form, the domestic modern Atlantic trout,

* Corresponding author. Fax: +33-467-144-554..
E-mail address: aurelle@crit.univ-montp2.fr (D. Aurelle)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00111-8

which does not originate from these rivers (Aurelle and Berrebi, 1998). The three forms may be found in the same river and can hybridise.

Nevertheless, the classification of individuals among the different forms is a prerequisite for the study of genetic interactions. Allozymes separate modern and ancestral forms, but no diagnostic markers are available to distinguish between domestic and wild modern Atlantic trout. However, microsatellites have shown that the distinction is justified as the populations of some rivers appear to be genetically different to hatchery strains (Aurelle and Berrebi, 1998). Because of microsatellite properties, distinction between individuals of the different forms remains difficult. These loci usually display a high mutation rate and are subject to retention of ancestral polymorphism and homoplasy phenomena (Jarne and Lagoda, 1996). There are numerous shared alleles between wild and domestic modern populations and only some differences in allelic frequencies. It is, therefore, necessary to use powerful statistical classification tools to appraise the genetic composition of the populations studied and at the same time to separate natural migration and human manipulations (stocking).

Artificial neural networks (ANNs) seem well-suited to the problem. They have already been used for a wide range of different studies and situations. They are commonly used in physics and chemistry but less so in ecology and population genetics. However, preliminary studies have shown that ANNs are suitable for these topics (Guégan *et al.*, 1998) and more effective than classic discriminant analysis Cornuet *et al.*, 1996; Mastrorillo *et al.*, 1997). Moreover, no particular assumptions are required concerning the data used for classification. ANNs have proven to be effective in population genetics, at several different taxonomic levels and with highly variable markers such as microsatellites (Cornuet *et al.*, 1996). They are, therefore, expected to be capable of classifying individuals in populations belonging to the same sub-species and genetically relatively similar (e.g. wild and domestic modern trout). Until now, neural networks have been tested with some well separated and genetically differentiated groups (such as bees in Cornuet *et al.*, 1996). In the work reported here, we applied them to mixed populations where samples may contain several genetic units; this raises the question of the reference samples necessary for training the network (see Section 2) and that of the validation procedures (how can we know if the result is right?). Several training and validation procedures were tested depending on the situation.

Analyses were performed with different purposes. Firstly, we wished to verify using independent markers (microsatellites), the distinction between modern and ancestral fishes which is shown by allozymes at only one locus (*LDH5**); this also enabled us to test the method in a clear, well known situation. We then sought wild modern populations (with no or almost no stocking influence). This enabled us to evaluate the genetic composition of the different populations analysed here. The importance of the different alleles in the classification (and their informative content) is also discussed for the different microsatellite loci used.

2. Materials and methods

2.1. The populations analysed

The populations from nine rivers and three hatchery strains were analysed. The origins and sizes of the samples are provided in Table 1. The numbers refer to Fig. 1, and the percentages of allele *LDH5**90 provide some information about the genetic composition of the populations. The ancestral form is characterised by allele 100 at this locus whereas the two modern forms possess allele 90. A population with 100% *LDH5**90 is then considered as modern, but we do not know whether these fishes are wild or domestic (there is no diagnostic allele for this distinction). Some populations consisting of only a few individuals were analysed because they were genetically and morphologically original (Andurentako) or because they seemed to be mixed (Marcadau which, according to local managers is heavily stocked; moreover hatchery fishes are often easy to recognise thanks to coloration) but we kept in mind the problems of small samples.

According to allozymic data (unpublished) some river samples consisted mainly of modern fishes (Chiroulet, Oussouet and Luz) and certain other samples were almost completely ancestral (Dancharia, Andurentako, Béhérékobentako and Bastan). According to local managers, these populations have not been stocked for several years. Moreover, the morphological characteristics would tend to show that Chiroulet, Oussouet and Luz fishes are mainly wild. Marcadau and Béhérobie contain both modern and ancestral fishes. The morphology of Marcadau fishes tend to show that the population is quite heavily restocked.

2.2. Microsatellite loci

Four microsatellite loci were analysed. *Strutta 58* has been cloned by Poteaux (1995). MST 73 and MST 15 have been cloned by Estoup (Estoup *et al.*, 1993). MSU 4 has been published in Genbank under accession number U43694; it was submitted directly by P.T. O'Reilly and has been

Table 1
Origin and characteristics of the samples; bold names refer to the samples names used in the text

No. (map)	Locality	River	Basin	Sample size	% *LDH5**90
	La Canourgue	hatchery		50	95
	Brassac	hatchery		30	100
	Suech	hatchery		36	99
1	Cauterets	**Marcadau**	Adour	15	33
2	Sare	**Beherekobentako**	Nivelle	24	0
3	**Dancharia**	Nivelle	Nivelle	30	2
4	Herboure	**Andurentako**	Untxin	5	0
5	Bidarray	**Bastan**	Adour	29	4
6	**Béhérobie**	Nive de Béhérobie	Adour	25	27
7	**Chiroulet**	Adour de Lesponne	Adour	86	89
8	Bagnères de Bigorre	**Oussouet**	Adour	86	82
9	Argeles	**Luz**	Adour	88	95

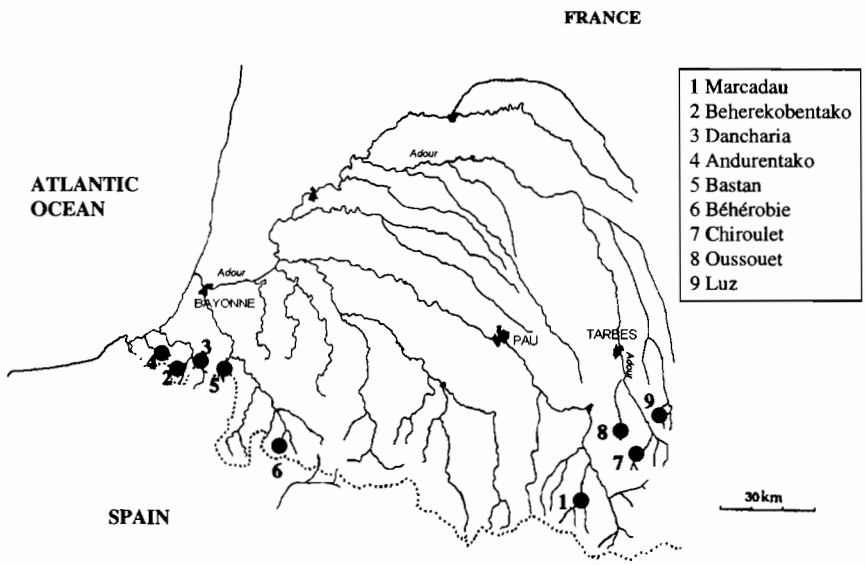

Fig. 1. Location of the sampling points.

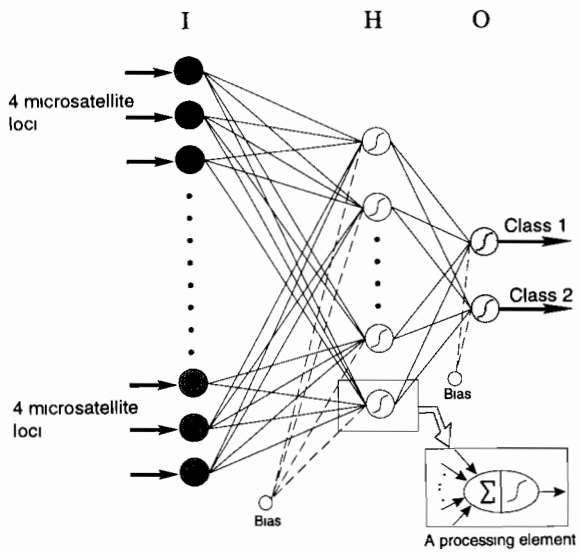

Fig. 2. Structure of an Artificial Neural Network (ANN).

identified in salmon (*Salmo salar*). Two of these four loci were highly variable (*Strutta 58* and MSU 4 with 38 and 18 alleles respectively). The two others displayed only a few alleles in comparison with the usual microsatellite variability (seven alleles for MST 73 and eight for MST 15). PCR and analyses procedures are described in Aurelle and Berrebi (1998).

2.3. Artificial neural networks

A classic *feed-forward* network (Rumelhart *et al.*, 1986) was used in the study. This network had three layers: an input layer, a hidden layer and an output layer. The input layer was connected with the variables used for discrimination; in our study, these variables were the 71 alleles coded as follows: for each allele, each individual was noted zero if it did not possess it, one if the fish was heterozygotic for the allele and two if it was homozygotic for it. The hidden layer was reduced to two neurones to avoid too large a number of parameters; this choice did not reduce the network efficiency beyond reasonable limits. The number of neurones in the output layer corresponds to the number of categories in which individuals should be classified (depending on the analyses, see Section 2.4).

Each neurone is connected with the neurones of the neighbouring layers; it receives and sends signals through these connections and always from input to output (Fig. 2). Each connection is weighted according to the signal intensity. Each neurone integrates the signals received from the former neurones and sends a new signal to the next ones. This signal is delivered according to a non-linear transfer function applied to the sum of the weighted signals of the former neurones (see Cornuet *et al.*, 1996; Mastrorillo *et al.*, 1997). Let w_i and x_i be the weight and the signal outgoing from the former neurone i (layer n); the incoming signal for one neurone in the layer $n+1$ will be:

$$z = \sum w_i * x_i \quad (1)$$

The outgoing signal for this neurone in layer $n+1$ will then be:

$$f(z) = [1 + \exp(-z)]^{-1} \quad (2)$$

For the input layer, incoming signals correspond to the variables used to classify individuals (the 71 alleles). The outgoing signals of the output layer designate the category where the studied individual will be assigned by the network. The decision is made in the light of the highest score. Nevertheless, as is mentioned in Section 2.4, absolute output values can and should be discussed. For example, individuals with a score of one in a group can be considered as quite accurately classified in this category but the interpretation of individuals with intermediate scores (0.5 for example) is not as easy. On the other hand, individuals with scores of zero to 0.1 in their original category can be considered to be incorrectly classified.

The network must be trained in order to classify individuals correctly. A training data set (randomly chosen in the global data set) is used to modify the weights of the different connections in order to maximise the percentage of well-classified individuals. We used a 'back-propagation' algorithm. First, the initial weights are randomly distributed. They are then modified iteratively depending on the differences between expected and observed output signals (assignation scores; see Cornuet *et al.*, 1996; Mastrorillo *et al.*, 1997).

Numerous iterations are usually necessary to obtain a good percentage of well-classified individuals without an over-fit to the training data set. Effectively, if the percentage of well classified individuals is much higher for the learning data than for the test data (see below), we can deduce that the network has learned the training data particularities and cannot be applied to a more general situation.

A hold-out procedure (Kohavi, 1995) can be used to test the validity of the network. For this purpose, a data set with some known categories is divided into two parts. The first part is used for training the network. When the training procedure has been completed, the network is then applied to the second part and we can evaluate the percentage of well classified individuals for data not used for learning. This second part is then used as a test. Once it has been verified that the network is well suited and does not over-fit the learning data, it can be applied to unknown data (application stage).

If the data set is too small to be divided into two parts or if its composition is not well known and possibly heterogeneous, one can use the leave-one-out procedure (Kohavi 1995). For example, for a data set with N individuals, training is performed with $N-1$ individuals (by assuming that their categories are known) and the network is applied to the Nth individual, which is then classified according to its proximity to one of the previously learned categories. This analysis is repeated for the N individuals which are all assigned to one group. Given the high number of training stages (N steps), the number of iterations for each training is limited to 500.

For analysis of the results, each individual was assigned to the category where it showed the highest score. At the population level, it is interesting to study the individual score distributions for the various categories. In order to analyse the contributions of the different alleles to classification, we used the Garson–Goh algorithm (Garson, 1991; Goh, 1995; Lek *et al.*, 1996a,b). This algorithm determines the relative importance of the various input variables by taking into account the weights of the hidden layer neurones connected with these input. Briefly, for each hidden neurone, the weight of the connection from one input variable to this neurone is multiplied by the weight of one output connection; these products are summed for all the output connections and then expressed relatively as a percentage for the comparison of all input variables. These percentages are intended to express the informative content of each variable.

2.4. Analysis protocols

(1) First, we tested the effectiveness of the method for a situation in which some genetic markers different from microsatellites were able to distinguish between several categories. Here, **modern and ancestral** individuals can be separated with allozymes (especially with the *LDH-5** locus). The training set consisted of four ancestral populations (Bastan, Béhérékobentako, Dancharia and Andurentako) versus four modern populations (the three hatcheries and Luz). This distinction was analysed using a hold-out (1a) and then a leave-one-out (1b) procedure.

2) We then analysed the **hatchery populations**. The different strains are assumed to be genetically quite similar so the sample analysed should be representative of the different hatchery strains used in the country. We tried to verify these assumptions by using a leave-one-out procedure (for the analysis of all individuals and because one strain may be heterogeneous) with three categories corresponding to the three strains analysed.

3) We also sought **wild modern Atlantic** populations. As a modern population may be heterogeneous and contain wild and domestic fishes, we decided to use a leave-one-out procedure with two classes comparing each modern river population to hatcheries which were pooled (according to the results of analysis two showing the genetic homogeneity of these strains). Three tests were performed: Chiroulet versus hatcheries, Oussouet/hatcheries and Luz/hatcheries.

4) The **other river populations** (ancestral and mixed) were also compared to hatcheries by the leave-one-out method to examine the potential influence of domestic fishes in these samples. The leave-one-out procedure is useful for this comparison because each fish is analysed individually

and the presence of a foreign fish (a domestic fish in a river) can theoretically be detected. We compared Bastan with the pooled hatcheries, Béhérobie versus hatcheries and Marcadau versus hatcheries.

3. Results

For each analyses we will give some percentages of so-called 'incorrectly classified individuals': this indicates individuals which were not classified by the network in the population where they were sampled. Nevertheless, they can either be classified in the population from which they originate (as for example some domestic fishes classified in the hatchery category but sampled in one river) or they can effectively correspond to some errors of the network.

3.1. The ancestral—modern distinction

(1a) The percentage of incorrectly classified individuals by leave-one-out is 2% in the global comparison between ancestral and modern. This proportion is 1% among supposedly modern individuals and 7% for populations expected to be ancestral. Analysis of the distribution of the scores within the ancestral category for ancestral populations (Fig. 3) shows that most individuals (65%) score between 0.8 and one; 26% are between 0.5 and 0.8, corresponding to less sharp and correct assignation, like the 2% scoring between 0.3 and 0.5. Finally, 7% should really be classified in the other group (score between 0.1 and 0.3). Conversely, more than 80% of modern individuals scored between zero and 0.1 and were then well classified in their original category. 1% scored between 0.9 and one and were assigned to the ancestral type whereas they were in a modern sample.

(1b) With the hold-out procedure, we observed 1% of incorrectly classified fishes in the learning stage and 5% in the test. When this network is applied to new populations, the percentage of modern individuals can be evaluated and compared to the frequency of the *LDH-5** modern allele (Table 2). There is a reasonably good correlation between the two sets of variables.

For this analysis, the contributions of the different alleles to the network are shown in Fig.

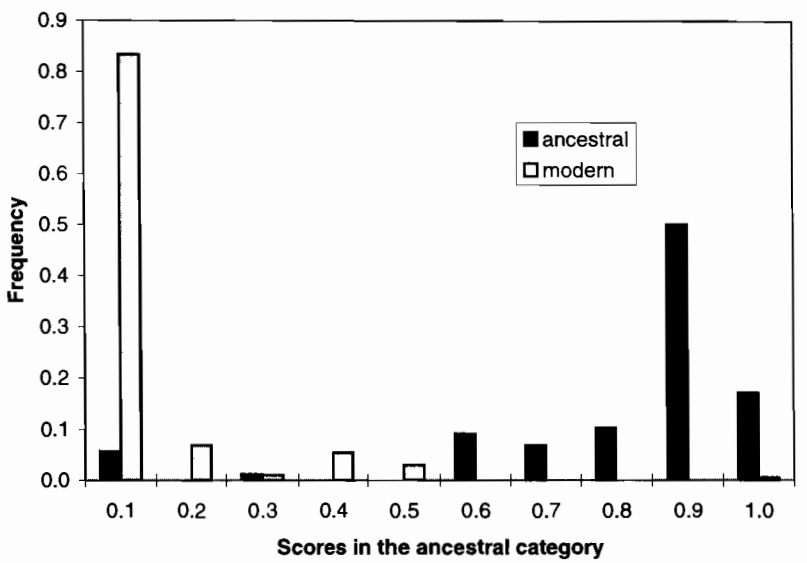

Fig. 3. Score distribution in the ancestral category for the leave-one-out comparison between ancestral and modern. The first category corresponds to scores of between zero and 0.1 and the second to scores of between 0.1 and 0.2,....

Table 2
Percentage of modern individuals in four populations as predicted by artificial neural network compared with the frequency of modern *LDH5** allele

Populations	Neural network predictions (% modern individuals)	Allozyme predictions (% modern alleles)
Béhérobie	32	27
Marcadau	53	33
Chiroulet	73	89
Oussouet	73	82

4. On the *x* axis, alleles are classified by increasing abundance in the overall data set. The more frequent alleles usually contribute more to analysis than those that are fairly rare, but with some exceptions. Some rare alleles can be useful or not, depending on the analysis.

3.2. The different hatchery strains

(2a) In the leave-one-out analysis with three categories corresponding to the three hatchery strains, 19% of the individuals were found to be incorrectly classified, which is high compared to the previous analyses. The score distribution of each strain in its corresponding category (Fig. 5) shows that only a very small proportion of individuals scored between 0.8 and one (2.6%; none for Brassac and Canourgue); most scores are be-

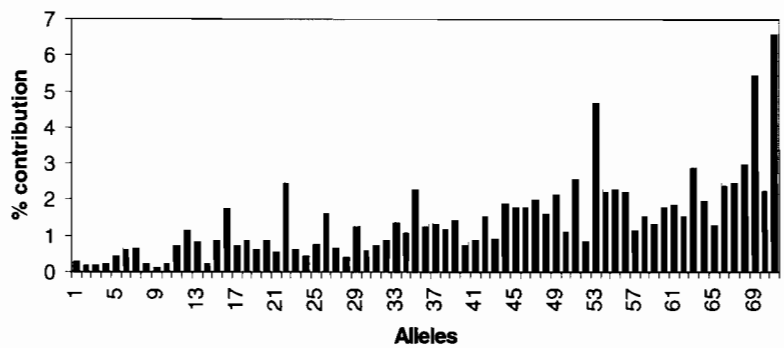

Fig. 4. Contributions of the different alleles to the leave-one-out ancestral/modern. Alleles are set out on the x axis according to their frequency in the overall data set (all loci are included). Contributions are computed with the Garson–Goh algorithm (Garson, 1991; Goh, 1995; Lek *et al.*, 1996a,b)

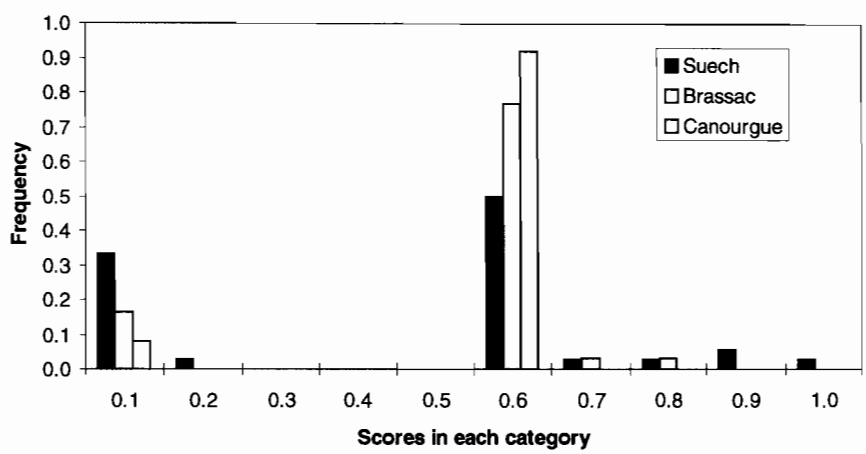

Fig. 5. Scores of the different hatcheries individuals in their own category for the leave-one-out with three groups corresponding to the three hatchery strains: Canourgue, Brassac and Suech.

Table 3
Percentage of incorrectly assigned individuals for each of the three comparisons between modern river and hatchery populations. Such individuals are assigned to the opposite category, e.g. 6% of Chiroulet individuals are classified in 'hatcheries'

Comparison	Chiroulet/hatcheries	Oussouet/hatcheries	Luz/hatcheries
river populations	6	8	5
hatcheries	5	3	3

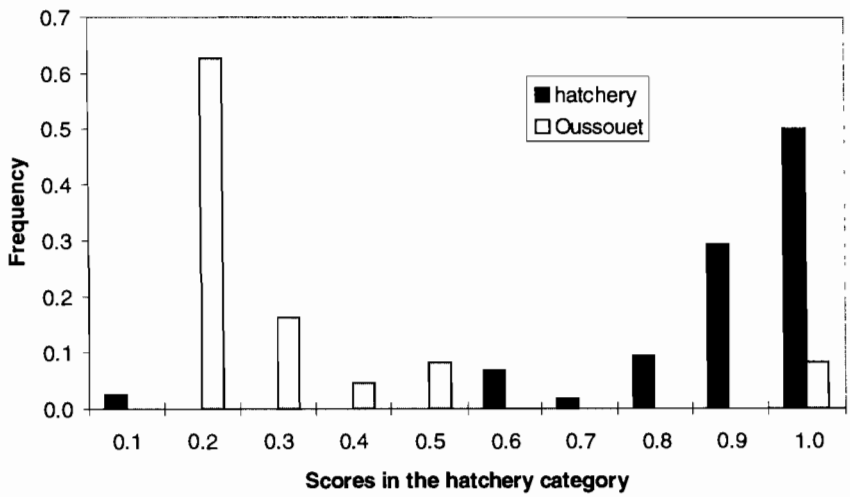

Fig. 6. Distribution of the scores in the hatchery class for the comparison between Oussouet (modern river population) and hatcheries (the three domestic strains analysed have been pooled). 1 = hatchery, 0 = Oussouet

tween 0.5 and 0.6 (75%) and a large proportion of fishes scored between zero and 0.1 (18%, corresponding to incorrectly classified individuals).

(2b) In the hold-out procedure, 2% of the individuals in the training set were not correctly classified, but the test showed 17% errors. This would tend to show that the network was suited to the features of the learning data set but not well suited to new data. There may be too small an overall difference between the different strains, preventing good application to new data.

3.3. The 'wild' modern populations

The percentages of incorrectly classified individuals for each of the three leave-one-out comparisons with hatchery samples (Luz, Chiroulet and Oussouet compared with domestic fishes) are given in Table 3. In the three river populations, the percentage of individuals assigned to domestic types varied from 5 (Luz) to 8% (Oussouet).

In the Oussouet/hatcheries comparison, the score distribution of Oussouet individuals in the hatchery category placed most individuals between 0.1 and 0.2 (Fig. 6), but with a large proportion between 0.2 and 0.5. Individuals with a result higher than 0.5 were all in the 0.9–one range and were then well assigned to hatcheries. Almost 80% of domestic individuals, scored between 0.8 and one. Individuals with a score lower than 0.5 were all in the 0–0.1 range and were then classified as Oussouet. It appeared to be more difficult to classify wild trout than domestic ones in this analysis.

3.4. Comparison of the other river populations with hatcheries

In the leave-one-out comparison between an ancestral population (such as Bastan) and hatcheries, we obtained 1% 'errors' in the domestic strains and 3% in the ancestral population. However, analysis of the score distribution in the hatcheries category (Fig. 7) shows that 97% of the Bastan fishes scored between 0.4 and 0.5; the remaining 3% corresponded to fishes classified as domestic (score between 0.9 and one in this group). In contrast, hatchery individuals are all well classified with scores between 0.6 and one in the hatchery category. The computation procedure may perhaps explain why no Bastan individual displayed a high score (between 0.8 and one) in its own category: as the time required by this technique is quite long, the number of iterations for the learning of each individual was limited to a maximum of 500. However, there must be a phenomenon making learning more difficult for this comparison than for the former ones. The same analysis was performed with Béhérobie (with some similar results to Bastan) and with Marcadau.

For Marcadau (Fig. 8), 40% of the individuals displayed a hatcheries category score of between 0.4 and 0.5; 60% scored between 0.9 and one and were then assigned to domestic type. These results agree well with morphological observations and with information from local managers, which tend to show that this population is quite heavily stocked. In this set of analyses, the scores of wild trout are limited to 0.5. However, if we agree that individuals with scores of between zero and 0.5 in the hatcheries category are wild trout, we can deduce that Marcadau is the population analysed that has been most modified by stocking.

4. Discussion

When the trout classes had been previously well defined using allozymes (comparison of ancestral and modern, tests (1a) and (1b)), the first analyses confirm that neural networks give good results when applied to microsatellite data despite all the problems usually associated with these markers, and especially the presence of rare alleles, ancestral polymorphism and homoplasy which means that some alleles of the same size are not always identical by descent (Jarne and Lagoda, 1996). Because of the high mutation rate of microsatellite loci (particularly for loci with a high number of alleles), and because of a possible relatively recent coancestry of the populations analysed (both natural and domestic), it is difficult to find diagnostic alleles separating wild and hatchery Atlantic populations and which could be used for

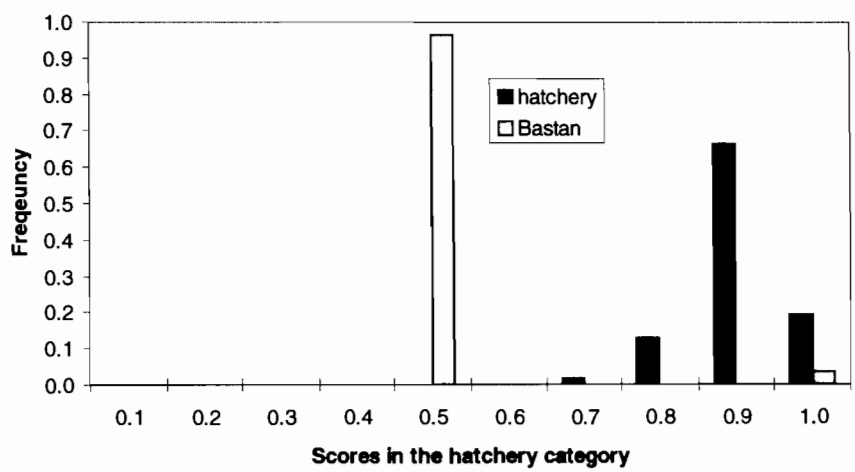

Fig. 7. Distribution of the scores in the hatchery category for the leave-one-out Bastan/hatcheries, 1 = hatcheries, 0 = Bastan.

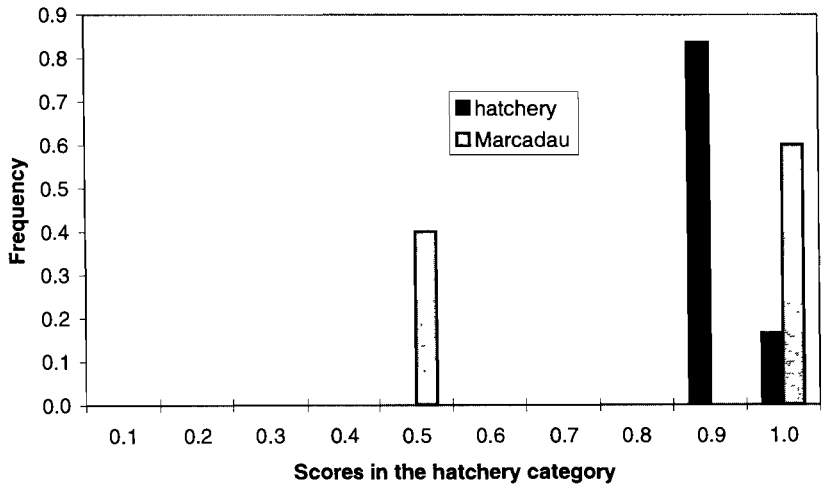

Fig. 8. Score distribution in the hatcheries category for the leave-one-out Marcadau/hatcheries. 1 = hatcheries, 0 = Marcadau.

several river drainage basins. For this reason, multilocus analysis is more useful and particularly ANN which can probably take into account quite small differences in allelic frequencies. Cornuet *et al.*, (1996) have already obtained good results in the classification of certain bee (*Apis mellifera*) lineages with microsatellite data and ANN.

4.1. Efficiency and utilisation of artificial neural networks

Homoplasy does not appear to drastically reduce the learning capacities of the neural network. For rare alleles, the graph showing the contributions of the different alleles according to their frequencies indicates that the most informative alleles are also often quite frequent; nevertheless, less frequent alleles may provide more information in some comparisons. In all cases, learning appears to be able to recognise the most discriminant information (for a particular comparison) among all the input variables, and this technique does not require any particular adaptation of the data. Neural networks gave some better results than classical discriminant analysis, as is shown by Cornuet *et al.* (1996).

For the first analysis (ancestral/modern comparison), the application of the network to populations other than those used for learning gave good results. The percentages of modern individuals predicted by the network agree well with the frequencies of modern alleles of *LDH-5**. The differences between these two parameters may be caused by different behaviour of the two markers, with randomly different introgression rates. The four supposed neutral microsatellite markers probably give a better description than a single allozymic (possibly selected) *LDH-5** marker. Moreover, one should keep in mind that these are a different type of information (allelic frequencies versus percentage of individuals).

Caution was required in this study because of the sample characteristics. Some samples (especially river populations) are or might be heterogeneous. Wild and domestic individuals may be found in the same sample of some of the 'modern' populations. For this reason, it was decided to test the leave-one-out procedure. It gave good results for the first comparison (ancestral/modern), and was then used for other comparisons. The technique appears well suited for the study of heterogeneous samples.

With both the leave-one-out and the hold-out procedures, neural networks associated with microsatellites confirm the distinction between modern Atlantic (wild or domestic) and ancestral Atlantic trout, which had previously only been analysed using allozymes.

4.2. Application to hatchery strains

Hatchery samples are needed as reference for assessing the proportion of domestic individuals in rivers. Analysis of these strains is necessary to evaluate the genetic diversity of domestic fishes; this shows whether the domestic samples analysed can be considered as representative of those used for stocking or if there is too much variability among hatcheries. Several studies have shown that these domestic strains were genetically quite similar (Guyomard, 1989; Garcia-Marin *et al.*, 1991), but we tried to verify this assumption using microsatellites and ANN. The high number of incorrectly classified individuals (both in the leave-one-out and hold-out results) underlines this homogeneity. The lack of differences may prevent good learning. It also shows that ANN can indicate when there is not enough differentiation between the categories used for learning as the network will not always give good percentages of correctly classified individuals, whatever is presented for learning. In our study, this homogeneity of domestic samples enabled us to pool them for the next analyses.

4.3. Characterisation of wild modern populations

Discriminating between wild and domestic modern Atlantic trout is an important objective. The identification of populations not or almost-not affected by stocking is useful for the protection and management of the genetic diversity of this species as this is threatened by stocking (Ferguson *et al.*, 1995). The use of the leave-one-out procedure for the comparison of each of the modern populations with hatchery populations gave low percentages of domestic individuals (from 5 to 8%) within the three modern populations (Chiroulet, Oussouet and Luz) which seemed to be mainly wild according to the morphological characteristics of their fishes. This would tend to show that these rivers are only modified by stocking slightly or not at all. Apart from this practical aspect, these results also show that neural networks are efficient even for genetically quite similar (but differentiated) entities.

The comparison of other samples with hatchery populations did not always give such clear results. For example, a large number of Bastan individuals had intermediate scores. This is probably linked with microsatellite properties and shared alleles, which in this case required more time for learning. However, individuals with intermediate scores could also be hybrid individuals and this raises the problem of how they are classified by the network. For example, in the Marcadau population (known to be heavily stocked), 40% of individuals displayed intermediate scores. This is probably the consequences of hybridisation of wild and domestic fishes; the strong impact of stocking on this population is confirmed by the percentage of individuals assigned to the domestic type (60%). This shows that when such individuals are present in a river population, the network is able to recognise them. There may be some hybrids in the Bastan population, (F1 or individuals resulting from backcrosses) even if allozymes indicate that it is a pure ancestral population; effectively, different markers can give different results because of selection and genetic drift. Moreover, as has already been explained, the training procedure may also cause this high proportion of intermediate scores. It should be noted that hardly any individuals in this population are clearly classified in the domestic category, as would have been expected in case of a high stocking impact (e.g. Marcadau). This population is probably not highly introgressed by domestic alleles.

Although the interpretation of these results is not as clear as for the former analyses, ANNs provided important information about the genetic composition of these populations.

5. Conclusion

From a technical point of view, our results confirm that ANNs are well suited to population genetics data. Effective analysis requires reference populations well chosen for the study, relatively balanced sample sizes and an appropriate validation procedure (hold-out or leave-one-out). For example, the leave-one-out procedure seems well

suited for mixed populations whereas the hold-out procedure gives a more precise idea of the prediction capability of the model. From a more fundamental point of view, this study confirms the presence in this area of several trout forms: two wild types (modern and ancestral) and one domestic form, which can coexist in the same river. Moreover, we identified certain pure or almost-pure wild populations. This raises the problem of their management and protection and is a new example of low stocking effectiveness. It is also an example of practical application of ANNs in ecology and population genetics.

Acknowledgements

This research was supported by the Bureau des Ressources Génétiques (grant No. 95011), the Conseil Supérieur de la Pêche (grant No. 9507127) and the Club Halieutique Interdépartemental. The field captures were performed by the local Fédérations de Pêche kindly assisted by scientists and students from ENSAT (Toulouse) and volunteers from Montpellier II University.

References

Aurelle, D., Berrebi, P., 1998 Microsatellite markers and management of brown trout *Salmo trutta fario* populations in south-western France. Génétique, Sélection, Evolution 30, S75–S90.

Cornuet, J.M , Aulagnier, S., Lek, S., Franck, P., Solignac, M., 1996. Classifying individuals among infra-specific taxa using microsatellites data and neural networks. C. R. Acad. Sci. Paris, Life sciences 319, 1167–1177.

Estoup, A., Presa, P , Krieg, F., Vaiman, D., Guyomard, R., 1993. CT)n and (GT)n microsatellites: a new class of genetic markers for *Salmo trutta* L. (brown trout). Journal of the Genetical Society of Great Britain 71, 488–496.

Ferguson, A., Taggart, J.B., Prodöhl, P.A., MacMeel, O., Thompson, C , 1995. The application of molecular markers to the study and conservation of fish populations, with special reference to *Salmo*. Journal of Fish Biology 47, 103–126.

Garcia-Marin, J.L., Jorde, P.E., Ryman, N., Utter, F., Pla, C., 1991. Management implications of genetic differentiation between native and hatchery populations of brown trout (*Salmo trutta*) in Spain. Aquaculture 95, 235–249.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artificial Intelligence Expert 6, 47–51.

Goh, A.T.C., 1995. Back-propagation neural networks for modelling complex systems. Artificial Intelligence Engineering 9, 143–151.

Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.

Guyomard, R., 1989. Diversité génétique de la truite commune. Bulletin Francais de Pêche et Pisciculture 314, 118–135.

Hamilton, K.E., Ferguson, A., Taggart, J B., Tomasson, T., Walker, A., Fahy, E., 1989 Post-glacial colonisation of brown trout, *Salmo trutta*. *Ldh*-5 as a phylogeographic marker locus. Journal of Fish Biology 35, 651–664.

Jarne, P., Lagoda, P.J.L., 1996. Microsatellites, from molecules to populations and back. Tree 11, 424–428.

Kohavi R., 1995. A study of cross-validation and bootstrap for estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., 1137–1143

Lek, S, Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquatic Living Resource 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Lauga, J , Aulagnier, S., 1996b. Application of neural networks to modelling non-linear relationships in ecology. Ecological Modelling 90, 39–52.

Mastrorillo, S., Lek, S , Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in river. Freshwater Biology 38, 237–246.

Poteaux, C., 1995 Interactions génétiques entre formes sauvages et formes domestiques chez la truite commune (*Salmo trutta fario* L.), Thesis of Université Montpellier II, France, 110 pp.

Rumelhart, D.E , Hinton, G.E , Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

ELSEVIER

Ecological Modelling 120 (1999) 325–335

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Predicting fish yield of African lakes using neural networks

Raymond Laë [a,*], Sovan Lek [b], Jacques Moreau [c]

[a] *Centre IRD de Brest, B.P. 70, 29280 Plouzané, France*
[b] *CNRS, UMR 5576, CESAC-Université Paul Sabatier, 118 Route de Narbonne 31062, Toulouse Cedex, France*
[c] *Labo d'Ingénierie Agronomique, ENSAT, INP, Av. de l'Agrobiopole, Auzeville-Tolosane, BP 107, 31326 Castanet-Tolosan Cedex, France*

Abstract

Artificial neural network (ANN) approaches to modelling and prediction of fish yield as related to the environmental characteristics were developed from the combination of six variables: catchment area over maximum area, fishing effort, conductivity, depth, altitude and latitude. For a total of 59 lakes studied, the correlation coefficients obtained between the estimated and observed values of abundance were significantly high with the neural network procedure (r adjusted = 0.95, $P < 0.01$). The predictive power of the ANN models was determined by the leave one out cross-validation procedures. This is an appropriate testing method when the data set is quite small and/or when each sample is likely to have 'unique information' that is relevant to the model. Fish yields estimated with this method were significantly related to the observed fish yields with the correlation coefficient reaching 0.83 ($P < 0.01$). Our study shows the advantages of the backpropagation procedure of the neural network in stochastic approaches to fisheries ecology. Using the specific algorithm, we can identify the factor influencing the fish yield and the mode of action of each factor. The limitations of the neural network approaches as well as statistical and ecological perspectives are discussed.

Keywords: Predictive modelling; Multiple regression; African lakes; Fish yield, Fisheries

1. Introduction

Understanding and predicting biological productivity is considered a key question by lake fisheries scientists. Several ecologists and fisheries managers have tried to determine the abundance of living stocks or the specific biodiversity in aquatic ecosystems using some of their characteristics, i.e. surface of the river drainage basin, surface area of lakes, flood plain areas, morpho-edaphic index, depth, coastal lines, primary production, etc. (Henderson and Welcomme, 1974; Ryder et al., 1974; Melack, 1976; Crul, 1992; Laë, 1992). In developing countries, the economical importance of fish and as a food source makes this topic particularly relevant.

Diverse multivariate techniques have been used to investigate how the various richness of fish is related to the environment, including several methods of ordination and canonical analysis, and univariate and multivariate linear, curvilin-

* Corresponding author. Fax: +33-2-98224514.
E-mail addresses: lae@ird.fr (R. Laë), lek@cict.fr (S. Lek), moreau@ensat.fr (J. Moreau)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00112-X

ear, and logistic regressions (Rawson, 1952; Hanson and Legget, 1982; Ryder, 1982; Schlesinger and Regier, 1982; Youngs and Heimbuch, 1982; Bernacsek and Lopes, 1984; Marshall, 1984; Welcomme, 1985, 1986; Payne and Harvey, 1989; De Silva et al., 1991; Moreau and De Silva, 1991; Payne et al., 1993). Complete and critical statistical methods reviewed by James and McCulloch (1990) assume that relationships are smooth, continuous, and either linear or involving simple polynomials. However, for quantitative analysis and more particularly for the development of predictive models of fish abundance, multiple linear regression and discriminate analysis have remained, the most frequently used techniques (Fausch et al., 1988; Jowett, 1993). These conventional techniques (based notably on multiple regression) are capable of solving many problems, but show sometimes serious shortcomings. This difficulty is that relationships between variables in sciences of the environment are often non-linear whereas methods are based on linear principles. Non-linear transformations of variables (logarithmic, power or exponential functions) allow to significantly improve results, even if it is still insufficient. However, the neural network, with the error backpropagation procedure, is at the origin of an interesting methodology which could be used in the same field as regression analysis particularly with the non-linear relations (Rumelhart et al., 1986). Nevertheless, few applications of this new technology in ecological sciences were published in contrast with the physical or chemical sciences (Smits et al., 1992; Lerner et al., 1994; Albiol et al., 1995; Faraggi and Simon, 1995).

Artificial neural networks (ANN) may be applied to different kinds of problems, e.g. pattern classification, interpretation, generalization or calibration. In this paper, neural networks have been used for multiple regression problems. The aim of this study was to analyze the level of relationships between some physical environmental parameters and the fish yield on African lakes, and also to propose the basis of the development of predictive tools using neural network methodology. We propose in order that, to analyze the level of relationships existing between some continuous physical environment variables and the fish yield.

2. Material and methods

2.1. Study sites and data

The 59 studied lakes are distributed all over Africa and Madagascar (Fig. 1). Currently available data on these lakes are insufficient. Most of them are old and/or just deal with survey periods sometimes less than 1 year. They came mainly from 'the source book for the inland fishery resources of Africa' (Burgis and Symoens, 1987; Bayley, 1988; Vanden Bossche and Bernacsek, 1990a,b, 1991; Crul, 1992; van der Knaap, 1994; Crul and Roest, 1995; Laë and Weigel, 1995a,b; Laë, 1997).

All data listed in the above quoted books have been used. When there were several annual surveys on one lake, we gave preference to the most recent data that had been controlled and updated. The choice of lakes focused on ecosystems the surface area of which was more than 10 km^2 in order to exclude too small or shallow water bodies that present specific modes of functioning and scanty data on fishing effort and catches.

For the 59 selected lakes, the characteristics were expressed in terms of latitude, altitude, morphometric parameters including catchment area/area ratio and average depth, physical and chemical parameters as conductivity. The productivity were expressed as annual fish yield (kg ha^{-1} $year^{-1}$) and the fishing effort as number of fishermen per km^2, that is the only relevant index for these lakes where fishing tackles and techniques can vary considerably.

2.2. Statistical analysis of data

Univariate, bivariate and multivariate analysis of data were performed by the SPSS Software® release 8 for Windows. The univariate analysis consisted of the determination of parametric (mean, standard deviation and coefficient of variation) and non-parametric (minimum, maximum, median and quartiles) statistical parameters. In the bivariate analysis, we studied the correlation between variables using Pearson's coefficients (values and probabilities of significance at 5 and 1% of confidence intervals). In the multivariate

analysis, the relationships between environmental characteristics and the fishing yield were studied with multiple regression analysis. Stepwise multiple linear regression procedures were applied. The diagnosis of the student residuals (normality and independence) was used to test the validity of the determination coefficient obtained (Weisberg, 1980; Tomassone et al., 1983).

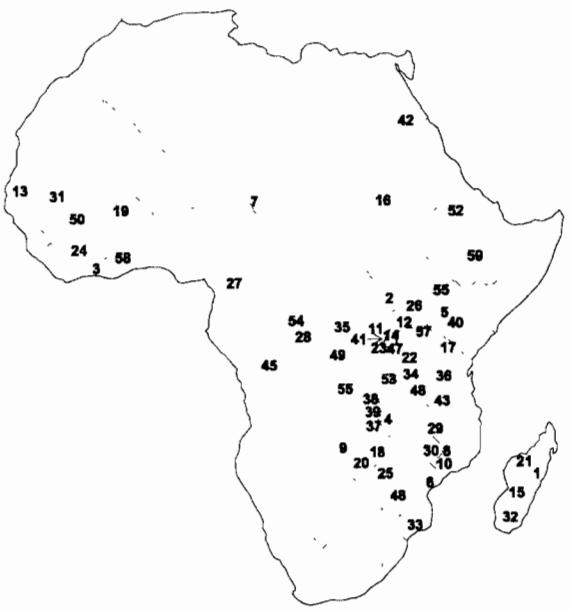

Fig. 1. Location of the 59 studied lakes, distributed in Africa and Madagascar. 1: Alaotra (Madagascar), 2: Albert (Zaire), 3: Ayame (Ivory coast), 4: Bangweulu (Zambia), 5: Baringo (Kenya), 6: Cahora Bossa (Mozambique), 7: Chad (Chad), 8: Chilwa (Malawi/Mozambique), 9: Chisi (Zambia), 10: Chiuta (Malawi/Mozambique), 11: Edward (Zaire), 12: George (Uganda), 13: Guiers (Senegal), 14: Ihema (Rwanda), 15: Itasy (Madagascar), 16: Jebel Aufia (Sudan), 17: Jipe (Kenya), 18: Kafue Flats/gorge (Zambia), 19: Kainji (Nigeria), 20: Kariba (Zambia), 21: Kinkony (Madagascar), 22: Kitangiri (Tanzania), 23: Kivu (Zaire), 24: Kossou (Ivory coast), 25: Xyle (Zimbabwe), 26: Kyoga (Uganda), 27: Lagdo (Cameroon), 28: Maji Ndombe (Zaire), 29: Malawi (Malawi), 30: Malombe (Malawi), 31: Manantali (Mali), 32: Mantasoa (Madagascar), 33: Massingir (Mozambique), 34: Mtera (Tanzania), 35: Mugesera (Rwanda), 36: Mujunju (Tanzania), 37: Mwadingusha (Zaire), 38: Mweru (Zaire), 39: Mweru wa Nt (Zaire), 40: Naivasha (Kenya), 41: Nasho (Rwanda), 42: Nasser (Egypt), 43: Nyumba Ya Mungu (Tanzania), 44: Nzilo (Zaire), 45: Pool Malebo (Congo/Zaire), 46: Robertson (Zimbabwe), 47: Rugwero (Burundi), 48: Rukwa (Tanzania), 49: Sake (Sake), 50: Selingue (Mali), 51: Sennar (Sudan), 52: Tana (Ethiopia), 53: Tanganyika (Zaire/Burundi), 54: Tumba (Zaire), 55: Turkana (Kenya), 56: Upemba (Zaire), 57: Victoria (Kenya), 58: Volta (Ghana), 59: Ziway (Ethiopia).

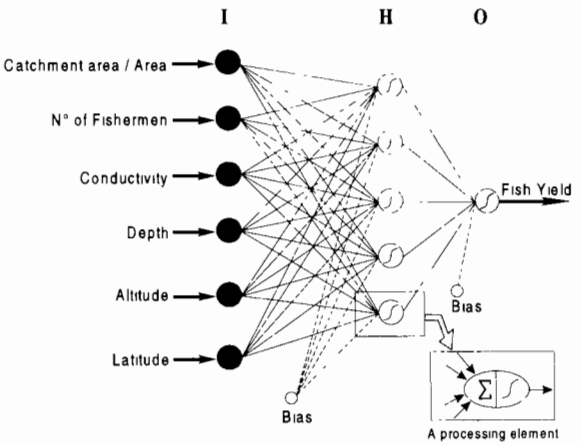

Fig. 2. Typical three-layered feedforward artificial neural network. Six input nodes corresponding to six independent environmental variables, five hidden layer nodes and one output node corresponding to the estimate of fish yield. Connections between nodes are shown by solid lines: they are associated with synaptic weights that are adjusted during the training procedure. The bias nodes are also shown, with 1 as their output value. The sigmoid activation functions are plotted within the node

2.3. *Artificial neural network (ANN) processing*

The multilayer feedforward neural network is one of the most popular network structures among all the ANN diagrams. The processing elements in the network are called neurons (or nodes or units). All the neurons in a multilayer feedforward neural network are arranged so that they have a layered structure. A typical three-layer feedforward ANN is shown in Fig. 2. The first layer connects with the input variables and is called the input layer. Here, it comprises six neurons (six independent variables). The last layer connects to the output variables and it is called the output layer of only one neuron (the dependent variable). Layers in-between the input and output layers are called hidden layers; there can be more than one hidden layer. The number of neurons of the hidden layer is an important parameter of the network. The empirical approach for the selection of the network consists of

a test for the number of different possible configurations and the selection of that which provides the best compromise between bias and variance (Geman et al., 1992; Kohavi, 1995), which is the training that gives a good generalization. In our study, a network with one hidden layer of five neurons has been retained (network with two hidden layers have also been tested, but the results do not differ significantly).

Each of the neurons is connected to the neurons of neighboring layers. The parameters associated with each of these connections are called weights. All connections are fed forward; that is, they allow information transfer only from an earlier layer to the next consecutive layers. No feed-back connections are permitted in these 'feed-forward' networks. Neurons within a layer are not interconnected, and neurons in nonadjacent layers are not connected. Considering an input vector $\boldsymbol{x}_i = (\boldsymbol{x}_{i0}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ip})$ for ith record, with $\boldsymbol{x}_{i0}$ always equal 1 which corresponds to the bias. The vector linking the input units to hidden units can be noted as $w_h = (w_{h0}, w_{h1}, \ldots, w_{hp})$. The incoming signal of the hidden layer for the hth neuron is the linear projection $z = w_h x_i$. The effective incoming signal z, is passed through a nonlinear activation function (called a transfer function or activation function) to produce the outgoing signal y^h of the hidden neuron, $y^h = f(w_h x_i)$ with f a transfer function $y^h = f(z) = 1/(1 + \exp(-z))$. In this study, the sigmoid function is preferred as compared to linear or threshold type functions. The same operation is repeated for the output layer, with values for the sigmoid function derived from the sum of the product of the outgoing signals from the hidden layer and the weight binding the hidden layer with the output layer. The outgoing signal of the output layer provides the predicted values of the network, i.e. the fish yield in this study.

ANNs are generally trained by the backpropagation algorithm (Rumelhart et al. 1986). The training is a method that determines values of network parameters which allow a good estimation of $\hat{y}$, values of the outgoing signals from the y network. The backpropagation algorithm assesses y repeatedly by a method of gradient descent. The training of the network starts with weights stemming from a random selection between -0.3 and 0.3. Adjustment of these weights is made according to the importance of the error $(y - \hat{y})$. Several repetitions of data are necessary to guarantee the convergence of estimated values (weak error as compared to observed values), without obtaining an overfit. The number of iterations was limited to 500. The compact form of feedforward ANN made the programming of the algorithm much easier, especially when using some matrix based software packages, e.g. Matlab® for Windows®.

In order to compare the results obtained with multiple linear regression and with neural network, an application was made on the whole database (59 units). Then, to justify the predictive quality of the ANN models, a leave one out procedure (Efron 1983; Jain et al. 1987) was used. The principle of this validation was to assess the assignment of each of the 59 individuals, the learning phase being performed with the other 58. It concerned in fact a cross-validation with the number of records reserved for the test limited to a unit at each time. This procedure is useful in cases where one has a weak quantity of observations.

2.4. Sensitivity of input variables

A disadvantage of ANN in comparison with MLR models is their lack of explanations regarding the relative importance of each independent variable considered. MLR analysis can identify the contribution of each individual input in determining the output and also can give some measures of confidence about the estimated coefficients. In addition, there is currently no theoretical or practical way of accurately interpreting the weights in ANN (Smith, 1994). For example, weights cannot be interpreted as a regression coefficient nor can difficulty be used to compute causal impacts or elasticity. Therefore, ANN are generally better suited for forecasting or predicting rather than for policy analysis. In ecology, however, it is necessary to know the impacts of each explanatory variable. Some authors have proposed methods which allow the determination of the impact of variables initially applied to the

Table 1
Statistical parameters of the variables studied[a]

	Min	Q1	Median	Q3	Max	Mean	SD	CV
Catchment area/area ratio	0.97	9 1	43.8	170	6813	337.2	983.2	292
Fishing effort	0.1	0.5	1.4	2 9	28.6	2.7	4.1	155
Conductivity	1	80	165	379	3300	358	588	164
Depth	0.3	3.0	5.0	15 7	570.0	29.3	94 9	324
Altitude	1	300	663	1160	1890	727	492	68
Latitude	0	2	8	14	24	8.5	6.2	73
Fish yield	1.2	22.4	52.1	77.3	252.9	59.1	51.8	88

[a] Q1, Q3, first and third quartile; SD, standard deviation; CV, coefficient of variation expressed as a percentage.

model (Dimopoulos et al. 1995; Garson 1991; Goh 1995; Lek et al. 1996a,b). In this work, an experimental approach has been used to determine the response of the model to each of the input variables separately by applying a typical range of variation of a single 'free' variable to the model, while the other ('blocked' variables) are held constant. The contribution of each environmental variable to fishing yield estimation was calculated using 12 values evenly spaced over the range between the minimum and the maximum that appeared in the set of data. The remaining 'blocked' variables were provisionally set at an arbitrary level. Because this level influenced the results, we set the remaining variables simultaneously together at their minimum value, first quartile, median, third quartile and maximum successively. Five responses were thus obtained for each of the 12 'free' variable values. They were further reduced to their median value. The operation was repeated for all of the environmental variables.

3. Results

3.1. Statistical parameters of variables

Table 1 shows a very large variability within the data. The coefficients of variation are high ranging from 100 to 200% for fishing effort and conductivity, 292% for the catchment area/area ratio, and 324% for mean depth. Among explanatory variables, the only ones that have coefficients of variation smaller than 100% are latitude and altitude and even these variables reach values of around 70%. These results confirm the heterogeneity and the diversity of the studied lakes.

The dependent variable (i.e. yield) varies from 1.2 to 253 kg ha^{-1} $year^{-1}$, with an average of 59 kg ha^{-1} $year^{-1}$. Such yields depend both on biotic capacities of the different ecosystems studied and fishing pressure. Low fishing effort mainly explains a low yield since the variable studied only gives information on the level of catches and not at all on the actual abundance of fish. The coefficient of variation (88%) confirms a large variability in yield. Fig. 3 shows that very high values of yield are rare, which is a very usual result in ecology (Verner et al. 1986).

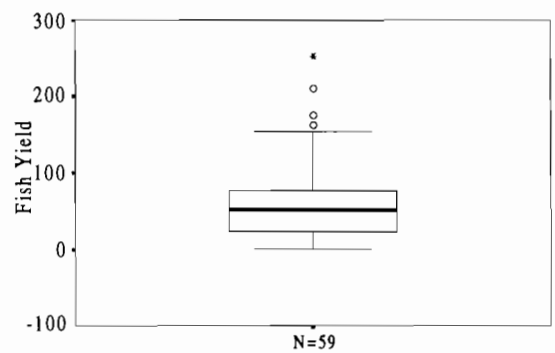

Fig. 3. Descriptive statistics of the variable Fish Yield: Boxplot representation. A circle designates an outlier values (values more than 1.5 box-lengths from 75th percentile), and an asterisk indicates extreme values (values more than three box-lengths from 75th percentile).

Table 2
Pearson correlation matrix between studied variable with two-tail significance of probability[a]

	Catchment area/area	Fishing effort	Conductivity	Depth	Altitude	Latitude	Fish yield
Catchment area/area		Ns	Ns	Ns	Ns	Ns	Ns
Fishing Effort	0.183		Ns	Ns	Ns	Ns	**
Conductivity	−0.139	−0.140		Ns	Ns	Ns	Ns
Depth	−0.085	−0 132	0.098		Ns	Ns	Ns
Altitude	−0.245	0.007	0.068	0.013		Ns	Ns
Latitude	−0.098	0 111	−0.208	−0.030	−0.107		Ns
Fish yield	0.043	0.569	−0.102	−0.212	−0.112	−0.037	

[a] Ns, not significant, $P>0.05$.
** Highly significant, $P<0.001$.

3.2. Relationship between fish yield and environmental variables

Fish yield was significantly related to only one variable (Table 2): Fishing Effort ($r=0.57$; $P<0.01$). With other variables, the correlation coefficient is weak, negative values with conductivity, depth, altitude, latitude ($|r|<0.21$; $P>0.05$) and positive only with the catchment area/area ratio ($r=0.04$; $P>0.05$). The relationship between yield and fishing effort explains only a low percentage of variance (32%). Among independent variables, the correlation was not significant for all of variables ($P>0.05$).

3.3. Multiple regression analysis

The comparison between MLR predictive power and ANN is not quite fair, unless the number of parameters (coefficients) of the MLR model is almost the same as ANN. A MLR was performed in order to check if a significant correlation could be obtained with this classical linear method. For the 59 samples, the stepwise procedure performed with SPSS selected only one variable at one step: Effort ($r=0.57$, $F_{1,57}=27.33$, $P<0.001$). With all of the six environmental variables, we obtained a correlation coefficient of only 0.62 ($F_{6,52}=5.45$, $P<0.001$). Low correlation coefficient testify the low percentages of explained variance (32% in stepwise regression). The supplementary variable addition as compared to the stepwise regression contributes only very little to the improvement of results (38% of explained variance).

In order to completely full file the requirement of MLR method (i.e. a normal distribution of variables considered) the fish yield and the six independent variables were transformed to their log10. The result of MLR show a correlation coefficient of 0.81, i.e. higher than before log transformation.

3.4. Neural network

In a first step, we developed a model with the 59 available lakes. In order to avoid possible overfitting, several tests were carried out with different configurations of the neural network (change in the number of neurons of the hidden layer). The configuration that had a minimal dimension and which gave satisfying results was retained. In this study, the number of neurons in the hidden layer of the network was fixed at five. To avoid again overfitting, the number of iterations was limited to 500, which is quite low in neural network modelling. The resulting correlation coefficient was 0.95 for the regression between observed and estimated values (Fig. 4), indicating that the ANN provided satisfactory results over the whole set of values for the dependent variable. The points are well aligned on the diagonal of the perfect fit line (co-ordinate 1:1). The linear adjustment between observed and estimated values gives a slope practically equal 1 ($y=0.8981x+4.82$). Although weakly repre-

sented, the strong values of the output variable are aligned around this same perfect fit line, with a few outliers (Fig. 4a). Some weak values were slightly overestimated.

Residuals have an average of 1.2 and a standard deviation of 16 with the minimum value of -55.7, and the maximum 39. In order to test the normality of model residuals, the statistical test of Lilliefors (1967) was applied. With 59 observations, the limit values of the test for the rejection of the hypothesis of normality were 0.115 for $\alpha = 0.05$ and 0.134 for $\alpha = 0.01$. Lilliefors test of normality gave a maximum difference of 0.099, $P = 0.15$. The study of the relationship between residuals and values estimated by the model showed complete independence (Fig. 4b). The coefficient of determination was negligible ($r^2 = 0.0004$) and the slope of correlation between estimated values and residuals close to 0 ($y = 0.0067x + 0.8171$); the residuals were well distributed on either side of the horizontal line (ordinate) representing the residual mean.

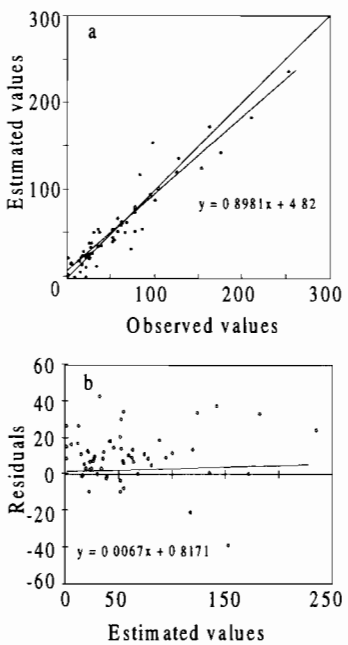

Fig. 4. Results of fitting the model with 59 observations and a 6-5-1 network. (a) Scatter plot of estimated values vs. predicted values. The solid line indicates the perfect fit line. (b) Relationship between residuals and estimated values.

3.5. Neural network sensitivity

The influence of the six independent environmental variables on the fish yield in the ANN modelling is illustrated by six curves (Fig. 5):

- Catchment area/area ratio (Fig. 5a): The relationship between yields and catchment area/area ratio is monotonously growing. It appears that smaller lakes situated in larger catchment areas are more productive.
- Number of fishermen (Fig. 5b): There is an increase of fishing yield in relationship with fishing effort. First, fish yield increases rapidly with the fishing Effort. After that, it stabilizes over level of 200 kg ha^{-1} $year^{-1}$ from 15 fishermen km^{-2} characterized by a practically horizontal line.
- Conductivity (Fig. 5c): there is an increase contribution: the fish yield increases rapidly when the value of the independent variable increases. Beyond 2000 µs cm^{-1}, it stabilizes for Conductivity. This profile is similar to the one of previous case with a lower amplitude.
- Depth (Fig. 5d): There is a linear decrease between fish yield and depth from 230 kg ha^{-1} $year^{-1}$ for very shallow lakes to 50 kg ha^{-1} $year^{-1}$ for deeper ones (500 m). The profile is represented practically by a line of almost constant slope.
- Altitude (Fig. 5e): Fish yield versus altitude displays a skewed-to-the-right profile. The maximum of contribution is situated at around 500 m of altitude, and decreases at higher altitudes. Altitude interacts weakly with fish yield despite the temperature differences which can reach 11°C between sea level and the highest lake.
- Latitude (Fig. 5f): Variations of fish yield with latitude are linearly growing. When the latitude increases from equator to 25° north or south, the increase in fish yield is only about 100 kg ha^{-1} $year^{-1}$.

3.6. Testing of the network

The predictive power of the ANN models was determined by the leave one out procedures. Leave-one-out cross-validation is appropriate when the data set is quite small and/or when each

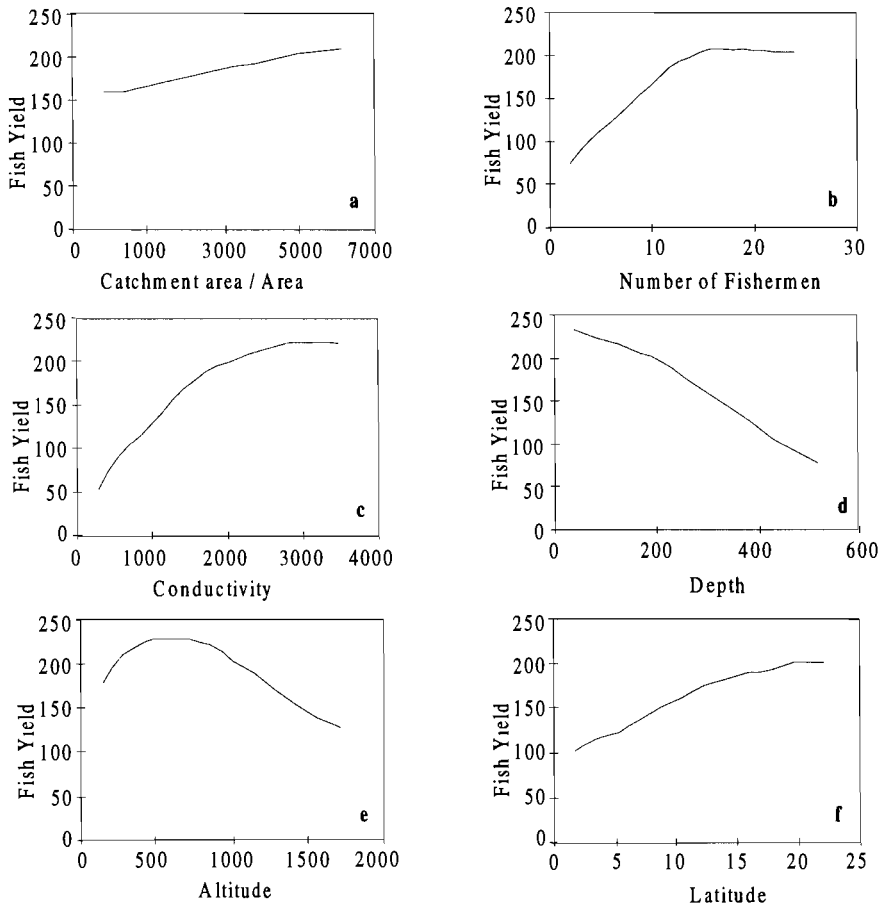

Fig. 5. Sensitivity profiles (or 'responses') of the predicted value of fish yield to each of the six independent variables. Each independent variable is tested versus the five other variables placed at one of five standard levels (minimum, 1st quartile, median, 3rd quartile, maximum).

sample is likely to have 'unique information' that is relevant to the regression model. For the leave one out procedure, the predictive performance was shown in Fig. 6a. By testing one record at each time on a model established from 58 remaining records, very good results were observed: the correlation coefficient was 0.831. This coefficient does not reflect entirely the result. The graph of correlation between observed and predicted values showed the majority of records were aligned on the diagonal of co-ordinate 1:1, despite the slope significantly different to 1 ($y = 0.6389x + 22.249$). Some overestimates of some weak values were possibly observed. The three high values were slightly underestimated. This was the consequence of the scarcity of high values in the database for an effective learning of the model.

Residuals have an average of -0.9 and a standard deviation of 29 with the minimum value of -92, and the maximum 100. Lilliefors test of normality gave a maximum difference of 0.337, $P < 0.001$. The study of the relationship between residuals and values estimated by the model showed complete independence (Fig. 6b). The coefficient of determination was negligible ($r^2 = 0.01$) with the slope of correlation coefficient between predicted values and residuals close to 0 ($y = 0.0806x - 5.7405$); the residuals were well distributed on either side of the horizontal line (ordinate) representing the residual mean.

4. Discussion and conclusion

Yield fish studied here have been reliably fitted to the easily measured environmental characteristics. Thus, variations in fish yield are strongly connected to a set of six environmental variables.

The theoretical advantage of conventional MLR models over ANN is that their parameters provide information about the relative importance of the independent variables (although this is not true when composite variables are used). However, the same results can be obtained by performing a sensitivity analysis of the ANN. Garson (1991), Goh (1995) have proposed the methods for interpreting neural networks connection weights to illustrate the explanatory variable importance inside the ANN. These studies demonstrated the potential of ANN approach for capturing non-linear interactions between variables in complex engineering systems and propose the procedure for partitioning the connection weights in order to determine the relative importance of the various input variables. Dimopoulos et al. (1995) propose the study of the first partial derivatives of the ANN's output with respect to each input is used to identify of the factors influencing the dependent variable and the mode of action of each factor. In ecology, Lek et al. (1995, 1996a,b) proposed an algorithm allowing the visualization of the profiles of explanatory variables. Aside from the predictive value of the model, an attempt was made to detect by a simple simulation method the sensitivity of the different variables.

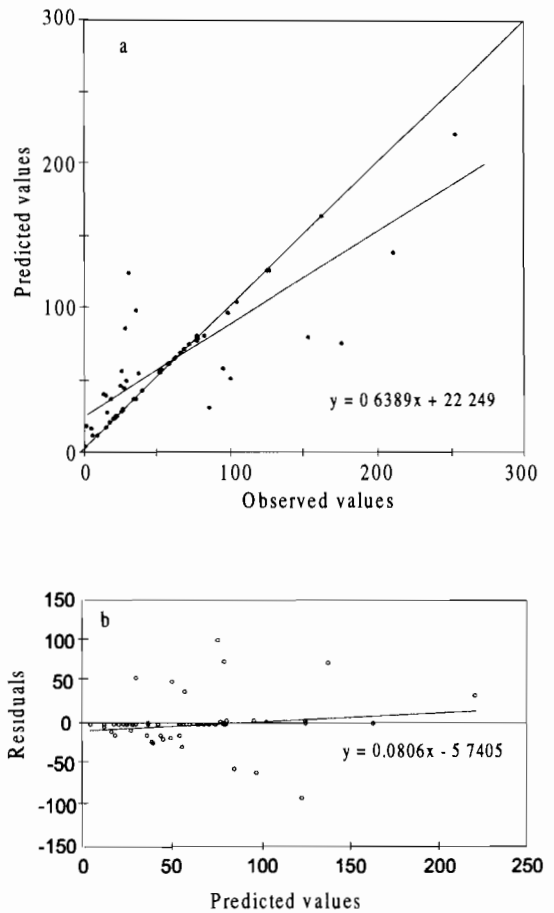

Fig. 6. Result of testing the model with 59 observations and a 6-5-1 network by the leave-one-out procedure. (a) Scatter plot of predicted values vs. observed values. The solid line indicates the perfect fit line. (b) Relationship between residuals and predicted values.

The main processes that determine biodiversity indices can be approximated by linear or simple non-linear (e.g. logarithmic) functions only to a limited extent. Therefore, such models are not able to reproduce the behaviour of real systems when very low or high values of the variables are considered (Lek et al. 1996b). In fish ecology, several models, based on MLR principle were proposed by several authors (Fausch et al. 1988). To improve the results, non-linear transformations of independent or/and dependent variables were frequently used. However, despite these transformations of variables, results obtained remained often insufficient. Moreover, ANN with only one hidden layer can model non-linear systems in ecology whatever is their complexity (Goh, 1995; Lek et al., 1996b; Scardi, 1996). Complex systems obviously need complex networks (more units in the hidden layer or more than one hidden layers), adequate training and a large data set to be modelled.

Multiple regression analysis and back propagation of the ANN were both used to develop stochastic models of fish yield prediction using habitat features on a macrohabitat scale (Lek et al. 1996b). This stochastic approach required an extensive database and care to obtain reliable

models. The selection of input variables, their ecological significance and the use of a test data set to assess the model precision and accuracy are important elements of this type of approach (Fausch et al. 1988). The advantage of ANN over MLR models is the ability of ANN to directly take into account any non-linear relationships between the dependent variables and each independent variable. Several authors have shown greater performances of ANN as compared to the MLR (Ehrman et al. 1996; Lek et al. 1996b; Scardi 1996). The backpropagation procedure of the ANN gave very high correlation coefficients comparing to the more traditional models, especially for the training calculation. In the test set, correlation coefficients were lower than in training but still remained clearly significant. This difference between training and testing sets is more amplified when the data set is small, and when each sample is likely to have 'unique information'; this is relevant to the model.

Through the present example taken in fish yield, we show that ANN models are viable when compared to traditional statistical methodologies. The ANN has demonstrated here a promising potential in ecology, as a tool to evaluate, understand, predict and manage African open fisheries. In any lakes, not already included in our database, the yield will be computed by introducing the six independent variables for these lakes in the model.

References

Albiol, J., Campmajo, C., Casas, C., Poch, M., 1995. Biomass estimation in plant cell cultures: a neural network approach. Biotechnol. Prog. 11, 88–92.

Bayley, P.B., 1988. Accounting for effort when comparing tropical fisheries in lakes, river–floodplains, and lagoons. Limnol. Oceanogr. 33, 963–972.

Bernacsek, G M., Lopes, S , 1984. Mozambique. Investigations into the fisheries and limnology of Cahora Bassa Reservoir seven years after dam closure. FAO Mozambique, GCP-006-SWE, Field Document. 9, Rome, p. 145.

Burgis, M.J., Symoens, J.J., 1987 African wetlands and shallow water bodies. Travaux et Documents 211, ORSTOM Paris, p. 651.

Crul, R.C.M., 1992. Models for estimating potential fish yields of African inland waters. FAO, CIFA Occasional Paper 16, p. 22.

Crul, R.C.M., Roest, F.C., 1995. Current status of fisheries and fish stocks of the four largest African reservoirs Kainji, Kariba, Nasser/Nubia and Volta. FAO, CIFA Technical Paper 30, p. 134

De Silva, S.S., Moreau, J., Amarasinghe, U.S., Chookajorn, T., Guerrero, R D., 1991. A comparative assessment of the fisheries in lacustrine inland waters in three Asian countries based on catch and effort data. Fish. Res. 11, 177–189.

Dimopoulos, Y , Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Process. Lett. 2 (6), 1–4.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–330.

Ehrman, J.M., Clair, T.A , Bouchard, A., 1996. Using neural networks to predict pH changes in acidified Eastern Canadian lakes. Artif Intell. Appl. 10, 1–8.

Faraggi, D., Simon, R , 1995. A neural network model for survival data. Stat. Med. 14, 73–82.

Fausch, K.D., Hawkes, C.L., Parsons, M.G., 1988. Models that predict the standing crop of stream fish from habitat variables. U.S. Forest Service General Technical Report PNW-GTR, p. 213.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artif Intell. Expert 6, 47–51.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4, 1–58.

Goh, A.T.C., 1995. Back-propagation neural networks for modelling complex systems. Artif. Intell. Eng. 9, 143–151.

Hanson, J.M., Legget, W.C., 1982. Empirical prediction of fish biomass and yield. Can J. Fish. Aquat. Sci. 39, 257–263.

Henderson, H F , Welcomme, R.L., 1974. The relationship of yield to morpho-edaphic index and numbers of fishermen in African inland fisheries. FAO, CIFA Occasional Paper 1, p. 19.

Jain, A.K., Dube, R.C , Chen, C., 1987. Bootstrap techniques for error estimation. IEEE Trans. Patt Anal. Mach. Intell. PAMI 9, 628–633.

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics. panacea or Pandora's box? Ann. Rev Ecol. Syst. 21, 129–166

Jowett, 1993. A method for objectively identifying pool, run, and riffle habitats from physical measurements N.Z.. J. Mar. Freshw. Res. 27, 241–248.

Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection Proceeding of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, pp. 1137–1143.

Laë, R., 1992. Influence de l'hydrologie sur les pêcheries du Delta Central du Niger de 1966 à 1989. Aquat. Living Resour. 5, 115–126.

Laë, R., 1997. Estimation des rendements de pêche des lacs Africains au moyen de modèles empiriques. Aquat. Living Resour. 10, 83–92.

Laë, R , Weigel, J.Y., 1995a. Diagnostic halieutique et propositions d'aménagement: l'exemple de la retenue de Sélingué (Mali). FAO-PAMOS, p. 73.

Laë, R., Weigel, J.Y., 1995b. La retenue de Manantali au Mali, diagnostic halieutique et propositions d'aménagement FAO-PAMOS, p. 65.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar Freshw. Res 46, 1229–1236.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquat. Liv. Res. 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996b. Application of neural networks to modelling nonlinear relationships in ecology. Ecol Model. 90, 39–52.

Lerner, B., Guterman, H., Dinstein, I., Romem, Y., 1994. Feature selection and chromosome classification using a multilayer perceptron neural network Proceedings of the IEEE International Conference on Neural Networks, Orlando, FL, pp. 3540–3545.

Lilliefors, 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown, J. Am. Stat. Assoc., 62. 399–402.

Marshall, B.E., 1984 Towards predicting ecology and fish yields in African reservoirs from pre-impoundment physico–chemical data. FAO, CIFA Technical Paper 12, p. 36.

Melack, J.M., 1976. Primary productivity and fish yields in tropical lakes. Trans. Am. Fish. Soc. 105, 575–580.

Moreau, J., De Silva, S.S., 1991. Predictive fish yield models for lakes and reservoirs of the Philippines, Sri Lanka and Thailand. FAO, Fisheries Technical Paper, 319, p. 42.

Payne, A.I., Harvey, M.J., 1989. An assessment of the *Prochilodus platensis* Holmberg population in the Pilcomayo river fishery, Bolivia using scale-based and computer-assisted methods. Aquac. Fish. Manag. 20, 233–248.

Payne, A.I., Crombie, J., Halls, A.S., Temple, S A., 1993. Synthesis of simple predictive models for tropical river fisheries, London, MRAG Ltd, p. 92.

Rawson, D.S., 1952. Mean depth and the fish production of large lakes. Ecology 33, 513–521.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

Ryder, R.A, Kerr, S.R., Loftus, K.H., Regier, H.A., 1974. The morpho-edaphic index, a fish yield estimator Review and evaluation J Fish. Res. Board Can. 31, 663–688.

Ryder, R.A , 1982 The morpho-edaphic index—use, abuse and fundamental concepts Trans Am. Fish Soc 111, 154–164.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Prog. Ser. 139, 289–299.

Schlesinger, D.A., Regier, H.A., 1982. Climatic and morpho-edaphic indices of fish yields from natural waters Trans. Am. Fish. Soc. 111, 141–150.

Smith, M., 1994. Neural networks for statistical modelling. Van Nostrand Reinhold, New York, p. 235.

Smits, J.R.M., Breedveld, L.W., Derksen, M W.J., Katerman, G., Balfoort, H.W., Snoek, J., Hofstraat, J.W., 1992. Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. Anal. Chim. Acta 258, 11–25.

Tomassone, R , Lesquoy, E., Miller, C., 1983. La régression, nouveaux regards sur une ancienne méthode statistique. INRA (Activités scientifiques et agronomique no. 13), Paris, France, p. 188.

van der Knaap, M , 1994. Status of fish stocks and fisheries of thirteen medium-sized African reservoirs FAO, CIFA Technical Paper, 26, p. 107.

Vanden Bossche, J P., Bernacsek, G.M., 1990a. Source book of the inland fishery resources of Africa, FAO, CIFA Technical Paper 18/1, p. 411

Vanden Bossche, J.P., Bernacsek, G.M , 1990b Source book of the inland fishery resources of Africa, FAO, CIFA Technical Paper 18/2, p. 240

Vanden Bossche, J.P. and Bernacsek, G.M., 1991. Source book of the inland fishery resources of Africa, FAO, CIFA Technical Paper 18/3, p. 219.

Verner, J., Morrison, M.L., Ralph, C J , 1986. Wildlife 2000: modelling habitat relationships of terrestrial vertebrates. Univ. Wisconsin Press, Madison, WI, p. 470.

Weisberg, S., 1980. Applied linear regression. Wiley, New York, p. 324

Welcomme, R L., 1985. River fisheries. FAO Fisheries Technical Paper 262, p. 330.

Welcomme, R.L., 1986. The effects of the Sahelian drought on the fishery of the central delta of the Niger river. Aquac Fish. Manag. 17, 147–154.

Youngs, W D., Heimbuch, D.G., 1982. Another consideration of the morpho-edaphic index Trans. Am. Fish Soc. 111, 151–153.

ELSEVIER

Ecological Modelling 120 (1999) 337–347

ECOLOGICAL MODELLING

www.elsevier com/locate/ecomodel

Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird

Stéphanie Manel [a,*], Jean-Marie Dias [b], Steve J. Ormerod [c]

[a] *UPRES 159, Université de Pau et des Pays de l'Adour, UFR Sciences et Technologie, 1 rue de Donzac, 64100 Bayonne, France*
[b] *UPRES-A-5033, Université de Pau et des Pays de l'Adour, UFR Sciences et Technologie, 1 rue de Donzac, 64100 Bayonne, France*
[c] *Catchment Research Group, School of Biosciences, Cardiff University, PO Box 915, Cardiff CF1 3TL, UK*

Abstract

We assessed the occurrence of a common river bird, the Plumbeous Redstart *Rhyacornis fuliginosus*, along 180 independent streams in the Indian and Nepali Himalaya. We then compared the performance of multiple discrimant analysis (MDA), logistic regression (LR) and artificial neural networks (ANN) in predicting this species' presence or absence from 32 variables describing stream altitude, slope, habitat structure, chemistry and invertebrate abundance. Using the entire data (= training set) and a threshold for accepting presence in ANN and LR set to $P \geq 0.5$, ANN correctly classified marginally more cases (88%) than either LR (83%) or MDA (84%). Model performance was assessed from two methods of data partitioning. In a 'leave-one-out' approach, LR correctly predicted more cases (82%) than MDA (73%) or ANN (69%). However, in a holdout procedure, all the methods performed similarly (73–75%). All methods predicted true absence (i.e. specificity in holdout: 81–85%) better than true presence (i.e. sensitivity: 57–60%). These effects reflect species' prevalence (= frequency of occurrence), but are seldom considered in distribution modelling. Despite occurring at only 36% of the sites, Plumbeous Redstarts are one of the most common Himalayan river birds, and problems will be greater with less common species. Both LR and ANN require an arbitrary threshold probability (often $P = 0.5$) at which to accept species presence from model prediction. Simulations involving varied prevalence revealed that LR was particularly sensitive to threshold effects. ROC plots (received operating characteristic) were therefore used to compare model performance on test data at a range of thresholds; LR always outperformed ANN. This case study supports the need to test species' distribution models with independent data, and to use a range of criteria in assessing model performance. ANN do not yet have major advantages over conventional multivariate methods for assessing bird distributions. LR and MDA were both more efficient in the use of computer time than ANN, and also more straightforward in providing testable hypotheses about environmental effects on occurrence. However, LR was apparently subject to chance significant effects from explanatory variables, emphasising the well-known risks of models based purely on correlative data.

Keywords: Neural networks; Logistic regression; Presence–absence data; River birds

* Corresponding author. Fax: + 33-2-59591936.
E-mail address: stephanie.manel@univ-pau.fr (S. Manel)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00113-1

1. Introduction

With clear relevance to resource assessment, environmental conservation, and biological monitoring, models of species presence and absence are of undoubted importance (Jongman et al., 1995; Fielding and Bell, 1997). Increasing focus on global and regional patterns of biodiversity prompt the need for modelling of this type at broad spatial scales, but methods are still evolving (Ricklefs and Schluter, 1993; Gaston, 1998). Traditionally, models used in ecology to predict species abundance have been based on linear relationships with environmental variables. Data in turn are assumed to have normal errors, appropriate for example in linear regression, multiple regression and multiple discriminant analysis. Difficulties in satisfying these assumptions have often raised statistical and theoretical concerns (Austin and Meyers, 1996; Lek et al., 1996a), so that new modelling paradigms are now being promoted (Venables and Ripley, 1997). They include linear methods such as logistic regression, which accommodates binomial error, and is already in wide use (Osborne and Tigar, 1992; Green et al., 1994; Austin and Meyers, 1996). By contrast, artificial neural networks, characterised by their ability to model non-linear relationships, are more novel in ecology (see Mastrorillo et al., 1997).

With such a range of approaches available for modelling, it is potentially difficult for practising ecologists to choose appropriate methods. Moreover, methods for comparing model performance are also evolving. This applies even to the relatively straightforward need to model species' presence or absence, where methods are often evaluated solely on prediction error–the number of cases in which species presence or absence is correctly assessed (e.g. Buckton and Ormerod, 1997; Fielding and Bell, 1997).

Clearly one of the greatest needs at present is for clear conclusions from comprehensive studies which compare model performance, but surprisingly few are available (e.g. Mastrorillo et al., 1997). In this paper, we therefore provide such a comparison illustrated from the distribution of one species of river bird, the Plumbeous Redstart, using data collected from a large area of the Himalayan mountains during 1994–96 (see below). We derived algorithms which modelled and predicted their distribution from sub-sets of 32 possible environmental variables, and compared the performances of multiple discriminant analysis (MDA), logistic regression (LR) and artificial neural networks (ANN). Our comparison largely follow recent protocols proposed by Fielding and Bell (1997).

The work is realistic, forming part of a larger study which aims to assess natural and anthropogenic influences on Himalayan river systems, in turn developing biological indicators of change (Ormerod et al., 1994, 1997; Jüttner et al., 1996; Rothfritz et al., 1997).

2. Materials and methods

2.1. Study area and sampling method

Our data came from seven regions of the Himalaya stretched over 1000 km between the Kumaon range (Uttar Pradesh) in the west and Kanchenjunga in eastern Nepal, in general an area recognised for its global significance to biodiversity. River birds here are more species rich than anywhere else on earth (Buckton and Ormerod, unpublished), but we have chosen one species-the Plumbeous Redstart-to investigate model performance. This species is a partial migrant, moving to lower altitudes in winter and higher altitudes during the summer monsoon. As a member of the diverse guild of Himalayan chats (Turdidae), it feeds by aerial flycatching directly in the river corridor, and often over the water surface. It is both abundant and conspicuous, being easily recorded where it occurs.

The field data were collected in winter (October–November, 1994–1996) from 180 study sites ($n = 19$–32 per region); all were second to fourth order streams in independent catchments. This regional pattern of visits was randomised as far as logistically possible to avoid spatio-temporal autocorrelation in the resulting data; streams in each region were sampled opportunistically when encountered by field teams trekking over long dis-

tances (< 200 km), thus representing as varied a range of altitude and physico–chemistry as possible (e.g. altitude 350–4695 m; channel width 0.4–60 m; slope 1–35°; conductivity 9–413 μS/cm).

At each stream, chemical samples were collected for full ionic analysis, and habitat structure was recorded over a 200 m reach using the UK Environment Agency's River Habitat Surveys (RHS). This survey records over 120 variables describing the stream channel, flow character, banks in addition to measurements of altitude and slope, respectively by altimeters and clinometers (Raven et al., 1997). Such a large array of variables is necessary to capture the complex structure of rivers that arises from local geomorphology, natural variations in vegetation, and river management. The results provide significant and meaningful correlates with the distribution of river birds (Buckton and Ormerod, 1997; Ormerod et al., 1997). Prior to any further analysis, habitat and chemical variables from RHS were reduced to major variates using principal components analysis on the correlation matrix (PCA). For RHS, this involved separate sets for variables describing flow character (FlowPC1-5), channel structure (ChanPC1-5) and riparian character (RiparPC1-5).

The presence of Plumbeous Redstarts was recorded using 8X or 10X binoculars over the same 200 m reaches involved in habitat surveys in the early morning (07.00–11.00) or late afternoon (15.00–18.00). This survey method had previously been validated along 46 streams in the Langtang region of central Nepal by comparing the detection of Plumbeous Redstarts on contiguous 200 and 400 m reaches; over 75% of occupied rivers were correctly detected using the 200 m reach alone (Buckton and Ormerod, unpublished data). As potential indicators of prey density, the abundance of benthic macroinvertebrates was assessed contemporaneously with the bird surveys.

2.2. MDA

In general, multiple discriminant analysis is well known, and often applied to ornithological data (e.g. Buckton and Ormerod, 1997; Buckton et al., 1998). Here, the procedure involved creating linear combinations of variables with normal errors that best discriminate between site groups defined *a priori* by the presence or absence of Plumbeous Redstarts. MDA was performed with SPLUS4 software release 3 (lda function, Mass library of Venables and Ripley, 1997), in which combinations of explanatory variables were selected to maximise the ratio of group means discriminant scores to within-group variance (Venables and Ripley, 1997).

2.3. Logistic regression

Presence and absence of Plumbeous Redstart were related to altitude, slope, transformed invertebrate abundance and to the habitat and chemical principal components using a generalized linear model: multiple logistic regression with a logit link and binomial error distribution (McCullagh and Nelder, 1989; Jongman et al., 1995). The logit transformation of the probability of presence/absence (p) was modelled as a linear function of thirty two possible explanatory variables (x_i, $i = 1, 32$):

$$\log it(p) = lop\frac{p}{1-p} = b_0 + \sum_{i=1}^{32} b_{1i}x_i \qquad (1)$$

in which b_0 and b_{1i} are the regression constants. Model were fitted using a maximum likelihood method (McCullagh and Nelder, 1989). We used backwards elimination to select the variables in the final model (Green et al., 1994; Austin and Meyers, 1996). The step function, used in the statistical package SPLUS4, provides a procedure for this purpose using Akaike's information criterion (AIC); this is a penalized version of the likelihood function in which the best model is given by the lowest value (Splus, 1997). Significant variables at each step had to significantly reduced the scaled deviance. The change in scaled deviance as each variable is eliminated is approximately distributed like χ_1^2 (McCullagh and Nelder, 1989; Collett, 1991). Although all explanatory variables are potential predictors, only those selected by these criteria were used in the final solutions.

2.4. Artificial neural networks

The presence or absence of Plumbeous Redstarts was predicted throughout the exercise using the back-propagation algorithm (Rumelhart et al., 1986) with a multi-layered feed-forward neural network of three layers (Fig. 1). This choice reflects the recognised quality of this method in fitting presence-absence data; it can approximate any continuous function from R^n (the departure set with a dimension of n) to [0; 1] (Comon, 1992).

The architecture of the layering has been described by other authors (Baran et al., 1996; Lek et al., 1996a) and is shown in Fig. 1. The first layer, called the input layer, comprises 32 cells representing each of the environmental variables. The second layer, or hidden layer, is composed of a further set of neurones whose number depends on the reliability required, and on the structure that best optimises bias and variance (Geman et al., 1992). We determined the number of second-layer neurones in our application through a series of iterations, in which the number of neurones varied between one and eight. In each case, we calculated the error sums of squares (Fig. 2A) and assessed model performance from good recognition (Fig. 2B). A network with one hidden layer of five neurones resulted in a stable fit and avoided overtraining. Each neurone in the hidden layer calculates the dot product between its weighting vector $Wj = [w_{,j}, j = 1, 5]$ and a data vector $X = [x_i, i = 1, 32]$ which is directly related to the magnitude of the observation at each site (see Fig. 1). This dot product provides a non-linear activation function which, if larger than a given threshold (b_j, see Fig. 1), produces an outgoing signal; in our application, the activation function was sigmoidal,

$$F(x) = \frac{1}{1 + \exp(-x)}. \quad (2)$$

This was necessary to allow use of the backpropagation algorithm and allow an output between 0 and 1. The third layer, or output layer, consists of one neurone responsible for prediction of presence or absence (y, see Fig. 1) from the explanatory variables.

2.5. Global modelling approach

2.5.1. Good recognition

For our first assessments of the performance of each of the three model types we used the entire data, and calculated the percentage of sites at which the presence or absence of Plumbeous Redstarts was correctly predicted. The entire matrix (180 sites × 32 environmental variables) was used to perform MDA, LR and ANN, with explanatory variables optimally selected as described above. In LR and ANN, the output variables for each case have a value within the range 0 and 1, and presence is usually accepted at a threshold of 0.5. For MDA, classification of each case is

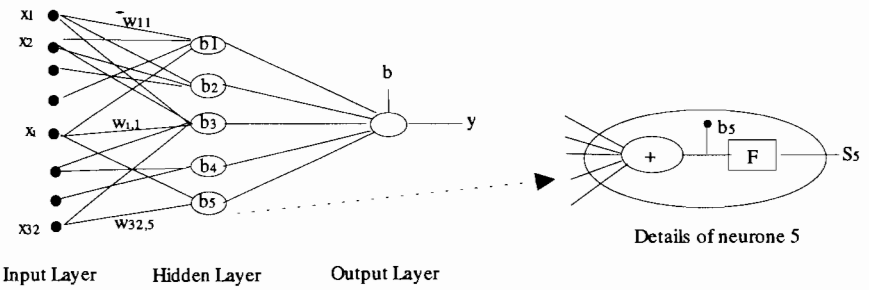

Fig. 1. The structure of the neural network used in this study. The input layer comprises 32 cells representing each of the 32 environmental variables x_i ($i = 1,32$). The hidden layer comprises 5 neurones which calculate the dot product between its vector of weights $Wj = [w_{ij}, i = 1, 32]$ and $X = [x_i, i = 1,32]$. This dot product is compared to a threshold (b_j) and then was passed through a non-linear activation function F, to produce an outgoing signal (S_j) (see detail of neurone 5) The output layer consists of one neurone, similar of the one of the hidden layer, responsible for prediction of presence or absence, y.

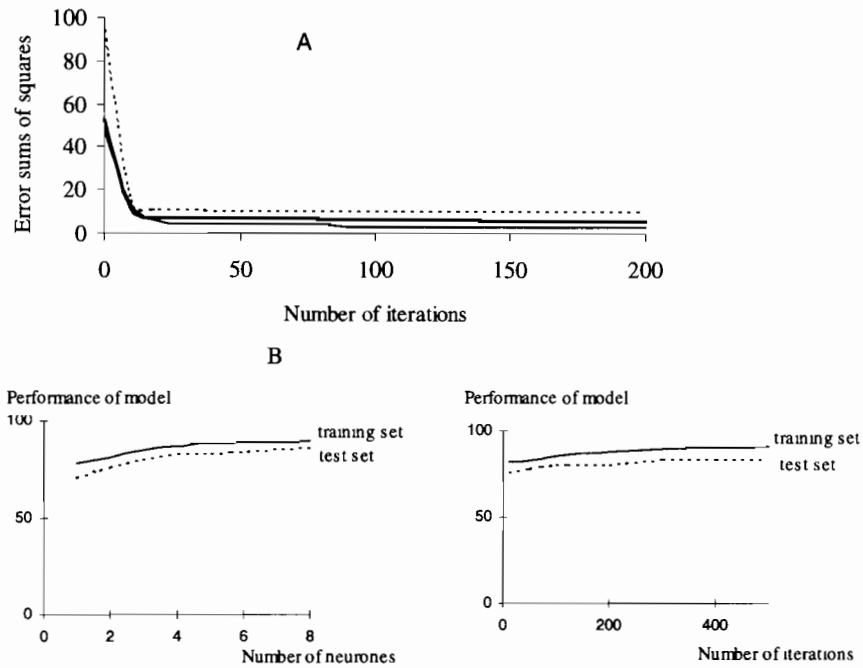

Fig. 2. Determination of number of neurones in the hidden layer. A. Variation of the error sum-of-squares with increasing numbers of iterations for 3 (broken line), 5 (bold) and 8 (thin line). B. Change in model performance of the network in relation to the number of neurones and iterations, illustrated for the training and test sets

derived from Euclidean distances to the centroids of the 'positive' and 'negative' groups.

In all approaches scores for correct assignment were expressed as percentages of the total number of cases. We also derived matrices of confusion, after Fielding and Bell (1997), in which true positive (a) and true negative (d) values were identified. These values could be used to give us, respectively, measures of sensitivity (= percentage of true presences correctly identified) and specificity (= percentage of true absences correctly identified).

2.5.2. Model testing: prediction performance

In addition to the assessment of good recognition, we needed to test each modelling procedure on independent case which were derived by partitioning data into training sets and test sets. For comparison, we chose two special cases of the k-fold partitioning technique, since there are currently discussions about how different methods of partitioning influence model error rates (Fielding and Bell, 1997). These were as follows:

- **Leave-one-out**: This jack-knife method allowed the separation of a test site from the entire suite of 180, so that 179 sites formed the training set. Presence or absence was predicted in the isolated site and compared with the true value. We iterated this operation for all 180.
- **Holdout partition**: In this 2-fold partitioning method (Kohavi, 1995), we made a random selection into a set of training sites (4/5 i.e. 144 sites), and an independent test set (1/5, i.e. 36 sites); selection was weighted so that it always reflected the true proportion of presences and absences. This entire operation was repeated five times to provide tests 1–5, and ANN, LR and MDA calibrated from the training set to predict the presence or absence of Plumbeous Redstarts in the test set. In each case, we compared predicted presence at each site with true presence, and calculated sensitivity and specificity as above.

2.5.3. The problem of threshold selection

Both LR and ANN require an arbitrary threshold probability (often $P = 0.5$) at which to accept species presence from model prediction, but the exact threshold chose will clearly influence model outcome. Moreover, the selection of a given threshold can interact with species' prevalence (i.e. frequency of occurrence) to influence positive and negative prediction error: decreasing frequency of occurrence can increase positive prediction error (Fielding and Bell, 1997). Some methods-notably logistic regression-are considered more sensitive to these effects than others.

We examined these possible influences on model comparison as follows. First, we used ROC plots (received operating characteristic) to compare LR and ANN across a range of thresholds (Zweig and Campbell, 1993). The plots are derived by plotting sensibility (i.e. the true positive fraction) against 1-specificity (i.e. false positive fraction) across different thresholds probabilities (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). Second, we simulated the effects of varying species prevalence on model performance by selectively removing at random increasing proportions (5, 10, 20% etc.) of sites at which Plumbeous Redstarts were present. We repeated the exercise by removing sites from which Plumbeous Redstarts were absent. In each case, we used only altitude as a predictor variable, and examined changes in the value of altitude at which $P = 0.5$. Our reasoning was that this exercise should assess threshold sensitivity to prevalence, but would not affect the relationship between presence/absence and altitude since site removal was made at random.

3. Results

3.1. Fitting and testing models

From the complete data set, mean percentage good recognition varied only slightly between models from 83% with LR, and 84% in MDA, to 88% with ANN; performance in all cases was clearly high. Jack-knife application to test data gave only a marginal average reduction in good recognition (75% correct), but in this case performance varied more strongly between methods: LR produced correctly classified more cases overall (82%) than either MDA (73%) or ANN (69%). As expected from the overall prevalence of Plumbeous Redstarts, occurring at 36% of sites, all the methods predicted true absences (i.e. specificity: 78–86%) better than true presences (i.e. sensitivity: 48–74%; Table 1). In this case, LR correctly classified substantially more positive cases (74%) than either of the other models which helps to explain its better performance overall. In the holdout procedure, general results for good recognition (73–75%) were similar to the jack-knife, but with the following important contrasts. First, all the modelling methods performed on average to near identical levels in good recognition, sensitivity (57–60%) and specificity (73–75%; Table 2). Second, there was marked variation in the results between random data sets (i.e. tests 1–5), particularly in sensitivity. Coefficients of variation between tests in sensitivity were much greater for LR (34%) than either ANN (26%) or MDA (9%).

In keeping with the apparently random variation in model performance between runs, there was also some variation in the detection of significant effects by different explanatory variables (Table 3). Significant effects on Plumbeous Redstart distribution always arose from altitude and channel PC1, while there were significant effects by flow PC5 and ephemeropteran abundance in 4 out of 5 test runs. By contrast, effects by riparian PC5, flowPC4, riparian PC1 and chemistry PC2 were less consistent.

Table 1
A Comparison of three methods (MDA = multiple discriminant analysis; LR = logistic regression, ANN = artifical neural networks), for predicting the presence–absence of Plumbeous Redstarts on 180 Himalayan rivers Sensitivity and specificity were estimated from a 'leave-one-out' jack-kniving repeated 180 times

	Sensitivity			Specificity		
Model	MDA	LR	ANN	DFA	LR	ANN
PR	44	74	48	78	86	82

Table 2
As for Table 1, but involving a holdout procedure repeated five times (test 1 to 5) [a]

Model	MDA			LR			ANN		
	Sn	Sp	PP	Sn	Sp	PP	Sn	Sp	PP
Test1	62	70	67	77	61	67	69	70	69
Test2	46	78	67	77	78	78	34	90	70
Test3	69	83	78	69	83	78	71	79	76
Test4	62	91	81	46	96	78	55	92	79
Test5	54	83	72	31	96	72	54	96	74
Mean	57	81	73	60	83	75	57	85	74
SD	8 8	7.7	6.4	20.6	14.5	4.9	14.8	10.6	9.4

[a] Overall good recognition (= prediction performance; PP in%) and sensitivity (Sn) and specificity (Sp) from MDA, LR and ANN were calculated using a model derived from a calibration set of 80% of the sites (23 absences and 13 presences) in turn applied to the remaining test sites. Mean and standard deviation (SD) were derived for the five tests.

3.2. Threshold effects on LR and ANN

ROC plots were drawn from the jack-knife results. They illustrated, for this application, that LR outperformed ANN in correctly classifying new cases irrespective of the threshold value of probability chosen to accept presence (Fig. 3). However, the sensitivity analysis confirmed the Fielding and Bell (1997) view that LR threshold probabilities are potentially at risk from varying prevalence. Artificial variations in prevalence showed that the altitudes predicted from LR at which $P = 0.5$ varied strongly with the size of both the 'absent' and 'present' group (Table 4). This is despite the selective removal of sites at random across the entire altitude range of the species. ANN was less sensitive. These results show that effects by threshold selection and prevalence must be treated with caution.

4. Discussion

Solely on the criteria of correctly predicting the presence or absence of Plumbeous Redstarts, all these modelling procedures performed well in all the tests we carried out: working either with the entire data, or with partitioned data sets, good recognition exceeded 69–88%. This result contrasts with recent studies that have suggested ANN out-perform more conventional methods of modelling ecological data (Baran et al., 1996; Lek et al., 1996a,b; Mastrorillo et al., 1997). Indeed, one of our major conclusions is that ANN do not currently have major advantages over logistic regression and discriminant analysis in modelling species distribution providing these latter methods are correctly applied. In fact, we found some clear disadvantages: with our optimisation procedure, neural networks require much more computing time than conventional statistical methods. At present, also, possible causal relationships between species distribution and environmental data are not immediately identified in ANN. Instead, such identification requires further procedures such as weight analysis (e.g. Roadknight et al., 1997), equation synthesis (Balls et al., 1996) or correlated activity pruning (Wiersma et al., 1995). Thus, the conventional linear methods allow more direct straightforward development of testable and falsifiable hypotheses. It should be noted that our comparison between modelling methods involved correctly applying logistic regression and discriminant analysis: we ensured, for example that explanatory variables were linearised and normalised by transformation and incorporation into principal components analysis prior to further analysis. We were also careful to collect our data from sites on independent rivers.

Although many ecologists assess species distribution models solely from 'good recognition' (i.e. overall predictive power), our study reaffirms the well-known value of testing models with partitioned data (Kohavi, 1995). More interesting in

Table 3
Significant explanatory variables indicated by logistic regression (standardised regression coefficient) during each run of the holdout procedure illustrated in Table 2 are given

	Alt	ChanPC1	Flowpc5	Ephem	RiparPC5	FlowPC4	RiparPC1	ChemPC2
Overall	−5.22	3.04	2.94	2.81	−2.78			
Train 1	−3.298	3.622	2.02				−2.34	
Train 2	−4.98	2.91		2.644				
Train 3	−4.62	3.364	2.26	2.61		−2.22		
Train 4	−4.74	2.52	2.45	3.89	−2.96			−2.40
Train 5	−4.68	2.42	2.21	2.77	−2.07			−2.34

view of recent discussion in the literature (Fielding and Bell, 1997), this work illustrates the importance of using a range of criteria in assessing performance – sensitivity and specificity, and performance across a range of probability thresholds. Finally, it illustrates the importance of considering a range of procedures for model testing-for example jack-knife sampling and holdout.

Particularly in the holdout method, our tests revealed how chance can be responsible for large variations in sensitivity and model accuracy, with ANN and LR apparently much more sensitive to these effects than MDA (see Table 2). Not only were ANN and LR prone to large variations between test runs in sensitivity, but also LR was prone to variation in detecting significant effects by different explanatory variables. Thus, not only can models appear well fitted by chance, but also they can produce potentially spurious explanations of distribution pattern. Chatfield (1995) questioned the use of data-partitioning for model testing, suggesting that splitting data arbitrarily is not the same as collecting new data. However, at the geographical scale over which this work was carried out—essentially the entire Himalayan mountains over much of the range of our model species—independent data sets would be precluded by time, cost and opportunity. Moreover, the lesson from this exercise was clear: any one of the five test runs in the holdout procedure might have represented a real attempt to model species distribution at new sites, with a wide array of possible outcomes apparent from the coefficient of variation in sensitivity: clearly there is a need for caution in interpreting real data and real model applications. We will return to this theme in another paper involving regional-scale applications to modelling the Himalayan distribution of a wider range of species (Manel, Dias and Ormerod, unpublished data).

With a data set comprising 180 independent cases-each an individual river-it was possible to simulate potentially important influences on models of species distribution. Fielding and Bell (1997) discussed potential effects by prevalence and probability thresholds on such models, and their effects were confirmed here. Effects by threshold probabilities did not appear to affect the comparison between ANN and LR, since LR outperformed ANN in one test across all thresholds (see Fig. 3). By contrast, LR was more sensitive than ANN to effects by species prevalence (see Table 4). However, prevalence affected all the modelling

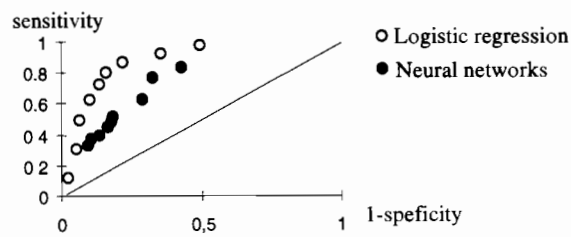

Fig. 3. ROC (Received operating characteristic) plot to compare LR and ANN across different threshold probabilities for the Plumbeous Redstart. The y-axis shows sensitivity defined as the fraction of [number of true positive]/[number of true positive + number of false negative]. The x-axis shows 1-specificity defined as [number of false positive results]/[number of true negative + number of false positive]. The relative position of the plots indicates the relative accuracy of the tests. The position of the points from LR above and to the left of the ANN plot indicates greater observed accuracy in the former.

Table 4
Simulating the effects of varying species prevalence, at a threshold $P = 0.5$, on the prediction of the presence or absence of Plumbeous Redstarts by ANN and LR [a]

Percentage of random reduced (%)	Positive occurrence		Negative occurrence	
	LR	ANN	LR	ANN
0	1595.2	1850	1595.2	
5	1561	1840	1644.7	1850
10	1521.4	1800	1665.7	1850
20	1441.5	1750	1775.2	1930
30	1317.6	1730	1829.5	1940
40	1204.9	1400	1954.9	2000
50	1072.4	1660	2080.6	2140
60	965.7		2206.6	2180
70	801.6		2321.3	2320
80	599.0		2604.2	2700
90	104.8		3572.7	2900

[a] The column headed 'positive occurrence' illustrates the effects on the value of altitude at which the probability of occurrence = 0.5 from reducing the number of sites with Plumbeous Redstarts; the column headed 'negative occurrence' illustrates the effects of the reducing the number of sites without Plumbeous Redstarts.

procedures by causing much lower sensitivity than specificity. This is despite the occurrence of Plumbeous Redstarts at 36% of our study sites-as one of the commonest species in the whole guild of Himalayan river birds. The prevalence effect on predictive power is often overlooked in distribution models developed by ecologists, but clearly it warrants careful consideration. It will be particularly important in instances where the distribution of scarce species is predicted for conservation purposes—for example in identifying areas for legal protection or species re-introduction.

5. Overview: aims of modelling determine the choice of models?

We began this work wishing to compare three different approaches for modelling species' distributions, and for assessing how distribution might be influenced by environmental features. In some respects, the recommendations that follow our results will depend on the aims of any particular programme (Fielding and Bell, 1997; Venables and Ripley, 1997). In instances where models are intended to be explanatory, any of the approaches used here might be suitable, since all produced good overall fit to the data. LR and MDA currently have clear advantages in developing testable hypotheses, since they provide the clearest indications of possible causal effects on distribution. For example, Edwin et al. (1998) recently illustrated the advantages of LR in describing the optimum habitat ranges, and hence suitability indices, for aquatic species.

The robust field testing of all model predictions-irrespective of the algorithm used-is a particularly important consideration given the well-known difficulties that arise when investigators rely solely on correlative data to interpret the causes of field pattern. All our modelling approaches require support from appropriate experimental tests but this, in turn, is a major challenge at the spatial scales involved in the work (Gaston, 1998). In many respects, the experimental validation of any large-scale model represents potentially greater problems than the choice between modelling methods. Due to the difficulties of experimentation at large scales, the testing of models by application in new locations provides one of the few robust procedures. In our work, the holdout procedure approximated such a testing method, revealing MDA to be more preferable over LR or ANN in some respects. Nevertheless,

in instances where there are complex or non-linear influences on species distribution, ANN may well turn out to be advantageous, but clear illustrations are required.

Acknowledgements

These data were collected under a programme funded by the Darwin Initiative for the Survival of Species co-ordinated by the UK Department of Environment Transport and the Regions. We thank Dr Alan Jenkins of the Institute of Hydrology (UK), Phil Brewin and Seb Buckton and Hem Sagar Baral, without whom the work would not have been possible. The analysis was funded by the Royal Society European Science Exchange Programme. We thank Professor Claude Mouchés for providing the important opportunity for this collaboration between the Université de Pau et des pays de l'Adour and Cardiff University.

References

Austin, M.P, Meyers, J.A., 1996. Current approaches to modelling the environmental niche of eucalyptus: implications for management of forest biodiversity. Forest Ecol. Manag. 85, 95–106.

Balls, G.R., Palmer-Brown, D., Sanders, G.E., 1996. Investigating microclimatic influences on ozone injury in clover (*Trofolium subterraneum*) using artificial neural networks. New Phyto. 132, 271–280.

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996. Stochastic models that predict trout population density or biomass on a mesohabitat scale. Hydrobiologia 337, 1–9.

Buckton, S.T., Ormerod, S.J., 1997. Use of a new standardised habitat survey for assessing the habitat preferences and distribution of upland river birds. Bird Study 44, 327–337.

Buckton, S.T., Brewin, P.A., Lewis, A., Stevens, P.A., Ormerod, S.J., 1998 The distribution of dippers *Cinclus cinclus* in the acid sensitive region of upland Wales, 1984–1995. Freshwater Biol. 39, 387–396.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. J. R Stat. Soc. 158, 419–466.

Collett, D., 1991. Modelling Binary Data. Chapman and Hall, London, p. 364.

Comon, P., 1992 Classification supervisée par réseaux multicouches. Traitement du signal 8, 387–407.

Edwin, T.H., Peeters, M., Gardeniers, J.P., 1998. Logistic regression as a tool for defining habitat requirements of two common gammarids. Freshwater Biol. 39, 605–615.

Fielding, A.H., Bell, J.F., 1997. A review methods for the assessment of prediction errors in conservation presence/absence models. Env. Cons. 24, 38–49

Gaston, K J., 1998. Some methodological issues in macroecology. Am Nat. 151, 68–83

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemna. Neural Computation 4, 1–58.

Green, R.E., Osborne, P.E., Sears, E.J., 1994. The distribution of passerines birds in hedgerows during the breeding season in relation to characteristics of hedgerows and adjacent farmlands. J. Appl. Ecol. 31, 677–692.

Jongman, R.H.G., Ter Braak, C.J.F., van Tongeren, O.F.R., 1995. Data Analysis in Community and Landscape Ecology, 2nd Edition. Cambridge University Press, Cambridge, p. 299

Jüttner, I., Rothfritz, H., Ormerod, S.J., 1996. Diatoms as indicators of river quality in the Nepalese Middle Hills with consideration of the effects of habitat-specific sampling. Freshwater Biol. 36, 101–112

Kohavi, R., 1995. A study of cross validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers, pp. 1137–1143.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I, Lauga, J., Aulagnier, S., 1996a. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 1634, 1–13.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996b. Rôle of some environmental variables in trout abundance models using neural networks. Aquat. Living Res 9, 23–29.

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997 The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biol. 38, 237–246.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd edition. Chapman and Hall, Monographs on Statistics and Applied Probability, London, p. 511.

Ormerod, S.J., Rundle, S.D.S., Wilkinson, M., Daly, G.P., Dale, K.M., Jüttner, I., 1994. Altitudinal trends in the diatoms, bryophytes, macroinvertebrates and fish of a Nepalese river system Freshwater Biol. 32, 309–322.

Ormerod, S.J., Baral, H.S., Brewin, P.A., Buckton, S.T., Jüttner, I., Rothfritz, H., Suren, A.M., 1997. River Habitat Surveys and Biodiversity in the Nepal Himalaya. In. Boon, P.J., Howell, D.L. (Eds.), Freshwater Quality: Defining the Indefinable. HMSO, Edinburgh, pp. 241–250.

Osborne, P.E., Tigar, B.J., 1992. Interpreting bird atlas data using logistic models. an example from Lesotho, Southern Africa. J. Appl. Ecol. 29, 55–62.

Raven, P.J., Fox, P., Everard, M., Holmes, N.T.H., Dawson, F.H., 1997. River habitat structure: a new system for classifying rivers according to their habitat quality. In: Boon, P.J , Howell, D.L. (Eds.), Freshwater Quality: Defining the Indefinable HMSO, Edinburgh, pp. 241–250.

Ricklefs, R.E., Schluter, D, 1993. Species Diversity in Ecological Communities. University of Chicago Press, Chicago, p. 414.

Roadknight, C.M., Balls, G.R., Mills, G.E., Palmer-Brown, D , 1997. Modelling complex environmental data' IEEE Trans. on Neural Networks 8, 852–862.

Rothfritz, H., Juttner, I., Suren, A., Ormerod, S.J., 1997. Epiphytic and epilithic diatom communities along environmental gradients in the Nepalese Himalaya: implications for the assessment of biodiversity and water quality. Archiv für Hydrobiologie 138, 465–482

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. Nature 323, 533–536.

S-PLUS4, 1997 Guide to Statistics Data Analysis Pruducts Division, Mathsoft, Seattle.

Venables, W.N., Ripley, B.D., 1997. Modern applied statistics with S-PLUS, 2nd edition. Springer, New York, p. 548.

Wiersma, F.R., Poel, M, Oudshoff, A.M., 1995. The BB neural network rule extraction method. In· Kappen, B., Gielen, S. (Eds.), Proceedings of 3rd annual SNN symposium on neural networks. Springer-Verlag, New York, pp 69–73.

Zweig, M.H., Campbell, G., 1993 Receiver-operating characteristic (ROC) plots. a fundamental tool in clinical medicine. Clin. Chem. 39, 561–577.

ELSEVIER

Ecological Modelling 120 (1999) 349–358

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Use of artificial neural networks for predicting rice crop damage by greater flamingos in the Camargue, France

Christophe Tourenq [a,b,*], Stéphane Aulagnier [c], François Mesléard [a], Laurent Durieux [a], Alan Johnson [a], Georges Gonzalez [c], Sovan Lek [d]

[a] *Station Biologique de la Tour du Valat, Le Sambuc, 13200 Arles, France*
[b] *Centre d'Ecologie Fonctionnelle Evolutive, Université de Montpellier II, route de Mende, 34000 Montpellier, France*
[c] *Institut de Recherche sur les Grands Mammifères, B.P. 27, 31326 Castanet-Tolosan Cedex, France*
[d] *CESAC, UMR 5576, CNRS-Univ. Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France*

Abstract

Since the 1980s, incursions of greater flamingo (*Phoenicopterus ruber roseus*) in rice fields have been reported almost every year in the Camargue, south-eastern France, and more recently in Spain. We assessed the performances of artificial neural networks (ANN) in predicting presence or absence of flamingo damages from 11 variables describing landscape features of rice paddies. The global matrix of 1978 records (276 with damage and 1702 without) for the 1993–1996 period was used to determine the suitable parameters: number of hidden layer nodes and number of iterations. In order to avoid particular inputs either in the training set or in the testing set, ten different randomly sampled training sets were available. A classic multilayer feed-forward neural network with back-propagation algorithm was used throughout these experiments. Data from 1993 to 1996 were used to predict data for 1997 (73 fields with damage and 1905 without) and 1998 (88 with damage and 1890 without). Three training set compositions were displayed: (I) the whole data set (1978 observations), (II) an equal number (276) of damaged and undamaged fields (552 observations), (III) a set with 1/3 of observations being damaged fields (276) and 2/3 undamaged (552). ANN faced some difficulty in predicting both presence and absence of damage. The number of each type record in the training set was particularly sensitive. ANN predicted the more frequent outcome, (i.e. absence of damage). Most often, better results were obtained when equilibrating the number of presences and absences. In this case, we obtained performances ranging from 64% up to 87% according to the presence and absence of data in the training set. When fitting ANN with the whole set of presences to predict damage 1 year later, these results stabilised at $\approx 79\%$ for 1997 and between 66 and 72% for 1998 when more than half of the damaged fields were never visited by flamingos during the period 1993–1997. Our performances are quite similar to the results obtained by previous authors and predictability from 1 year to the following one also supports that ANN can be an alternative or a supplement to actual scaring methods in identifying potential damaged fields and propose agricultural management plans or concentrate scaring actions on these high-risk areas.

Keywords: Flamingos; Rice; Damage; Artificial neural networks; Prediction; Landscape features; Camargue; France

* Corresponding author. Fax: + 33-(0)490-972-019.
E-mail address: tourenq@tour-du-valet.com (C. Tourenq)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00114-3

1. Introduction

Rice-crop damage by shorebirds, ducks and/or passerines has been studied mainly in North and South America, Africa and Australia, where rice is cultivated over very large areas. Damage by these 'pests' has been estimated at millions of dollars annually (Berryman, 1966; Wilson *et al.*, 1989; Decker *et al.*, 1990) and huge efforts have been made to find solutions (*e.g.* Meanley, 1971; Elliot, 1979; Ward, 1979; Holler *et al.*, 1982; Avery and Decker, 1994; Avery *et al.*, 1995; Katondo, 1996).

In Europe, rice cultivation is restricted to parts of the Mediterranean region and this phenomenon has received less attention. However, in spring 1978, greater flamingos (*Phoenicoterus ruber roseus*) began to feed in rice fields of the Camargue, the delta of the River Rhone in south-eastern France. Scaring campaigns have been carried out every year since 1981, and crop losses from flamingos have been reduced. This habit spread in 1993 to the Ebro delta, north-eastern Spain, and Spanish farmers now face the same problem as the French (Jimenez and Soler, 1996; Johnson and Mesléard, 1997).

Scaring programs, begun in 1981, involve use of gas exploders, rotating firing devices and Very pistols (André and Johnson, 1981; Hoffmann and Johnson, 1991). Even if these techniques are efficient in scaring or keeping away flamingos from some rice fields, they are costly and time consuming. Monitoring of flamingo movements and behaviours must occur over a wide foraging range (over 60 km from the breeding site at the Etang du Fangassier; Johnson, 1989). We based our study on the hypothesis that some plots were more attractive than others, e.g. that landscape features may influence the flamingo's choice of plots in which to forage (André and Johnson, 1981; Sourribes, 1993; Rogers, 1995; Jimenez and Soler, 1996; Durieux, 1997).

A model identifying the most vulnerable plots could be helpful to farmers and wildlife managers by helping to evaluate the risk of crop damage in problem areas. Due to the non-linearity of most of the variables in ecology and the use of qualitative traits in the data set, we computed ANN to propose predictive models for the damage caused by flamingos in rice fields and to characterize the explicative landscape variables.

2. Study area

The Camargue delta of the River Rhône, lies on the Mediterranean Sea coast. Rice was introduced into the area in the early 1940s and today paddies cover some 24 000 ha (16% of the total surface area of the Camargue and 46% of the agricultural land, Chauvelon, 1996). Our study was carried out in the Fumemorte Basin, one of six independent drainage basins of the delta. This sector is in the eastern part of the delta proper and comprises ≈ 70 km^2. Rice fields represent some 31% of the total surface of the basin and 61% of the agricultural land. There are also extensive areas of natural land (32%) and abandoned farm lands (23.2%). The agricultural land is subdivided into small cultural units, 75% being less than 3 ha (Chauvelon, 1996). The southern part of the basin is 2 km from the unique breeding site of the greater flamingo in France (16.5 km for the northern part). The Etang du Fangassier is the only breeding site of the greater flamingo in France and one of the most important in the Mediterranean area (Rendon Martos and Johnson, 1996).

Flamingos frequent rice fields between sunset and sunrise from the end of April to the beginning of June. This period corresponds to the critical germination period of rice in the Mediterranean region (Fasola and Ruiz, 1996; Barbier and Mouret, 1992). Damage to crops is caused in four ways (Hoffmann and Johnson, 1991): (i) trampling which prevents germination; (ii) disturbance of the grain, causing it to float to the surface where it is blown to the downwind shore; (iii) seedlings destroyed by trampling and (iv) ingurgitation of rice seeds. Whether flamingos visit the fields in search of invertebrates or to feed on the rice grain, or both, is not known. It has been shown, however, that flamingos prefer some paddies to others and visit the same fields on consecutive nights and from 1 year to the next (Rogers, 1995; Jimenez and Soler, 1996).

3. Methods

3.1. Monitoring damage

We analysed occurrence of rice-crop damage by flamingos for the period 1993–1998. From 1993 to 1995, data were taken from internal reports of the Parc Naturel Régional de Camargue, and by interviewing landowners. Only ascertained flamingo damaged paddies were considered. For the period 1996–1998, three methods of monitoring rice crop damage were used (Durieux, 1997):

1. a bi-weekly aerial survey (at 400 ft) of the Fumemorte basin in the morning. Each field with turbid water or with tracks was visited the same day to confirm that flamingos were responsible for these tracks (presence of feathers, footprints).
2. daily observations at dusk and at night in strategic places on farmlands considered vulnerable. Information gathered by this method was scarce due to the darkness and size of the area surveyed.
3. interviews with farmers who plotted on a map the distribution of fields frequented by flamingos and the number of birds involved. This inquiry was carried out at the end of June, but farmers telephoned the 'French Rice Centre' or the 'Tour du Valat Biological Station' immediately when they noticed groups of flamingos in their fields.

The presence or absence of damage was coded (1) and (0) respectively.

3.2. Environmental variables

We considered 11 environmental variables for each of 1978 rice fields of the Fumemorte Basin. These were: surface area; distance from natural marshes; distance from the breeding site; distance from the closest wooded hedge or copse; distance from power lines; distance from habitations; distance from principal roads; distance from secondary roads; height of hedges surrounding the paddy; number of wooded sides; adjacent (1) or not (0) to damaged field.

Surface area was measured in ha and distances were considered from the geometric centre of the field (in m or km). The height of hedges was assigned to one of five classes according to the main vegetation occurring in the Camargue (Durieux, 1997): < 50 cm (herbaceous plants or absence of vegetation); 50 cm–150 cm (mostly Reed, *Phragmites australis*); 150 cm–3 m (hedges composed of Reed, Tamarisk, *Tamarix gallica*, Hawthorn, *Crataegus monogina*, Phillyrea, *Phillyrea angustifolia*, Elderberry, *Sambucus nigra*), 3 m–15 m (Narrow-leaved Ash, *Fraxinus excelsior*, Laurel, *Laurus nobilis*; Oleaster, *Eleagnus angustifolia*); > 15 m (Common Alder, *Alnus glutinosa*, Downy Oak, *Quercus pubescens*, Italian Cypress, *Cupressus sempervirens*, Elm, *Ulmus campestris*, White Poplar, *Populus alba*, False Acacia, *Robinia pseudacacia*).

3.3. ANN modelling

3.3.1. Fitting and testing

The global matrix of 1978 records (276 with damage and 1702 without) for the 1993–1996 period was used to train the ANN and to determine the suitable parameters: number of hidden layer nodes (HN) and number of iterations. In order to test the classification quality of the model, the data matrix was randomly decomposed into two sets. The first set was used to train the neural networks (training sets). The remaining individuals (testing sets) were used to evaluate the quality of their assignment in a hold-out procedure (Kohavi, 1995). Due to the larger number of absences of damage, three set compositions were sampled: sets A, B and C (Table 1). In order to avoid particular inputs either in the training set or in the testing set, ten different training sets C were randomly sampled (C1–C10).

We used a classic multilayer feed-forward neural network with back-propagation algorithm (Rumelhart *et al.*, 1986) throughout these experiments. We trained networks with one hidden layer of one to 15 neurons. The output variables were: 0 = absence of damage, 1 = damage.

Training the network consisted of using a training data-set to adjust the connection weights in order to obtain the maximum number of individuals correctly classified. The connection weights, initially taken at random in the range $[-0.3, 0.3]$,

Table 1
Three randomly sampled training and testing sets used for fitting ANN models

Set	Training sets			Testing sets		
	Damage	No damage	**Total**	Damage	No damage	**Total**
A	207	1277	**1484**	69	425	**494**
B	207	414	**621**	69	1288	**1357**
C	207	207	**414**	69	1495	**1564**

were iteratively adjusted by a method of gradient descent based on the difference between the observed and expected outgoing signals. The number of iterations (necessary to guarantee the convergence of estimated values toward their expectations) was first limited to 500, then to 400 in order to avoid an overfit (see Gallant, 1993). Training was performed first on sets A, B and C with six, eight, ten, 12 and 15 hidden neurons, second on three sets C with one, two, three, four, five, six, seven, eight, ten, 12 and 15 hidden neurons, third on ten sets C with six hidden neurons.

3.3.2. Predicting

Data from 1993 to 1996 were used to train a model and predict data from 1997 (73 with damage and 1905 without) and 1998 (88 with damage and 1890 without). Three training set compositions were displayed: (I) the whole data set (1978 observations), (II) an equal number (276) of damaged and undamaged fields (552 observations), (III) a set with 1/3 of the observations being damaged fields (276) and 2/3 undamaged (552). ANNs were trained with ten sets of each composition.

Note that all the paddies used by flamingos in 1997 were previously visited by birds while 48 paddies (out of 88) were first visited by the flamingos in 1998. The contribution of each environmental variable was determined from trainings of ten type II data sets using the Goh procedure (Garson, 1991; Goh, 1995).

4. Results

4.1. Fitting and testing models

The larger number of 'absences' in the set A induced the learning of absences far better than presences (Fig. 1). Training with the set B was good for both absence and presence, but presence was poorly predicted ($\approx$ 50% of correct classification). The best results were obtained when equilibrating the 'presences' and 'absences' (set C). After 180 iterations, performances fluctuated, according to the number of neurons of the hidden layer (HN = 6–15), between 80 and 99% for training, and between 61 and 84% for testing. For the next steps of analysis, we considered 400 iterations when the correct classification percentage was between 84 and 99% for training, and between 61 and 77% for testing, in order to avoid anoverfit. When training the ANN with three sets C, there was little variation in the testing scores according to the number of hidden layer nodes (Fig. 2). However, predictions seemed to be more balanced with an intermediate number of hidden layer nodes (HN = 6): 69 to 75% for absences and 68 to 77% for presences. Training ANN with an equal set of presences and absences gave the best correct classification percentages after 400 iterations when using a model with 6 hidden layer nodes. This configuration was used for the following analysis. The correct prediction, repeated 5 times, for ten randomly sampled testing sets, associated with equilibrate training sets, varied from 64% (set 4) to 87% (set 2) for presences (Fig. 3), and from 65% (set 1) to 79% (set 3) for absences. However the scores were balanced and quite homogeneous for all the sets.

4.2. Prediction

A model (I) with an intermediate number of hidden layer nodes (HN = 6) and trained with the whole 1993–1996 data set predicted more ab-

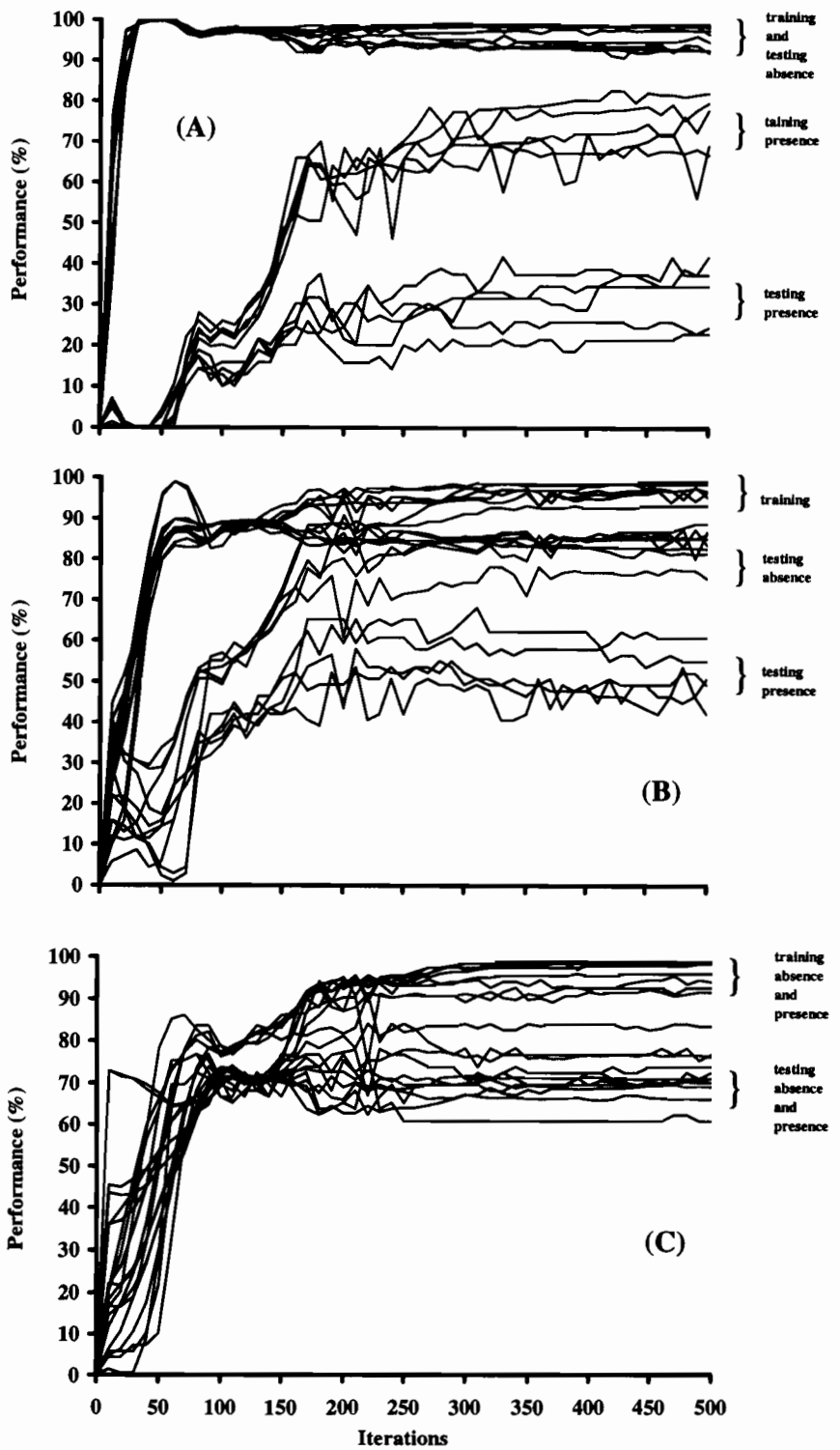

Fig. 1 Number of iterations and performance (percentage of correct classification) obtained for three set compositions (A, B, C; see text) by ANN model in training and testing. Five configurations of hidden layer nodes are represented (HN = 6, 8, 10, 12, 15).

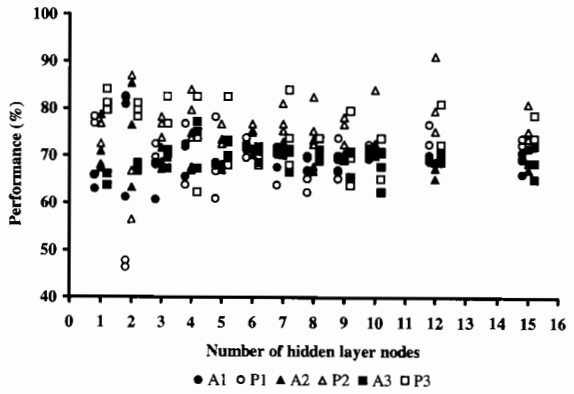

Fig. 2. Performances of three networks with an equilibrate number of presences and absences (207) according to the number of hidden layer nodes. For each network, training was proceeded three times and tested three times with the rest of the observations. A = absences, P = presences.

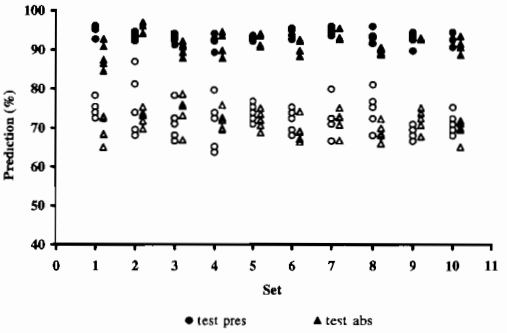

Fig. 3. Predictive power of ANN models (HN = 6) determined from five trainings of ten sets with an equilibrate number of presences (pres) and absences (abs).

sences (92.4%) than presences of flamingos in rice fields (53.4%) in 1997. A model (II) with similar number of HNs (6) but with an equal number of presences and absences predicted more the presences (93.2%) than the absences (65.7%). A similar model (III) with 1/3 of observations being presences and 2/3 being absences gave a balanced prediction for 1997. Predictive scores of model III were quite stable for ten trainings $\approx 79\%$ for both presences and absences (Table 2).

Predictions differed slightly for 1998 (Table 3). A model using a type III training set and same training scores as 1997, predicted more absences than presences ($\approx 76\%$ vs 60%). There were no such differences between accuracy of classification using a type II training set. Despite higher training scores, the predictive scores were lower than in 1997, mainly for absences. These results can be easily related to the somewhat different location of damage in 1998 compared with the previous years.

Table 2
Predictions for 1997 of ten type III (1/3 observations being presences and 2/3 being absences) ANN models with an intermediate number of hidden layer nodes (HN = 6)

Training	Testing	Training		Testing	
(1993–1996)	(1997)	Presences (%)	Absences (%)	Presences (%)	Absences (%)
III.1	1905 *A* + 73 *P*	80 8	94.7	83.5	79 6
III.2	id	79 3	89.3	80 8	77.6
III.3	id.	85.9	91.8	78.1	77.6
III.4	id.	79	93.1	72 6	80.5
III.5	id.	83	94.2	79.4	79.7
III.6	id.	75 3	95.5	71 2	80 4
III 7	id.	80	94.5	80.8	79.5
III.8	id	81 2	92.4	78 1	77.9
III 9	id.	80	91.3	78.1	77.7
III.10	id	81.5	94	84 9	81.3

Table 3
Predictions for 1998 of ten type II (equal number of absences and presences) and ten type III (1/3 observations being presences and 2/3 being absences) ANN models with an intermediate number of hidden layer nodes (HN = 6)[a]

	Model II				Model III			
	Training		Testing		Training		Testing	
Set	Presence (%)	Absence (%)	Presence (%)	Absence (%)	Presence (%)	Absence (%)	Presence (%)	Absence (%)
1	92.03	93.84	69.32	66 19	85.15	95.29	60.23	76 88
2	93.84	88.04	76.14	65.34	82.61	92.75	57.96	75.19
3	91.67	93 84	73.86	65 19	79.71	92.39	51.14	77 73
4	89.13	93.12	68.18	68.36	86.96	93.3	63.64	74.79
5	92 03	88 04	71.59	66.14	81.88	95.11	56.82	78 73
6	92.75	89 49	70.46	64.55	84.78	93.3	64.77	78.25
7	93.12	90.22	73.86	65.93	80.8	93.12	55 68	76.51
8	92.39	90.58	68.18	65.87	84.78	93.48	60.23	77.78
9	92 39	92.03	73.86	67.19	86.23	93.3	56 82	76 24
10	93.48	91.3	70.46	64.07	85.51	92.57	63.64	76.35
Mean	92.283	91 05	71.591	65.883	83.841	93.461	59.093	76.845
S.D.	1.301	2.169	2.731	1.240	2.435	0.983	4.252	1.285

[a] S.D. = standard deviation.

4.3. Contributions of environmental variables

From one model to another, all variables displayed high contributions (Table 4). However the contributions of the surface of rice fields (SUP), and also of the distance from the colony (DCO), was often weak, while the distance from natural marshes (DNM) and the distance from the closest wooded hedge (DWO) exhibited high contributions in most of the models. Note that contributions of input variables varied considerably among models. For example, model four attributed a huge contribution to the distance from natural marshes (DNM), the number of closed sides (NWS) exhibited also a heavy contribution, while these variables were weakly implicated in model seven.

5. Discussion

Artificial neural networks faced some difficulties in predicting both presence and absence of damage. The number of each type record in the training set was particularly sensitive. As previously observed by Spitz *et al.* (1996), Mastrorillo *et al.* (1997), Manel *et al.* (1999), ANNs delivered better prediction for the largest occurrence. Better results were obtained when equilibrating the number of presences and absences. This is a problem, because in ecology absences are often far more frequent than presences, and obviously, information is lost by decreasing the number of absences in training sets. The weak improvement of the performance of ANN with the increasing number of hidden layer neurons could be related to close relative input variables, but we can hardly conceive that it is the case with environmental variables such as distance to the natural marshes and number of wooded sides to the field.

When equilibrating correct predictions of presence and absence of damage, we obtained performances ranging from 64% up to 87% according to the sampled data in the training set. When fitting ANN with the whole set of presences to predict damage 1 year later, these results stabilized $\approx$ 79% for 1997 and between 66 and 72% for 1998 when more than half of the damaged fields were never visited by flamingos during the period 1993–1997. These performances are quite similar to the results obtained by Spitz *et al.* (1996) in predicting the impact of Wild Boar (*Sus scrofa*)

on cultivated fields (approximately 80% for presence, but only 42% for absence).

Damage of rice fields by flamingos may be a trivial problem on an international or on a national scale, but, at a regional or local scale the situation is more critical. Flamingo damage for the Camargue has been estimated at approximately $153 000 annually (Johnson and Mesléard, 1997). Even if crop losses attributable to flamingos has no perceptible impact on farming in terms of national crop production, like other bird problems in Europe (O'Connor and Shrubb, 1986; Edgell and Williams, 1991), the same fields can be visited on consecutive nights and over consecutive years (Rogers, 1995) and crop losses can be important for a single farmer.

Until now, several non-lethal or lethal techniques were advanced to prevent damages to rice paddies by birds (Meanley, 1971; Elliot, 1979; Ward, 1979; Wilson *et al.*, 1989; Decker *et al.*, 1990; Hoffmann and Johnson, 1991; Avery *et al.*, 1995). However, the effectiveness of these operations is shown to be conditioned by the number of birds and by the mobility and behaviour of the species concerned (O'Connor and Shrubb, 1986; Brugger *et al.*, 1992). Rather than searching for short-term methods of control which are not necessarily efficient, nor ethical (Morrisson, 1975; Van Vessem *et al.*, 1985; Caughley and Sinclair, 1994), long-term solutions to this particular problem should be sought. Predictability from 1 year to the next supports the idea that ANN can be an alternative or a supplement to actual scaring methods in identifying vulnerable fields. This would enable agricultural management plans to be established or scaring actions to be concentrated on these high-risk areas.

The next step of our study is to extend predictions to the whole of the Camargue and to accurately identify vulnerable fields in order to concentrate scaring methods or propose management actions on these high-risk areas. This study interestingly revealed the ability of ANN to predict damage by greater flamingos from a small set of environmental variables which it is easy to collect. However, before extending the model, some new analyses are needed to improve the predictions, and also to find a method of identifying the most relevant environmental variables for modelling the prediction (discriminant analysis,

Table 4
Relative importance of input variables for ten type II sets[a,b]

Set	SUP	DNM	DCO	DWO	DTL	DHA	DPR	DSR	HHS	NWS	CON
II.1	9.22	9.85	9.86	10.18	8.23	6.30	9.20	9.43	8.35	11 84	7.55
II.2	9.18	13.30	10.80	11.3	9.28	8.54	4.11	5.41	8.99	9.59	9.41
II.3	10.4	13 40	6.48	13.7	8.7	10.40	8.32	9.14	6 80	3 84	8.85
II.4	4.31	19.20	5.85	9.77	5.8	9.91	8.14	9.24	7.43	12.00	8.32
II.5	5.17	12.50	5.41	12.5	7.74	12.80	8.90	7.39	9.52	9.63	8.44
II 6	11.30	7.13	8.93	11.8	8.57	14.2	6.71	12.30	7.68	4.47	6.85
II 7	3.69	9.14	5.38	11.8	9.92	8.96	10.90	9.91	8.93	13.70	7.64
II.8	5.03	12.00	8.70	9.48	9.36	9.21	9.84	11.00	7.35	8.26	9.75
II.9	9.81	11.20	5.79	10.20	5.27	7.46	10.20	10 20	8.18	13.60	8.17
II.10	7.87	10.80	9.83	10.20	11.50	11.80	6.23	8.53	7.87	5.75	9.63
Mean	7.6	11.85	7.7	11.09	8.44	9.96	8.25	9.26	8.11	9.27	8.46
S.D.	2.79	3.23	2.12	1.36	1.85	2.42	2.06	1.9	0.85	3.63	0.96

[a] SUP: surface area, DNM: distance from natural marshes, DCO: distance from the breeding site, DWO: distance from the closest wooded hedge or copse, DTL: distance from power lines, DHA: distance from habitations, DPR: distance from principal roads, DSR: distance from secondary roads, HHS: height of hedges surrounding the paddy, NWS: number of wooded sides, CON: adjacent (1) or not (0) to damaged field.

[b] S.D. = standard deviation

logistic regression . . .), as in ANNs usually all the variables contribute to the models. However, the use of qualitative traits, which is possibly responsible for the important variation of contributions between different trainings, can be a problem for other classification methods. While keeping a small set of input variables, the temporal structure of damage should be usefully investigated if flamingos exhibit more site-fidelity than proximate response to environmental factors.

Acknowledgements

The authors are grateful to Dr L. Hoffmann, President of the Tour du Valat Foundation, for his support to this study. We are indebted to J. Toutain and A. Sanche, Aéroclub de Montpellier, C. Pin, D. Vernet and V. Heurteaux, Station Biologique de la Tour du Valat, for their help in the field and aerial surveys. The authors thank R. Viannet, Parc Naturel Régional de Camargue, M. Roux-Cuvellier, Centre Français du Riz, Dr. S. Mañosa i Rifé, University of Barcelona, Spain, for their availability and all the numerous and interesting discussions they had. This study was funded by the Sansouïre Foundation, the MAVA Foundation, and the Centre Français du Riz.

References

André, P., Johnson, A R , 1981. Le problème des flamants roses dans les rizières de Camargue et les résultats de la campagne de dissuasion du printemps. Courrier du Parc Naturel Régional de Camargue 22–23, 20–35.

Avery, M L , Decker, D.G., 1994. Field tests of a copper-based fungicide as a bird repellent rice seed treatment. in: Halverson, W.S., Crabb, A.C. (Eds.), Proceedings of the 16th Vertebrate Pest Conference, University of California, Davis, pp. 250–254.

Avery, M.L., Decker, D.G., Humphrey, J.S., Aronov, E., Linscombe, S.D., Way, M.O., 1995. Methyl anthralinate as a seed treatment to deter birds. Journal of Wildlife Management 59, 50–56.

Barbier, J.M., Mouret, J.C., 1992. Le riz et la Camargue. Histoire et Recherche. INRA mensuel 64/65, 39–51.

Berryman, J.H., 1966. Statement for Bird Problem Meeting Sponsored by the American Farm Bureau. October 27, 1966. Division of Wildlife Services, Bureau of Sport Fisheries and Wildlife, Washington DC, 6 pp.

Brugger, K.E., Labisky, R.F., Daneke, D.E , 1992. Blackbird roost dynamics at Millers Lake, Louisiana: implications for damage control in rice. Journal of Wildlife Management 56, 393–398.

Caughley, G. and Sinclair, A.R.E. 1994. Wildlife Ecology and Management. Blackwell Scientific Publishing, Boston, 334 pp.

Chauvelon, P., 1996. Hydrologie quantitative d'une zone humide méditérranéenne aménagée: le bassin du Fumemorte en Grande Camargue, delta du Rhône. PhD, University of Montpellier, 254 pp.

Decker, D.G., Avery, M.L. and Way, M.O., 1990. Reducing blackbird damage to newly planted rice with a nontoxic clay-based seed coating. In: Davis, L R., Marsh, R.E. (Eds.), Proceedings of the 14th Vertebrate Pest Conference, University of California, Davis, pp. 327–331.

Durieux, L., 1997. Le paysage rizicole camarguais et son impact sur l'avifaune MSc, University of Aix-Marseille, 124 pp.

Edgell, J., Williams, G., 1991 The financial and economic valuation of goose grazing in the European Community. In: Van Roomen, M., Madsen, J. (Eds.), Waterfowl and Agriculture: Review and Future Perspective of the Crop Damage Conflict in Europe, Special Publication 21. IWRB, Slimbridge, pp. 79–80.

Elliot, C.C.H., 1979. The harvest time method as a means of avoiding *Quelea* damage to irrigated rice in Chad/Cameroun. Journal of Applied Ecology 16, 23–35.

Fasola, M., Ruiz, X., 1996. The value of rice fields as substitutes of natural wetlands for waterbirds in the Mediterranean Region (Special Publication 1). Colonial Waterbirds 19, 122–128.

Gallant, S I., 1993. Neural network learning and expert systems, MIT Press, London, pp. 365.

Garson, G.D., 1991. Interpreting neural network connection weights. Artificial Intelligence Expert 6, 47–51.

Goh, A.T.C , 1995. Back-propagation neural networks for modeling complex systems. Artificial Intelligence Engineering 9, 143–151.

Hoffmann, L., Johnson, A R., 1991. Extent and control of flamingo damage to rice crops in the Camargue (Rhône delta, southern France). In: Pintos Martin, M.R., Pietro Ojeda, S., Rendon Martos, M., Johnson, A.R. (Eds.), Reunión Técnica Sobre la Situación y Problemática del Flamenco Rosa (*Phoenicopterus ruber roseus*) en el Mediterráneo Occidental y Africa Noroccidental. Agencia de Medio Ambiente, Sevilla, pp. 119–127.

Holler, N R , Naquin, H.P., Lefebvre, P.W., Otis, D.L., Cunningham, D.J., 1982. Mesurol® for protecting sprouting rice from blackbird damage in Louisiana. Wildlife Society Bulletin 10, 165–170

Jimenez, X. and Soler, E., 1996. El Conflicto Entre los Flamencos y los Arroceros en el Delta del Ebro; Estudio Preliminar, Limonium S.C.P., Tarragona, 48 pp.

Johnson, A.R., 1989. Movements of greater flamingos (*Phoenicopterus ruber roseus*) in the Western Palearctic. Revue d'Ecologie (Terre Vie) 44, 75–94.

Johnson, A.R., Mesléard, F. 1997.Les flamants et la rizicul-ture. In: Clergeau, P., Oiseaux à Risques en Ville et à la Campagne, INRA, Paris, pp. 53–60.

Katondo, J.J.M., 1996. Ecology of *Anatidae* and their damage to rice crops at Lower Moshi irrigation scheme, northern Tanzania. In: Birkan, M., van Vessem, J., Havet, P., Madsen, J., Trolliet, B., Moser, M. (Eds.), Proceedings of the *Anatidae* 2000 Conference, 5–9 December 1994, Strasbourg, France,Gibier Faune sauvage, 13· 737–750.

Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Manel, S., Dias, J.M. and Ormerod, S.J , 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecological Modelling, (in press).

Mastrorillo, S , Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biology 38, 237–246.

Meanley, B., 1971. Blackbirds and the southern rice crop. U.S. Department of the Interior, Fish and Wildlife Service. Research Publication 100, Washington, 64 pp.

Morrisson, K., 1975. War on birds. Defenders of Wildlife News 50, 17–19.

O'Connor, R.J. and Shrubb, M., 1986. Farming and Birds. Cambridge University Press, Cambridge, 290 pp.

Rendon Martos, M , Johnson, A.R., 1996 Management of nesting sites for greater flamingos. Colonial Waterbirds 19, 167–183

Rogers, N.H.L. 1995. Distribution and impact of vertebrate pests in rice fields of the Camargue, southern France. MSc, University of Reading, 61 pp.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536

Sourribes, V.C., 1993. Etude du phenomène d'incursion des flamants roses dans les rizières de Camargue. Parc Naturel Régional de Camargue/Station Biologique Tour du Valat, Arles, 33 pp.

Spitz, F., Lek, S., Dimopoulos, I., 1996. Neural network models to predict penetration of wild boar into cultivated fields Journal of Biological Systems 4, 433–444.

Van Vessem, J., Draulans, D., de Bont, A.F., 1985. The effects of killing and removal on the abundance of grey herons at fish farms In: Proceedings of the XVIIth Congress of the International Union of Game Biologists, Brussels, Belgium, September 1985, pp. 337–343.

Ward, P.W., 1979 Rational strategies for the control of *Quelea* and other migrant bird pests in Africa. Philosophical Transactions of the Royal Society of London 287, 289–300

Wilson, E.A., LeBoeuf, E.A., Weaver, K.M., LeBlanc, D.J., 1989. Delayed seeding for reducing blackbird damage to sprouting rice in south-western Louisiana. Wildlife Society Bulletin 17, 165–171.

ELSEVIER

Ecological Modelling 120 (1999) 359

ECOLOGICAL MODELLING

www elsevier.com/locate/ecomodel

Announcement

2nd International Conference on Applications of Machine Learning to Ecological Modelling Adelaide, Australia, 27 November–1 December 2000

Organisational sponsors

University of Adelaide
University Paul Sabatier Toulouse
CNRS, Department des Sciences de la Vie, Paris
International Society for Ecological Modelling
Modelling and Simulation Society of Australia and New Zealand

Organising committee

F Recknagel, School of Ecology, University of Adelaide
Glen Osmond, SA 5064, Australia
Tel: +61-8-83036787; fax: +61-8-83036511
E-mail: Friedrich.Recknagel@adelaide.edu.au

S Lek, National Research Centre for Aquatic Ecosystems, University of Toulouse
118 Route de Narbonne, 31062 Toulouse, France
Tel: +33-5-61558687; fax: +33-5-61556096
E-mail: lek@cict.fr

Objectives

The conference will provide a forum for the presentation and discussion of recent research on machine learning such as artificial neural networks and genetic algorithms, and its application to ecological modelling. Ecosystems are characterised by high non-linearities and complexity, which artificial neural networks and genetic algorithms seem to be suited to. Therefore modelling by machine learning is expected to improve the understanding and prediction of aquatic and terrestrial ecosystems.

The aim of the conference is to encourage and facilitate interdisciplinary communication and research amongst professionals in machine learning, ecological modelling and ecosystem management.

A number of specialised sessions will be organised to focus on following themes:

- Ecological Applications of Artificial Neural Networks
- Ecological Applications of Genetic Algorithms
- Hybrid Modelling of Ecosystems by Machine Learning
- Elucidation, Monitoring and Forecasting of Ecosystems by Machine Learning
- Analysis and Synthesis of Ecological Data by Machine Learning

Call for papers

Papers are invited on the topics outlined above and others which fall within the general scope of the conference. Abstracts should be submitted to the Conference Secretariat by 30 November 1999 by electronic mail to Friedrich.Recknagel@adelaide.edu.au or lek@cict.fr.

ELSEVIER

Ecological Modelling 120 (1999) 361

ECOLOGICAL MODELLING

Author index of volume 120

PII: S0304-3800(99)00153-2

ELSEVIER

Ecological Modelling 120 (1999) 363–364

ECOLOGICAL MODELLING

Subject index of volume 120

PII: S0304-3800(99)00154-4

ELSEVIER

Ecological Modelling 120 (1999) 365–366

ECOLOGICAL MODELLING

Contents of volume 120