Journal of Applied Ecology

British Ecological Society

Journal of Applied Ecology 2010, **47**, 290–298 doi: 10.1111/j.1365-2664.2010.01782.x

Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion

Dean R. Paini[1,2]*, Susan P. Worner[1,3], David C. Cook[1,2,4], Paul J. De Barro[1,2] and Matthew B. Thomas[1,2,5]

[1]*Cooperative Research Centre for National Plant Biosecurity, Bruce, ACT 2617, Australia;* [2]*CSIRO Entomology, Canberra, ACT 2601, Australia;* [3]*National Centre for Advanced Bio-Protection Technologies, Lincoln University, Canterbury, New Zealand;* [4]*Fenner School of Environment and Society, The Australian National University, Canberra, ACT 2000, Australia; and* [5]*Department of Entomology and Center for Infectious Disease Dynamics, Penn State University, PA 16802, USA*

Summary

1. Predicting which species are more likely to invade a region presents significant difficulties to researchers and government agencies. Methods for estimating the risk of establishment are often qualitative and rely on consultation with experts and stakeholders. The inherent subjectivity of this process can lead to ambiguities in any estimate of a species' risk of establishment.

2. Using global presence/absence data of insect crop pests employed a self-organizing map (SOM) to categorize regions based on similarities in species assemblages. This technique enabled them to generate a list of species and rank them based on an index of the risk of establishment. However, the sensitivity of this risk list to errors in the presence/absence data has never been tested.

3. We evaluated the sensitivity of the SOM method by altering the original presence/absence data by increasing amounts and compared estimates of risk with those generated by a national coordinating body (Plant Health Australia) utilizing expert stakeholder opinion.

4. The risk list was unaffected by alterations of up to 20% of data over all regions. The error rate we detected in the data was within these limits.

5. Comparison with the expert stakeholder methodology revealed significant differences in the estimates of establishment risk. Further analysis of the Australian data revealed that a number of regions with strong trade links to Australia supported species assemblages similar to those in Australia, suggesting they are possible sources of pest species with high probability of establishment.

6. *Synthesis and applications.* This analysis confirms that the SOM methodology is a robust tool in the quantification of risk of establishment. In addition, SOMs can deliver a level of objectivity, which can complement current consultative processes employed by many biosecurity agencies around the world, providing a better overall assessment of invasion risk. This assessment can inform research and development funding decisions and incursion management plans for both government and host industries. While SOMs are utilized in this work for the prioritization of pest insects they can potentially be applied to any taxa (pest or native) or at any scale in which the data are available.

Key-words: artificial neural networks, Australia, biosecurity, establishment, exotic pest, invasion, non-indigenous species, species assemblages

Introduction

There are hundreds, perhaps thousands, of insect species that have the potential to invade and establish in any particular region or country. Identifying which species are more likely

*Correspondence author. Dr Dean R. Paini, CRC-NPB, CSIRO Entomology, Black Mountain, Clunies Ross Rd, Acton ACT 2601, Australia. E-mail: dean.paini@csiro.au

than others to invade and establish is extremely difficult, yet the capacity to do so is vitally important to the biosecurity of a nation. Currently, government agencies consult industry stakeholders and technical experts, as well as published data to generate a risk assessment for a particular insect pest. Although a valid and often informative practice, this approach requires each species to be evaluated individually, making it extremely time intensive to rank and prioritize hundreds of species. Further, the reliance on expert testimony leaves the process susceptible to framing, context dependence and motivational bias (Burgman 2005), which may result in misleading prioritization. This is especially the case given that expert testimony can only reflect upon the knowledge invested in the experts, which is seldom all encompassing. As a consequence, it is difficult to assess all potential invasive pest species, and key threats might be overlooked, particularly those where available knowledge and information is lacking.

In addition, any estimate of establishment risk, though often based on a significant amount of information, is ultimately subjective. Government organizations are, therefore, continually searching for additional methods to prioritize pest lists and generate more objective establishment estimates. One such approach is the use of climate envelope models (Stephens *et al.* 2007) where the climatic parameters of a pest's native range are used to predict its likely exotic range. However, problems in the accuracy of these predictions have been highlighted by several authors (Hulme 2003; Sax *et al.* 2007). Other statistical methods include generalized additive models (Bunnell *et al.* 2009), boosted regression trees (Jacobs & Macisaac 2009), maximum entry method (Brown *et al.* 2008) and mechanistic niche modelling (Kearney & Porter 2009), but to rank and prioritize hundreds of species using any of these approaches would again mean evaluating each species individually. With so many species the time required would be prohibitive and costly.

An alternative is to take a community ecology approach that studies the species assemblage of a region. In such a way a large number of species can be analysed simultaneously and ranked according to their 'risk' of establishing in a particular region based on species associations (i.e. any species that is commonly found with a particular species assemblage is more likely to establish in a region where that species assemblage is found). This would enable an initial screening of potential pests to a more manageable number that could then be further analysed using any of the modelling methods mentioned above. A self-organizing map (SOM), which is an unsupervised artificial neural network, can be used to generate values that indicate the strength of association of a species with a species assemblage, which can be used as a risk index. Worner & Gevrey (2006) utilized invasive insect pest data from the CABI Crop Protection Compendium (CABI, 2003) and constructed a SOM to classify and group 459 regions of the world into clusters, based on their insect assemblages drawn from a global pool of 844 known insect pest species. The insect assemblage present in a particular region captures a significant proportion of biological, ecological, and abiotic factors that cannot be measured. Given the assumption that regions with similar assemblages provide similar niches, Worner & Gevrey were able to identify those regions with similar species assemblages to New Zealand and hence those regions that may be of highest risk as a source of invasive insects. In addition they were able to utilize the SOMs neuron weights to generate a quantitative estimate of the risk of establishment in New Zealand for all 844 insect species.

This technique provides an innovative way of predicting likelihood of establishment of a large number of species, and so adds valuable information to the currently employed methods. However, the CABI Crop Protection Compendium used in Worner & Gevrey (2006) is known to contain errors such that some species recorded as present in a particular region are actually absent, while some species recorded as absent are present. The rate of these errors is largely unknown, but given the methodology relies on drawing similarities between species assemblages, the occurrence of false positives and false negatives could be problematic. The question then becomes, how sensitive are these risk lists to errors in the data?

In the current paper we address this question using Australia as a case study. We first generate a risk list for Australia using the same global pest dataset as that used by Worner & Gevrey (2006). We then alter this dataset by progressively introducing errors (essentially altering species from present to absent, and absent to present) and compare the resultant risk lists with the unmodified original to determine the sensitivity of the approach. In addition, having generated a risk list, we identify those regions most closely clustered with Australia and its states and hence most likely to act as future sources of insect pests (both known and unknown). Finally, we compare the establishment risk values estimated using the SOM approach with estimates based on expert stakeholder opinion, which is currently used to inform Australian biosecurity policy and resource prioritization. Together, these analyses point to the considerable strengths of the SOM approach.

Materials and methods

DATA

We used the same data set as Worner & Gevrey (2006), which was extracted with permission from the CABI Crop Protection Compendium (CABI 2003). This data set is comprised of the presence and absence of 844 insect pests within 459 geographical regions. These regions are political countries with many of the larger countries further subdivided into their states or provinces. The result was a 459×844 matrix comprising 459 vectors each with 844 elements, where each element of a vector represented the presence (1) or absence (0) of an insect species in a region.

SOM MODEL

A SOM is an artificial neural network capable of converting high dimensional data into a two dimensional map in which data points that are found close together on the map are more similar than those that are further away. A SOM consists of two layers of artificial neurons (or nodes), the input layer and the output layer. In the SOM, the input layer is essentially the raw data and comprises 844 neurons (one neuron for each insect species) with each neuron connected to all

459 regions. The output layer is the two dimensional map comprising a suitable number of neurons, laid out in a hexagonal grid. For this data set, a map of 108 neurons with dimensions of 12 rows by 9 columns was used (see Worner & Gevrey 2006).

Further details describing a SOM analysis can be obtained from (Kohonen 2001; Worner & Gevrey 2006), but essentially, each of the 459 regions occupies a particular point in space of 844 dimensions. Each region's position in this space is determined by the 844 element vector that is the presence or absence of all 844 insect pests in that region. The SOM projects its 108 neurons into this space via neurone weight vectors. As with the region vectors, these neuron weight vectors are comprised of 844 elements. In effect, each SOM neuron is occupying a point in the same multidimensional space as the regions, thereby allowing them to 'interact' with the regions (see below for further explanation).

These neuron weight vectors can be initially projected randomly into the multidimensional space, but we use a linear initialization that distributes the neuron weight vectors corresponding to the first two eigenvalues of a principle component analysis. This linear initialization distributes the neuron weight vectors in a way that is more representative of the raw data and significantly reduces the time taken to train the network and complete the analysis (Kohonen 2001).

When the analysis is initiated, each raw data point is assessed and the neuron that is closest to this data point in this multidimensional space is deemed to be the best matching unit (BMU). The neurone weight vector of the BMU is adjusted so that it moves closer to the data point. Because all neurons are connected together similar to a large elastic net, the process of one neuron moving exerts a gravitational force that drags other neurons in the SOM with it. While each data point can be assessed individually, doing so means the learning is highly dependent on the order in which data points are assessed (Worner & Gevrey 2006). Assessing data points simultaneously, using a batch algorithm solves this problem and was used in this analysis.

Data points are repeatedly assessed and over time the neurons spread out to occupy approximately the same area that the data points occupy in the multidimensional space. When the analysis is complete each data point or region will have a BMU, which is its closest neuron. Regions that have very similar pest assemblages will be located close together in the multidimensional space and will have the same BMU. Each neuron therefore occupies a point in the multidimensional space, which is described by its neuron weight vector.

In this study the neuron weight vector is composed of 844 elements with each element having a value between 0 and 1. Each element corresponds to one of the 844 insect species and can be interpreted as a risk index or an index of how strongly that species is associated with other species in that neuron and hence the species assemblage of any region associated with that neuron (BMU). For Australia, the risk list generated would be the neuron weight of its BMU. The analysis was performed using Matlab (Mathworks 2007) and the SOM Toolbox (version 2·0) developed by the Laboratory of Information and Computer Science Helsinki University of Technology (http://www.cis.hut.fi/projects/somtoolbox/).

Because this SOM analysis does not give 'crisp edges' to clustered regions of the world, further analysis was performed by a conventional cluster analysis of the neurone weights. As no one clustering algorithm is recommended, we repeated the cluster analysis using a different clustering algorithm each time (single link, nearest neighbour, complete link, furthest neighbour, average link, median sorting, and group average clustering) and compared the results to determine which neurons were consistently clustered together. This analysis was performed using GenStat (2007).

Once a risk list for Australia was generated, we classified all those species that are recorded as absent from Australia into risk categories similar to the categories utilized by Biosecurity Australia (the Australian government agency that undertakes science-based risk assessment, and provides quarantine policy advice) to obtain a semi-quantified estimate of risk of establishment (Table 1a). However, we condensed the lower four categories of Biosecurity Australia's scheme into one category (Table 1b) as making fine grade distinctions between low risk level pests is considered relatively unimportant.

While these three risk categories (Table 1b), which are a function of the risk scores allotted to each species, follow the established Biosecurity Australia model, other agencies might be more interested in a basic ranking of species and identifying, for example, the top 100 threats. This approach has been used by the Global Invasive Species Program in a database, which lists 100 of the world's worst invasive alien species (http://www.issg.org/database/species/search.asp?st = 100ss). In addition, biosecurity agencies may wish to filter all the possible invasive species into a more manageable list for which they would seek advice from experts or stakeholders. Generating a top 100 list would be analogous to this filtering process. Such a list would not be directly dependent on a risk value but rather the relative ranking. In line with this we also generated a top 100 list of species posing the highest establishment risk for Australia.

LIST COMPARISON

To determine the sensitivity of any risk list to errors in the data set, we deliberately altered the data set by increasing amounts to simulate error rates. After the data were altered, a new SOM was generated and a subsequent risk list for Australia extracted. This list was then categorized into the three risk categories described above (Table 1b) and the top 100 insect pests were also extracted into a second list. These two lists were then compared to the original lists. List fidelity was assessed by recording the proportion of species present in each of the three risk categories that stayed in those same risk categories after the data were altered. For the top 100 insect pests, the proportion of these insects that stayed in the top 100 was recorded. In addition, to get an assessment of the fidelity of the overall list we performed Spearman's rank correlations on the entire list before and after data alteration.

DATA ALTERATION

We utilized Impact Risk Assessments (IRAs) generated by the Australian Government's Department of Agriculture, Forestry, and

Table 1. Risk categories used by (a) Biosecurity Australia (Biosecurity Australia, 2001) and (b) the three categories used in this analysis

Likelihood	Probability range
(a)	
High	0·7–1·0
Moderate	0·3–0·7
Low	0·05–0·3
Very low	0·001–0·05
Extremely low	0·000001–0·001
Negligible	0–0·000001
(b)	
High	0·7–1·0
Moderate	0·3–0·69
Low	0–0·29

Fisheries (http://www.daff.gov.au/ba/ira/final-plant) to estimate the error rate in a sample of the CABI data and determine the range of data alteration required. These IRAs assess the risk of importing a particular product from another country and identify the known insect pest species present in the exporting country and associated with that product. These lists of pest species are generated from published sources and are independent of the CABI database. We compared these lists with the CABI database and found error rates for 58 countries ranging from 0% to 38%, and averaging 8·54% (see Table S1 in Supporting Information). Included in this mean was an estimate for the error rate for Australia, which was calculated by randomly selecting 200 species and comparing the CABI database with the Australian Plant Pest Database (http://www.planthealthaustralia.com.au/our_projects/display_project.asp?category=4&ID=1) and by consulting taxonomy experts in the Australian National Insect Collection (CSIRO ANIC). The error rate for these 200 species was 2·5%.

Initially, data from all regions (459) was altered by increasing amounts (5%, 10%, 20%, and 30%). To do this, a set percentage of species were randomly selected from each region and their presence/absence score reversed. For example, for 5% alteration, 42 of the 844 species were randomly selected and their presence/absence score reversed (i.e. any species that was present was made absent and vice versa). This was done separately for all regions so that no two regions had exactly the same species data altered.

The error rate in the CABI database varied across countries and we therefore wanted to determine what variation in error rate could be tolerated by the SOM. Because the analysis indicated that an error rate of 20% across all countries could be accommodated (see Results) we subsequently separated the 59 countries in Table S1 into two groups; those with an error rate of 20% or less and those above 20%. There were 55 countries (93% of the 59 countries) with error rates equal to or less than 20% and four countries above 20%. The mean error rate for the 55 countries was 7·15%. We assumed that if 93% of countries tested against the IRAs averaged 7·15% then the same proportion of regions in the CABI database (428 regions) would have a similar error rate. To be conservative, however, we set the alteration rate for these regions at the slightly higher value of 10%. The remaining 31 regions were subsequently tested for increasing alteration rate to determine how the list was affected.

COMPARISON WITH EXPERT STAKEHOLDER OPINION

To determine the differences in ranking that can occur between a SOM analysis and the expert or stakeholder consultation currently employed by many biosecurity agencies, we compared the risk estimates obtained in this analysis with those generated by Plant Health Australia (PHA), a national coordinating body addressing the biosecurity of Australia's plant industries. As part of its role this organization generates industry biosecurity plans (http://www.planthealthaustralia.com.au/site/Industry_Biosecurity_Plan_Mainpage.asp). Within these plans, the risk of establishment for insect pests has been estimated using a process of qualitative risk assessment, which consults expert opinion. These lists and rankings for each industry are not intended as definitive or actionable lists for the purposes of quarantine arrangements but are compiled for the purpose of determining biosecurity threats for each plant industry. Of the 567 insect pests in our data set that are absent from Australia, 226 (39·8%) were also evaluated by PHA. The risk of establishment for these 226 insect pests were classified into four categories (high, medium, low, and negligible) making direct comparison with our list relatively simple, if the 'negligible' and 'low' categories are grouped together.

Many of the species (21·9%) had multiple risk categories attributed to them by PHA, depending on which crop was being considered. We counted species agreements if the SOM categorization agreed with any one of the classifications made by PHA. We also calculated Cohen's kappa statistic (Cohen 1960), to test the level of agreement between PHA and SOM after taking into account any agreement that could occur by chance. However, the kappa statistic assumes a species is classified into only one category and PHA classified many species into more than one category. For a species with multiple classifications, we determined if one of these classifications matched the SOM classification and if it did, we took that classification. If the classifications did not match we used the highest classification PHA had given that species.

Results

DATA ALTERATION

When data from all regions were altered there was a marked decrease in the fidelity of species to their original categories after 20% alteration (Fig. 1). This was confirmed by the top 100 list, which showed a similar large decrease in fidelity of species after 20% data alteration (Fig. 2). However, Spearman's rank correlation showed significant correlations between the original list and lists generated up to 30% alteration (Table 2).

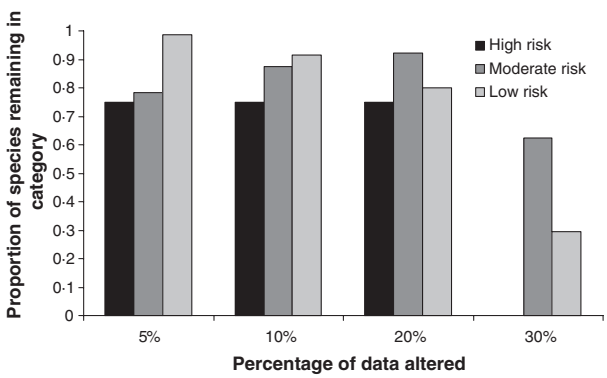

Fig. 1. The proportion of species remaining in each risk category in response to an increasing level of data alteration.

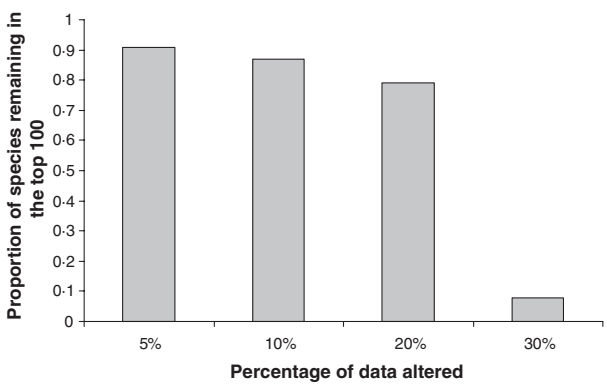

Fig. 2. The proportion of species remaining in the top 100 list in response to an increasing level of data alteration.

When regions were separated into two groups (the larger group of regions with an alteration rate of 10%) we found that using the three categories, an alteration rate of 30% in the smaller group of regions maintained category fidelity (Fig. 3). Using the top 100 list, this alteration rate could be increased to 40% before significant reductions in fidelity occurred (Fig. 4). The Spearman's rank correlation analysis revealed that the alteration rate for the smaller group of regions could be 100% and still maintain a significant correlation over the entire list (Table 2).

COMPARISON WITH EXPERT STAKEHOLDER OPINION

Comparing the groupings obtained by the SOM analysis with those estimated by PHA we found that overall only 22·1% of species had the same risk estimate by the two methods. The category with the lowest level of agreement was the low risk category where only 14·0% of species classified as low by the SOM analysis were also classified as low by PHA. For the medium and high categories there was 53·3% and 50·0% agreement respectively. Further, while the SOM analysis only classified two species as high risk, PHA classified 92 species as high risk and of these 92 species, 73 (79·3%) were classified as low risk by the SOM analysis (Tables S2 and S3). Cohen's kappa statistic was calculated to be 0·032, which meant there was little correspondence between the lists.

AUSTRALIA'S RISK LIST

The risk list for the top 100 insect pests of threat to Australia generated from the unaltered data is shown in Table 3 with the species separated into the three risk categories outlined in Table 1b (for the full list see Table S4). Australia's best matching unit (BMU), in other words the neuron to which Australia is closest in the multidimensional space, was the same neuron with which Papua New Guinea is associated. In addition, two other neurons were always clustered with Australia's BMU regardless of which clustering algorithm was used (Table 4).

Table 2. Spearman's rank correlations comparing the original list generated by SOM with lists generated from altered data

Data alteration	r_s (adjusted for ties)
All regions at 5%	0·87†
All regions at 10%	0·82†
All regions at 20%	0·69†
All regions at 30%	0·44†
All regions at 40%	0·05
31 regions at 0%‡	0·79†
31 regions at 10%‡	0·82†
31 regions at 20%‡	0·79†
31 regions at 30%‡	0·78†
31 regions at 40%‡	0·79†
31 regions at 50%‡	0·68†
31 regions at 60%‡	0·70†
31 regions at 70%‡	0·69†
31 regions at 80%‡	0·81†
31 regions at 90%‡	0·82†
31 regions at 100%‡	0·74†

‡428 regions at 10%.
†$P < 0{\cdot}001$.

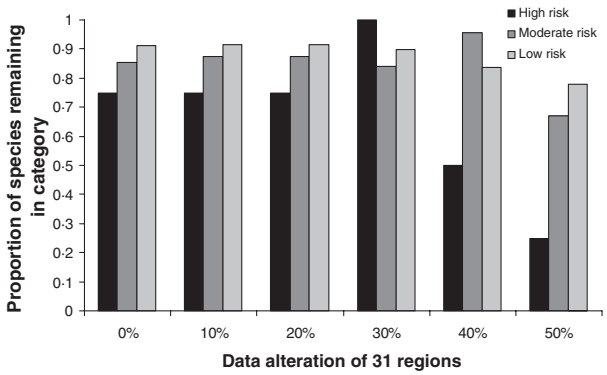

Fig. 3. The proportion of species remaining in each risk category in response to an increasing level of data alteration for 31 randomly selected regions. The alteration rate for the remaining regions was maintained at 10%.

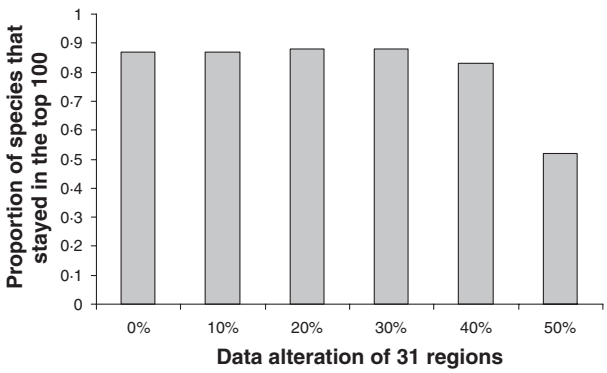

Fig. 4. The proportion of species remaining in the top 100 list in response to an increasing level of data alteration for 31 randomly selected regions. The alteration rate for the remaining regions maintained at 10%.

Discussion

As current methods for quantifying the risk of establishment of invasive species can be relatively subjective, identifying and evaluating new methods that could help in generating more accurate invasive risk lists is paramount, and Worner & Gevrey (2006) have introduced the novel idea of using a SOM to generate these risk lists. As with any modelling technique, it is important to obtain high quality data and/or test model sensitivity to data error. We found that an error rate in species distribution lists of up to 20% across all regions will still generate risk lists of relatively high fidelity.

Although an overall error rate of 20% would appear to be adequately handled by any SOM analysis, our estimation of the error rate showed that a small proportion of regions had higher error rates (Table S1). The error rate for these countries averaged 27% and the subsequent analysis that separated the 459 regions into two groups indicated that either a categorical list or a top 100 list could both be insensitive to errors at this

Table 3. Top 100 risk list for insect pests from the SOM analysis. Lines indicate the three risk categories utilized in the data analysis (see Table 1). For a full list see Table S4

rank	Insect pest species	risk index	rank	Insect pest species	risk index	rank	Insect pest species	risk index
1	Scirpophaga incertulas	0·7924	35	Schizaphis graminum	0·4895	69	Helopeltis bradyi	0·3634
2	Oryctes rhinoceros	0·7722	36	Aproaerema modicella	0·4841	70	Phyllotreta striolata	0·3604
3	Sesamia inferens	0·7695	37	Xylotrechus quadripes	0·4835	71	Minthea rugicollis	0·36
4	Scrobipalpa heliopa	0·7128	38	Attacus atlas	0·4821	72	Idioscopus niveosparsus	0·3589
5	Marasmia exigua	0·6856	39	Ceratovacuna lanigera	0·4774	73	Scotinophara coarctata	0·3589
6	Diaphorina citri	0·6807	40	Bactrocera latifrons	0·4651	74	Naranga diffusa	0·3445
7	Aleurocanthus woglumi	0·6801	41	Hypothenemus hampei	0·4606	75	Liriomyza trifolii	0·3431
8	Leucinodes orbonalis	0·6748	42	Henosepilachna pusillanima	0·4605	76	Urentius hystricellus	0·342
9	Stephanitistypica	0·6675	43	Idioscopus clypealis	0·4589	77	Megymenum brevicorne	0·3326
10	Xylosandrus compactus	0·6632	44	Sternochetus frigidus	0·4586	78	Elaeidobius kamerunicus	0·3311
11	Pelopidas mathias	0·6446	45	Trichoplusia ni	0·4573	79	Aulacaspis tegalensis	0·3262
12	Dicladispa armigera	0·6378	46	Dialeurodes citri	0·4532	80	Bombyx mori	0·3194
13	Acherontia styx	0·6356	47	Pyrilla perpusilla	0·4499	81	Rhynchophorus vulneratus	0·3189
14	Chilo auricilius	0·6338	48	Chilo sacchariphagus	0·4491	82	Aulacophora foveicollis	0·3156
15	Nephotettix virescens	0·6279	49	Atherigona soccata	0·4395	83	Tessaratoma papillosa	0·3147
16	Planococcus lilacinus	0·594	50	Sinoxylon conigerum	0·4335	84	Statherotis discana	0·3145
17	Batocera rubus	0·5836	51	Cricula trifenestrata	0·4329	85	Cydia leucostoma	0·3132
18	Aulacophora lewisii	0·5702	52	Chilo partellus	0·4206	86	Bactrocera umbrosa	0·3105
19	Toxoptera odinae	0·5694	53	Orthezia insignis	0·4138	87	Pseudococcus jackbeardsleyi	0·3105
20	Orseolia oryzae	0·5676	54	Rhipiphorothrips cruentatus	0·4108	88	Perkinsiella vastatrix	0·3052
21	Odoiporus longicollis	0·5669	55	Batocera rufomaculata	0·41	89	Artona catoxantha	0·3043
22	Chilo infuscatellus	0·5559	56	Omiodes indicata	0·4097	90	Heterobostrychus aequalis	0·3037
23	Zeuzera coffeae	0·551	57	Plocaederus obesus	0·4047	91	Poecilocoris latus	0·3037
24	Helopeltis theivora	0·5425	58	Bactrocera tau	0·4008	92	Opisina arenosella	0·303
25	Hypomeces squamosus	0·5395	59	Adoretus versutus	0·3917	93	Chondracris rosea	0·2994
26	Orgyia postica	0·5343	60	Hieroglyphus banian	0·3888	94	Prays endocarpa	0·2905
27	Pinnaspis strachani	0·529	61	Rastrococcus iceryoides	0·3879	95	Chromatomyia horticola	0·2827
28	Rastrococcus invadens	0·519	62	Fulmekiola serrata	0·3834	96	Medythia suturalis	0·2746
29	Parasa lepida	0·5177	63	Phyllotreta chotanica	0·3803	97	Acherontia lachesis	0·2735
30	Papilio polytes	0·5119	64	Bactrocera zonata	0·3802	98	Aphis fabae	0·2704
31	Bactrocera dorsalis	0·5008	65	Melanagromyza obtusa	0·3747	99	Tetramoera schistaceana	0·2684
32	Hydrellia philippina	0·4957	66	Liriomyza huidobrensis	0·3743	100	Rhynchocoris poseidon	0·2682
33	Omphisa anastomosalis	0·4933	67	Tarophagus colocasiae	0·372			
34	Erionota thrax	0·4931	68	Chilo polychrysus	0·3658			

level. In addition, rank correlations were maintained at significant levels up to 100% (i.e. all species in this small proportion of countries could be incorrectly classified as present or absent), indicating that, when considering all species, the list still maintained constant rankings.

If a biosecurity agency wished to use SOM as an initial filtering device that would enable a large number of species to be reduced to a smaller manageable number that could be utilized in any expert solicitation, then the Spearman rank correlations indicate that error rates up to 30% could be accommodated. Considering the error rate we found in the CABI CPC data averaged only 8·5%, SOM would appear to be very robust.

Another important aspect of this SOM analysis is how it compares to risk assessment techniques that rely on expert and stakeholder consultation. Keller *et al.* (2007) demonstrate that the benefits generated by the adoption of a risk assessment strategy based on expert judgement are potentially large. But, the perceived lack of accuracy of this approach perhaps explains why the vast majority of countries have not mandated risk analysis for non-indigenous species (Keller *et al.* 2007). On this basis, a SOMs-supplemented method of risk assessment could generate substantial economic gains over time if it produces higher confidence, and therefore greater adoption of pre-import risk assessments as a key instrument of biosecurity policy.

Comparing the PHA risk classification method with SOMs predictions reveals a relatively low level of agreement between the two methodologies, with PHA classifying significantly more species (108) as a high risk of establishment compared to the SOM methodology (2). In addition, many of the species (21·9%) had multiple risk categories attributed to them by PHA, depending on which crop was being considered. As we counted any species agreement if the SOM categorization agreed with any one of the classifications made by PHA, any

Table 4. Regions in the same BMU as Australia and its states and territory, and the regions associated with the neurones that were most often clustered with the Australian BMU. For details of the cluster analyses see Table S5. For the full list of neurons and associated regions see Table S9

Target region and regions placed in the same BMU	Regions placed in the neurons most often clustered with the target region BMU
1. Australia, Papua New Guinea	Bangladesh, China, Taiwan (China), Indonesia, Java (Indonesia), India, Japan, Sri Lanka, Myanmar, Malaysia, Peninsular Malaysia (Malaysia), Philippines, Pakistan, Singapore, Thailand, Vietnam
2. Western Australia	South Australia, Tasmania (Australia), Victoria (Australia), New Zealand, Azores (Portugal), Saudi Arabia, St Helena
3. Northern Territory (Australia)	Delhi (India), Gujarat (India), Indian Punjab (India), Rajasthan (India), Andhra Pradesh (India), Bihar (India), Maharashtra (India), Madhya Pradesh (India), Orissa (India), Uttar Pradesh (India), Northern Mariana Islands
4. South Australia, Tasmania (Australia), Victoria (Australia), New Zealand, Azores (Portugal)	Western Australia, Saudi Arabia, St Helena
5. New South Wales (Australia), Queensland (Australia), Fiji, New Caledonia, Solomon Islands	Assam (India), Karnataka (India), Kerala (India), Tamil Nadu (India), West Bengal (India), Brunei Darussalam, Guangdong (China), Hong Kong (China), Sumatra (Indonesia), Cambodia, Laos, Sabah (Malaysia), Sarawak (Malaysia)

percentage agreement between PHA and SOM may be inflated. Finally, Cohen's kappa statistic indicated very low agreement in the classification by SOM and PHA. Values for kappa range from -1 (complete disagreement) to 1 (complete agreement). A value close to zero, as reported here, indicates that any agreement between PHA and SOM can only be attributed to chance.

The large disparity between SOM and expert or stakeholder consultation may indicate the inclination of people, even those with a significant level of biological and ecological knowledge, to be risk averse and classify a species as a high risk when perhaps it is not. These experts or stakeholders also may not have an in depth knowledge of all potential pests and may therefore confuse the risk of establishment for a pest with the potential impact of that pest (Gary Fitt, CSIRO, personal communication). In addition, expert or stakeholder solicitation has often been found susceptible to a range of cognitive biases such as the format of the question(s), past experience, overconfidence, motivational bias, lack of independence, and cultural, political or philosophical context (see Burgman 2005 for review). Despite this, there are a range of methods a facilitator can utilize to improve any estimates. One of these methods is to give the expert feedback on their estimates and allow them the opportunity to alter them (Burgman 2005). Any species ranking list generated by a SOM analysis would not only be independent of the 'human' biases mentioned above, but could be used by a facilitator as additional information that could serve as feedback for the experts to consider in their final estimates.

The placing of a pest species into different categories by PHA can also complicate the pest risk analysis and the second advantage therefore of using a SOM analysis is only one risk estimate is provided and this can be utilized by biosecurity agencies without having to consider the multiple risk categories that could be obtained in an expert or stakeholder consultation process.

Finally, while classification into the three categories is possible in both methodologies, this gives an equal 'value' of risk to all species within the same risk category. The SOM analysis however, gives quantitative estimates of risk, which can allow further prioritization within each risk group and a more refined list.

One of the important reasons for using expert and stakeholder consultation is to ensure that factors associated with specific production concerns are met. The ability therefore of stakeholders to provide input and be part of the risk assessment procedure will ensure these stakeholders accept a shared responsibility for managing biosecurity concerns. For this reason, the SOM methodology should not replace the process of consultation, but can provide a framework and guide to the consultative process, enabling consultants access to more analytical assessments, which can better inform their recommendations regarding a pest's risk of establishment. Considering the cognitive biases mentioned above that are inherent in any consultative process, the addition of the SOM methodology and the information it provides can only improve the subsequent estimates of likelihood experts and stakeholders will produce.

It should be noted that SOM estimations of establishment likelihood are based on current distributions of species, which is inherently a function of historical pathways. If new trade pathways become established, some species that have a restricted range due to pathway limitation may invade new regions and species assemblages would be altered, thereby altering SOM predictions. However, analyses conducted using simulations in a virtual world of invasive pests indicate that SOM is able to predict even those species with restricted ranges (unpublished data) and we maintain our confidence in SOMs predictive powers.

Once a species' risk of establishment is determined, further analysis of host availability and distribution as well as possible entry pathways would be appropriate to assess overall risk. That is, a species may have a high likelihood of establishment but if the pathway is absent, then the likelihood of entry is low. In addition, species at the top of the list could be further

analysed using climate or niche matching models to identify specific regions within a country at greatest risk from a pest species. Finally, information on economic costs of a particular pest should also be considered. For example, the third highest pest species at risk of establishing in Australia (Table 3) is *Sesamia inferens* (Lepidoptera: Noctuidae), and while this species is a pest of rice, sugarcane, maize, sorghum and wheat, it is considered the least destructive of the stem borer pests (CABI 2003). An economic analysis might suggest that despite this pest having a high likelihood of establishing in Australia, it may not be considered a serious economic threat.

In contrast, *Chromatomyia horticola* (Diptera, Agromyzidae) is a very serious pest in almost all countries in which it is found, causing serious damage to tomatoes, legumes, lettuce, cruciferous crops and cucurbits, among others (CABI 2003). Although this pest was only ranked 95th in the top 100 risk list (Table 3) and has a low risk of establishment, its potential to cause significant economic damage may motivate government authorities to treat this pest as a more serious threat than indicated by establishment risk alone.

In addition to the pest rankings there is additional information provided by the SOM analysis that can be utilized by biosecurity agencies. The first is determining which regions have been allocated to the same neuron as the target region. For Australia, the only other region associated with the same neuron is Papua New Guinea (PNG), indicating these two regions have a significant percentage similarity in insect assemblage (48·4%). As such, they may share similar climatic, biological and ecological characteristics and insect pests that are established in PNG may therefore have a high risk of establishing in Australia.

Neurons that are neighbouring the BMU of the target region also provide information. The regions belonging to these neighbouring neurons, though not such a close match to the target region as those regions allocated to its BMU will have similar insect pest assemblages and hence also represent a potential source of insect pests. Most neurons in a SOM will have six neighbouring neurons unless it is on the edge of the map, where it will have only four neighbours, or in the corner of the map, where it will have only three neighbouring neurons (Fig. S1). These neurons occupy a point in the multidimensional space and are not necessarily evenly distributed throughout this space. Some neighbouring neurons may therefore be closer than others to a BMU. A conventional cluster analysis can reveal which of these neighbouring neurons are closest to a BMU and hence which regions are more similar to the target region. Australia's BMU was on the edge of the map and it therefore had only four neighbouring neurons. Of these four neurons, two were consistently clustered with Australia's BMU regardless of the clustering algorithm used (Table S5).

As with PNG, the regions associated with these closely clustered neurons are also possible sources for the species already in the risk list. In addition, because of the significant percentage similarity in insect assemblages (Table S6) and therefore the potential similarities in climatic, biological and ecological conditions, these regions could also be sources for insect pests that have not been included in the database used in this analysis and extra caution should be taken when inspecting imports from these regions. As five of these regions are in the top ten for merchandize imports into Australia (China, Japan, Singapore, Thailand, and Malaysia) (Table S7) entry pathway may not provide a constraint to their arrival.

For example, the CABI Crop Protection Compendium (CABI 2003) provides distributional data on 21 of 30 (70%) of the world's worst invasive crop and forest insect pests as determined by the IUCN–World Conservation Union (Table S8) (http://www.issg.org/database). One of these species not found in the Compendium and hence not included in the analysis is *Aulacaspis yasumatsui* (Hemiptera: Diaspididae), a major pest of cycads. This insect is present in China, Thailand, Singapore and Taiwan (Germain & Hodges 2007), all countries closely clustered with Australia (Table 4) and all exporting significant quantities of commodities into Australia (Table S7). It follows that this pest, though not included in the database, might present a high risk of establishing in Australia. Further analysis utilizing habitat suitability modelling (Kriticos *et al.* 2003), impact simulation modelling (Cook *et al.* 2007), and benefit costs analysis (Cook 2008) may reveal more about the risk this species poses to Australia.

While this methodology has been used to generate a risk list for Australia, a list for any country in the database could be generated and utilized by that country's biosecurity agency in the ways discussed above. However, risk lists generated for developing regions should be checked for errors. An error rate for such regions greater than 20% could result in inaccurate risk list and a test of list fidelity similar to the one presented here would be appropriate.

In addition to risk lists for whole countries, the CABI Crop Protection Compendium also has data for the states or provinces within many of the larger countries. Countries in this database that have been divided up into states or territories include Australia, Brazil, Canada, China, India, Indonesia, Japan, Malaysia, Russia, and USA. This can give predictions at a finer scale and also reveal which insect pests found in a state also present a high risk of establishment in a neighbouring state. Government agencies could then identify not only those threats from outside the country but also those from within.

Generating more accurate estimates for the risk of invasion and establishment of species is vital for informed biosecurity. Government agencies require such lists and estimates of risk so as to allocate resources in such a way that will efficiently prioritize pest detection methods. These quantitative estimates can also feed into economic models used in import risk assessments (Cook *et al.* 2007) that can affect policy decisions. Importantly, the use of SOMs can be extended to any taxa, such as weeds, marine pests, or even organisms of threat to natural systems. As long as the worldwide distributional data is available, this methodology can be utilized by any agency or researcher in which prioritization or prediction of establishment likelihood is required. In addition, while this SOM approach uses species assemblages, it may

be possible to include abiotic factors to further improve predictions (Ferrier *et al*. 2005).

Our analysis suggests that SOMs can provide important information for the evaluation and prioritization of pest lists, with species rankings appearing relatively robust to quite large errors in species distribution data. Given such errors are inevitable, these findings illustrate the practical utility of this approach and the utility of SOMs as a method, which can complement the current approaches used by biosecurity agencies. The addition of such a tool has the potential to allow a better overall assessment of invasion risk and we encourage further evaluation and adoption by researchers and stakeholders.

Acknowledgements

The authors would like to acknowledge the support of the Cooperative Research Centre for National Plant Biosecurity, established and supported under the Australian Government's Cooperative Research Centres Program. We thank CAB International for use of the data included in the Crop Pest Compendium (2003). We also thank Sharyn Taylor of Plant Health Australia for the establishment data from their industry biosecurity plans and for comments on the manuscript. Andy Sheppard, Nadiah Kristensen, and Simon Barry made helpful comments on early versions of the manuscript. John LaSalle, Marianne Horak, Rolf Oberprieler, and Peter Gillespie assisted in clarifying presence/absence data. Bob Forrester gave statistical advice. Two anonymous reviewers made helpful comments on the manuscript also. This work was done in collaboration with the EU 7th Framework project, PRATIQUE.

References

Biosecurity Australia (2001) *Guidelines for Import Risk Analysis, Draft Report*, September, Department of Agriculture, Fisheries and Forestry, Canberra, Australia.

Brown, K.A., Spector, S. & Wu, W. (2008) Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. *Journal of Applied Ecology*, **45**, 1639–1648.

Bunnell, D.B., Madenjian, C.P., Holuszko, J.D., Adams, J.V. & French, J.R.P. (2009) Expansion of *Dreissena* into offshore waters of Lake Michigan and potential impacts on fish populations. *Journal of Great Lakes Research*, **35**, 74–80.

Burgman, M. (2005) *Risks and Decisions for Conservation and Environmental Mangement*. Cambridge University Press, Cambridge.

CABI (2003) *Crop Protection Compendium, Global Module*, 5th edn. CAB International, Wallingford, UK.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cook, D.C. (2008) Benefit cost analysis of an import access request. *Food Policy*, **33**, 277–285.

Cook, D.C., Thomas, M.B., Cunningham, S.A., Anderson, D.L. & De Barro, P.J. (2007) Predicting the economic impact of an invasive species on an ecosystem service. *Ecological Applications*, **17**, 1832–1840.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2005) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264.

GenStat (2007) *GenStat*, 10th edn, version 10.1. Lawes Agricultural Trust, Hemel Hempstead, UK.

Germain, J.F. & Hodges, G.S. (2007) First report of *Aulacaspis yasumatsui* (Hemiptera : Diaspididae) in Africa (Ivory Coast), and update on distribution. *Florida Entomologist*, **90**, 755–756.

Hulme, P.E. (2003) Biological invasions: winning the science battles but losing the conservation war? *Oryx*, **37**, 178–193.

Jacobs, M.J. & Macisaac, H. J. (2009) Modelling spread of the invasive macrophyte *Cabomba caroliniana*. *Freshwater Biology*, **54**, 296–305.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–350.

Keller, R.P., Lodge, D.M. & Finnoff, D.C. (2007) Risk assessment for invasive species produces net bioeconomic benefits. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 203–207.

Kohonen, T. (2001) *Self-Organizing Maps*, 3rd edn. Springer, Berlin, Germany.

Kriticos, D.J., Sutherst, R.W., Brown, J.R., Adkins, S.W. & Maywald, G.F. (2003) Climate change and the potential distribution of an invasive alien plant: *Acacia nilotica* ssp. *indica* in Australia. *Journal of Applied Ecology*, **40**, 111–124.

Mathworks (2007) *MATLAB*, version 7.4. The Mathworks, Natick, MA.

Sax, D.F., Stachowicz, J.J., Brown, J.H., Bruno, J.F., Dawson, M.N., Gaines, S.D., Grosberg, R.K., Hasting, S.A., Holt, R.D., Mayfield, M.M., O'Connor, M.I. & Rice, W.R. (2007) Ecological and evolutionary insights from species invasions. *Trends in Ecology & Evolution*, **22**, 465–471.

Stephens, A.E.A., Kriticos, D.J. & Leriche, A. (2007) The current and future potential geographical distribution of the oriental fruit fly, *Bactrocera dorsalis* (Diptera : Tephritidae). *Bulletin of Entomological Research*, **97**, 369–378.

Worner, S. P. & Gevrey, M. (2006) Modelling global insect pest species assemblages to determine risk of invasion. *Journal of Applied Ecology*, **43**, 858–867.

Received 25 August 2009; accepted 12 January 2010
Handling Editor: Brendan Wintle

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1. Error rates present in the CABI CPC for 58 countries.

Table S2. A confusion matrix comparing the classification into three levels of likelihood of establishment by the SOM analysis and Plant Health Australia (PHA).

Table S3. Comparison of species risk rankings between the SOM analysis and PHA.

Table S4. The full risk list for insect pest species absent from Australia.

Table S5. The neurones clustered with Australia and its states and territory in cluster analyses.

Table S6. Species similarity indices between Australia and closely clustered regions.

Table S7. The top 20 countries exporting goods into Australia based on the mean monthly percentage of total imports for 2007.

Table S8. List of 58 of the world's most invasive insect pests.

Table S9. Full list of neurones and associated regions.

Figure S1. SOM map with numbered neurones.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.