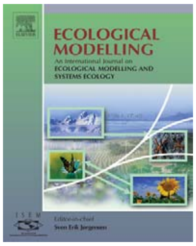
ECOLOGICAL MODELLING 197 (2006) 361–372

available at www.sciencedirect.com

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



Estimating risk of events using SOM models: A case study on invasive species establishment

Muriel Gevrey [a,*], *Sue Worner* [a], *Nikola Kasabov* [b], *Joel Pitt* [a], *Jean-Luc Giraudel* [c]

[a] *National Centre for Advanced Bio-Protection Technologies, Lincoln University, PO Box 84, Canterbury, New Zealand*
[b] *Knowledge Engineering and Discovery Research Institute (KEDRI), Auckland University of Technology (AUT), Technology Park, 581 Great South Road, Penrose, Auckland, New Zealand*
[c] *Équipe Périgourdine de Chimie Appliquée (EPCA), LPTC, Université Bordeaux 1, CNRS (UMR 5472) BP 1043, 24001 Périgueux Cedex, France*

ARTICLE INFO

Article history:
Received 22 June 2005
Received in revised form 22 March 2006
Accepted 28 March 2006
Published on line 12 May 2006

Keywords:
Artificial neural networks
Self-organising map
Fuzzy sets
Pest control
New Zealand

ABSTRACT

The use of advanced modelling methods in ecology expands as ecological data accumulates and increases in complexity. Artificial neural networks (ANN), and in particular, the self-organising map (SOM), has become very popular for analysing particular kinds of ecological datasets. As SOM have become more utilised, it has become increasingly clear that the results of SOM models must be interpreted carefully.

SOM have been used in a number of ecological studies to investigate the spatial distribution of species. When using presence–absence data of species distributions at given locations, the input vectors to a SOM are binary and the connection weights after learning are between 0 and 1. Using fuzzy set theory, we present an approach to the interpretation of these weights. Taking an example from invasive species research, we show that in the case of presence/absence data, a connection weight can be interpreted as a risk that an event will occur at a given location.

A SOM was used to model the worldwide distribution insect pests to determine geographic patterns and define the species assemblages. The SOM weights were used as a measure of the risk of invasion for each species such that its potential to invade a geographic area could be evaluated.

This paper shows that while there are limitations concerning the interpretation of a model parameter, it is still possible to obtain relevant information when such limits are recognised. We emphasise however, that the interpretation of SOM weights must be appropriate to the particular study of interest.

1. Introduction

As the amount and complexity of ecological data increases, powerful modelling methods are required for their analysis. Examples of classical statistical methods that have been used for many years to analyse ecological data are: linear regression (Ricker, 1975); multiple linear regression (Binns and Eiserman, 1979; Faussh et al., 1988); polar ordination (Whittaker et al., 1979); canonical correspondence analysis (ter Braak, 1987); principal component analysis (Grossman et al., 1991); multiple

---

[*] *Corresponding author at*: National Centre for Advanced Bio-Protection Technologies, Ecology and Entomology Group, Lincoln University, PO Box 84, Canterbury, New Zealand. Tel.: +64 3 325 3696x8382; fax: +64 3 325 3864.
E-mail address: gevrey@cict.fr (M. Gevrey).

0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.03.032

dimensional scaling (Guilherme and Cintra, 2001). However, new methods that are better adapted to complex data, are now available and are increasingly being applied in ecological research. They are genetic algorithms (Chaves et al., 2003), classification and regression trees (Gregor et al., 2002), artificial neural networks (ANN) (Brosse et al., 1999; Maier and Dandy, 2000; Michaelides et al., 2001; Cereghino et al., 2001), fuzzy neural networks and evolving connectionist systems (Kasabov, 1996, 2002).

The unsupervised ANN algorithm called the self-organising map (SOM) (Kohonen, 1982, 2001) is a method well adapted to complex ecological data analysis. This method is widely used in a variety of disciplines for vector quantisation and for classification. Some examples of the application of SOM in ecology are: to determine the relationship between complex ecological communities (Chon et al., 1996); to model microsatellite data (Giraudel et al., 2000); to study fish assemblages (Brosse et al., 2001); to detect pattern in aquatic macroinvertebrate diversity (Cereghino et al., 2003); to define conservation strategies for threatened endemic fish species (Park et al., 2003b).

The SOM algorithm is a method used to project high dimensional data vectors onto a lower dimensional space preserving the similarity and the difference between the data vectors. To assist visualisation of patterns and relationships in high dimensional data, the SOM is often used to project the dataset in a non-linear way onto a topological rectangular grid arranged as a hexagonal lattice that is called a map. In addition to this visualisation there is also an underlying SOM model whose outputs can be displayed in several different ways to reveal different types of information.

While the classical result from a SOM is a map (where the objects to be classified are mapped into map nodes), the internal parameters of the SOM that describe the relationship of each descriptor variable to the object to which it belongs, provide important ecological information. The internal parameters or the connection weights are the main features of the network and result from the SOM training on a data set. However, the meaning of the connection weights is not always clear and therefore their interpretation must be carefully considered. A particular case is their interpretation when the input vectors representing the presence/absence of species at locations are binary. We illustrate how SOM weights can be interpreted to give useful information in an ecological setting using an example from invasive species research. The particular example concerns global insect pest invasions and the potential of certain species to invade New Zealand (Worner et al., 2004).

2. Self-organising maps as models of risk of species establishment

Much research in pest risk assessment involves the assessment of the potential for establishment of invasive species and centres around using the existing geographic distribution of a particular species to determine its potential for establishment in areas where it is not normally found. A SOM is used here to detect pattern in global pest species assemblages associated with geographic areas. The input data to the SOM consisted of binary data indicating the presence and absence of an invasive species at certain geographic locations.

2.1. *The SOM algorithm*

Self-organising maps belong to the unsupervised artificial neural network modelling methods (Kohonen, 1982). The model typically projects a high dimensional dataset on to a lower dimensional space. The SOM network consists of two layers: the input and the output layers. The dataset presented to the network is comprised of samples characterised by p descriptors—variables. Each sample is represented by a vector that includes all p descriptors and there are as many *sample vectors* as samples.

The input layer is comprised of p nodes (neurons). The output layer forms a d-dimensional map, where $d<p$. In this study, the map is in the form of a rectangular 2D grid with l by m neurons laid out on a hexagonal lattice ($C=l\times m$ neurons in the output layer). Each neuron $\mathbf{c}_j$ of the output layer, also called a cell, is linked to the neurons $i=1, 2, \ldots, p$ of the input layer by connections that have weights $\boldsymbol{w}_{ij}$ associated with them, forming a vector $\boldsymbol{w}_{ij}$. These weights represent the virtual values for each descriptor in each output neuron such that each output neuron or cell of the output layer $\mathbf{c}_j$ stores a *virtual vector* of connection weights $\boldsymbol{w}_{ij}$. These virtual vectors represent the co-ordinates of centres of groups of similar input vectors, where similarity is measured in terms of Euclidean distance:

$$D(\boldsymbol{x}, \boldsymbol{w}_j) = \left[\sum_{i=1,\ldots,p} (\boldsymbol{x}_i - \boldsymbol{w}_{ij})^2\right]^{1/2},$$

for all neurons (cells) $\mathbf{c}_j$ and with $\boldsymbol{x}$ a sample vector (1)

The aim of the SOM algorithm is to organise the distribution of sample vectors in a d-dimensional space (in our case, two-dimensional) using their relationship to the virtual vector distribution thus preserving the similarity and the difference between the input vectors. Similar input vectors are allocated to the same virtual vector and the virtual vector changes with the addition of new input vectors to it. The virtual vectors that are neighbours on the map (neighbouring neurons) are expected to represent neighbouring groups (clusters) of sample vectors; consequently, sample vectors that are dissimilar are expected to be distant from each other on the map.

Two different learning algorithms could be used in a SOM: sequential or batch. The first one is an incremental algorithm that is commonly used but learning is highly dependent on the order of input. The batch algorithm overcomes this drawback. Furthermore, the batch algorithm is significantly faster (Kohonen and Somervuo, 1998) and was chosen for this study. The process involves presenting the whole sample vectors as input to the SOM at once. Using a distance measure, the sample vectors are compared to the virtual vectors that have been randomly assigned to the output neurons at the beginning of the algorithm (Fig. 1). Each sample vector is assigned to the nearest virtual vector according to the distance results and the virtual vectors are modified to the mean of the sample vectors that are assigned to it. Details about the algorithm can