

Modelling global insect pest species assemblages to determine risk of invasion

S. P. WORNER* and MURIEL GEVREY†

*National Centre for Advanced Bio-Protection Technologies, Lincoln University, PO Box 84, Canterbury, New Zealand; and †Univerity Paul Sabatier, UMR-CNRS LADYBIO, Toulouse, France

Summary

1. The many thousands of potential invasive species pose one of the greatest threats to global biodiversity world-wide. In this study we propose that assemblages of well-known global invasive pest species, irrespective of whether they arise by anthropogenic means, are non-random species groupings that contain hidden predictive information. Such information can assist the identification and prioritization of species that have the potential to pose an invasive threat in regions where they are not normally found.

2. Data comprising the presence and absence of 844 insect pest species recorded over 459 geographical regions world-wide were analysed using a self-organizing map (SOM), a well-known artificial neural network algorithm. The SOM analysis classified the high dimensional data into two-dimensional space such that geographical areas that had similar pest species assemblages were organized as neighbours on a map or grid.

3. The SOM analysis allowed each species to be ranked in terms of its risk of invasion in each area based on the strength of its association with the assemblage that was characteristic for each geographical region. A risk map for example species was produced to illustrate how such a map can be compared with the species' actual distribution and used with other information, such as the species' biotic characteristics and interactions with the abiotic environment, to improve pest risk assessments further.

4. *Synthesis and applications.* This study presents a new approach to the identification of potentially high-risk invasive pest species based on the hypothesis that global insect pest assemblages are non-random species groupings that can be subjected to traditional community analysis. A well-known data mining and knowledge discovery method for high dimensional data, SOM, was used to determine pest species assemblages for global regions. Species were ranked according to their potential for establishment based on their strength of association with the species assemblage that characterizes a particular region. Such an analysis can then be used to support additional risk assessment of potential invasive species, giving invasive species researchers, conservation managers, quarantine and biosecurity scientists a means for prioritizing species as candidates for further research.

Key-words: artificial neural networks, exotic pest, invasive species, pest species assemblage, New Zealand, self-organizing map

Journal of Applied Ecology (2006) **43**, 858–867
doi: 10.1111/j.1365-2664.2006.01202.x

Introduction

Species assemblages are groupings of species that co-occur in the same place and at the same time. Such assemblages are not instantly created but come into being through the progressive invasion of species such that the assemblage is built up sequentially from a

simple starting point (Begon, Harper & Townsend 1996). In this study we investigated the largely anthropogenic global pest assemblages of large-scale geographical regions to determine whether such species groupings can provide predictive knowledge and insight into exotic pest species invasions. We used the term assemblage in a broad sense, in line with authors such as Keddy (1992) and Diaz, Cabido & Casanoves (2001), who apply it to any regional species pool that has come about as a result of 'a filter of any kind'.

Jones & Kitching (1981) define a pest as an organism that damages crops, destroys products, transmits or causes disease, is annoying or in other ways conflicts with human needs or interests. International concerns about preserving biodiversity further extend the definition of a pest to a species that can either cause native species decline or alter the structure and function of natural ecosystems (Worner 2002).

Exotic animals can be introduced into new areas either accidentally or on purpose. While many have little impact on the lives of humans or on local native flora and fauna, some, as a consequence of the absence of natural enemies, undergo an explosive increase in numbers and rapid spread. Such species soon come to occupy large areas, inflicting dramatic damage on crops, stock and native ecosystems that can cause considerable economic and environmental harm (Elton 1958; Sailer 1983; Pimentel 1986; Simberloff 1986, 1989; Worner 1994). More recently, pest invasions have been recognized as an important cause of loss of biodiversity. One of the best ways to reduce the likelihood of exotic species invasions is to prevent their establishment, but this is dependent on identifying potential invasive species in advance of their establishment.

Clearly, for any species given the opportunity to invade a new area, a source of food and a place to live and reproduce are fundamental to establishment success. But, to evaluate fully the invasive potential of a species, the conditions required for invasion, the characteristics of the invasive species and the ecosystems that are susceptible to invasion, need to be examined (Worner 2002).

Successful establishment of a species arriving in a new environment depends on both biotic and abiotic factors, including climate and the specific environmental conditions of the habitat in the area of invasion (Mooney & Drake 1989). The species assemblages investigated in this study were unusual because they comprise groupings of phytophagous pest species that occur largely as a result of human actions. Our hypothesis was that geographical areas with similar pest assemblages share similar biotic and abiotic conditions that allow particular pest species to invade the area. Biotic conditions can include the particular assemblage of crops and garden plants growing in a region, which in turn reflect abiotic, mainly climatic, conditions that affect species distribution and abundance. In addition, abiotic factors can include many anthropogenic factors, such as a similar history of invasion or pathways of entry,

the amount and type of trade, the amount of protected cultivation, the effectiveness of quarantine procedures and the resources available to find and report pest species in any country. We hypothesized that the particular combination of pest species in a region integrates these complex factors and their interactions. In other words, pest species assemblages are non-random species groupings that contain hidden predictive information that can be analysed using ecological community analysis techniques.

Despite species invasions having been studied for decades, little progress has been made such that advances towards arresting biological invasions will not be made unless the limitations of standard methods are addressed (Hulme 2003). Furthermore, for pest insects in particular, research has usually focused on individual species. More recently, Levine & D'Antonio (2003) have used species accumulation models to estimate the percentage increase in species invasion with increasing international trade. Such curves are based on the concept that there is not an infinite pool of invasive species from which to sample. To our knowledge, there has been no investigation of potential species invasion taking into account pest assemblages. This approach is particularly relevant as a basis for further pest risk assessment studies for conservation of biodiversity and the development of national quarantine and biosecurity strategies.

Traditional analysis tools for measuring and understanding species communities and relating community patterns to changes in spatial and temporal ecosystem conditions include the simple metrics of species richness, diversity and similarity. The description of complex species assemblage structures using a single attribute has been criticized because valuable information may be lost (Begon, Harper & Townsend 1996). Recent research in the study of communities and complex systems has shown that the newer analytical methods of non-linear artificial intelligence (Park *et al.* 2003a,b) can characterize species assemblages as components of ecological communities extremely well. Examples of such studies are the analysis of the Trichoptera assemblage in Danish streams (Wiberg-Larsen *et al.* 2000), the assessment of the Luxembourg river water quality using diatom assemblages (Gevrey *et al.* 2004), the investigation of macroinvertebrate assemblages in Brazil (Buss *et al.* 2004) and the prediction and spatial mapping of New Zealand freshwater fish and decapod assemblages (Joy & Death 2004).

A self-organizing map (SOM), which is an artificial neural network (ANN) model (Kohonen 1982), was used to identify pest species assemblages and potential invasive insects based on a comprehensive database of the global presence or absence of pests (CABI 2003). The SOM is an efficient method for analysing systems ruled by complex non-linear relationships and provides an alternative to traditional statistical methods for classifying complex data (Lek *et al.* 1996; Lek & Guegan 2000; Park *et al.* 2003a). More generally, SOM are used

widely for knowledge discovery, pattern recognition, clustering and visualization of large multidimensional data sets. Potentially useful but hidden information can be revealed and further examined using more traditional techniques. Successful results in aquatic community ecology using such models have been well documented (Chon *et al.* 1996; Cereghino, Giraudel & Compin 2001; Giraudel & Lek 2001; Park *et al.* 2003b). This study highlights how these novel tools could improve risk assessments of potential insect invaders.

The objectives of this study were to (i) classify global geographical areas based on assemblages of exotic insect pest species associated with each area or cluster, and (ii) identify and quantify the potential risk of invasion of these insect species, based on their strength of association with other species within the geographical clusters. New Zealand was used as a particular example.

Materials and methods

DATA

The data used in this analysis were extracted with permission from the Crop Protection Compendium (CABI 2003). This database encompasses a wide range of different types of information on all aspects of crop protection (e.g. pests, diseases, weeds, natural enemies and crops) associated with most areas in the world. The geographical areas represented in the compendium include countries, regions and states of countries such that all continents are represented (with the exception of the Arctic and Antarctic regions). The full compendium includes many species for which only partial information is available. To ensure that we only included species with adequate distributional information, we selected those that occur in more than 2% of geographical areas (Waite 2000). Of those species, only those for which information was confirmed by experts were used. This comprised 844 mainly phytophagous insect pests for 459 geographical areas (Table 1). The presence (1) and the absence (0) of each species in each geographical area resulted in a database comprising a $[844 \times 459]$ matrix.

Table 1. Summary of the pest species represented in the database

Order	Family number	Species number
Lepidoptera	33	257
Hemiptera	24	228
Coleoptera	23	203
Diptera	12	83
Thysanoptera	2	39
Orthoptera	2	16
Hymenoptera	7	11
Isoptera	2	3
Psocoptera	1	3
Collembola	1	1

SOM MODEL

A conventional cluster analysis comprising single and complete linkage clustering using the Jaccard coefficient and Euclidean distance as similarity metrics (Krebs 2001) failed to organize the data matrix. The result was a series of long drawn-out clusters with a large number of branches that we were unable to interpret. A SOM, however, is able to analyse such high dimensional data, performing a non-linear projection of the multidimensional data space onto two-dimensional space. The SOM neural network consists of two layers of elements or neurones: the input layer and the output layer. The output layer is represented by a map or a rectangular grid with l by m neurones (or cells), laid out in a hexagonal lattice.

It is possible to use two different learning algorithms in a SOM. An incremental algorithm is commonly used but learning is highly dependent on the order of input; a batch algorithm overcomes this fault and was used here. It is significantly faster and does not require specification of any learning rate factor (Kohonen & Somervuo 1998). The batch SOM algorithm can be summarized as follows. (i) Initialize the values of the virtual vectors (VV_i , $1 \leq i \leq c$) using random values. (ii) Repeat steps (iii) to (vi) until convergence. (iii) Read all the sample vectors (SV) one at a time. (iv) Compute the Euclidean distance between SV and VV . (v) Assign each SV to the nearest VV according to the distance results. (vi) Modify each VV with the mean of the SV that were assigned to it. Details of the SOM algorithm and its theoretical basis can be found in Kohonen (1995), Kohonen & Somervuo (1998) and Kohonen (2001).

In our study the input layer comprised 844 neurones (one for each pest species) connected to the 459 sites (geographical areas) such that the representation of the presence-absence of 844 species for each site formed 459 sample vectors. The output layer comprised 108 neurones organized in an array with 12 rows and nine columns. The number of neurones or cell in the output layer was first defined using the formula $c = 5\sqrt{n}$ proposed by the Laboratory of Computer and Information Science (CIS), Helsinki University of Technology (Espoo, Finland), where c is the number of cells and n is the number of training samples (sample vectors). Each cell of the output layer is linked to the input neurones (i.e. 844) by connections that have weights associated with them, forming a virtual vector. During the learning process, a virtual pest species assemblage is then computed for each cell (Fig. 1). While initially the number of output neurones was decided using the formula cited previously, several maps were created using different sizes. Classification results were very similar. We used the topographic and quantification errors to determine final map size (Kohonen 2001). For the final map size (108 neurones or cells), the errors were, respectively, 0.022 and 0.5832. A smaller number of output neurones increased both error values; a larger

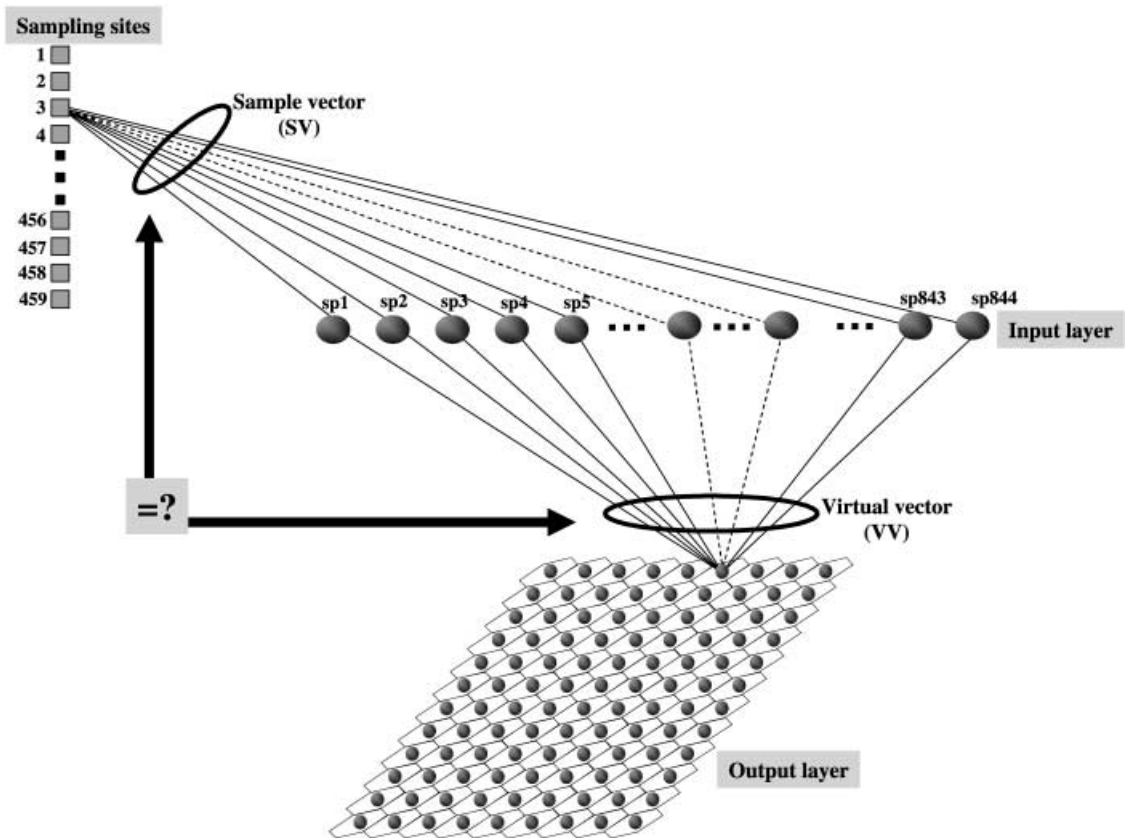


Fig. 1. Self-organizing map architecture. The input layer is linked to the cells of the output layer by connections called weights that define the virtual assemblages of the species.

number of neurons decreased the errors slightly but the limitation of computer memory was quickly reached. In addition, a larger map size made interpretation difficult.

The aim of the SOM algorithm is to assist organization and visualization of data by arranging the distribution of the sample sites onto a two-dimensional space represented by the map cells. With 459 sites a traditional approach would require $0.5 \times 459(459 - 1) = 105\,111$ similarity indices to be sorted, which is clearly unmanageable. The virtual vectors that are neighbours on the map represent neighbouring clusters of sample sites. Various distance measures can be used to organize and project similar sample vectors onto the map. In this study we selected the Euclidean distance, advised by Kohonen & Somervuo (1998). After the SOM had finished learning, each cell of the output layer had a virtual vector that could be interpreted as a virtual pest species assemblage. These vectors were composed of values over the interval $[0,1]$, each of which could be interpreted as a risk index that indicates a species' potential to be present or to be associated with the sites within each cell. All the sites that are associated with the same cell have a similar pest species assemblage composition both in regard to species presence as well as absence at each site. Therefore, a species that is present in one site but not in another in the same cell can be considered to pose a high risk of invasion to that

site. Using a grey colour gradient, the weights associated with each virtual site vector can be used to display the strength of association of each species with the assemblage in each cell. The darker the cell, the higher the risk the species might invade the sites included in the cell.

The SOM simulator used in this study was programmed using the Matlab programming language (Mathworks 2001) and SOM toolbox (version 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/documentation/somalg.shtml>, accessed 21/6/06). The geographical areas were obtained using ArcView 3.2 (ESRI Corporation, Redlands, CA).

CLUSTER ANALYSIS

Sites that are neighbours on the grid are expected to be more similar to each other, whereas sites distant from each other (according to their pest species assemblage) are expected to be distant in the feature space. To detect the cluster boundaries on the map, a cluster analysis is applied to the SOM model output (Park *et al.* 2003a,b). Some authors also use the unified-matrix (U-matrix) approach (Ultsch & Siemon 1990), where the U-matrix displays the distances between the virtual sites and provides a landscape formed by light plains

separated by dark ravines. However, this method does not give crisp boundaries to the clusters. In this study a hierarchical cluster analysis with a Ward linkage method was applied to the SOM results to identify the edges of each cluster of sample vectors (Park *et al.* 2003a,b). The Davies–Bouldin index (DBI; Davies & Bouldin 1979) was then calculated to justify the choice of the number of clusters. To check the validity of non-intuitive groupings of geographical sites, the Jaccard similarity index and percentage similarity (Krebs 2001) calculated between selected sites were examined.

Results

The non-linear projection of presence–absence data onto two-dimensional space allowed us to classify global geographical areas according to the similarity of their pest species assemblages (see Figure S1 and Appendix S1 in the supplementary material). A hierarchical cluster analysis was applied to the SOM results (Fig. 2a). The optimum number of clusters according to the DBI value (0.981) was six clusters (Ia, Ib1, Ib2, IIa, IIb1 and IIb2), as shown in Fig. 2b.

Cluster Ia represented geographical areas in northern latitudes that have low numbers of pest species recorded in the database. Most of the locations (85%) in this cluster had only 2% of the (844) species present and were mainly small islands, desert areas and colder areas of Greenland, Alaska and parts of Russia. The USA and most of Canada comprised cluster Ib1. Cluster Ib2 included New Zealand, several regions of Australia (south Australia, Tasmania, Victoria, western Australia), a large part of Europe, Chile and some Mediterranean countries such as Sardinia, Sicily, Cyprus, Algeria, Lebanon, Libya, Morocco and Tunisia. Cluster IIa was a specific cluster that included larger geographical areas, for example the whole of the USA and Australia. This cluster was an artefact of the original database construction and could be ignored. Cluster IIb1

included the countries of South and Central America; Cluster IIb2 comprised many African and Asian countries. These clusters were plotted on a map of the world using different grey scales and effects to visualize each cluster (Fig. 3).

For each cell of the SOM map, and therefore each geographical area, it was possible to define a virtual species assemblage. The strength of the association of each of the 844 species with each assemblage could be visualized on the SOM map, creating one map per species. The virtual assemblage inside each cell could be considered an index of the risk of invasion of each species in the countries associated with each cell. For New Zealand the risk values of the most invasive species were calculated (Table 2) and their presence or absence in New Zealand recorded. Risk and distribution maps of two high profile species, the Mediterranean fruit fly *Ceratitis capitata* (Wiedemann) (Fig. 4) and the gypsy moth *Lymantria dispar* L. (Fig. 5) are shown. Both species are absent in New Zealand and, while this country is currently more threatened by the Asian form of the gypsy moth, the European form is also a potential threat to the country. The analysis of the distribution of risk for *C. capitata* indicated that this species had a high risk (risk index 0.73) of invasion whereas the gypsy moth *L. dispar* had a lower risk index of 0.31, indicating that risk of this species establishing in New Zealand is in the medium range. Clearly, however, this information must be interpreted within the context of a full risk assessment for each species.

Discussion

Invasive species research usually focuses on individual species to determine appropriate management strategies in response to a significant threat to a region or country. Currently, there is no objective scientific approach to prioritize and identify species that should be subject to more detailed risk assessments. In our study we used the

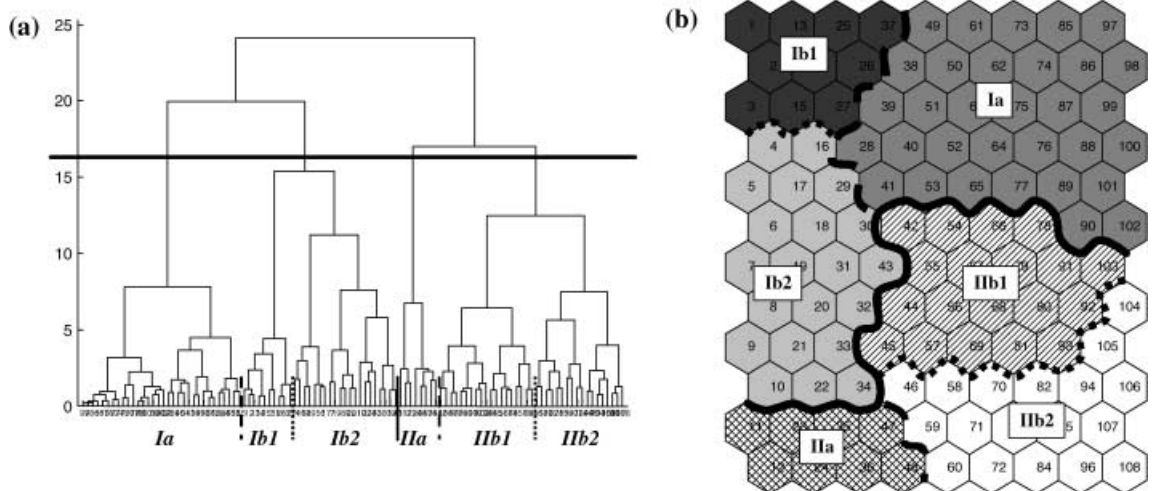


Fig. 2. (a) Dendrogram of the cluster analysis; (b) self-organizing map with the clusters defined by the cluster analysis: Ia, Ib1, Ib2, IIa, IIb1 and IIb2 (see text).

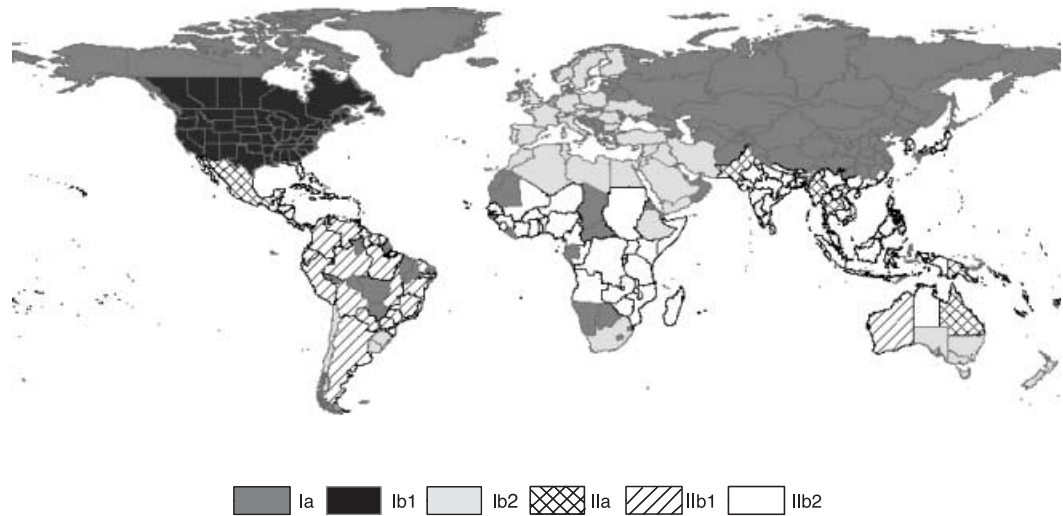


Fig. 3. World map showing the 459 geographical areas represented by different shades of grey and patterns according to the cluster to which they belong.

information available on the geographical distribution of a wide range of global insect pest species to investigate whether the analysis of entire pest species assemblages can help to rank the many potential invasive insect species that threaten New Zealand. This approach can provide similar information for many countries and geographical regions.

The analysis of complex ecological data requires more advanced tools than those currently available if valuable information is not to be lost (Begon, Harper & Townsend 1996). Simple metrics are often used because of the paucity of methods that can handle large amounts of data (Giske, Huse & Fiksen 1998). More complex modelling methods, such as canonical correspondence analysis (ter Braak 1986) and ANN, can greatly assist interpretation by retaining more information. ANN are particularly tolerant to noisy data (Hepner *et al.* 1990) and are better able to handle outliers than more traditional statistical analysis methods (Lippman 1987). Furthermore, they can predict non-linear data and represent even more complex relationships between variables (Rumelhart, Hinton & Williams 1986). These are clearly advantages appropriate for this study.

SOM are particularly good at detecting outliers that can be confined to part of the map without affecting the other parts (Cereghino, Giraudel & Compin 2001). In this study, areas with very low numbers of species were grouped into cluster Ia and large geographical areas that essentially repeat the information contained over smaller geographical scales (a peculiarity of the compendium database design) were grouped into cluster IIa. This problem of scale, where larger geographical areas should not normally be compared with smaller geographical areas, could be interpreted as a limitation of the SOM analysis. We emphasize, however, that the results of an analysis such as that carried out here should not be interpreted in isolation and do not constitute a full risk assessment for any species. The utility

of the analysis depends entirely on the questions asked of the data. For example, the risk analyst might only be interested in the combinations of species that occur in different regions.

A large number of factors can influence phytophagous insect pest species distribution and therefore pest species assemblages at any geographical location (Baker *et al.* 2005). The presence of a host-plant and a suitable climate are crucial, but other significant factors can include trade routes and plant or produce importations that either historically and/or currently provide invasion pathways, as well as agricultural and quarantine practices associated with the region. We hypothesize that the interplay of all these factors will result in a characteristic mix of exotic species in any given region. Much research has been focused on the climatic suitability of a location as an indicator of the potential for establishment or invasion of an exotic species. If we examine the countries that share the same cell as New Zealand, it is not immediately apparent why New Zealand's pest assemblage is also in the same cluster as a number of Mediterranean countries. However, similarity indices showed that New Zealand shares a significant percentage of its pest species with many countries in this region, for example Italy and France (59% and 58%, respectively), Turkey and Morocco (46% and 48%, respectively). Moreover, if the history of New Zealand's agriculture is examined, the vast majority of that country's cropping and pasture species originated from the Mediterranean region. In addition, the region possesses many transitional climates that are similar to parts of New Zealand. Finally, many garden plants have been imported into New Zealand from Mediterranean regions (A. Stewart, personal communication).

Our method of analysis could provide biodiversity managers and quarantine authorities of any country with a list of species that are ranked according to the risk they pose. This could help to prioritize more

Table 2. List of pest species that have the highest potential risk of invasion in New Zealand. For each species the risk and their presence or absence (P or A) in New Zealand is noted

Name	Risk index	P or A	Name	Risk index	P or A
<i>Planococcus citri</i>	0.93	0	<i>Toxoptera aurantii</i>	0.49	1
<i>Icerya purchasi</i>	0.92	1	<i>Taylorilygus pallidulus</i>	0.49	0
<i>Myzus persicae</i>	0.87	1	<i>Aleurothrixus floccosus</i>	0.48	0
<i>Cydia pomonella</i>	0.86	1	<i>Pseudaulacaspis pentagona</i>	0.48	0
<i>Nezara viridula</i>	0.85	1	<i>Pieris rapae</i>	0.47	1
<i>Brevicoryne brassicae</i>	0.83	1	<i>Hadula trifolii</i>	0.47	0
<i>Delia platura</i>	0.80	1	<i>Ephestia elutella</i>	0.47	1
<i>Phthorimaea operculella</i>	0.79	1	<i>Rhopalosiphum rufiabdominale</i>	0.46	1
<i>Pseudococcus longispinus</i>	0.79	1	<i>Liriomyza trifolii</i>	0.46	0
<i>Aphis spiraeicola</i>	0.77	1	<i>Sitona discoideus</i>	0.46	1
<i>Saissetia oleae</i>	0.77	1	<i>Spodoptera exigua</i>	0.46	0
<i>Coccus hesperidum</i>	0.77	1	<i>Sitobion avenae</i>	0.45	0
<i>Aonidiella aurantii</i>	0.76	1	<i>Therioaphis trifolii</i>	0.45	1
<i>Eriosoma lanigerum</i>	0.76	1	<i>Locusta migratoria</i>	0.45	1
<i>Aphis gossypii</i>	0.76	1	<i>Prays citri</i>	0.43	0
<i>Viteus vitifoliae</i>	0.75	1	<i>Hippotion celerio</i>	0.43	1
<i>Ceratitis capitata</i>	0.73	0	<i>Pantomorus cervinus</i>	0.43	1
<i>Agrotis ipsilon</i>	0.73	1	<i>Schizaphis graminum</i>	0.42	0
<i>Bemisia tabaci</i>	0.70	1	<i>Oulema melanopus</i>	0.42	0
<i>Helicoverpa armigera</i>	0.70	1	<i>Scolytus rugulosus</i>	0.42	0
<i>Acyrtosiphon pisum</i>	0.70	1	<i>Drosophila melanogaster</i>	0.42	0
<i>Thrips tabaci</i>	0.69	1	<i>Sitona lineatus</i>	0.42	1
<i>Saissetia coffeae</i>	0.68	1	<i>Mythimna unipuncta</i>	0.41	0
<i>Rhopalosiphum maidis</i>	0.68	1	<i>Pectinophora gossypiella</i>	0.41	0
<i>Plutella xylostella</i>	0.68	1	<i>Hellula undalis</i>	0.41	1
<i>Chrysomphalus dictyospermi</i>	0.67	0	<i>Peridroma saucia</i>	0.41	0
<i>Aspidiotus nerii</i>	0.67	1	<i>Parlatoria ziziphi</i>	0.41	0
<i>Frankliniella occidentalis</i>	0.61	1	<i>Gonipterus gibberus</i>	0.40	1
<i>Rhopalosiphum padi</i>	0.61	1	<i>Acanthoscelides obtectus</i>	0.40	1
<i>Hyperomyzus lactucae</i>	0.61	1	<i>Ceroplastes floridensis</i>	0.40	1
<i>Agrius convolvuli</i>	0.60	1	<i>Parasaissetia nigra</i>	0.40	1
<i>Diaspidiotus perniciosus</i>	0.60	1	<i>Lixus juncei</i>	0.40	0
<i>Aphis fabae</i>	0.60	0	<i>Sminthurus viridis</i>	0.40	1
<i>Phoracantha semipunctata</i>	0.59	1	<i>Diaspidiotus ostreaeformis</i>	0.39	1
<i>Heliothrips haemorrhoidalis</i>	0.59	1	<i>Henosepilachna elaterii</i>	0.39	0
<i>Macrosiphum euphorbiae</i>	0.59	1	<i>Lepidosaphes ulmi</i>	0.39	0
<i>Phyllocnistis citrella</i>	0.58	0	<i>Scrobipalpa ocellatella</i>	0.38	0
<i>Ceroplastes rusci</i>	0.57	0	<i>Siphoninus phillyreae</i>	0.38	1
<i>Chrysomphalus aonidium</i>	0.57	0	<i>Antigastra catalaunalis</i>	0.38	0
<i>Parthenolecanium persicae</i>	0.56	1	<i>Unaspis citri</i>	0.38	0
<i>Trichoplusia ni</i>	0.55	0	<i>Mythimna loreyi</i>	0.37	0
<i>Cadra cautella</i>	0.54	0	<i>Thrips simplex</i>	0.37	1
<i>Lepidosaphes beckii</i>	0.54	0	<i>Bactrocera oleae</i>	0.37	0
<i>Aphis craccivora</i>	0.54	1	<i>Lipaphis erysimi</i>	0.37	1
<i>Lampides boeticus</i>	0.54	1	<i>Spodoptera littoralis</i>	0.36	0
<i>Agrotis segetum</i>	0.54	0	<i>Orthezia insignis</i>	0.36	0
<i>Sitophilus zeamais</i>	0.53	0	<i>Prays oleae</i>	0.36	0
<i>Pieris brassicae</i>	0.53	0	<i>Listroderes costirostris</i>	0.36	1
<i>Hemiberlesia lataniae</i>	0.52	1	<i>Liriomyza huidobrensis</i>	0.35	0
<i>Toxoptera citricida</i>	0.52	1	<i>Cryptoblabes gnidiella</i>	0.35	1
<i>Parthenolecanium corni</i>	0.50	1	<i>Sesamia cretica</i>	0.35	0
<i>Grapholita molesta</i>	0.50	1	<i>Acronicta rumicis</i>	0.34	0
<i>Metopolophium dirhodum</i>	0.49	1	<i>Dialeurodes citri</i>	0.34	0
<i>Hemiberlesia rapax</i>	0.49	1	<i>Gynaikothrips ficorum</i>	0.34	0

general risk assessment efforts. For example, in 2002 the melon thrip *Thrips palmi* Karny attracted attention as a possible invasive species to New Zealand and significant resources were invested in assessing the risk of establishment (Dentener, Whiting & Connolly 2002). However, in the analysis reported here this species is not strongly associated with the New Zealand assemblage (risk index 0.06) and, to date, it has not

established in New Zealand. Furthermore, our analysis predicted New Zealand's most recent invasive insect pest *Chrysomphalus aonidium*. This species is among the 12 most highly ranked non-established species.

By grouping countries with similar pest assemblages, the SOM can indicate countries where a new invasion should alert local biosecurity authorities. For example, because of New Zealand's proximity to south Australia,

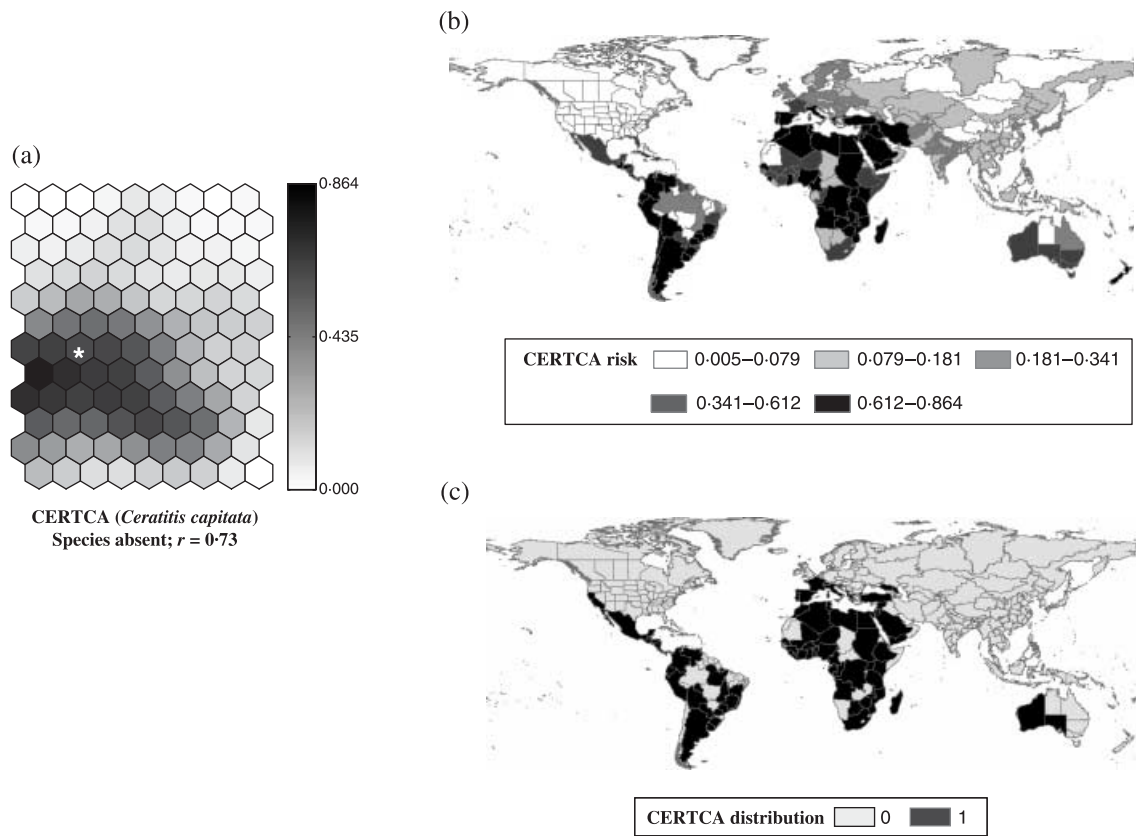


Fig. 4. Risk distribution of the Mediterranean fruit fly *Ceratitis capitata*, which is not present in New Zealand, (a) represented on the SOM map (the white * shows the cells where New Zealand is located) and (b) represented on a world map. (c) The actual distribution (presence and absence) of the species on a world map.

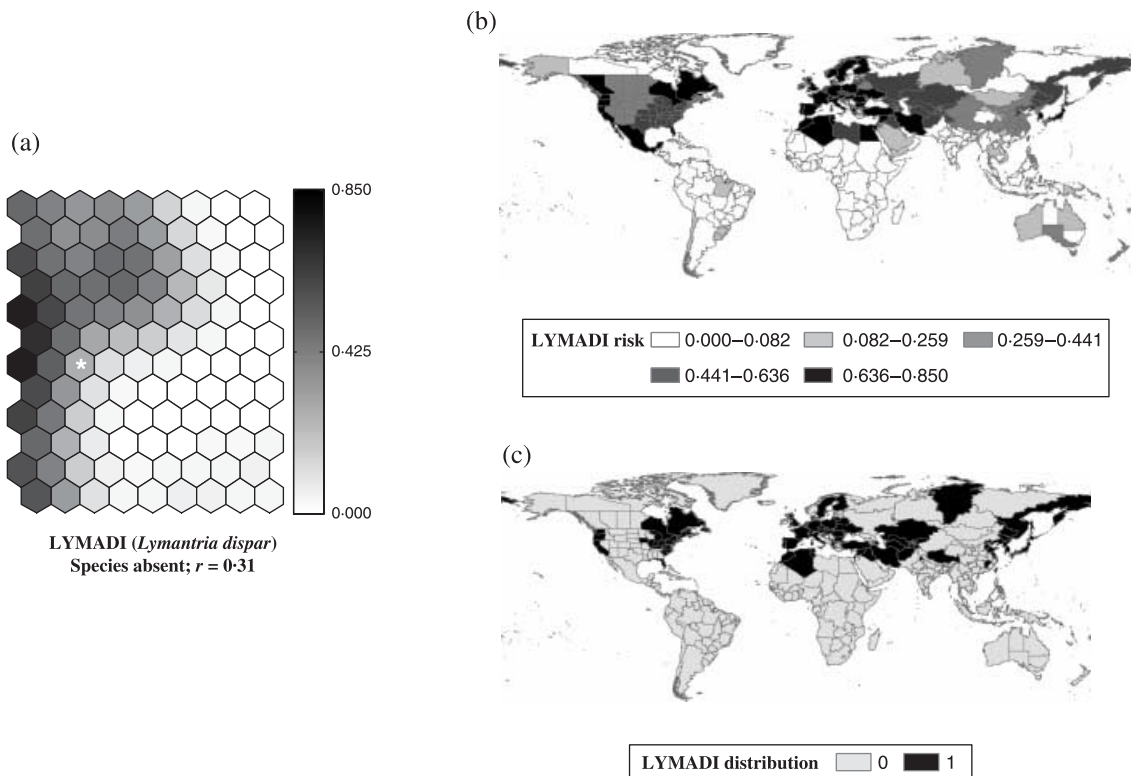


Fig. 5. Risk distribution of the gypsy moth *Lymantria dispar*, which is not present in New Zealand, (a) represented on the SOM map (the white * shows the cells where New Zealand is) and (b) represented on the world map. (c) The actual distribution (presence and absence) of the species on a world map.

with which it shares a similar pest assemblage, any new invasion in that region should put New Zealand biosecurity/quarantine authorities on high alert.

Several species with comparatively low ranks in the risk index are established in New Zealand (Table 2); such species are not strongly associated with New Zealand's exotic pest assemblage. Some endemic species that have become pests in their native country and have not yet spread widely will be given a low rank. In such situations, it may be more appropriate to study non-native invasive species on their own. Furthermore, a species may have a global distribution but be limited to specific environmental conditions and therefore not strongly associated with any pest assemblage, or it may be a newly emergent pest. At the other extreme, *Planococcus citri* (Risso) has the highest risk of establishing in New Zealand according to our results but is considered absent. In fact, this species was recorded as established in the 1980s but has not been recorded since (R. Henderson, personal communication). Such examples illustrate that this type of approach to prioritizing invasive species will have weaknesses under certain circumstances and that such an analysis is best used to support expert knowledge.

A limitation of the SOM approach to data classification is that every SOM is different and will find slightly different similarities among the sample vectors each time the initial conditions are changed. Recent developments involve bootstrapping the data used in this analysis at least 1000 times to test internal cluster quality and robustness (S. P. Worner & M. Watts, unpublished data). Each time the data are resampled and a new SOM model created, the change in rank of each species is noted to determine the degree of confidence in the SOM ranks. Not only will this indicate the sensitivity of the method to the presence or absence of particular species, but it will help to indicate those species for which data are insufficient or poor. Preliminary investigations show that the initial maps are very stable and the high ranked non-established species (the species in which we are most interested) show little change in their average rank.

A further limitation of the SOM analysis is that the geographical regions examined in this study may have indeterminate boundaries. That limitation is not confined to this study and is a problem in general ecological community studies. In some respects, the regional boundaries of the data used in this study may be more clearly defined than is usual because of the existence of border controls and trade practices specific to many regions.

In this study we found the SOM was able to reduce very high dimensional data into patterns that could be usefully interpreted. In addition to the fact that the SOM analysis can perform significant data reduction, it is the only method apart from *k*-means clustering that can give information about individual species. While *k*-means clustering is a more conventional method used for the analysis of high dimensional data,

the SOM is an unsupervised learning algorithm able to preserve topology of the clusters. With more than 3000 potential global insect pests, a method that can rank these species in terms of their strength of association with species assemblages, and also indicate their risk of invasion, will help focus the resources and research effort of conservation, quarantine and biosecurity scientists and managers. Particular species can be easily targeted for more detailed investigation. Furthermore, while similarity between assemblages has been emphasized as containing hidden predictive information, dissimilarity or species absence may also increase our understanding of species invasions and the invasion process. We are confident that analyses similar to the one presented here will also add value to the large amounts of invasive species distribution data that are currently collected globally.

Acknowledgements

We thank CAB International for use of the data included in the Crop Compendium – Global Module, 5th edition, © CAB International, Wallingford, UK (2003). This research was funded in part by the Centre of Research Excellence-funded postdoctoral fellowship (<http://bioprotection.lincoln.ac.nz/>). We also thank Dr Alan Stewart, Plant Breeder, Ceres Research Farm, Christchurch, for sharing his extensive knowledge regarding the origins of New Zealand's non-native plant species, Joel Pitt for his help extracting data and formatting the database, Brad Case for his valued assistance preparing the maps and several anonymous referees for their helpful comments.

References

- Baker, R., Cannon, R., Bartlett, P. & Barker, I. (2005) Novel strategies for assessing and managing the risks posed by invasive alien species to global crop production and biodiversity. *Annals of Applied Biology*, **146**, 177–191.
- Begon, M., Harper, J.L. & Townsend, C.R. (1996) *Ecology: Individuals, Populations and Communities*, 3rd edn. Blackwell Science, Oxford, UK.
- ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.
- Buss, D.F., Baptista, D.F., Nessimian, J.L. & Egler, M. (2004) Substrate specificity, environmental degradation and disturbance structuring macroinvertebrate assemblages in neotropical streams. *Hydrobiologia*, **518**, 179–188.
- CABI (2003) *Crop Protection Compendium, Global Module*, 5th edn. CAB International, Wallingford, UK.
- 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. *Ecological Modelling*, **146**, 167–180.
- Chon, T.S., Park, Y.S., Moon, K.H. & Cha, E.Y. (1996) Patterning communities by using an artificial neural network. *Ecological Modelling*, **90**, 69–78.
- Davies, D.L. & Bouldin, D.W. (1979) A cluster separation measure. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, **1**, 224–227.
- Dentener, P.R., Whiting, D.C. & Connolly, P.G. (2002) *Thrips*

- palmi* Karny (Thysanoptera: Thripidae): could it survive in New Zealand? *New Zealand Plant Protection*, **55**, 18–22.
- Diaz, S., Cabido, M. & Casanoves, F. (2001) Functional implications of trait–environment linkages in plant communities. *Ecological Assembly Rules* (ed. E. Weiher), pp. 338–362. Cambridge University Press, New York, NY.
- Eilton, C.S. (1958) *The Ecology of Invasion by Animals and Plants*. Methuen, London, UK.
- Gevrey, M., Rimet, F., Park, Y.S., Giraudel, J.L., Ector, L. & Lek, S. (2004) Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshwater Biology*, **49**, 208–220.
- Giraudel, J.L. & Lek, S. (2001) A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*, **146**, 329–339.
- Giske, J., Huse, G. & Fiksen, O. (1998) Modelling spatial dynamics of fish. *Reviews in Fish Biology and Fisheries*, **8**, 57–91.
- Hepner, G.F., Logan, T., Ritter, N. & Bruant, N. (1990) Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, **56**, 469–473.
- Hulme, P.E. (2003) Biological invasions: winning the science battles but losing the conservation war? *Oryx*, **37**, 178–193.
- Jones, R.E. & Kitching, R.L. (1981) Why an ecology of pests? *The Ecology of Pests, Some Australian Case Histories* (eds R.L. Kitching & R.E. Jones), pp. 254. CSIRO Australia, Melbourne, UK.
- Joy, M.K. & Death, R.G. (2004) Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology*, **49**, 1036–1052.
- Keddy, P.A. (1992) Assembly and response rules: two goals for predictive community ecology. *Journal of Vegetation Science*, **3**, 157–164.
- Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59–69.
- Kohonen, T. (1995) *Self-Organizing Maps*. Springer-Verlag, Heidelberg, Germany.
- Kohonen, T. (2001) *Self-Organizing Maps*. Springer, Berlin, Germany.
- Kohonen, T. & Somervuo, P. (1998) Self-organizing maps of symbol strings. *Neurocomputing*, **21**, 19–30.
- Krebs, C.J. (2001) *Ecology. The Experimental Analysis of Distribution and Abundance*, 5th edn. Benjamin Cummings, San Francisco, CA.
- Lek, S. & Guegan, J.-F. (2000) *Artificial Neuronal Networks, Application to Ecology and Evolution*. Springer-Verlag, Heidelberg, Germany.
- Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. & Aulagnier, S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, **90**, 39–52.
- Levine, J.M. & D'Antonio, C.M. (2003) Forecasting biological invasions with increasing international trade. *Conservation Biology*, **17**, 322–326.
- Lippman, R.P. (1987) An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Magazine*, **4**, 4–22.
- The Mathworks (2001) *MATLAB*, Version 6.5. The Mathworks, Natick, MA.
- Mooney, H.A. & Drake, J.A. (1989) Biological invasions: a SCOPE program overview. *Biological Invasions. A Global Perspective. SCOPE 37* (eds J.A. Drake, H.A. Mooney, F. di Castri, R.H. Groves, F.J. Kruger, M. Rejmanek & M. Williamson), pp. 525. John Wiley & Sons, Chichester, NY.
- 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. *Ecological Modelling*, **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. *Conservation Biology*, **17**, 1748–1758.
- Pimentel, D. (1986) Biological invasions of plants and animals in agriculture and forestry. *Ecology of Biological Invasions of North America and Hawaii* (eds H.A. Mooney & J.A. Drake), Vol. 58, pp. 149–162. Springer Verlag, New York, NY.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (ed. D.E.J.M. Rumelhart), Vol. 1, pp. 318–362. MIT Press, Cambridge, MA.
- Sailer, R.J. (1983) History of insect introductions. *Exotic Plant Pests and North American Agriculture* (eds C.L. Wilson & C.L. Graham), pp. 15–38. Academic Press, New York, NY.
- Simberloff, D. (1986) Introduced insects: a biogeographic and systematic perspective. *Ecology of Biological Invasions of North America and Hawaii* (eds H.A. Mooney & J.A. Drake), Vol. 58, pp. 3–26. Springer Verlag, New York, NY.
- Simberloff, D. (1989) Which insect introductions succeed and which fail? *Biological Invasions: A Global Perspective* (eds J.A. Drake, H.A. Mooney, F. di Castri, R.H. Groves, F.J. Kruger, M. Rejmanek & M. Williamson), pp. 61–75. Wiley, New York, NY.
- Ultsch, A. & Siemon, H.P. (1990) Kohonen's self organizing feature maps for exploratory data analysis. *INNC'90, International Neural Network Conference*, pp. 305–308. Kluwer, Dordrecht, the Netherlands.
- Waite, S. (2000) *Statistical Ecology in Practice. A Guide to Analysing Environmental and Ecological Field Data*. Prentice Hall, London, UK.
- Wiberg-Larsen, P., Brodersen, K.P., Birkholm, S., Gron, P.N. & Skriver, J. (2000) Species richness and assemblage structure of Trichoptera in Danish streams. *Freshwater Biology*, **43**, 633–647.
- Worner, S.P. (1994) Predicting the establishment of exotic pests in relation to climate. *Quarantine Treatments for Pests of Food Plants* (eds J.L. Sharp & G.J. Hallman), pp. 11–32. Westview Press, Boulder, CO.
- Worner, S.P. (2002) Predicting the invasive potential of exotic insects. *Invasive Arthropods and Agriculture: Problems and Solutions* (eds G. Halman & C.P. Schwalbe), pp. 119–137. Science Publishers Inc., Enfield, NH.

Received 4 August 2005; final copy received 30 April 2006
Editor: Phil Hulme

Supplementary material

The following supplementary material is available as part of the online article (full text) from <http://www.blackwell-synergy.com>.

Appendix S1. Geographic areas with similar pest assemblages.

Figure S1. The cells of the self-organizing map.