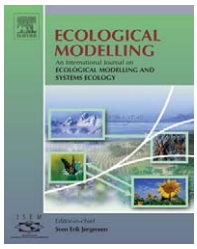
ECOLOGICAL MODELLING 195 (2006) 83–93

available at www.sciencedirect.com

SCIENCE DIRECT®

journal homepage: www.elsevier.com/locate/ecolmodel



Methodological issues in building, training, and testing artificial neural networks in ecological applications

Stacy L. Özesmi[a,*], *Can O. Tan*[b], *Uygar Özesmi*[a]

[a] *Environmental Science Chair, Department of Environmental Engineering, Erciyes University, 38039 Kayseri, Turkey*
[b] *Department of Biology, Middle East Technical University, 06531 Ankara, Turkey*

ARTICLE INFO

Keywords:
Artificial neural networks
Back-propagation
Modelling
Model
Model generalizability
Ecology

ABSTRACT

We evaluate the use of artificial neural networks, particularly the feedforward multilayer perceptron with back-propagation for training (MLP), in ecological modelling and make suggestions on its use. In MLP modelling, there are no assumptions about the underlying form of the data that must be met as in standard statistical techniques. Instead, researchers must clarify the process of modelling, as this is most critical to how the model performs and is interpreted. Overfitting on the data, a potential problem, can be avoided by limiting the complexity of the model and by using techniques such as weight decay, training with noise, and limiting the training of the network. Methods on when to stop training include: (1) early stopping based on cross-validation, (2) stopping after a analyst defined error is reached or after the error levels off, and (3) use of a test data set. The third method is not ideal as the test data set is then not independent of model development and the resulting model may have little generalizability. The importance of an independent data set cannot be overemphasized as we found dramatic differences in model accuracy assessed with prediction accuracy on the training data set, as estimated with bootstrapping, and from use of an independent data set. The comparison of the artificial neural network with a general linear model (GLM) as a standard procedure is recommended because a GLM may perform as well or better than the MLP. In such cases, there are no interactions or non-linear terms that need to be modelled and it will save time to use the GLM. Techniques such as sensitivity analyses, input variable relevances, neural interpretation diagrams, randomization tests, and partial derivatives should be used to make MLP models more transparent, and further our ecological understanding, an important goal of the modelling process. Based on our experience we discuss how to build an MLP model and how to optimize the parameters and architecture. The process should be explained explicitly to make the MLP models more readily accepted by the ecological research community at large, as well as to make it possible to replicate the research.

1. Introduction

Ecological data are typically highly complex and non-linear. For example, species exhibit variability in both space and time with changing environmental conditions, including historical conditions. Species composition and abundance are also affected by predators, competitors, and parasites. Many species are rare and consequently ecological data can contain

* *Corresponding author*. Tel.: +90 352 437 4937/32800; fax: +90 352 437 4404.
E-mail address: stacy@erciyes.edu.tr (S.L. Özesmi).

0304-3800/$ – see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2005.11.012

many zeroes. In contrast, some species can be observed very frequently and/or at very high densities. Because of these difficulties ecologists are continually searching for new modelling paradigms.

Ecologists started using artificial neural networks for modelling in the 1990s. Artificial neural networks were reported to have advantages for ecological studies where data rarely meet parametric statistical assumptions and where non-linear relationships are prevalent. They were also reported to perform better than linear models and generalize well to new data. However, artificial neural networks also have disadvantages. They are computationally intensive. Many parameters must be determined with few guidelines and no standard procedure to define the architecture. No global method exists for determining when to stop training and thus overtraining is problematic. Neural networks are sensitive to composition of the training data set and to initial network parameters. Finally, they are perceived as black box models.

Perhaps because it is one of the easiest neural networks to understand, the feedforward multilayer perceptron, with back-propagation for training, has been the most commonly used neural network in ecology. More details on how this type of neural network works can be found elsewhere (i.e., Lek and Guegan, 1999; or in texts such as Anderson, 1995; Weiss and Kulikowski, 1991; Bishop, 1995; or Ripley, 1996).

In this article, we review the use of the MLP, or feedforward multilayer perceptron with back-propagation for training, in ecological modelling and how it is practiced. Based on our experience we discuss how to build MLP models and how to optimize the parameters and architecture. We make recommendations for use of the MLP, which include the importance of avoiding overfitting, use of an independent test data set, and use of sensitivity analyses, neural interpretation diagrams, input variable relevances, and other methods to open up the black box model. Although in this article we focus on the MLP, some of our recommendations are also relevant to other types of artificial neural networks.

2. Literature review

Early papers on the use of MLP for ecological applications showed that MLP was a viable technique and had advantages over linear models. These included Brey et al. (1996), who predicted benthic invertebrate production/biomass ratios; Levine et al. (1996), who classified soil structure from soil sample data; Tan and Smeins (1996), who predicted changes in the dominant species of grassland communities based on climatic input variables; and Poff et al. (1996), who modelled streamflow response based on average daily precipitation and temperature inputs. Paruelo and Tomasel (1997) predicted normalized difference vegetation index (NDVI) used in remote sensing.

Later papers expanded the range of ecological applications for MLP. Phytoplankton production (Scardi, 1996, 2001; Scardi and Harding, 1999) and phytoplankton occurrence and succession (Recknagel et al., 1997; Karul et al., 2000) have been modelled with the MLP. Other modelling studies have included fish abundance based on habitat variables (Baran et al., 1996; Lek et al., 1996), fish yield (Lae et al., 1999), and fish and micro-habitat use (Reyjol et al., 2001). The MLP has been used to predict presence or absence, based on habitat variables, of macro-invertebrates (Hoang et al., 2001), birds (Manel et al., 1999), golden eagle nest sites (Fielding, 1999b), nests of interacting marsh-breeding birds (Özesmi and Özesmi, 1999), and cyanobacteria (Maier et al., 1998). The MLP has been used to predict damage to agricultural fields by flamingo (Tourenq et al., 1999) and wild boar (Spitz and Lek, 1999). Bird abundance (Lusk et al., 2001) and macro-invertebrate abundance and species richness (Lek-Ang et al., 1999) has been modelled.

In addition, later papers have started to addresss the particular problems associated with typical ecological data. For example, Scardi (2001) discussed constrained training and metamodelling as techniques to improve training of networks when training data is limited. Hoang et al. (2001) used sensitivity analyses to select relevant input variables, for each macro-invertebrate taxa, from a total 37 habitat variables. They recommended putting macro-invertebrate data in ecological or taxonomical groups to avoid difficulties in training neural networks when species either rarely or frequently occur. Tourenq et al. (1999) found that their model predicted most accurately in the test data when an equal number of presence and absence records were used in the training data. Their data had many more absence records than presence records. Dimopoulos et al. (1999), who used MLP to predict lead concentrations in grasses, demonstrated the use of partial derivatives to determine the sensitivity of predictions to input variables. This technique was also used by Reyjol et al. (2001).

2.1. *Criticisms of modelling with MLP*

From a literature review, we saw a few problems with the reporting on the use of MLP. Sometimes the modelling process was not clearly described. For example, some research did not report why certain variables were chosen for a final model. Others did not tell how the parameters were set or how the architecture, the number of hidden units, was determined. The number of samples used to train, validate and test the model was not always given.

2.2. *Overtraining*

However, the major problem was overtraining (overfitting) on data or giving vague statements on how it was avoided. An exception is Paruelo and Tomasel (1997), who provide a discussion of their experience with overtraining. Unfortunately, it seems that often studies do not make sufficient effort to avoid overfitting. To avoid overfitting on the data, the complexity of the model must be limited.

The best generalization performance is usually achieved with a model whose complexity is neither too small nor too large (Bishop, 1995). As a simple example, a first-order polynomial (straight line) may be too simple to identify the relationship between the independent and dependent variable but a 15th-order polynomial may be fitting noise. This is also known as the bias-variance trade-off. An oversimplified model will have a large bias while a too complex model will have a large variance. The best generalizabiltiy is achieved with the best compromise between bias and variance. One way to reduce