Article

Using artificial neural networks to predict the distribution of bacterial crop diseases from biotic and abiotic factors

Michael J. Watts[1], Susan P. Worner[2]

[1]School of Earth and Environmental Sciences, University of Adelaide, North Terrace, SA 5005, Australia

[2]Bio-Protection Research Centre, PO Box 84, Lincoln University, Lincoln 7647, New Zealand

E-mail: mjwatts@ieee.org

Received 6 November 2011; *Accepted 10 December 2011*; *Published online 1 March 2012*

IAEES

Abstract

Constructing accurate computational global distribution models is an important first step towards the understanding of bacterial crop diseases and can lead to insights into the biology of disease-causing bacteria species. We constructed artificial neural network models of the geographic distribution of six bacterial diseases of crop plants. These ANN modelled the distribution of these species from regional climatic factors and from regional assemblages of host crop plants. Multiple ANN were combined into ensembles using statistical methods. Tandem ANN, where an ANN combined the outputs of individual ANN, were also investigated. We found that for all but one species, superior accuracies were attained by methods that combined biotic and abiotic factors. These combinations were produced by both ensemble and cascaded ANN. This shows that firstly, ANN are able to model the geographic distribution of bacterial crop diseases, and secondly, that combining abiotic and biotic factors is necessary to achieve high modelling accuracies. The work reported in this paper therefore provides a basis for constructing models of the distribution of bacterial crop diseases.

Keywords artificial neural network; biotic and abiotic factors; bacterial crop disease; prediction.

1 Introduction

The rising rate of global trade is rapidly increasing the threat to the agricultural and horticultural production of many countries by unintended introductions of exotic crop diseases, including bacterial diseases. There is therefore a pressing need to develop methods that have a higher level of prediction accuracy to assist the risk assessment process.

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 crop plants and species of bacterial crop diseases. It is desirable to be able to accurately predict their distribution, so that the threat they pose to the agriculture and biodiversity of various regions can be more accurately assessed.

The factors affecting the establishment within a geographical region of a particular bacterial species can be divided into two general groups: firstly, biotic factors, which include the presence of potential host 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 bacterial crop diseases in regions where they are not normally found.

Artificial neural networks (ANN) have previously been used for many applications in ecology (Lek et al, 1996; Gevrey and Worner, 2006; Zhang and Wei, 2009; Zhang, 2010, 2011), including modelling the relationship between cities and the levels of contaminants in grasses (Dimopulos et al, 1999) and the presence of certain species of freshwater fish (Joy and Death, 2004). However, no work has yet come to light that predicted the distribution of bacterial crop diseases from both biotic and abiotic factors. The goal of the research reported here was to investigate the use of ANN, specifically multi-layer perceptrons (MLP), to model the distribution of several bacterial crop diseases by predicting their presence and absence in world-wide geographical regions. These predictions were made in three ways: Firstly, from climate (abiotic) factors; Secondly, from biotic factors, in the form of regional host plant assemblages; Thirdly, by using ensembles of MLP and a cascaded or tandem MLP architecture that combined predictions made from both climate and host plant assemblages.

2 Method

2.1 Data

The data set used in this study consisted of data describing the climate for each of 459 worldwide geographic regions , the presence or absence in each of these regions of 114 host crop plants, and the presence or absence in each of those regions of 130 crop disease-causing bacteria species. The species presence data were sourced from the CABI Crop Protection Compendium (CABI, 2003). The climate data were compiled from Internet sites maintained by recognised meteorological organisations. None of this data contained any explicit information about the links between climate in a region and the presence or absence of any particular disease.

Forty-five climate variables were available and are listed in Table S1. Although data was available describing the climate on a monthly basis for each region, only seasonal data was included. This is because organisms follow a seasonal cycle, rather than a calendar cycle, and seasons are reversed between the Northern and Southern hemispheres. Any generalisations made about species establishment drawn from Southern hemisphere monthly data would not be applicable to the Northern hemisphere, and vice versa, even for the same species. To represent the range of each variable within a region the minimum, mean and maximum of each variable was calculated. There was therefore a total of one hundred and thirty five input variables describing the climate in each region. The data for each variable was linearly normalised to the range of zero to unity and comprised the input to the MLP models.

The second set of data was the presence and absence of 114 host plants in the same geographic regions as above. Only those hosts that were recorded as being present in more than 5% of the regions were retained in the data set.

To verify that the species assemblages were non-random, both the bacterial crop disease and host plant assemblages were subjected to a null model analysis (Gotelli, 2000). The software used for this analysis was EcoSim version 7 (Gotelli and Entsminger, 2006). Ten thousand iterations were performed and the C-score, V-ratio, number of checker boards and number of species combinations were all evaluated. The default setting of retaining degenerate matrices was retained. The results of each run showed that both assemblages were significantly non-random (p=0.001).

Six target species were selected, all of which were ranked as significant threats according to the method described in (Worner and Gevrey, 2006; Watts and Worner, 2009; Watts, 2011). Although this technique was

originally developed to estimate the risk of invasion of insect pest species, there appears to be no reason why it should not be applicable to other threatening organisms. The six species identified are presented in Table S2.

The data was randomly split into two major sets. The first, containing 80% of the data, was the training and test set, from which samples were randomly drawn to form training and test data sets for each trial. The second comprised the independent validation set, which was used to perform an independent evaluation of the prediction accuracy for each target species.

The division was done on the basis of regions, that is, the 459 regions were randomly assigned into two groups in an 80%/20% ratio, and the corresponding climate data and host assemblages found. Thus, the validation set contained data corresponding to the same regions for both the climate and host assemblages.

2.2 Training and evaluation of MLP

Standard three neuron-layer multi-layer perceptrons (MLP) were used in these experiments, and the learning algorithm used was unmodified back-propagation with momentum (Rumelhart et al, 1986). The parameters of the MLP and learning algorithm were three hidden neurons, a learning rate and momentum of 0.03 and either 500 or 750 training epochs. Table S3 lists the epochs selected for each MLP used to model each species, for each data set. These parameters were found via experimentation to yield the best balance of training and generalisation errors: more than three hidden neurons consistently caused over-training, while less meant the MLP were unable to learn. Similar parameters were previously found to be effective for modelling the establishment of insect pest species (Watts and Worner, 2008). While the target data in this work was different (presence and absence of bacterial diseases, rather than insect pests) the basic problem was the same: predicting the presence or absence of a species from the same set of regional environmental factors.

The method of training and evaluating the MLP (and also selecting the parameters) was similar to that suggested in Flexer (1996) and Prechelt (1996). A total of one thousand trials were performed over each species. For each trial, the training and test data set (consisting of 80% of the total data available) was randomly divided into a training set and a testing set in a two-thirds/one-third ratio. A MLP was then created with randomly initialised connection weights and trained over the training division. The accuracy of the MLP over the training division was then evaluated to determine how well the network had learned, and again over the testing division to determine how well the network generalised. Accuracy was measured using Cohen's Kappa statistic (Cohen, 1960), where a kappa of less than 0.2 is considered poor accuracy, 0.2 to 0.4 fair, 0.4 to 0.6 moderate, 0.6 to 0.8 good and over 0.8 very good, with 1 being perfect accuracy. Kappa was used because it is a simple and well-known statistic (Manel et al., 2001) that is not biased by different proportions of presences or absences, and gives results that are qualitatively similar to more complex measures such as area under the curve (Elith et al., 2006; Graham et al., 2008). At the completion of the one thousand trials, the MLP with the highest kappa over the test data (that is, the MLP with the best generalisation performance) was selected as the winner for that species. The accuracy of this winning network was then evaluated over the validation data set, so that an unbiased estimate of the generalisation capability of the MLP could be obtained.

2.3 Ensembles of artificial neural networks

Ensembles of ANN are a way of combining the predictions made by several ANN into a single prediction that incorporates the "knowledge" of all of the members of the ensemble (Battiti and Colla, 1994; Costa et al, 1996; Filippi et al., 1994; Hansen and Salamon, 1990; Maqsood, Khan and Abraham, 2004; Perone and Cooper, 1993; Sharkey, 1996; Sharkey and Sharkey, 1997). Ensemble methods statistically or algorithmically combine the outputs of several ANN and can yield improved performance. ANN ensembles can perform better than

single ANN because each member of the ensemble was trained on a different part of the problem space. Combining the predictions of several ANN means that a larger part of the problem space can be modelled.

Several ensemble methods were used in this work. First the final prediction of the ensemble was calculated as the minimum, maximum, mean and median, respectively, of all of the individual predictions within the ensemble. The majority vote method was also used. In this technique, the final prediction is taken to be positive if the majority of predictions within the ensemble are above 0.5 (which is taken here to be a positive prediction), and negative if vice versa.

Ensembles were constructed using the top ten networks for each species, that is, networks in the top 1% with respect to generalisation accuracy. The number 1% is fairly arbitrary, as any value up to 100% could have been selected. However, as the number of networks in the ensemble increases, the ensemble obviously becomes less efficient, as more networks require more computational power to combine and some of the networks would perform so poorly that they could jeopardise the performance of the entire ensemble.

Accuracy of the ensemble was assessed by resampling the data set. Only test sets were evaluated, as no additional training of the ensemble members was carried out, nor was it possible to train the ensembles. One thousand resamplings were carried out, and the accuracies used to select the optimal ensemble method for each species. The methods selected for each species are listed in Table S4.

Ensembles were constructed for climate and host assemblage networks. Ensembles were also constructed that combined climate and host predictions. These ensembles thus combined biotic and abiotic factors to predict establishment. The top ten most accurate ANN for climate and host assemblages were included in these ensembles and therefore there were twenty ANN in these “combination” ensembles.

2.4 Cascading neural networks

To produce the training data for the cascaded networks, the winning climate networks and the winning host assemblage networks for each species were selected. The relevant data for all regions were then propagated through each network. The outputs from these networks were used as the inputs for the cascaded networks, which were then trained to predict the presence or absence of the target disease species. The cascaded networks therefore combined predictions made from climate and host assemblages into one final prediction.

The training and testing procedure for the cascaded networks was the same as the climate and host networks above. A wider range of hidden-layer sizes were found to be useful for these networks, as shown in Table S3, while 500 training epochs and learning rate and momentum of 0.03 were found to be optimal.

For each trial of the cascaded networks the contributions of each input neuron to the output of the network was determined. While many methods have been proposed for determining the importance of each of the input neurons of an MLP, the work of Olden, Joy and Death (2004) has shown that the method of Olden and Jackson (2002) is the least biased, and it has also been previously used in ecological modelling applications (Joy and Death, 2004; Watts and Worner, 2008a, b).

A sensitivity analysis was also performed over each input variable of the winning cascaded network. This was carried out to illustrate the response of the network to variations of each variable so that the influence of strongly contributing inputs (as determined above) could be visualised. The sensitivity analysis was performed by setting each input, except the one being investigated, at its mean value for the data set. The values for the input being investigated were then varied across the range of zero to unity and the network recalled for each step.

3 Results

The accuracies for each species over each data set, as measured by Cohen's Kappa statistic (Cohen, 1960), are presented in Table 1. These results show that there was a wide range of mean accuracies over the testing sets, ranging from a low of 0.09 for *Spiroplasma kunkelii* to a high of 0.41 for *Rhizobium radiobacter*. Climate networks, however, had an even larger range of variation over the testing data set, ranging from a low of 34% of the mean for *R. radiobacter* to a high of 140% of the mean for *Xanthomonas campestris pv. campestris*. Accuracy over the independent validation data set, however, was generally higher, with a minimum validation accuracy of 0.32 for *R. radiobacter* and *S. kunkelii* and a maximum validation accuracy of 0.68 for *Xanthomonas axonopodis pv. dieffenbachiae*.

Table 1 Mean and standard deviation of accuracies (as Cohen's Kappa). "Train" is the accuracy over the training data sets. "Test" is the accuracy over the testing data sets. "Validate" is the accuracy over the validation data set. Rows labelled "Climate" present the results for the networks trained over the climate data set. "Hosts" present the accuracies of the networks trained over the host assemblage data set. "Climate+Host" presents the results of ensembles combining predictions made from climate and host assemblages. "Cascade" presents the accuracies of cascaded networks that combine predictions made from climate and host assemblages. "Original" rows are the accuracies over single networks. "Ensemble" rows present the accuracies over the ensemble of networks. The best validation accuracies for each species are shown in bold and underlined.

		R. radiobacter			*X. campestris pv. aberrans*		
		Train	Test	Validate	Train	Test	Validate
Climate	Original	0.57/0.16	0.41/0.14	0.32	0.38/0.23	0.19/0.13	0.34
	Ensemble		0.62/0.08	0.34		0.45/0.08	0.30
Hosts	Original	0.64/0.33	0.32/0.17	0.42	0.71/0.14	0.24/0.11	0.44
	Ensemble		0.77/0.06	0.47		0.70/0.06	**0.49**
Climate+Host			0.71/0.07	**0.61**		0.62/0.07	0.43
Cascade		0.70/0.03	0.71/0.07	0.49	0.60/0.05	0.59/0.07	**0.49**
		E. carotovora subsp.			*X. campestris pv. campestris*		
		Train	Test	Validate	Train	Test	Validate
Climate	Original	0.53/0.22	0.37/0.13	0.35	0.18/0.25	0.15/0.21	0.48
	Ensemble		0.53/0.08	0.27		0.69/0.15	0.48
Hosts	Original	0.66/0.37	0.24/0.17	0.52	0.0/0.04	0.0/0.05	0.56
	Ensemble		0.87/0.04	**0.55**		0.74/0.13	0.74
Climate+Host			0.72/0.07	0.47		0.65/0.13	**0.79**
Cascade		0.85/0.03	0.84/0.05	0.45	0.82/0.13	0.76/0.18	0.56
		S. kunkelii			*X. axonopodis pv. dieffenbachiae*		
		Train	Test	Validate	Train	Test	Validate
Climate	Original	0.12/0.16	0.09/0.10	0.32	0.42/0.26	0.28/0.18	0.68
	Ensemble		0.47/0.1	0.17		0.59/0.07	0.67
Hosts	Original	0.10/0.27	0.05/0.12	0.29	0.71/0.05	0.36/0.08	0.60
	Ensemble		0.91/0.05	**0.46**		0.68/0.06	0.55
Climate+Host			0.73/0.09	0.39		0.68/0.06	0.65
Cascade		0.80/0.04	0.78/0.07	0.47	0.71/0.03	0.70/0.06	**0.72**

Climate ensembles improved test accuracy for all species over climate. Climate ensemble variation ranged from 11.9% of the mean for *X. axonopodis pv. dieffenbachiae* to 21.7% of the mean for *X. campestris pv. campestris*, which is much less than the original networks, although this is to be expected since the ensembles used only the top ten networks. However, validation accuracy was only improved for *R. radiobacter*.

The accuracy of the host networks ranged from 0.0 for *X. campestris pv. campestris* to 0.36 for *X. axonopodis pv. dieffenbachiae*. The variation of accuracy of the host networks was also very large, ranging from a low of 22.2% for *X. axonopodis pv. dieffenbachiae* to a high of 240% for *S. kunkelii*. With the exception of *X. axonopodis pv. dieffenbachiae*, the mean testing accuracy of the host networks was less than that of the climate networks. The validation accuracy over the host networks ranged from 0.29 for *S. kunkelii* to 0.60 for *X. axonopodis pv. dieffenbachiae*. Host networks, however, had a higher validation accuracy than climate networks, with the exception of *S. kunkelii*. Again, this was not surprising, given the large degree of variation in accuracy for the host networks.

The testing accuracies over the host data sets were improved by the use of ensembles for all species, most notably for *X. campestris pv. campestris*, which improved from a mean of zero to 0.74, and for *S. kunkelii*, which improved from 0.05 to 0.91. Host ensemble variation ranged from 4.6% for *Erwinia carotovora subsp. atroseptica* to 17.6% for *X. campestris pv. campestris*. The validation accuracies for all species were improved by the use of ensembles, and the best validation accuracy of 0.55 for *E. carotovora subsp. atroseptica* was provided by host ensembles.

Ensembles that combined both climate and host predictions had disappointing performance, as for each species the accuracy was less than that of the host ensembles for all species but *X. axonopodis pv. dieffenbachiae*. Also, the validation accuracies were inferior for *Xanthomonas campestris pv. aberrans*, *E. carotovora subsp. atroseptica* and *S. kunkelii*, although it did yield the best validation accuracy for *R. radiobacter* of 0.61. It appears that the performance of these climate networks was simply too low, and this dragged down the performance of the combined ensembles.

There was little sign of over-training for the cascaded networks, with the most severe gap between training and testing accuracies being 0.06 for *X. campestris pv. campestris*. Variation over the testing accuracies ranged from 8.6% of the mean for *X. axonopodis pv. dieffenbachiae* to 23.7% for *X. campestris pv. campestris*. The testing accuracies were greater than those of either the single climate and host networks as well as the climate network ensembles, but less than those of the host network ensembles for *R. radiobacter*, *X. campestris pv. aberrans* and *S. kunkelii*. There were significant differences between the accuracies of the Climate+Host ensembles and cascaded networks for all species except *R. radiobacter*, but not between host ensemble and cascaded networks for *X. campestris pv. campestris* (two-tailed *t*-test, p=0.001). The validation accuracies of the cascaded networks were superior for *S. kunkelii* and *X. axonopodis pv. dieffenbachiae*, and best-equal for *X. campestris pv. aberrans*, which tied with host ensembles. The validation accuracies for *R. radiobacter*, *E. carotovora subsp. atroseptica* and *X. campestris pv. campestris* were all inferior to other methods.

The results of the contribution analysis of the cascaded networks are presented in Table 2. It can be seen in this table that the contribution of the host networks was greater than that of the climate networks for all species except *X. axonopodis pv. dieffenbachiae*. The differences between these contributions were quite small with the exception of *E. carotovora subsp. atroseptica* and *S. kunkelii*, where the contributions of the climate networks were much smaller than those of host networks. The results of the sensitivity analysis are presented in Fig. 1. This figure supports the interpretation of the contribution analysis results, as the curves for climate for *E. carotovora subsp. atroseptica* and *S. kunkelii* are much lower than the curves for hosts. The curve for

climate is higher than that of host only for *X. axonopodis pv. dieffenbachiae*, which was the only species where the contribution of climate was higher. Overall it is clear that the host network contributed more to the outputs of the cascaded networks than the climate networks did, although for most species the difference between the two was not large.

Table 2 Input contributions for cascaded networks

Species	**Climate**	**Host**
R. radiobacter	20.82/3.43	29.82/3.57
E. carotovora subsp. atroseptica	6.43/4.49	44.91/4.37
X. campestris pv. aberrans	20.93/3.77	28.55/3.48
X. campestris pv. campestris	25.43/5.04	32.03/7.35
S. kunkelii	16.25/4.45	44.49/4.01
X. axonopodis pv. dieffenbachiae	31.743/3.97	26.84/2.84

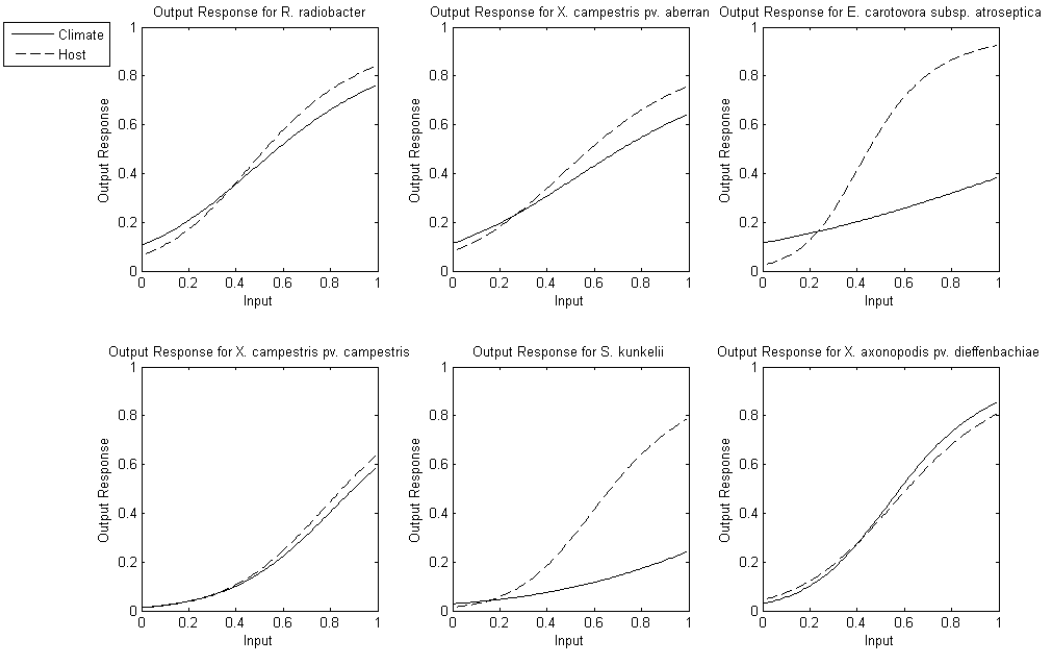

Fig. 1 Results of sensitivity analysis of cascaded network combining climate and host predictions.

4 Discussion

The results presented in this paper show that MLP were able to predict the presence or absence of bacterial crop diseases from regional climate variables and host plant assemblages. Combining predictions made by

several MLP using ensemble methods significantly improved the accuracy of predictions made from either climate or host assemblages. This was however not as effective as methods that combined predictions made from both abiotic (climate) and biotic (host assemblages) as these methods yielded superior performance. The most effective method of combining abiotic and biotic variables was a cascaded MLP architecture, whereby the outputs of the single most accurate climate and host networks were combined by a following network. Other methods that combined multiple climate and host networks together yielded slightly less accurate predictions, although such predictions were still more accurate than the original, single data set, models and varied by species.

Although species that are more prevalent tend to be easier to model with ANN, which as a data-driven model can be sensitive to the proportions of different classes within the data, this effect was not strongly apparent in these results, as species with relatively low prevalences such as *X. campestris pv. campestris* had validation accuracies as high or higher than more prevalent species such as *X. axonopodis pv. dieffenbachiae*. The amount of variation over the accuracies of the climate and host MLP indicates that the learning process was fragile, that is, while some networks learned very well, others learned very poorly, even with the same training parameters. This indicates that the learning process was very sensitive to the quality of the data given to it, which is not surprising given the status of ANN as data-driven models. Ensembles had much less variation, but this was because only the best, that is, most accurate, MLP were used to create them.

Input contribution of the cascaded networks showed that the host variables contributed the most strongly to the output of the network for the majority of species. However sensitivity analysis of these networks showed that the effects of the host networks were in general not much greater than those of the climate networks. This reinforces the need to perform a sensitivity analysis on significant variables so that their effects can be visualised.

No published work has yet come to light on predicting the global distribution of bacterial species. The results presented in this paper are thus significant as they point towards the use of ANN as a useful predictive tool.

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 environment 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. Alternatively, the disease may have once been established, but since been eradicated.

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, it is desirable, for future work, to be able to identify which of the input variables to the climate and host plant assemblage MLP are the most significant (Olden and Jackson, 2002; Olden, Joy and Death, 2004). A PCA transformation of the input data would remove the correlations, but may complicate the task of identifying the contribution of the original variables during the analysis of the MLP.

5 Conclusion

The results presented in this paper show that artificial neural networks are a useful tool for predicting the global distribution of bacterial crop diseases. Moderate to good accuracies, as measured by Cohen's Kappa, were achieved over the independent testing data sets for each species. Overall, predictions from abiotic variables were less accurate than those from biotic variables, while combining these predictions using either an ensemble method or a cascaded neural network method yielded accuracies that were superior to both for the majority of species.

Acknowledgements

The authors wish to acknowledge the work of Muriel Gevrey and Joel Pitt who prepared the climate data used in this research. This study was funded by the Centre of Research Excellence, Bio-protection, at Lincoln University, New Zealand. Data from the Crop Protection Compendium was used with permission of CAB International, Wallingford, UK.

References

Battiti R, Colla AM. 1994. Democracy in neural nets: Voting schemes for classification. Neural Networks, 7(4): 691-707

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

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46

Costa M, Gay P, Palmisano JD, et al. 1996. A Neural Ensemble for Speech Recognition. PROC of ISCAS 1996, IEEE International Symposium on Circuits and Applied Systems, Atlanta(GE), USA, 12-15

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

Filippi E, Costa M, Pasero E, 1994. Combining Multi-Layer Perceptrons in Classification Problems. European Symposium on Artificial Neural Networks. ESANN

Elith J, Graham CH, Anderson RP, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129-151

Flexer A. 1996. 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

Gevrey M, Worner SP. 2006. Prediction of global distribution of insect pest species in relation to climate by using an ecological informatics method. Journal of Economic Entomology, 99: 979 - 986.

Gotelli N, Entsminger G. 2006. EcoSim: Null models software for ecology. Version 7. Acquired Intelligence Inc. & Kesey-Bear,. Jericho, VT 05465, USA

Gotelli NJ. 2000. Null model analysis of species co-occurrence patterns. Ecology, 81(9): 2606-2621

Graham CH, Elith J, Hijmans RJ, et al. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45: 239-247

Hansen LK, Salamon P. 1990. Neural Network Ensembles. IEEE Transactions on Pattern Analysis and Machine Learning, 12(10): 993-1001

Joy MK, Death RG. 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology, 49: 1036-1052

Lek S, Delacoste M, Baran P, et al. 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling, 90:39-52

Manel S, Williams HC, Ormerod SJ. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology, 38: 921-931

Maqsood I, Khan MR, Abraham A. 2004. An Ensemble of Neural Networks for Weather Forecasting. Neural Computing and Applications, 13: 112-122

Olden JD, Jackson DA. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling, 154: 135-150

Olden JD, Joy MK, Death RG. 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling, 178: 389-397

Perone MP, Cooper LN. 1993. When networks disagree: Ensemble methods for hybrid neural networks. In: Neural Networks for Speech and Image Processing (Mammone RJ ed). Chapman-Hill, USA

Prechelt L. 1996. A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. Neural Networks, 9(3): 457-462

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. Nature, 323: 533-536

Sharkey AJC. 1996. On Combining Artificial Neural Nets. Connection Science, 8: 299-313

Sharkey AJC, Sharkey NE. 1997. Combining diverse neural nets. The Knowledge Engineering Review, 12(3): 231-247

Watts MJ, Worner SP. 2008a. Using artificial neural networks to determine the relative contribution of abiotic factors influencing the establishment of insect pest species. Ecological Informatics, 3: 64-74

Watts MJ, Worner SP. 2008b. Comparing ensemble and cascaded neural networks that combine biotic and abiotic variables to predict insect species distribution. Ecological Informatics 3, 354–366

Watts MJ, Worner SP. 2009. Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. Ecological Modelling, 220(6): 821-829

Watts MJ. 2011. Using data clustering as a method of estimating the risk of establishment of bacterial crop diseases. Computational Ecology and Software, 1(1): 1-13

Worner SP, Gevrey M. 2006. Modelling global insect pest species assemblages to determine risk of invasion. Journal of Applied Ecology, 43: 858-867

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore

Zhang WJ. 2011. Simulation of arthropod abundance from plant composition. Computational Ecology and Software, 1(1):37-48

Zhang WJ, Wei W. 2009. Spatial succession modeling of biological communities: a multi-model approach. Environmental Monitoring and Assessment, 158: 213-230