Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Using MLP to Determine Abiotic Factors Influencing the Establishment of Insect Pest Species

Michael J. Watts, *Member, IEEE*, and S. P. Worner

***Abstract*— The use of multi-layer perceptrons (MLP) to determine the significance of climatic variables to the establishment of insect pest species is described. Results show that the MLP are able to learn to accurately predict the establishment of a pest species within a specific geographic region. Analysis of the MLP yielded insights into the contribution of the individual input variables and allowed for the identification of those variables that were most significant in either encouraging or inhibiting establishment.**

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. There is therefore a desperate need to develop methods that have a higher level of prediction to assist the pest risk assessment process and that are able to mine the vast quantities of data in existence for useful information. This mining includes assessment of the risk of establishment of exotic species and analysis of the available data with respect to the significance of the numerous individual variables present. The factors affecting the establishment within a geographical region of a particular species can be divided into two general groups: firstly, biotic factors, which includes the presence of food or host species, as well as competing species; and abiotic factors, which is essentially the climate of the region in question.

A number of models and approaches have been designed to predict the establishment of species in regions where they are not normally found. 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.

Artificial neural networks (ANN) have previously been used for many applications in ecology [7], [4], including modelling the relationship between cities and the levels of contaminants in grasses [2] and the presence of certain species of freshwater fish [5]. ANN have also been used to detect significant features in ecological data [6].

A large amount of data exists that describes many features of the climate in numerous geographical locations, as well as the presence or absence of numerous species of insect pests. It is desirable to be able to identify the abiotic factors that affect the establishment of particular species, so that the threat they pose to the agriculture and biodiversity of various regions can be more accurately assessed.

The goal of the research reported here was to investigate the use of ANN, specifically MLP, in identifying abiotic factors that are important for the establishment of insect pest species.

II. Method

A. Data

The complete data set used in these experiments consisted of data about the climate in each of 459 geographic regions and the presence in or absence from each of those regions of 844 insect pest species. The species presence data were sourced from the CABI Crop Protection Compendium [1]. The climate data were compiled from Internet sites maintained by recognised meteorological organisations.

These species were divided into two groups, those species that are recorded as being present in New Zealand, and those that are not. Each of these sets were ordered according to the threat posed by the species, according to the method described in [11], [12]. The top two species of each set were selected as case studies for this paper and are listed in Table I. The column in Table I labelled "Prevalence" lists the number of geographic regions in which each particular species is recorded as being present.

TABLE I
Target Species.

Name	Prevalence
Myzus persicae	234
Brevicoryne brassicae	210
Sitophilus zeamais	127
Drosophila melanogaster	109

M. persicae is commonly known as the green peach aphid, while *B. brassicae* is the cabbage aphid. The common name of *S. zeamais* is the maize weevil, and *D. melanogaster* is a fruit fly.

A total of forty five climate variables were available. For each of these variables, the maximum, mean and minimum were used as separate inputs. This is because a single geographical region had to be represented as a single input vector, while including the maximum and minimums was necessary for describing the range of the climate variables within that region. This yielded a total of one hundred and thirty five input variables. The data for each of these variables was linearly normalised to the range of zero to unity. The output targets were the presence or absence of the target species in each region.

The data was split into two major sets. The first, containing 80% of the data, was the training and test set, from which

National Centre for Advanced Bio-Protection Technologies, PO Box 84, Lincoln University, Canterbury, New Zealand (phone: 64-3-325-3696; fax: 64-3-325-3864; email: wattsm2,worner@lincoln.ac.nz).

0-7803-9490-9/06/$20.00/©2006 IEEE

samples were randomly drawn to form training and test data sets. The second was the validation set, which was used only once for each species.

B. Training and Evaluation of MLP

Standard three neuron-layer MLP were used in these experiments, and the learning algorithm used was unmodified back-propagation with momentum. The parameters of the MLP and learning algorithm are presented in Table II. These parameters were found via experimentation to yield the best balance of training and generalisation errors.

TABLE II

TRAINING PARAMETERS

Parameter	Value
Hidden neurons	3
Learning rate	0.03
Momentum	0.03
Epochs	750

The method of training and evaluating the MLP (and also selecting the parameters) was similar to that suggested in [3], [10]. A total of one thousand runs were performed over each species. For each run, the training and test data set was randomly divided into a training set, consisting of two-thirds of the available data, and a test set consisting of the remaining one-third. A MLP was then created with randomly initialised connection weights and trained over the training data set. The accuracy of the MLP over the training set was then evaluated to determine how well the network had learned the training data. The accuracy of the MLP was then evaluated over the testing data set to determine how well the network generalised. Accuracy was measured as both the percentage of examples correctly classified and using Cohen's Kappa statistic. Whereas percentage accuracy is easily interpreted, it is also easily biased by unbalanced numbers of classes. That is, percentage correct may be misleadingly high when the data set in question has only a small number of examples from one class. The Kappa statistic takes the number of examples of each class into account and thus yields a less biased measure of accuracy than percentages.

For each run the contributions of each input neuron to the output of the network was also determined, using the method of Olden and Jackson as described in [8]. This method has been experimentally determined to give the least-biased estimate of the contribution of each input neuron [9] and has been used previously in ecological modelling applications [5].

At the completion of the one thousand runs, the MLP with the highest kappa over the test set was selected as the winner for that species. The accuracy of this winning network was then evaluated over the validation data set. A sensitivity analysis was also performed over each input variable of the winning network. This was to illustrate the response of the network to variations in each variable so that the influence of strongly contributing inputs (as determined above) could be investigated.

III. RESULTS

A. Accuracies

The accuracies for each species are presented in Table III. For the accuracies over the training and test sets, the results are presented as the mean and standard deviation of both the overall percentage and the Kappa.

TABLE III

ACCURACIES

Species		Train	Test	Validate
M. persicae	%	84.52/2.32	74.72/3.64	68.48
	κ	0.69/0.05	0.49/0.07	0.37
B. brassicae	%	84.30/2.20	75.01/3.45	67.39
	κ	0.68/0.05	0.49/0.07	0.36
S. zeamais	%	81.53/6.42	74.54/3.95	78.26
	κ	0.44/0.28	0.26/0.18	0.40
D. melanogaster	%	82.56/2.73	73.65/3.54	75.00
	κ	0.44/0.14	0.18/0.10	0.22

These results show that the MLP were able to imperfectly learn the relationships between the climate variables and the presence of the target species. The low testing and validation kappas over *D. melanogaster* suggests that a degree of over-training occurred, although the lower training accuracies for both *S. zeamais* and *D. melanogaster* indicates that the less prevalent species were harder for the MLP to learn. The low Kappa values for these two species, compared to their higher percentage accuracies, suggests that a large number of false negative predictions were made.

B. Contribution of Input Variables

The three inputs that positively contributed the most to the networks are listed for each species in Table IV. The mean and standard deviation of the contributions are listed, along with the name of each input variable.

TABLE IV

MOST POSITIVELY CONTRIBUTING INPUTS

Species	Variable Name	Contribution
M. persicae	Max RSprr1	18.755/5.763
	Mean TAut1	9.85/3.520
	Mean TAut2	9.42/3.379
B. brassicae	Max RSprr1	9.986/5.353
	Max TAut1	9.666/3.306
	Max TAut2	9.279/3.163
S. zeamais	Max RWinr2	8.906/4.956
	Max RWinr1	7.779/4.497
	Max Im300	5.569/3.644
D. melanogaster	Max AnnualDayLength	7.350/3.456
	Max Im300	7.188/3.235
	Max Mi	7.123/3.143

The variable "RSprr1" is the rainfall during the first month of spring. "TAut1" and "TAut2' are the temperatures during the first and second month of autumn, respectively. "RWinr1" and "RWinr2" variables describe the rainfall during the first and second month of winter, while "Im300" is the moisture index of the soil at a depth of 300mm. "AnnualDayLength" is the length of the day from sunrise to sunset. The variable "Mi" is a moisture index. The fact that both *M. persicae* and

TABLE V

COMMON VARIABLES FROM TOP TEN RANKED VARIABLES FOR *M. persicae* AND *B. brassicae*

	Contribution	
Species	*M. persicae*	*B. brassicae*
Max RSprr1	18.755/5.763	9.986/5.323
Max TAut2	8.840/3.755	9.280/3.123
Mean RSprr1	8.340/3.403	6.552/3.113
Max TAut1	8.296/3.726	9.666/3.306
MaxDD5	7.704/3.682	7.831/3.405

B. brassicae had the same variables most highly ranked is informative, although perhaps unsurprising since both species are types of aphids. Further investigation of the similarities between the input rankings of *M. persicae* and *B. brassicae* revealed that five of the top ten ranked variables were the same for both species: these variables are listed in Table V, where "DD5" is the "degree days" variable, a measure of the temperature accumulation above a 5 degrees Celsius threshold across a certain time period.

That two of the variables are spring rainfalls and another two are autumn temperatures suggests a link to the aphid life cycles. During spring the aphids hatch from over-wintering eggs and their host plants undergo a spring flush. High rainfall during that time could potentially improve the development of the host plants and thus provide more bountiful food supplies for the aphids. During autumn the aphids lay their eggs, which hibernate over winter and hatch in spring. Higher temperatures in autumn could potentially allow the aphids to survive to migrate back to their primary over-wintering hosts.

The inputs that negatively contributed the most to the networks are listed for each species in Table VI. The mean and standard deviation of the contributions are listed, along with the name of each input variable.

TABLE VI

MOST NEGATIVELY CONTRIBUTING INPUTS

Species	Variable	Contribution
M. persicae	Min DD15	-18.098/5.018
	Min PEannual	-11.002/3.881
	Min Climate values	-9.686/4.971
B. brassicae	Min RAutr3	-13.276/3.985
	Min RSprr3	-11.766/4.577
	Max RSprr2	-8.368/4.858
S. zeamais	Mean RAutr3	-9.182/4.534
	Max AEannual	-6.845/4.308
	Min DD15	-6.820/3.817
D. melanogaster	Min Climate Values	-7.255/3.651
	Max AEannual	-7.143/3.938
	Max RSprr3	-5.268/3.801

The variable "PEannual" is the potential evapotranspiration while "AEannual" is the actual evapotranspiration. "RAutr3" is the rainfall during the third month of autumn. "Climate Values" are the most extreme minimum of all climate values.

The final analysis of the input contributions involved examining the top ten inputs for each species and the bottom ten inputs for every other species. The goal of this was to identify input variables that contribute positively to the establishment of one species, but negatively to another. This analysis showed that the variable "Max RSprr3" contributed positively to *M. persicae* and negatively to *D. melanogaster*, and that the variable "Mean Climate values" contributed positively to *B. brassicae* but negatively to *S. zeamais*. These results are summarised in Table VII, where the mean and standard deviation of the contributions are presented.

TABLE VII

OPPOSING VARIABLES

Variable	Species	Contribution
Max RSprr3	*M. persicae*	8.827/4.946
	D. melanogaster	-5.267/3.801
Mean Climate values	*B. brassicae*	5.316/4.275
	S. zeamais	-6.116/4.732

C. Sensitivity Analysis

Sensitivity analysis is a way of visualising the response of an ANN to variations of a single variable. To perform a sensitivity analysis over variable n, all other input variables are set to their mean values, while the values of n are varied across the range of n, and the output of the ANN recorded.

The advantage of a sensitivity analysis is that it allows for a more detailed investigation of the importance of a particular variable. Whereas an analysis of the importance of each input will yield a single overall value for the contribution of each input, a sensitivity analysis shows how the network reacts to that variable across its range. When plotted for positively and negatively contributing climate variables, sensitivity analysis allows for the visualisation of the probability of a species establishing as the climate varies.

In this subsection, the results of a sensitivity analysis of the most positively and negatively contributing variables are plotted.

Figures 1 and 2 display the results for *M. persicae* and *B. brassicae*. In Figure 1 it can be seen that the output of the network increases rapidly with respect to increases in the "Max RSprr1" variable, and decreases even more rapidly with respect to increases in the "Min DD15" variable. The same effect, with respect to "Max RSprr1", is seen in Figure 2, while the effect of the "Min RAutr3" variable is less pronounced. Referring to the values in Table IV and VI shows that the magnitude of the contribution of the "Min RAutr3" variable is significantly less than that of the "Min DD15" variable for *M. persicae*. This supports the conclusion that the magnitude of the calculated input contribution is truly indicative of the importance of the input variable.

The results of the sensitivity analysis of *S. zeamais* and *D. melanogaster* are presented in Figures 3 and 4, respectively. Again, the plots show the steady increase and decrease of output values as the positively and negatively contributing variables are increased. The curves for the output responses for the negatively contributing variables are substantially flatter for *S. zeamais* and *D. melanogaster* than they are

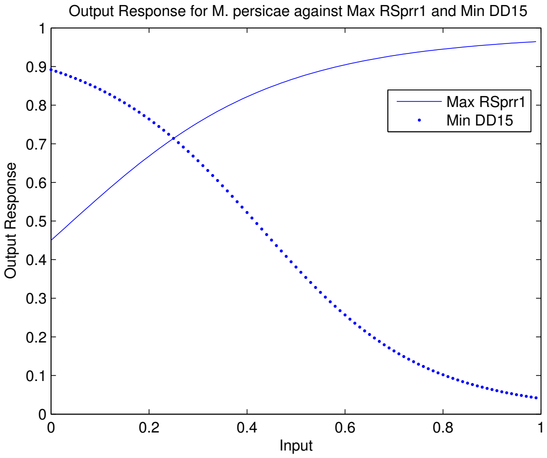

Fig. 1. Output response of *M. persicae* network against Max RSprr1 and Min D15 variables

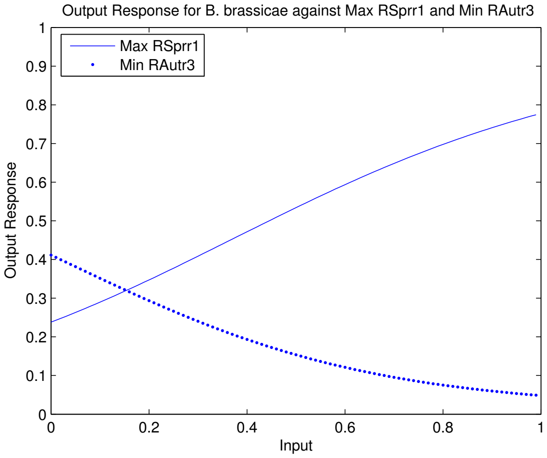

Fig. 2. Output response of *B. brassicae* network against Max RSprr1 and Min RAutr3 variables

for *M. persicae* and *B. brassicae*. It seems that for these species, the high number of false negative predictions was the result of a lower overall activation of the output neuron, which also depresses the curves for the sensitivity analyses of the negatively contributing variables. The curves for the positively contributing variables are lower than for *M. persicae* and *B. brassicae*, but still show an upward swing as the values of the relevant variables are increased.

In Table VII two variables are identified that have contradictory effects for specific pairs of established and unestablished species. Figure 5 presents the results of the sensitivity analysis over these variables for the species *M. persicae* and *D. melanogaster*, while Figure 6 presents the results for the species *B. brassicae* and *S. zeamais*.

The curves in Figure 5 are unique in that the curves do not cross. That is, at all points the activation for the *M. persicae* is greater than the activation for *D. melanogaster*. This means that the probability of *M. persicae* establishing is always greater than the probability of *D. melanogaster* establishing,

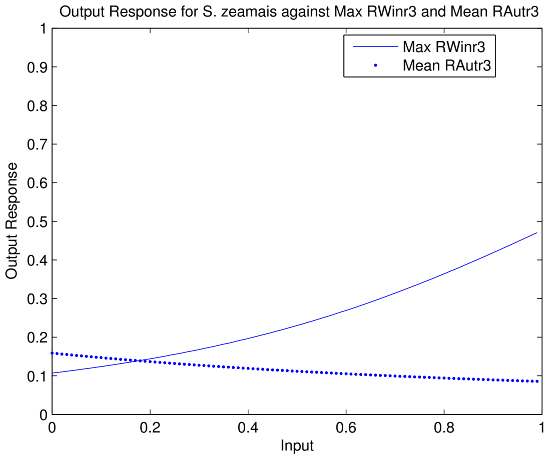

Fig. 3. Output response of *S. zeamais* network against Max RWinr3 and Mean RAutr3 variables

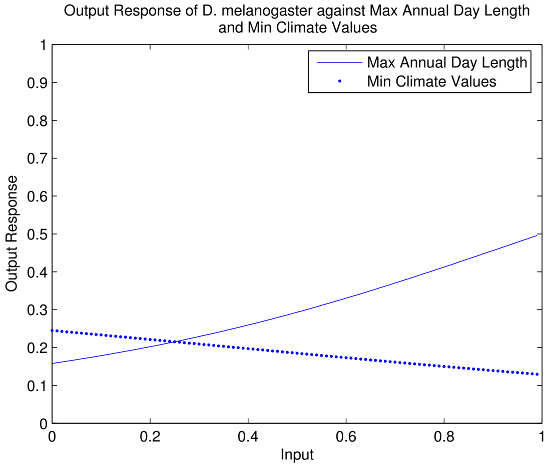

Fig. 4. Output response of *D. melanogaster* network against Max Annual Day Length and Min Climate Values variables

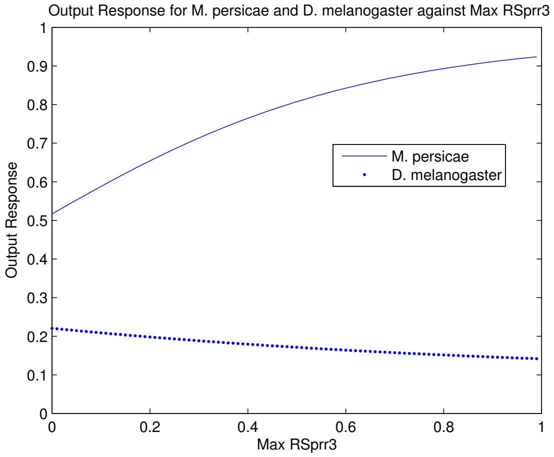

Fig. 5. Output responses of *M. persicae* and *D. melanogaster* networks against Max RSprr3 variable

no matter the value of the "Max RSprr" variable. However, a depressive effect for *D. melanogaster* is shown by the plot.

The curves in Figure 6 show that at extremely low values of the target variable the probability of *S. zeamais* establishing is higher than that of *B. brassicae*, with the probability dropping well below *B. brassicae* as the values increase.

Again, the "flatness" of the curves for the negatively contributing variables seems to be related to the magnitude of the contribution, as listed in Table VII.

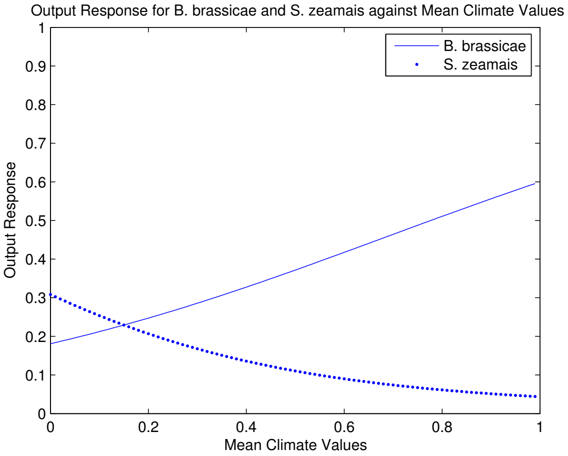

Fig. 6. Output responses of *B. brassicae* and *S. zeamais* networks against Mean Climate Values variable

IV. Discussion

As is the case with many ecological data sets, the data used in this study is likely to be very noisy. For example, while the climate in a particular region may be conducive to the establishment of a species, the species may never have gained access and therefore not established in the region. Alternatively, while a species may be listed as being absent from a particular geographic region, this may be because it has never been officially recorded in that region, as opposed to being truly absent. Conversely, a species may be falsely recorded as being present in a region due to misidentification of a specimen. There is nothing that can really be done about this, other than to admit that it may be detrimental to the performance of the ANN.

While use of the maximum, minimum and mean of the climate variables provides useful information, in terms of providing the range of the variables for a region, there is a high degree of correlation between the mean and the other two statistics. There is also likely to be correlation between the climate variables themselves. This could be reduced by performing a principal components analysis (PCA) over the data and using only the top few principal components. However, the issue of identifying the contribution of the original variables during the analysis of the networks would have to first be resolved.

An investigation of the biology of the target species would be of great use in interpreting the results of this study. For example, both *M. persicae* and *B. brassicae* are aphids, and both are positively affected by the same variable. On the other hand, *S. zeamais* is a weevil and *D. melanogaster* is a fruit fly, and there is less similarity between the importance of the variables for these species. Whether the similarities and differences are due to fundamental differences in the biology of the species, or whether they are simply artifacts of noisy data, is a question that can only be answered by a study of the biology of the species involved. The work reported in this paper, however, does at the very least, provide a starting point for proposing appropriate questions to be investigated.

V. Conclusion

The paper has presented an investigation into the use of MLP in determining the importance of different climatic variables to the establishment of several species of insect pests. The results show that the MLP are able to learn the relationship between the climate within a geographic region and the establishment of pest species. Analysis of the trained MLP was able to identify those input variables that contribute the most to both encouraging and discouraging establishment. While the most important features were generally idiosyncratic to each species, there were also some similarities between species of the same taxon.

Problems encountered were the high number of false negative predictions made by the MLP for some species. This is related to the amount of noise that is inherent in ecological data, the correlations between the input variables and the unbalanced nature of the data, that is, the relatively smaller number of positive examples that exist for some species.

Future work will focus on alternative training methods, a more detailed analysis of the biological effect of the climate variables identified as important, and data processing methods such as PCA that will reduce the amount of correlation between variables.

Acknowledgements

The authors wish to acknowledge the work of Muriel Gevrey and Joel Pitt 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] Crop Protection Compendium - Global Module, 5th Edition. ©CAB International, Wallingford, UK 2003.

[2] Dimopulos, I., Chronopoulos, J., Chronopoulou-Sereli, A., and Lek, S. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece). Ecological Modelling 120:157-165. 1999.

[3] Flexer, A. Statistical Evaluation of Neural Network Experiments: Minimum Requirements and Current Practice. In: Trappl, R., Cybernetics and Systems '96, Proceedings of the 13th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, 1005-1008. 1996.

[4] Gevrey, M. and Worner, S.P. Prediction of global distribution of insect pest species in relation to climate using an ecological informatics method. Environmental Entomology (Accepted). 2006.

[5] Joy, M.K. and Death, R.G. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49:1036-1052. 2004.

[6] Kimes, D.S., Nelson, R.F. and Fifer, S.T. Predicting ecologically important vegetation variables from remotely sensed optical/radar data using neural networks. In: Artificial Neuronal Networks: Application to Ecology and Evolution. S. Lek and J-F. Guegan, eds. 2000.

[7] Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. and Aulagnier, S. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90:39-52. 1996.

[8] Olden, J.D. and Jackson, D.A. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154:135-150. 2002.

[9] Olden, J.D., Joy, M.K. and Death, R.G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178:389-397. 2004.

[10] Prechelt, L. A Quantitative Study of Experimental Evaluations of Neural Network Learning Algorithms: Current Research Practice. Neural Networks 9(3) 457-462. 1996.

[11] Watts, M.J. and Worner, S.P. Comparison of Artificial Neural Networks Models for Predicting Insect Pest Establishment. In: Proceedings of 2005 International Conference on Intelligent Computing (ICIC 2005), Hefei, PRC 520-529. 2005.

[12] Worner, S.P. and Gevrey, M. Modelling global insect pest species assemblages to determine risk of invasion. Journal of Applied Ecology. (Accepted) 2006.