Comparison of a Self Organising Map and Simple Evolving Connectionist System for Predicting Insect Pest Establishment

Michael J. Watts and S. P. Worner

National Centre for Advanced Bio-Protection Technologies
PO Box 84
Lincoln University
Canterbury
New Zealand

{wattsm2,worner}@lincoln.ac.nz

Abstract

A comparison of two artificial neural network methods for predicting the risk of insect pest species establishment in regions where they are not normally found is presented. The ANN methods include a well-known unsupervised learning algorithm and a relatively new supervised constructive method. A New Zealand pest species assemblage as an example was used to compare model predictions. Both methods gave similar results for already established and non-established species.
Keyword: Self-Organising Maps, Evolving Connectionist Systems, pest invasion prediction.

I. Introduction

The rising rate of global tourism and trade is rapidly increasing the threat to human health, agricultural and horticultural production and biodiversity of many countries by unintended introductions of exotic species. While each nation has regulatory methods that are intended to prevent exotic or invasive species establishing there is a desperate need to develop methods that have a higher level of prediction to assist the pest risk assessment process.

A number of models and approaches have been designed to predict the establishment of species in regions where they are not normally found. These models are usually based on a combination of biotic and environmental predictors. Such methods range from classical statistical approaches that relate species presence and absence at localities to environmental factors, to process models that describe species response to the environment. A neurocomputing approach that has often been applied to ecological questions is the unsupervised ANN algorithm, the Kohonen Self-Organising Map (SOM) [6]. Kohonen SOMs have been applied to classification problems in ecology, particularly to detect patterns in communities of species (e.g. [1,3]). SOMs have recently been used to investigate the assemblages of insect pest species at global locations to predict their establishment in new areas [10]. In this study the species were ranked with regard to their risk of establishment. It is clear however that such species ranking will change depending on the initial conditions of the learning algorithm. In this study we compare the SOM rankings together with those produced by a relatively new ANN method.

II. Predicting Pest Establishment with ANN

The two types of ANN used in this study, Kohonen Self-Organising Maps (SOM) [6] and Simple Evolving Connectionist System (SECoS) [7], are both types that perform spatial clustering of training examples during learning. By determining the cluster to which a geographic region represented by its species assemblage belongs to, it is possible to predict which pest species may become established in that region.

The rationale behind this approach is that regions that have similar assemblages are likely to have climatic or other environmental properties in common that allow the species to establish. If a particular region in a cluster does not have a pest species present, while other regions within the cluster do, it is hypothesised that it is likely that the region possesses an environment conducive to the establishment of that species if it were introduced.

Central to the prediction of pest establishment is the meaning of the connection weights in the ANN. Each input neuron in the ANN represents one pest species. As the networks train, species that are seen more commonly in a particular cluster are assigned larger weights by the training algorithm. The weights are therefore a reflection of the frequency of the species in the training data for that particular cluster. Since each cluster represents geographic regions that are similar in terms of pest assemblages, the weights of each species are a reflection of the probability of each species being found in a geographic region within that cluster. The weights are thus an indicator of the risk [4] of each species establishing in regions that are present in that cluster.

The following subsections describe how these risk weightings are extracted from the two types of ANN. For each type of network, a list of species is extracted where each species has a risk weighting or risk of establishment associated with it.

A. *Predicting Risk of Establishment with Kohonen SOM*

At the conclusion of training each cell in the output map represents the centre of a cluster of training examples. A list of risk weightings for a particular geographic region is created from a SOM by propagating through the network the vector that represents that region. The winning neuron is then identified and its weight vector taken as the list of species risk weightings.

B. *Predicting Risk of Establishment with SECoS*

Evolving Connectionist Systems (ECoS) [5] are a class of constructive ANN that are similar in the way in which they are constructed and learn. Advantages of ECoS are that they learn very quickly and are resistant to catastrophic forgetting.

The SECoS ANN [7] is a minimalist implementation of the ECoS principles and consists of three layers of neurons. The first is the input layer. The second is the evolving or constructive layer: it is this layer to which neurons are added and in which learning takes place. The activation of the evolving layer neurons is based on the distance between the input to evolving layer weight vectors and the current input vector. The third layer is the output layer.

Learning in the evolving layer takes place as a mixture of adding neurons and weight adjustment. Neurons are added if the current training vector is above a certain threshold level of dissimilarity. Connection weights in the input to evolving layer connections are adjusted such that the connection weights vector of the most highly activated (winning) evolving layer neuron are adjusted to be closer in space to the current training vector. Weight adjustment in the evolving to output layer is done via a simple error minimization algorithm. The details of the SECoS ANN and its learning algorithm are presented in the following section.

C. *The SECoS Algorithm*

The activation *A* of an evolving layer neuron *n* is determined by Equation 1.

$$A_n = 1 - D_n \quad (1)$$

where:
A_n is the activation of the neuron *n*, and D_n is the distance between the input vector and the incoming weight vector for that neuron.
Since SECoS networks are fully connected, the number of connections coming into an evolving layer neuron from the input layer is the same as the number of input neurons. Thus, the incoming weight vector has the same dimensionality as the vector input to the evolving layer. It is therefore possible to directly measure the distance in Euclidean space between the two vectors. Although the distance can be measured in any way that is appropriate for the inputs, this distance function must return a value in the range of $[0,1]$. For this reason, the SECoS algorithm assumes that the input data will be normalised, as it is far easier to formulate a distance function that produces output in the desired range if it is normalised to the range $[0,1]$.
Thus, examples which exactly match the exemplar stored within the neurons incoming weights will result in an activation of unity, while examples that are entirely outside of the exemplars region of input space will result in an activation of near zero.
Whereas most ANN propagate the activation of each neuron from one layer to the next, in SECoS only the activation of the winning (most highly activated) evolving layer neuron is propagated to the following neuron layers.

D. *SECoS Training*

The SECoS Learning algorithm is based on accommodating within the evolving layer new training examples, by either modifying the weight values of the connections attached to the evolving layer neurons, or by adding a new neuron to that layer. The algorithm employed is described below:

- Propagate the input vector *I* through the network.
- Find the most highly activated (winning) neuron *j* and its activation A_J.
- IF A_J is less than the sensitivity threshold S_{thr} ..
    - Add a neuron.
- ELSE
    - Evaluate the errors between the calculated output vector O_c and the desired output vector O_d.
    - IF the absolute error over the desired output is greater than the error threshold E_{thr}
        - Add a neuron.
    - ELSE
        - Update the connections to the winning evolving layer neuron.
- Repeat for each training vector.

When a neuron is added, its incoming connection weight vector is set to the input vector *I*, and its outgoing weight vector is set to the desired output vector O_d.
The weights of the connections from each input *i* to the winning neuron *j* are modified according to Equation 2.

$$W_{i,j}(t+1) = W_{i,j}(t) + \eta_1 I_i - W_{i,j}(t) \quad (2)$$

where:
$W_{i,j}(t)$ is the connection weight from input *i* to *j* at time *t*, η_1 is the learning rate parameter for the input to evolving layer connections, I_i is the *i*th component of the input vector *I*, The weights from neuron *j* to output *o* are modified according to Equation 3.

$$W_{j,o}(t+1) = W_{j,o}(t) + \eta_2 A_j E_o \quad (3)$$

where:
$W_{j,o}(t)$ is the connection weight from *j* to output *o* at time *t*, η_2 is the learning rate parameter for the evolving to output layer connections, A_j is the activation of *j*, E_o is the signed error at *o*, as measured according to Equation 4.

$$E_o = O_d - A_o \quad (4)$$

where:
O_d is the desired activation value of *o* and A_o is the actual activation of *o*.

E. Extracting Risk Weightings from SECoS

It is the spatial learning in the input to evolving layer connection weights that is most useful in the clustering of insect pest assemblages. At the completion of training over a data set, each evolving layer neuron represents the centre of a cluster of training examples, where the weights represent the coordinates of this centre.
The way in which a list of risk weightings is created from a trained SECoS is similar to the way in which one is created from a SOM: the vector representing the geographic region of interest is propagated through the network and the winning evolving layer neuron identified. The incoming connection weights are then taken as the species weights for risk of establishment.

III. Method

The data used was taken from the Crop Protection Compendium (CPC) [2] and consisted of representations of 459 geographic regions. Each example consisted of 844 elements, where each element indicated the presence or absence in that region of a particular insect pest species.
The SOMs were trained using the data described above. The SOM output maps had a twelve-by-nine neuron hexagonal topology and Euclidean distances were used [9]. Each SOM was trained for one thousand epochs. The training parameters were as in **Table 1**.

Table 1 Kohonen SOM learning parameters

Parameter	Value
Ordering phase learning rate	0.9
Tuning phase learning rate	0.02

Since SECoS is a supervised training algorithm, target values were needed for each of the species-presence vectors. For this work, the target values were unity minus the Jaccard similarity coefficient of each vector as compared to the New Zealand vector. Thus, vectors that were identical to the New Zealand vector (that is, geographic regions that have the same pest assemblage as New Zealand) had a target value of unity, while regions that were mostly different to New Zealand had a target value of close to zero. The New Zealand vector was omitted from the training set for the SECoS networks to ensure that a separate evolving layer neuron was not added for New Zealand. The training parameters used for SECoS are presented in **Table 2**.

Table 2 SECoS learning parameters

Parameter	Value
Sensitivity threshold	0.5
Error threshold	0.1
η_1	0.5
η_2	0.5

One thousand runs were carried out for both experiments. For each trained network created, a list of risk weightings for each species (a risk listing) was generated where each risk weight is a representation of the risk of establishment of the pest species.

IV. Method

The species in a particular risk list were ranked according to the mean risk weightings determined by the ANN. While each SOM had 108 neurons in the output map, the SECoS had an average of forty-two neurons in the evolving layer, which corresponds to forty-two clusters formed in the SECoS evolving layer. The mean performance of the SECoS over the learning data set was a mean-squared error (MSE) of 0.0018 with a standard deviation of 4.2×10^{-4}.

For the pest risk lists produced by each method, the highest ranks in the lists were overwhelmingly occupied by species that were already established in New Zealand, while the bottom ranks in the lists were overwhelmingly occupied by species that are recorded as absent. **Table 3** presents a list of the top 20 ranked species for SOM, while **Table 4** presents a list of the top twenty for SECoS. While the exact weightings and orderings of the species were different, fifteen of the species were common to both lists, while eight of those were in the top ten positions of both lists. This indicates that both SOM and SECoS were learning to cluster similar assemblages together. Of the species that were present in both lists, only one, *Sitophilus zeamais* is not recorded as being present in New Zealand.

Table 3 List of top twenty ranked species for SOM

Bayer Code	Species	Weight	Bayer Code	Species	Weight
MYZUPE	Myzus persicae	0.913	PSECAD	Pseudococcus longispinus	0.779
BRVCBR	Brevicoryne brassicae	0.886	HYLEPL	Delia platura	0.777
PSECCI	Planococcus citri	0.859	COCCHE	Coccus hesperidum	0.773
APHIGO	Aphis gossypii	0.856	SAISHE	Saissetia coffeae	0.772
NEZAVI	Nezara viridula	0.843	PHTOOP	Phthorimaea operculella	0.767
ERISLA	Eriosoma lanigerum	0.825	AONDAU	Aonidiella aurantii	0.766
PLUTMA	Plutella xylostella	0.822	SAISOL	Saissetia oleae	0.756
RHOPMA	Rhopalosiphum maidis	0.799	THRITB	Thrips tabaci	0.754
ICERPU	Icerya purchasi	0.794	RHOPPA	Rhopalosiphum padi	0.744
HELIAR	Helicoverpa armigera	0.792	APHICR	Aphis craccivora	0.744

Table 4 List of top twenty ranked species for SECoS

Bayer Code	Species	Weight	Bayer Code	Species	Weight
MYZUPE	Myzus persicae	1	PHTOOP	Phthorimaea operculella	1
BRVCBR	Brevicoryne brassicae	1	THRITB	Thrips tabaci	1
APHIGO	Aphis gossypii	1	SAISOL	Saissetia oleae	1
PSECCI	Planococcus citri	1	RHOPPA	Rhopalosiphum padi	1
NEZAVI	Nezara viridula	1	TOXOCI	Toxoptera citricida	1
ERISLA	Eriosoma lanigerum	1	HELTHA	Heliothrips haemorrhoidalis	1
PLUTMA	Plutella xylostella	1	PANMCE	Pantomorus cervinus	1
RHOPMA	Rhopalosiphum maidis	1	LISTCO	Listroderes costirostris	1
PSECAD	Pseudococcus longispinus	1	AONDAU	Aonidiella aurantii	0.999
SAISHE	Saissetia coffeae	1	CALAZM	Sitophilus zeamais	0.999

The twenty highest ranked species that are recorded as absent from New Zealand as identified by the Kohonen SOM are presented in Table 5, while those identified by SECoS are presented in Table 6. It is interesting to note that *S. zeamis*, which was the only species in the top twenty not already present in New Zealand, was at the top of both lists. Ten of the species were common to both lists, of which seven were in the top ten of both lists. Again, this supports the interpretation that both SOM and SECoS were learning to cluster similar assemblages together.

Table 5 List of top twenty species, that are recorded as absent from New Zealand, for SOM

Bayer Code	Species	Weight	Bayer Code	Species	Weight
CALAZM	Sitophilus zeamais	0.562	APHIFA	Aphis fabae	0.337
DROSME	Drosophila melanogaster	0.49	PHYNCI	Phyllocnistis citrella	0.333
MELHSA	Melanaphis sacchari	0.469	AGROSE	Agrotis segetum	0.324
LAPHEG	Spodoptera exigua	0.465	LEPSBE	Lepidosaphes beckii	0.314
CERTCA	Ceratitis capitata	0.445	TAYLPA	Taylorilygus pallidulus	0.301
CHYSFI	Chrysomphalus aonidum	0.425	PSEAPE	Pseudaulacaspis pentagona	0.300
FRANSC	Frankliniella schultzei	0.412	TRIPNI	Trichoplusia ni	0.300
AONDCI	Aonidiella citrina	0.403	OTIOCR	Otiorhynchus cribricollis	0.284
UNASCI	Unaspis citri	0.369	PECTGO	Pectinophora gossypiella	0.282
CHYSDI	Chrysomphalus dictyospermi	0.352	LEPSUL	Lepidosaphes ulmi	0.278

V. Discussion

A difficulty with this work is validating the results. It is encouraging that the top positions in the list are mostly filled by species that are present in New Zealand, while the bottom slots are filled with species that are recorded as absent.

The interpretation of the weights also requires some caution. Risk weightings are not the same as probabilities. If a species is established in a region, its probability of establishment is unity, yet the ANN may assign a weight to that species that is not unity.

Despite this, the results were encouraging, as the top positions in the risk lists were mostly filled by species that are present in New Zealand, while the bottom slots were filled with species that are recorded as absent. However it is clear that it is difficult to say which of the two ANN models used performed best. The advantage of SOM is that it is easy to visualise the clusters that are formed. The disadvantages are that training is extremely slow and the number of clusters is fixed. The advantages of SECoS are the number of clusters is determined automatically during training and the speed with which training occurs. Methods also exist for extracting fuzzy rules from SECoS [8]. However, since there is no quantisation in SECOS, it is much more difficult to visualise the clusters that form.

Table 6 List of top twenty species, that are recorded as absent from New Zealand, for SECoS

Bayer Code	Species	Weight	Bayer Code	Species	Weight
CALAZM	Sitophilus zeamais	0.999	EPILVG	Epilachna vigintioctopunctata	0.562
MELHSA	Melanaphis sacchari	0.941	CHYSFI	Chrysomphalus aonidum	0.552
DROSME	Drosophila melanogaster	0.924	OTIOCR	Otiorhynchus cribricollis	0.460
FRANSC	Frankliniella schultzei	0.645	LEPSUL	Lepidosaphes ulmi	0.425
AONDCI	Aonidiella citrina	0.632	ALEDDE	Aleurodicus destructor	0.408
LAPHEG	Spodoptera exigua	0.586	COSPFL	Anomis flava	0.386
CERPRB	Ceroplastes rubens	0.585	HYALPR	Hyalopterus pruni	0.365
IPSXGR	Ips grandicollis	0.585	CARHHU	Carpophilus humeralis	0.326
PAPLDE	Papilio demoleus	0.585	THEROL	Theretra oldenlandiae	0.326
UNASCI	Unaspis citri	0.573	PERGMA	Peregrinus maidis	0.314

VI. Conclusion

Models for predicting the risk of insect pest species establishing in a particular geographic region are presented. The models comprise two types of ANN, Kohonen SOMs and a constructive network, ECoS. Both models were found to produce ranked lists of species that appear to be plausible and both gave similar clusters of assemblages at the higher rankings. The use of both methods for pest risk assessment has value as each method confirms the output of the other adding weight to an overall assessment for a particular species.

Acknowledgements

The authors wish to acknowledge the work of Muriel Gevrey who prepared the data used in this research. This study was funded by the Centre of Research Excellence, Bio-protection, at Lincoln University.

References

[1] Chon, T.S., Park, Y.S., Moon, K.H., and Cha, E.Y. Patternizing Communities by using an Artificial Neural Network. Ecological Modelling, 90, (1996) 69-78.

[2] Crop Protection Compendium - Global Module, 5th Edition. CAB International, Wallingford, UK (2003).

[3] Gevrey, M., Rimet, F., Park, Y.S., Giraudel, J.L., Ector, L., and Lek, S. Water Quality Assessment using Diatom Assemblages and Advanced Modelling Techniques. Freshwater Biology, 49, (2004) 208-220.

[4] Gevrey M., Worner S.P, Kasabov, N., Pitt, J. and Giraudel, J-L. Estimating Risk of Events Using SOM Models: A Case Study on Invasive Species Establishment. Journal of Ecological Modelling. Accepted (2006)

[5] Kasabov, N. The ECOS Framework and the ECO Learning Method for Evolving Connectionist Systems. Journal of Advanced Computational Intelligence, 2(6) (1998) 195-202.

[6] Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43, (1982) 59-69.

[7] Watts, M.J. and Kasabov, N. Simple Evolving Connectionist Systems and Experiments on Isolated Phoneme Recognition. In Proceedings of the First IEEE Conference on Combinations of Evolutionary Computation and Neural Networks, San Antonio, May 2000 (IEEE Press, 2000) 232-239.

[8] Watts, M.J. Fuzzy Rule Extraction from Simple Evolving Connectionist Systems. International Journal of Computational Intelligence and Applications. 4(3) (2004) 1-10.

[9] Worner, S.P. and Gevrey, M. Modelling Global Insect Pest Species Assemblages to Determine Risk of Invasion: a New Zealand Example. Journal of Applied Ecology (Submitted 2005)

Dr Michael J. Watts is a postdoctoral fellow in the National Centre for Advanced Bio-Protection Technologies at Lincoln University, New Zealand. He earned the degree of Bachelor of Science with First Class Honours in Information Science from the University of Otago in 1996. Dr Watts started work as a teaching fellow in the Department of Information Science, University of Otago, in February 2000, and was promoted to the grade of senior teaching fellow in 2002. He gained his PhD degree in 2004, on the topic of the characterisation, simplification, formalisation, explanation and optimisation of evolving connectionist systems. He commenced his postdoctoral work in October 2004. His research interests include ecological modelling, bioinformatics, artificial neural networks, evolutionary computation and knowledge discovery.

Dr Susan P. Worner is a senior lecturer and researcher at Lincoln University, New Zealand. Her experience is in ecological data analysis and modelling, particularly the prediction of insect population timing and abundance Other research has involved the analysis and modelling of climatic influences on invasive insect populations to predict potential distribution and abundance. Dr Worner's recent research interests have extended to the use of geostatistics to model insect dispersion for spatial analysis, but particularly the application of new developments in artificial neural networks and machine learning to modelling and predicting ecological data. Dr Worner is a project leader in the National Centre for Advanced Bio-Protection Technologies, a Centre of Research Excellence (CoRE) hosted by Lincoln University. Dr Worner's research within the CoRE, involves the application of neurocomputing to the development of intelligent systems for the prediction and detection of pest invasions.