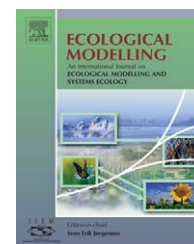


available at www.sciencedirect.comjournal homepage: www.elsevier.com/locate/ecolmodel

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.

© 2006 Elsevier B.V. All rights reserved.

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).

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 c_j of the output layer, also called a cell, is linked to the neurons $i = 1, 2, \dots, p$ of the input layer by connections that have weights w_{ij} associated with them, forming a vector 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 c_j stores a *virtual vector* of connection weights 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:

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

for all neurons (cells) c_j and with \mathbf{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

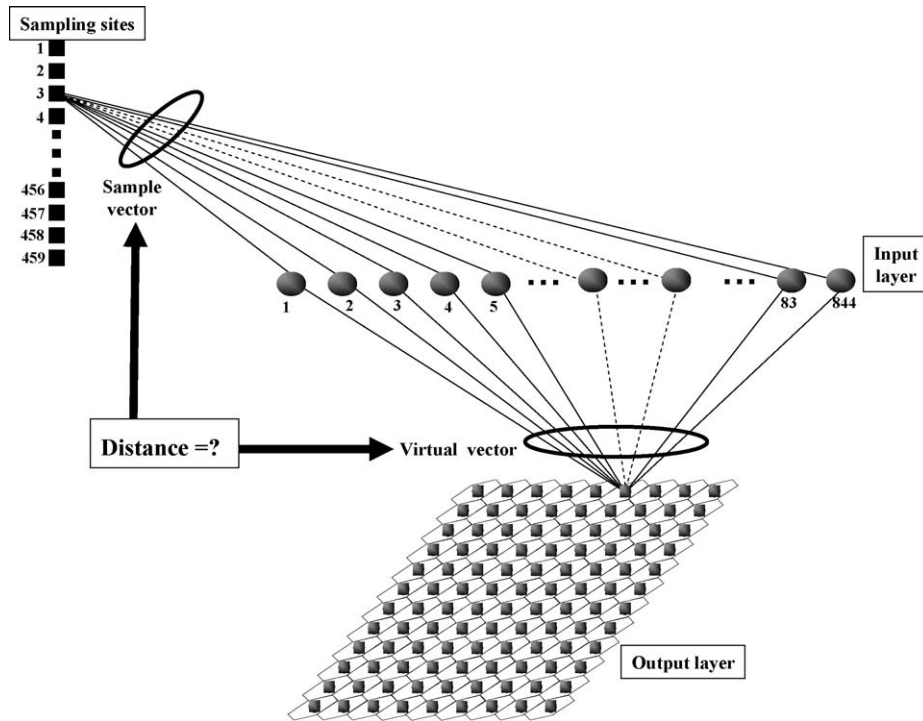


Fig. 1 – Self-organising map architecture. The input layer is linked to the cells of the output layer by connections called weights which defined the virtual assemblages of the species. The sample vector composed of as much element as descriptors is compared to the virtual vectors associated to each neuron of the output using a distance measure.

be found in Kohonen (1995, 2001) and Kohonen and Somervuo (1998).

At the end of the training, an output neuron has been determined for each sample vector such that each sample is then assigned to a neuron or cell of the map, and the virtual values of the descriptors are known for each neuron of the map.

2.2. SOM output

2.2.1. Sample distribution

A direct result of the SOM algorithm is a distribution of the samples on the SOM topological map. According to the properties of the algorithm, the samples that are in the same cell are very similar, and similar to those in neighbouring cells. They are less similar to samples that are in distant cells. Each cell unit is at this stage, a cluster and the SOM training procedure constitutes a clustering method, clustering the samples in cells and similar cells together.

The approximate number of cells in the output layer can be defined using the formula $C = 5\sqrt{n}$ (<http://www.cis.hut.fi/projects/soomtoolbox/documentation/somalg.shtml>, 2000) where C is the number of cells and n is the number of training samples (sample vectors).

2.2.2. Clustering information

Despite that a SOM clusters samples onto the cells of the map, it is of interest to define larger clusters by regrouping the neighbouring cells that contain similar samples. The definition of larger clusters can be achieved using several methods. A method well-known to experienced SOM users is the

unified-matrix (U-matrix) approach (Ultsch and Siemon, 1990). The U-matrix displays the distances between the virtual sites and provides a landscape formed by light plains separated by dark ravines.

A second method is a classical clustering analysis of SOM output using any of the classical distance measures and linkages, or the k-means method (Park et al., 2003a,b). These methods are applied to the results of the SOM model, or more precisely on the virtual vector for each neuron of the output layer.

2.2.3. Visualisation of input variables

To analyse the contribution of input variables to cluster structures of the trained SOM, each input variable and the connection weight of its associated descriptor calculated for each virtual vector during the training process, can be visualised in grey scale on the SOM map.

Remembering that each cell of the map is represented by a virtual vector, and also that each virtual vector is composed of as many weight values as descriptors, it is possible to visualise each descriptor's weight values associated with each neuron or cell of the trained SOM map. A map can be visualised separately for each descriptor.

2.2.4. Relationship between multiple descriptors

It may be important to investigate the relationship between sets of descriptors (input variables) and try to find meaningful patterns of their values in combination. For example, it might be important to investigate the relationship between biological and environmental variables across the samples.

A second set of descriptors for each sample can be introduced into the SOM and trained along with the first set of descriptors. Initially, each descriptor set is submitted to the trained SOM, and then, the mean value of each descriptor in the descriptor sets in each output of the trained SOM is calculated. If a neuron was not occupied by input vectors, the value is replaced with the mean value of neighbouring neurons. These mean values assigned on the SOM map can once again be visualised in grey scale and then compared with the map of the samples as well as other descriptor maps.

2.2.5. The connection weights

As described previously, the two layers of the SOM network are linked by connections that are called weights. The set of weight values for each output neuron comprises a virtual vector for that neuron. These weights represent the co-ordinates of each output neuron in a multidimensional space with as many dimensions as descriptors. The values of the weights taken independently for each descriptor in all the output neurons can be used to obtain the map discussed below. But in case of binary data, because the observed or real values are 0 or 1, the virtual values are constrained between 0 and 1. It should be immediately obvious that these values can be used as some sort of measurement, evaluation, gradient, or as an index depending on the content of the data set and the meaning of the descriptors. One interpretation theory is given in the next section.

2.3. The fuzzy set theory

In the classical approach of the set theory, if a subset A of a set E is considered, the characteristic function is χ_A . χ_A is a two-valued function taking its values in $\{0, 1\}$ and defined as follows:

$$\text{if } x \in A, \quad \chi_A(x) = 1, \quad \text{if } x \notin A, \quad \chi_A(x) = 0$$

But in some cases, the membership of an element to a given subset is imprecise. Zadeh (1965) proposed fuzzy set theory to explicitly account for this. In this case, the characteristic function is replaced by the membership function f_A where f_A is a real-valued function taking its values in $[0; 1]$. Now the function $f_A(x)$ gives an indication of the degree of truthfulness for x to be member to the fuzzy subset A .

2.4. Evaluation of risk of species establishment using SOM weights

In our model, each input vector x (sample) represents a geographical location (site) and each descriptor (input variable) x_i in it represents the presence (1) or absence (0) of a particular species at this site. In this way, a geographical site is represented by its vector of species and sites are compared for their similarity based on the occurrence of these species. The more similar the assemblage of species that are present (or not present) at two locations x and y , the more similar these locations are considered to be (see formula (1)) as it is assumed that they share similar conditions for them to establish viable populations at these locations (or not to establish).

When a SOM model is trained on such data, a virtual vector c_j will represent a center of many similar sites (similarity is measured across all variables (species)), that is also represented as a node with its location in the output topological SOM map (e.g. 2D).

A virtual vector c_j is represented in the p -dimensional space by its weight vector w_j as a geometrical center of similar vectors (sites) and each co-ordinate w_{ij} is a representation of the similarity of all vectors $x_{j1}, x_{j2}, \dots, x_{jk}$ allocated to the node c_j in respect only to the variable x_i . A value w_{ij} represents an index of the risk for the species x_i to occur at a site vector mapped into the cell c_j based on the overall similarity of all sites (mapped at this cell). The value w_{ij} is not based on the number of sites that have the species x_i present, among all sites mapped into c_j (this would be a probability measure based on a single variable). Generally speaking, probability is a ratio expressing the chances that a certain event will occur. The simple mathematical definition would be that the probability of an event is a numerical measure of the likelihood or degree of predictability that the event will occur (Glover and Mitchell, 2002). The mathematical formulation is: the total number of occurrences of a selection/event a , divided by the total number of occurrences (a) plus the number of failures of those occurrences b (i.e. total possible outcomes). This leads to the basic formula:

$$p = \frac{a}{a + b} \quad (2)$$

If we cluster all samples from a data set into clusters (using some statistical clustering methods, e.g. k-means, Fuzzy C-means clustering, etc.) and for each cluster C_j , that has a statistically sufficient number of samples in it, we count the number of the samples with a particular event x_i present (1) and not present (0) respectively as a_j and b_j , then we can calculate the probability of an event x_i to occur (within the samples of this cluster only) as

$$p_i = \frac{a_i}{a_i + b_i} \quad (3)$$

But the SOM procedure does not do this exactly. It maps samples into a (2D) map of N neurons (cells) and allocates similar samples to the same neuron depending on the number N . The connection weights represent the respective co-ordinates of the neurons in the input space and have the meaning of geometrical centres. Generally speaking, the values of the input variables in a SOM model can be 0 (event not present) and 1 (event present), but they can also be any number of discrete values (e.g. 0–4, representing, for example, not present for sure, not likely to be present, may be present, likely to be present, present for sure). So w_{ij} is not a probability but a measure of the truthfulness of the statement: “the species x_i is present”. If A_i is the fuzzy set of the sites that have the species x_i present, w_{ij} can be seen as the value of the membership function f_{A_i} .

If a new vector x (representing a site) is mapped into a trained SOM and x maps into a node c_j then the connection weight w_{ij} can be seen as $f_{A_i}(x)$ and this connection weight would indicate a risk for an event x_i to occur at the site.

3. Case study: the risk of insect pest invasion in New Zealand

3.1. The data

The data used in this analysis were extracted from the Crop Protection Compendium (Global Module, 5th ed.[©], CAB International, Wallingford, UK, 2003). This science-based tool encompasses a wide range of different types of information on all aspects of crop protection (e.g. pests, diseases, weeds, natural enemies, crops) associated with all countries in the world. The geographic areas represented in the compendium consist of countries, regions or states of countries distributed worldwide. All continents are represented. The full compendium includes many species for which only partial information is available. In order to ensure that we include species with adequate distribution information, we decided to select only species which occur in more than 2% of sample units (Waite, 2000). A total of 844 mainly phytophagous insect pests for 459 geographic areas were used in our analysis. Representing the presence (1) and the absence (0) of each species at each site, resulted in a database comprised of a [844 species × 459 sites] matrix.

3.2. The model

The 459 sample sites were used to train a SOM model and were consequently mapped into the SOM 2D map. The input vari-

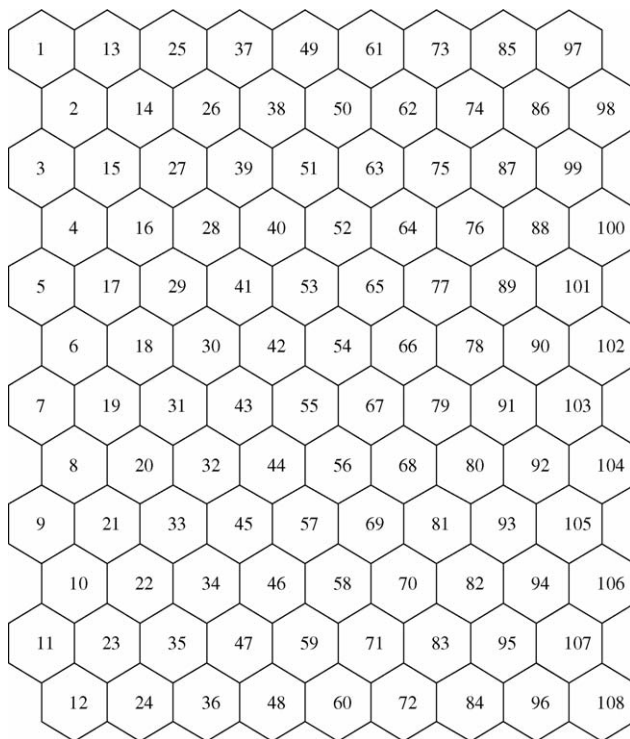


Fig. 2 – [12 × 9] Self-organising map (108 cells). The number inside each cell corresponds to the number of the output neuron and can be linked to the Appendix A to define which geographic area has been plotted in each cell.

ables were the establishment (presence and absence) of 844 pest species. The input layer was therefore composed of 844 neurons. The output layer consisted of 108 neurons organised in an array with 12 rows and 9 columns.

The Euclidean distance was the distance measurement method that provided the most accurate data representation on the map. A cluster analysis to detect cluster boundaries on the trained map also used the Euclidean distance measure and the Ward linkage method. The Davies–Bouldin Index (DBI) (Davies and Bouldin, 1979) was calculated to identify the optimum number of clusters.

4. Results and discussion

The non-linear projection of presence-absence data on to two-dimensional space allowed us to classify the geographic areas according to the similarity of their pest species assemblages (Fig. 2 and Appendix A). The results of the clustering method applied to the SOM results are shown in Fig. 3a. The optimum cluster number was six (Worner et al., 2004) (Ia, Ib1, Ib2, IIa, IIb1 and IIb2; Fig. 3b) as determined by the DBI which for this cluster number was 0.981.

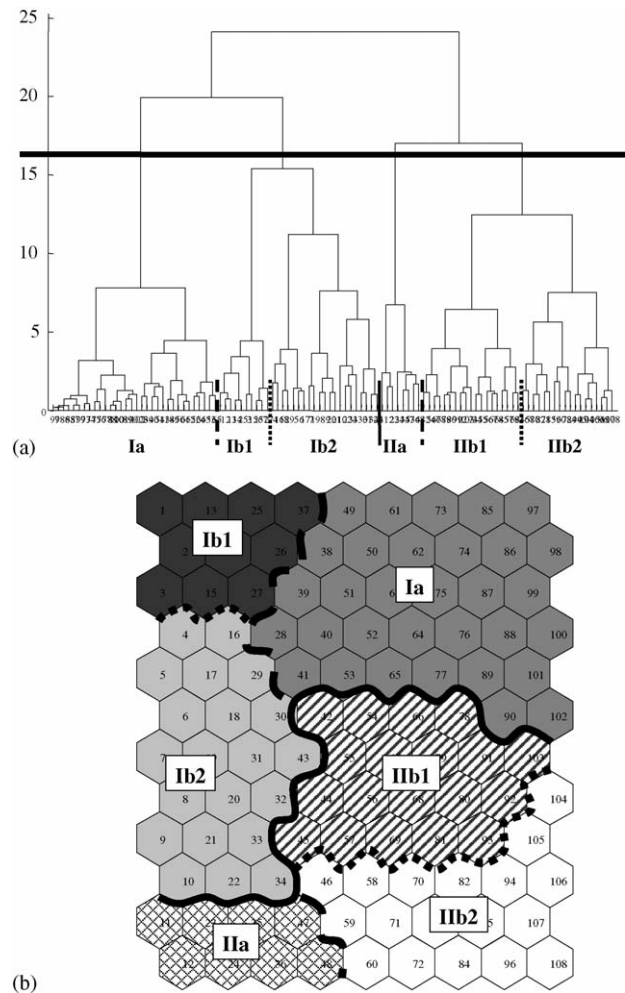


Fig. 3 – (a) Dendrogram of the cluster analysis; (b) self-organising map with the six clusters defined by the cluster analysis: Ia, Ib1, Ib2, IIa, IIb1 and IIb2 (see text).

The cluster Ia represents geographic areas that have low pest occurrence values or at least have low numbers of pests recorded in the database. In fact these countries can be considered outliers. Around 85% of those geographic areas have only 2% of the species present. These geographic areas are

mainly small islands, desert areas or colder areas of Greenland, Alaska and parts of Russia. The United States and most of the Canadian regions comprise cluster Ib1. Cluster Ib2 includes New Zealand as well as several regions of Australia (South Australia, Tasmania, Victoria, Western Australia), and

Table 1 – List of the pest species out of a potential 844 which have the highest potential risk of invasion in New Zealand based on the SOM analysis

Name	code	risk	p/a	Name	code	risk	p/a
<i>Planococcus citri</i> *	PSECCI	0,93	1	<i>Toxoptera aurantii</i>	TOXOAU	0,49	1
<i>Icerya purchasi</i>	ICERPU	0,92	1	<i>Taylorilygus pallidulus</i>	TAYLPA	0,49	0
<i>Myzus persicae</i>	MYZUPE	0,87	1	<i>Aleurothrixus floccosus</i>	ALTHFL	0,48	0
<i>Cydia pomonella</i>	CARPPPO	0,86	1	<i>Pseudaulacaspis pentagona</i>	PSEAPE	0,48	0
<i>Nezara viridula</i>	NEZAVI	0,85	1	<i>Pieris rapae</i>	PIERRA	0,47	1
<i>Brevicoryne brassicae</i>	BRVCBR	0,83	1	<i>Hadula trifolii</i>	SCOOTR	0,47	0
<i>Delia platura</i>	HYLEPL	0,80	1	<i>Ephestia elutella</i>	EPHEEL	0,47	1
<i>Phthorimaea operculella</i>	PHTOOP	0,79	1	<i>Rhopalosiphum rufiabdominale</i>	RHOPRU	0,46	1
<i>Pseudococcus longispinus</i>	PSECAD	0,79	1	<i>Liriomyza trifolii</i>	LIRITR	0,46	0
<i>Aphis spiraeola</i>	APHISI	0,77	1	<i>Sitona discoideus</i>	SITODI	0,46	1
<i>Saissetia oleae</i>	SAISOL	0,77	1	<i>Spodoptera exigua</i>	LAPHEG	0,46	0
<i>Coccus hesperidum</i>	COCCHE	0,77	1	<i>Sitobion avenae</i>	STOBAB	0,45	0
<i>Aonidiella aurantii</i>	AONDAU	0,76	1	<i>Therioaphis trifolii</i>	THAPTR	0,45	1
<i>Eriosoma lanigerum</i>	ERISLA	0,76	1	<i>Locusta migratoria</i>	LOCUMI	0,45	1
<i>Aphis gossypii</i>	APHIGO	0,76	1	<i>Prays citri</i>	PRAYCI	0,43	0
<i>Viteus vitifoliae</i>	VITEVI	0,75	1	<i>Hippotion celerio</i>	HPPOCE	0,43	1
<i>Ceratitis capitata</i>	CERTCA	0,73	0	<i>Pantomorus cervinus</i>	PANMCE	0,43	1
<i>Agrotis ipsilon</i>	AGROYP	0,73	1	<i>Schizaphis graminum</i>	SCZAGR	0,42	0
<i>Bemisia tabaci</i>	BEMITA	0,70	1	<i>Oulema melanopus</i>	LEMAME	0,42	0
<i>Helicoverpa armigera</i>	HELIAR	0,70	1	<i>Scolytus rugulosus</i>	SCOLRU	0,42	0
<i>Acyrtosiphon pisum</i>	ACYRON	0,70	1	<i>Drosophila melanogaster</i>	DROSME	0,42	0
<i>Thrips tabaci</i>	THRITB	0,69	1	<i>Sitona lineatus</i>	SITOLI	0,42	1
<i>Saissetia coffeae</i>	SAISSE	0,68	1	<i>Mythimna unipuncta</i>	PSEDUN	0,41	0
<i>Rhopalosiphum maidis</i>	RHOPMA	0,68	1	<i>Pectinophora gossypiella</i>	PECTGO	0,41	0
<i>Plutella xylostella</i>	PLUTMA	0,68	1	<i>Hellula undalis</i>	HLULUN	0,41	1
<i>Chrysomphalus dictyospermi</i>	CHYSDI	0,67	0	<i>Peridroma saucia</i>	PERISA	0,41	0
<i>Aspidiotus nerii</i>	ASPDNE	0,67	1	<i>Parlatoria ziziphi</i>	PARLZI	0,41	0
<i>Frankliniella occidentalis</i>	FRANOC	0,61	1	<i>Gonipterus gibberus</i>	GONPSC	0,40	1
<i>Rhopalosiphum padi</i>	RHOPPA	0,61	1	<i>Acanthoscelides obtectus</i>	ACANOB	0,40	1
<i>Hyperomyzus lactucae</i>	HYPELA	0,61	1	<i>Ceroplastes floridensis</i>	CERPFL	0,40	1
<i>Agrius convolvuli</i>	HERSCO	0,60	1	<i>Parasaissetia nigra</i>	SAISNI	0,40	1
<i>Diaspidiotus perniciosus</i>	QUADPE	0,60	1	<i>Lixus juncii</i>	LIXUJU	0,40	0
<i>Aphis fabae</i>	APHIFA	0,60	0	<i>Sminthurus viridis</i>	SMINVI	0,40	1
<i>Phoracantha semipunctata</i>	PHOASE	0,59	1	<i>Diaspidiotus ostreaeformis</i>	QUADOS	0,39	1
<i>Heliothrips haemorrhoidalis</i>	HEALTHA	0,59	1	<i>Henosepilachna elaterii</i>	EPILCH	0,39	0
<i>Macrosiphum euphorbiae</i>	MACSEU	0,59	1	<i>Lepidosaphes ulmi</i>	LEPSUL	0,39	0
<i>Phylloxera vitifoliae</i>	PHYNCI	0,58	0	<i>Scrobipalpa ocellatella</i>	PHTOOC	0,38	0
<i>Ceroplastes rusci</i>	CERPRU	0,57	0	<i>Siphoninus phillyreae</i>	SPNNPH	0,38	1
<i>Chrysomphalus aonidium</i>	CHYSFI	0,57	0	<i>Antigastra catalaunalis</i>	ANTICA	0,38	0
<i>Parthenolecanium persicae</i>	PTLCPE	0,56	1	<i>Unaspis citri</i>	UNASCI	0,38	0
<i>Trichoplusia ni</i>	TRIPNI	0,55	0	<i>Mythimna loreyi</i>	MYTHLO	0,37	0
<i>Cadra cautella</i>	EPHECA	0,54	0	<i>Thrips simplex</i>	TAETSI	0,37	1
<i>Lepidosaphes beckii</i>	LEPSBE	0,54	0	<i>Bactrocera oleae</i>	DACUOL	0,37	0
<i>Aphis craccivora</i>	APHICR	0,54	1	<i>Lipaphis erysimi</i>	LIPAER	0,37	1
<i>Lampides boeticus</i>	LAMDBO	0,54	1	<i>Spodoptera littoralis</i>	SPODLI	0,36	0
<i>Agrotis segetum</i>	AGROSE	0,54	0	<i>Orthezia insignis</i>	ORTHIN	0,36	0
<i>Sitophilus zeamais</i>	CALAZM	0,53	0	<i>Prays oleae</i>	PRAYOL	0,36	0
<i>Pieris brassicae</i>	PIERBR	0,53	0	<i>Listroderes costirostris</i>	LISTCO	0,36	1
<i>Hemiberlesia lataniae</i>	HEBELA	0,52	1	<i>Liriomyza huidobrensis</i>	LIRIHU	0,35	0
<i>Toxoptera citricida</i>	TOXOCI	0,52	1	<i>Cryptoblabes gnidiella</i>	CRYBGN	0,35	1
<i>Parthenolecanium corni</i>	PTLCCO	0,50	1	<i>Sesamia cretica</i>	SESACR	0,35	0
<i>Grapholita molesta</i>	LASPMO	0,50	1	<i>Acronicta rumicis</i>	ACRNRU	0,34	0
<i>Metopolophium dirhodum</i>	METODR	0,49	1	<i>Dialeurodes citri</i>	DIALCI	0,34	0
<i>Hemiberlesia rapax</i>	HEBERA	0,49	1	<i>Gynaikothrips ficorum</i>	GYNAFI	0,34	0

Table 1 – (Continued)

Name	code	risk	p/a	Name	code	risk	p/a
Lobesia botrana	POLYBO	0,33	0	Trialeurodes vaporariorum	TRIAVA	0,24	1
Aphis pomi	APHIPO	0,33	0	Opogona sacchari	OPOGSC	0,24	0
Bemisia tabaci (B biotype)	BEMIAR	0,33	1	Spodoptera frugiperda	LAPHFR	0,24	0
Epidiaspis leperii	EPDALE	0,33	0	Helicoverpa zea	HELIZE	0,24	0
Dysmicoccus brevipes	DYSMBR	0,32	0	Heliolithis virescens	HELIVI	0,24	0
Philaenus spumarius	PHILSU	0,32	1	Deilephila elpenor	DEILEL	0,23	0
Pissodes castaneus	PISONO	0,32	0	Sitona humeralis	SITOHU	0,23	0
Euproctis chryorrhoea	EUPRCH	0,32	0	Mayetiola destructor	MAYEDE	0,23	1
Limothrips cerealium	LIMITCE	0,32	1	Caliroa cerasi	ERICLI	0,23	1
Chromatomyia horticola	PHYYHO	0,32	0	Mamestra brassicae	BARABR	0,23	0
Listronotus bonariensis	LISRBO	0,31	1	Leptinotarsa decemlineata	LEPTDE	0,23	0
Spoladea recurvalis	HYMERE	0,31	1	Oryzaephilus surinamensis	ORYZSU	0,23	0
Etiella zinckenella	ETIEZI	0,31	0	Saturnia pyri	STURPY	0,23	0
Frankliniella schultzei	FRANSC	0,31	0	Parlatoria pergandii	PARLPE	0,22	0
Brucephagus roddi	BRPHRO	0,31	1	Saccharicoccus sacchari	PSECSA	0,22	0
Meligethes aeneus	MELIAE	0,31	0	Elasmopalpus lignosellus	ELASLI	0,22	0
Lymantria dispar	LYMADI	0,31	0	Phenacoccus madeirensis	PHENMD	0,22	0
Anarsia lineatella	ANARLI	0,30	0	Brachycaudus helichrysi	ANURHE	0,21	1
Sitobion fragariae	STOBFK	0,30	1	Maruca vitrata	MARUTE	0,21	0
Pseudococcus calceolariae	PSECGA	0,30	1	Tipula paludosa	TIPUPA	0,21	0
Otiorynchus sulcatus	OTIOSU	0,30	1	Manduca sexta	PROTSE	0,21	0
Tribolium castaneum	TRIBCA	0,29	0	Delia antiqua	HYLEAN	0,21	0
Operophtera brumata	OPRPBM	0,29	0	Phyllotreta cruciferae	PHYECR	0,21	0
Hypera postica	HYRPO	0,29	0	Cryptolestes pusillus	CRYLPU	0,21	1
Liposcelis bostrychophila	LIPOBO	0,29	0	Grapholita funebrana	LASPFU	0,21	0
Parlatoria oleae	PARLOL	0,29	0	Lepidosaphes gloverii	LEPSGL	0,21	0
Myzus cerasi	MYZUCE	0,29	1	Atherigona orientalis	ATHEOR	0,21	0
Acanthophilus helianthi	ACAIHE	0,29	0	Pterochloroides persicae	PTCHPE	0,21	0
Acyrtosiphon kondoi	ACYRKO	0,29	1	Araecerus fasciculatus	ARAEFA	0,21	1
Nipaeoccus nipae	NIPANI	0,28	0	Liriomyza bryoniae	LIRIBO	0,20	0
Melanaphis sacchari	MELHSA	0,28	0	Sitotroga cerealella	SITTCE	0,20	0
Laodelphax striatellus	CALGMA	0,28	0	Delia radicum	HYLEBR	0,20	0
Malacosoma neustria	MALANE	0,27	0	Metopolophium festucae	METOFK	0,20	1
Earias insulana	EARIIIN	0,27	0	Eupoecilia ambiguella	CLYSAM	0,20	0
Nesidiocoris tenuis	CRTOTE	0,27	0	Liriomyza brassicae	LIRIBC	0,20	1
Pentalonia nigronervosa	PENLNI	0,27	0	Aulacophora foveicollis	AUACFO	0,20	0
Bedellia somnulentella	BDLLSO	0,27	1	Neotalitrus tenellus	CIRCTE	0,20	0
Ostrinia nubilalis	PYRUNU	0,27	0	Sesamia nonagrioides	SESANO	0,20	0
Thysanoplusia orichalcea	TRIPOR	0,27	1	Xyleborus perforans	XYLBPE	0,19	0
Scolytus multistriatus	SCOLMU	0,26	1	Zeuzera pyrina	ZEUZPY	0,19	0
Anaphothrips obscurus	ANAPOB	0,26	1	Brachycerus muricatus	BRCCMU	0,19	0
Hercinothrips bicinctus	HERCBI	0,26	1	Orthosia cerasi	ORTSST	0,19	0
Rhyacionia buoliana	EVETBU	0,26	0	Xyleborus ferrugineus	XYLBFE	0,19	0
Thrips angusticeps	THRIAN	0,26	0	Rhagoletis cerasi	RHAGCE	0,19	0
Diuraphis noxia	BRAYNO	0,26	0	Diatraea saccharalis	DIATSA	0,19	0
Thaumetopoea pityocampa	THAUPI	0,25	0	Apate monachus	APATMO	0,19	0
Aspidiotus destructor	ASPDDE	0,25	0	Sitophilus granarius	CALAGR	0,19	0
Naupactus leucoloma	GRAGLE	0,25	1	Jacobiasca lybica	EMPOLY	0,19	0
Eulecanium tiliae	LECATI	0,25	0	Erinnyis alope	ERINAL	0,18	0
Rhyzopertha dominica	RHITDO	0,25	0	Dysaphis plantaginea	DYSAPL	0,18	0
Aonidiella citrina	AONDCI	0,25	0	Acrolepiopsis assectella	ACROAS	0,18	0
Otiorynchus crabricollis	OTIOCR	0,24	0	Anastrepha fraterculus	ANSTFR	0,18	0
Cacoecimorpha pronubana	TORTPR	0,24	0	Orgyia antiqua	ORGYAN	0,18	0
Cosmopolites sordidus	COSMSO	0,24	0	Rhynchophorus palmarum	RHYCPA	0,18	0

For each species the risk and their presence or absence in New Zealand is noted. Those species not already established have been highlighted. ^aP. citri (Risso) is problematic as it is considered by New Zealanders as not present yet it is recorded in the compendium as present. This species was considered established in the 1970s but has not been recorded since. That however does not rule out the presence of relict populations.

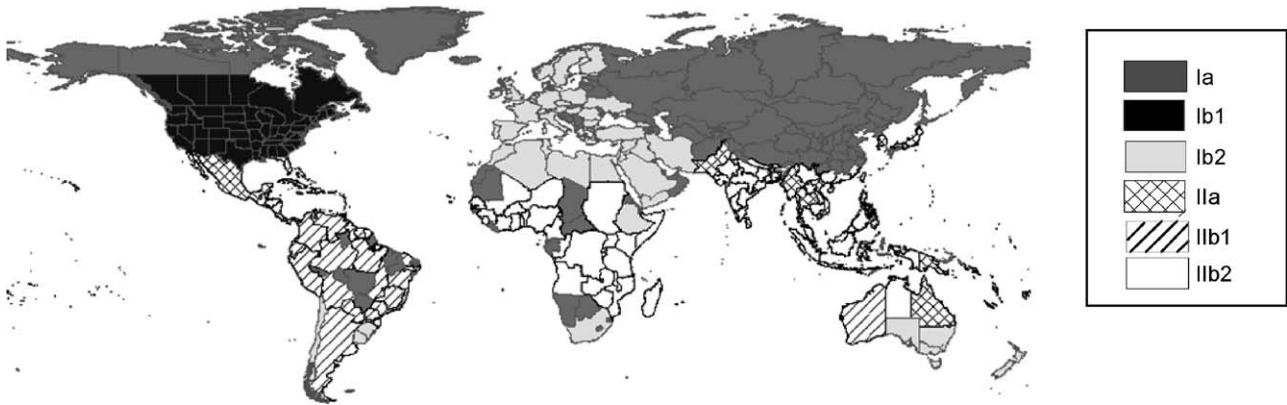


Fig. 4 – World map showing the 459 geographic areas represented by different greyscale and patterns according to the cluster they belong to.

also a large part of Europe. Chile, and some Mediterranean countries as Sardinia, Sicily, Cyprus, Algeria, Lebanon, Libya, Morocco, Tunisia are also included. The cluster IIa is again a specific cluster that repeats the more detailed regional information over larger geographic areas, for example the whole of the United States or Australia. This cluster is an artefact of database construction and can be ignored because it represents a summation of the same information at the regional level, the latter being of more interest. Cluster IIb1 included countries of South and Central America. Cluster IIb2 is composed of a large number of African and Asian countries.

Then, these clusters were plotted on a geographic map of the world using GIS and different patterns and greyscale to better visualise the clusters (Fig. 4).

The connection weights for each species associated with each cell of the map can be interpreted as an index of the importance of that species to the species assemblages in each cell and therefore is an index of the risk that the species may establish a geographic area if it has the opportunity. For example, in the case of New Zealand, the species that have the highest risk index are global pest species that are considered to have a high potential to establish in that country and indeed have already done so (Table 1) (Worner et al., 2004). But of

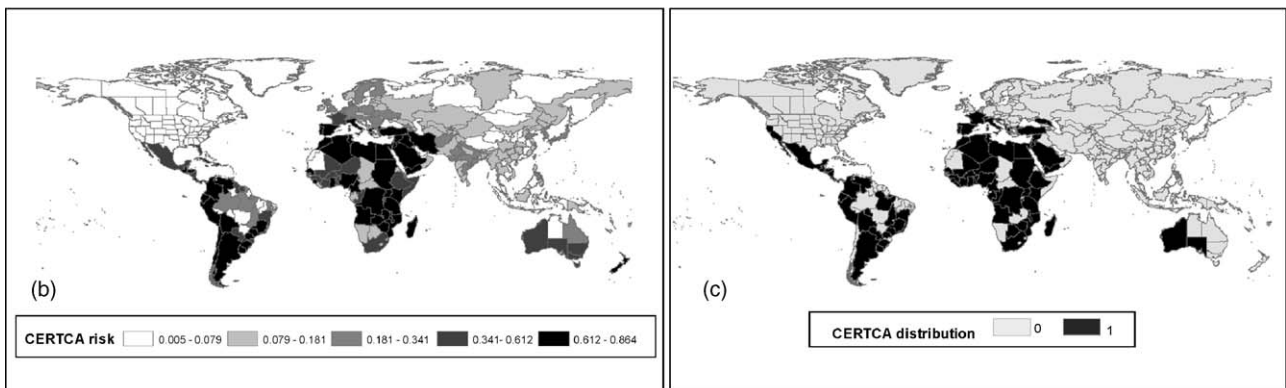
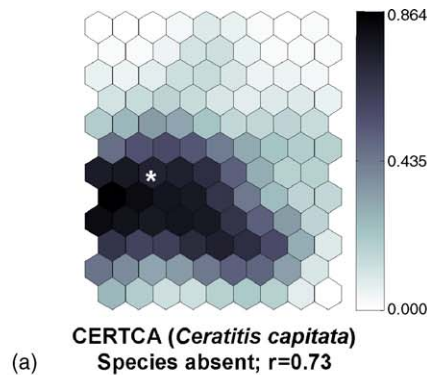


Fig. 5 – Distribution of *C. capitata* (Weidemann), the Mediterranean Fruit Fly risk on the SOM map (a) and on the world map (b). (c) The real distribution (presence and absence) of the species in the world.

most interest are the species that have a high risk of establishment but are not currently present in New Zealand. They are therefore species to which special attention should be given.

For example, the Mediterranean Fruit Fly, *Ceratitis capitata* (Weidemann), is not present in New Zealand, yet, the SOM model predicts the risk of establishment is high (0.73). This species is an interesting example where it has often been intercepted by monitoring traps near major airports in New Zealand. Authorities considered *C. capitata* established in 1996 however it was quickly eradicated. Moreover, other modelling attempts have indicated the high risk of establishment of this species in New Zealand (Worner, 1988; Baker, 1996) adding weight to the present study. Furthermore, *C. capitata* is established in: South Australia, Rio Grande do Sul, Switzerland, Cyprus, Algeria, Egypt, Spain, Canary Islands, Ethiopia, France, Corsica, Greece, Israel, Italy, Sicily, Sardinia, Jordan, Lebanon, Libya, Morocco, Malta, Portugal, Azores, Madeira, Russian Federation, Saudi Arabia, Syria, Tunisia, Turkey, Uruguay, Yemen, Yugoslavia, South Africa. All these countries are mapped into the same node (cell) of the SOM map as New Zealand.

The distribution of each of the 844 species (one map per species) can be visualised on the SOM map and the SOM allows a risk map for each species to be evaluated. The map for *C. capitata* is shown in Fig. 5a along with a geographic representation of global risk using GIS (Fig. 5b). The current distribution (presence and absence) of *C. Capitata* is shown in Fig. 5c.

While in the case of binary data, a SOM connection weight cannot be used as probability, the weights are values of the membership function for a fuzzy set. The method proposed here is applicable when both binary and non-binary values are available in the data to represent the occurrence of an event, that includes unknown values represented as 0.5 fuzzy membership degree, or other degrees suggested by experts. This study is a good example of a possible broader interpretation of SOM weights. Using them as an index of risk is especially relevant in biosecurity research where it is difficult to determine which new species are likely to invade a country. Previous studies have highlighted the threat of some species invasions using climate similarities, but climate is not the only explanation of species establishment. Similarities with the original habitat other than its climate can also be reasons of potential invasions (Worner, 2002). For example, similar pathways of arrival, host plants and the characteristics of invaded areas such as high disturbance and absence of specific diseases and natural enemies may help explain the similarity of species assemblages in particular areas.

It is clear that the interpretation of SOM weights would need to be adapted according to the particular ecological study of interest. For example the weights could be used to evaluate a pollution gradient or be used as a measure of the impact or risk of predation.

2.0 beta compatible with Matlab 6.5) developed by the Laboratory of Information and Computer Science, Helsinki University of Technology (<http://www.cis.hut.fi/projects/somtoolbox>). Data was reproduced with permission from the Crop Compendium, Global Module, 5th ed.®, CAB International, Wallingford, UK, 2003. We acknowledge the Centre for Bioprotection CoRE-funded post-doctoral fellowship (<http://bioprotection.lincoln.ac.nz/>).

Appendix A

Cell	Geographic area name	Cluster
1	Alabama, Connecticut, Georgia, Illinois, Indiana, Louisiana, Massachusetts, Maryland, Michigan, Missouri, Mississippi, North Carolina, New Jersey, New York, Ohio, Pennsylvania, South Carolina, Tennessee, Texas, Virginia	Ib1
3	British Columbia, Nova Scotia, Ontario, Quebec, California, Oregon, Washington	Ib1
4	Norway	Ib2
5	Austria, Belgium, Bulgaria, Switzerland, Germany, Denmark, Finland, United Kingdom, Hungary, The Netherlands, Poland, Romania, Sweden	Ib2
6	France	Ib2
7	Spain, Greece, Italy, Portugal, Turkey	Ib2
8	Cyprus, Algeria, Lebanon, Syria, Tunisia	Ib2
9	Egypt, Israel, Iran, Morocco	Ib2
11	Mexico, Florida	Ia
12	Honshu, Republic of Korea	Ia
13	Arkansas, Delaware, Iowa, Kansas, Kentucky, Oklahoma, Wisconsin, West Virginia	Ib1
14	Arizona, Colorado, Idaho, New Mexico, Utah	Ib1
16	Czech Republic, Ireland	Ib2
17	Ukraine	Ib2
18	Corsica	Ib2
19	Malta	Ib2
20	Canary Islands, Jordan	Ib2
21	Iraq, Libya	Ib2
23	Hawaii	Ia
24	Taiwan	Ia
25	Manitoba, New Brunswick, Maine, Minnesota, Montana, North Dakota, Nebraska, New Hampshire, Nevada, Rhode Island, South Dakota, Wyoming	Ib1

Acknowledgments

The SOM simulator used in this study was programmed using the Matlab programming language and SOM toolbox (version

Appendix A (Continued)			Appendix A (Continued)		
Cell	Geographic area name	Cluster	Cell	Geographic area name	Cluster
28	Armenia, Azerbaijan, Russian Far East	Ia	61	Bosnia and Herzegovina, Channel Islands, Mongolia, Western Siberia, Alaska	Ia
29	Georgia (Republic), Yugoslavia	Ib2	63	Nei Menggu, Xinjiang	Ia
30	South Australia, Tasmania, Victoria, Sicily, Sardinia, Azores	Ib2	64	Anhui, Henan, Jiangxi, Shanxi, Shaanxi, Xizhang	Ia
31	New Zealand, Madeira	Ib2	65	Hubei, Hunan, Jiangsu, Sichuan, Zhejiang	Ia
32	Chile	Ib2	66	Para	Ib1
33	Saudi Arabia, Yemen	Ib2	67	Bahia, Rio de Janeiro, Nicaragua, Paraguay, El Salvador	Ib1
34	New South Wales, Ethiopia, South Africa	Ib2	68	Barbados, Guyana, Haiti, Suriname, Trinidad and Tobago	Ib1
35	Queensland	IIa	70	Angola, Cameroon, Ghana, Mozambique, Sudan, Sierra Leone, Senegal, Zambia, Congo Democratic Republic	Ib2
36	Java, Papua New Guinea, Philippines, Thailand, Vietnam	IIa	72	Andhra Pradesh, Bihar, Madhya Pradesh, Orissa, Indian Punjab, Nepal	Ib2
37	Alberta, New foundland, Prince Edward Island, Saskatchewan, Vermont	Ib1	73	Andorra, Northwest Territories, Yukon Territory, Faroe Islands, Gibraltar, Iceland, Liechtenstein, Monaco, Eastem Siberia, Northern Russia	Ia
38	Lithuania, Latvia	Ia	74	Qinghai, Western Sahara, Chandigarh, Damman, Dadra and Nagar Haveli, Diu, Mizoram, San Marino	Ia
39	Albania, Kazakhstan, Moldova, Central Russia, Slovakia, Uzbekistan	Ia	75	Gansu, Ningxia	Ia
40	Hokkaido	Ia	76	Moluccas	Ia
41	Afghanistan	Ia	77	Guizhou, Meghalaya	Ia
42	Western Australia, Saint Helena	Ib1	78	Amazonas, Ceara, Espirito Santo, Goias, Pernambuco, Belize, Cayman Islands	Ib1
43	Rio Grande do Sul, Uruguay	Ib2	79	Bahamas, French Guiana, United States Virgin Islands	Ib1
44	Argentina, Bermuda, Sao Paulo, Colombia, Ecuador, Peru	Ib1	80	Antigua and Barbuda, Dominica, Grenada, Guadeloupe, Saint Kitts and Nevis, Saint Lucia, Martinique, Montserrat, Saint Vincent and the Grenadines	Ib1
46	Kenya, Mauritius, Zimbabwe	Ib2	82	Burkina Faso, Burundi, Benin, Congo, Côte d'Ivoire, Guinea, Mali, Niger, Rwanda, Somalia, Togo	Ib2
48	Bangladesh, Hong Kong, Sri Lanka, Myanmar, Pakistan	IIa	84	Delhi, Gujarat, Haryana, Rajasthan	Ib2
49	Estonia, Siberia	Ia	85	Greenland, Russia (Asia), Serbia	Ia
50	Belarus, Balearic Islands, Luxembourg, Macedonia, Slovenia	Ia			
51	Croatia, Kyrgyzstan, Southern Russia, Tajikistan, Turkmenistan	Ia			
52	Hebei, Heilongjiang, Jilin, Liaoning, Jammu and Kashmir	Ia			
53	Shandong, Himachal Pradesh, Kyushu, Shikoku, Korea-DPR	Ia			
54	Santa Catarina	Ib1			
55	Bolivia, Minas Gerais, Parana	Ib1			
56	Costa Rica, Cuba, Dominican Republic, Guatemala, Honduras, Jamaica, Panama, Puerto	Ib1			
58	Madagascar, Malawi, Nigeria, Reunion, Tanzania, Uganda	Ib2			
60	Assam, Karnataka, Maharashtra, Tamil Nadu, Uttar Pradesh, West Bengal	Ib2			

Appendix A (Continued)

Cell	Geographic area name	Cluster
86	Qatar	Ia
87	Bahrain, Djibouti, Arunachal Pradesh, Goa, Lakshadweep, Nagaland, Kuwait	Ia
88	Botswana, Equatorial Guinea, Crete, Lesotho, Namibia	Ia
89	The Netherlands Antilles, Manipur, Swaziland	Ia
91	British Virgin Islands	Iib1
96	Australian Northern Territory, Guangdong, Fiji, Kerala	Iib2
97	Lord Howe Is., Bonaire, Acre, Amapa, Fernando de Noronha, Rondonia, Roraima, Beijing, Shanghai, Falkland Islands, Northern Ireland, Line Islands, Kanton and Enderbury, Kermadec Islands, Marquesas, Bismarck Archipelago, Bougainville, Saint Pierre and Miquelon, Severo-Osetinskaya Respublika, Turks and Caicos Islands, East Timor, Krymskaya Oblast, US Minor Outlying Islands, Midway Islands, District of Columbia, Socotra, Mayotte	Ia
98	Anguilla, Aruba, Alagoas, Piaui, Sergipe, Macau, Curacao, Galapagos Islands, Nauru, Aldabra, Wake	Ia
99	Island Matto Grosso do Sul, Matto Grosso, Rodriguez Island, Pitcairn Islands, Ascension, Tokelau, Johnston	Ia
100	United Arab Emirates, Maranhao, Paraiba, Rio Grande do Norte, Cocos Islands, Easter Island, Christmas Island (Indian Ocean), Guinea-Bissau, British Indian Ocean Territory, Mauritania	Ia
101	Eritrea, Tripura, Norfolk Island, Oman	Ia
102	Central African Republic, Gabon, Gambia, Comoros, Liberia, Maldives, Chad, Tuvalu	Ia
103	Sao Tome and Principe, Wallis and Futuna	Iib1
104	American Samoa, Cook Islands, Cape Verde, Kiribati, Marshall Islands, Niue, French Polynesia	Iib2

Appendix A (Continued)

Cell	Geographic area name	Cluster
105	New Caledonia, Belau, Seychelles, Tonga, Vanuatu, Samoa	Iib2
106	Bhutan, Guangxi, Hainan, Yunnan, Federated states of Micronesia, Guam, Kalimantan, Andaman and Nicobar Islands, Sikkim, Northern Mariana Islands	Iib2
107	Fujian, Ryukyu Archipelago	Iib2
108	Fujian, Ryukyu Archipelago, Brunei Darussalam, Irian Jaya, Nusa Tenggara, Sulawesi, Sumatra, Cambodia, Laos, Peninsular Malaysia, Sabah, Sarawak, Solomon Islands, Singapore	Iib2

REFERENCES

Baker, R.H.A., 1996. Developing a European pest risk mapping system. *Bull. OEPP* 26, 485-494.

Binns, N.A., Eiserman, J.P., 1979. Quantification of fluvial trout habitat in Wyoming. *Trans. Am. Fish. Soc.* 198, 215-228.

Brosse, S., Guegan, J.-F., Tourenq, J.-N., Lek, S., 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecol. Model.* 120, 299-311.

Brosse, S., Giraudel, J.L., Lek, S., 2001. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecol. Model.* 146, 159-166.

Cereghino, R., Giraudel, J.L., Compin, A., 2001. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. *Ecol. Model.* 146, 167-180.

Cereghino, R., Park, Y.S., Compin, A., Lek, S., 2003. Predicting the species richness of aquatic insects in streams using a restricted number of environmental variables. *J. N. Am. Benthol. Soc.* 22, 442-456.

Chaves, P., Kojiri, T., Yamashiki, Y., 2003. Optimization of storage reservoir considering water quantity and quality. *Hydrol. Proces.* 17, 2769-2793.

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

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 224-227.

Faussh, K.D., Hawkes, C.L., Parsons, M.G., 1988. Models that Predict the Standing Crop of Stream Fish from Habitat Variables: 1950-1985, PNW-GTR-213 US. Department of Agriculture, Forest Service, Pacific North Research Station, Portland, OR.

Giraudel, J.-L., Aurelle, D., Lek, S., 2000. Application of the self-organizing mapping and fuzzy clustering microsatellite data: how to detect genetic structure in brown trout (*Salmo Trutta*) populations. In: Lek, S., Guegan, J.-F. (Eds.), *Artificial Neuronal Networks: Application to Ecology and Evolution*. Springer-Verlag, Heidelberg, pp. 187-201.

- Glover, T., Mitchell, K., 2002. An introduction to Biostatistics. Mc Graw Hill.
- Gregor, J., Garrett, N., Gilpin, B., Randall, C., Saunders, D., 2002. Use of classification and regression tree (CART) analysis with chemical faecal indicators to determine sources of contamination. *N.Z. J. Marine Freshwater Res.* 36, 387–398.
- Grossman, G.D., Nickerson, D.M., Freeman, M.C., 1991. Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. *Ecology* 72, 341–347.
- Guilherme, E., Cintra, R., 2001. Effects of intensity and age of selective logging and tree girdling on an understory bird community composition in Central Amazonia, Brazil. *Ecotropica* 7, 77–92.
- Kasabov, N., 1996. Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering. The MIT Press, CA, MA.
- Kasabov, N., 2002. Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines. Springer-Verlag, Heidelberg, London, NY.
- Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69.
- Kohonen, T., 1995. Self-Organizing Maps. Springer-Verlag, Heidelberg.
- Kohonen, T., 2001. Self-Organizing Maps. Springer, Berlin.
- Kohonen, T., Somervuo, P., 1998. Self-organizing maps of symbol strings. *Neurocomputing* 21, 19–30.
- Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resource variables: a review of modelling issues and applications. *Environ. Model. Software* 15, 101–124.
- Michaelides, S.C., Pattichis, C.S., Kleovoulou, G., 2001. Classification of rainfall variability by using artificial neural networks. *Int. J. Climatol.* 21, 1401–1414.
- Park, Y.S., Cereghino, R., Compin, A., Lek, S., 2003a. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecol. Model.* 160, 265–280.
- Park, Y.S., Chang, J.B., Lek, S., Cao, W.X., Brosse, S., 2003b. Conservation strategies for endemic fish species threatened by the Three Gorges Dam. *Conserv. Biol.* 17, 1748–1758.
- Ricker, W.E., 1975. Computation and interpretation of biological statistics of fish populations. *Bull. Fish. Res. Board Can.* 191, 1–382.
- ter Braak, C., 1987. The analysis of vegetation–environment relationships by canonical correspondence analysis. *Vegetatio* 69, 69–77.
- Ultsch, A., Siemon, H.P., 1990. Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the INNOC'90 International Neural Network Conference, Kluwer, Dordrecht, The Netherlands, pp. 305–308.
- Waite, S., 2000. Statistical Ecology in Practice. A Book to Analysing Environmental and Ecological Field Data. Prentice Hall, London.
- Whittaker, R.H., Gilbert, L.E., Connell, J.H., 1979. Analysis of two-phase pattern in a mesquite grassland, Texas. *J. Ecol.* 67, 935–952.
- Worner, S.P., 1988. Ecoclimatic assessment of potential establishment of exotic pests. *J. Econ. Entomol.* 81, 973–983.
- Worner, S.P., 2002. In: Halman, G., Shalbe, C.P. (Eds.), Predicting the Invasive Potential of Exotic Insects. In *Invasive Arthropods and Agriculture: Problems and Solutions*. Science Publishers, Inc., Enfield, NH, pp. 119–137.
- Worner, S.P., Gevrey, M., Peacock, L., Pitt, J., 2004. The contribution of artificial intelligence to the analysis of the invasion potential of exotic pests. In: Proceedings of the XXII International Congress of Entomology, Brisbane, Queensland, Australia.
- Zadeh, L.A., 1965. Fuzzy sets. *Inform. Control* 8, 338–353.