2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Evolving Connectionist Systems can Predict Outbreaks of the Aphid *Rhopalosiphum padi*

Michael J. Watts*

***Abstract*— Modeling of insect pest outbreaks is important for the protection of economically significant crops. This paper describes an attempt to model the outbreaks of the aphid *Rhopalosiphum padi* in the Canterbury region of New Zealand. Outbreaks were predicted using two representations of weather variables: Firstly, from moving time windows over the variables; Secondly, from the gradient or rate of change of the variables, which is presented here for the first time. Two artificial neural network types were used in this modeling, Multi-Layer Perceptrons (MLP) and Simple Evolving Connectionist Systems (SECoS). The results show that while SECoS are able to predict outbreaks of *R. padi* from either approach, MLP are unable to do so. Also, the results show that there is no significant difference in the modeling accuracy of SECoS between either modeling approach. These results indicate that the rate of change of weather variables is as important to the prediction of aphid outbreaks as the values of those variables. This work represents the first steps towards an outbreak prediction system that can assist with the management of these crop pests.**

I. Introduction

APHIDS of the species *Rhopalosiphum padi* are a pest of economically significant crop plants. *R. padi* are winged phytophagous insects that damage crops in two ways: Firstly, by consuming the sap of the host plants, weakening the plant; Secondly, *R. padi* carries the barley yellow dwarf virus, which stunts the growth of the infected plant and severely reduces its ability to produce the grains for which it is cultivated.

The Canterbury region of New Zealand is well-suited to growing cereals, due to its fertile soil and mild weather patterns. Wheat and other cereal crops are economically significant to the region, which makes the adequate control of pests such as aphids an important activity. The most significant of the aphid species in Canterbury is *R. padi*. While aphids can be controlled with pesticides, it is simply not feasible to apply pesticide every week in the hope of preventing aphid outbreaks. This is due to the economic cost of the pesticides and the negative environmental impacts associated with pesticide over-use, which includes the build-up of toxic chemicals in the food chain and the destruction of otherwise beneficial species. These negative impacts can be mitigated by the targeted application of pesticides, as determined by the accurate prediction of when aphid numbers are large enough to require control. These increases in aphid abundance are known as outbreaks.

It is known that the flight of aphids is correlated with weather [6]. Previously published work [4], [5], [9], [10], [11], [12] has focused on using weather variables to predict the abundance of aphids. Abundances are also influenced by factors such as predators, parasites, diseases and crop size. These complex factors make the prediction of aphid populations an ideal application for artificial neural networks (ANN).

While the previous work has focused on predicting aphid abundance as a time-series, precise prediction of aphid numbers is not only difficult but unnecessary. Assuming that the goal of modeling aphid numbers is to decide when to utilize control measures (such as spraying crops with pesticide), it is sufficient simply to predict when the numbers of aphids is going to substantially increase. That is, for the purposes of control, it is enough to predict the occurrence of outbreaks of aphids, rather than the magnitude of the outbreaks. The aphid prediction problem can therefore be simplified to a classification problem of predicting when these outbreaks, or spikes in abundance, are going to occur.

Interestingly, some of the previous work [12] has found that the changes in weather variables were just as effective predictors than the actual values of the variables. In other words, the numbers of aphids were influenced by the changes in the weather that occurred over time. From this it can be hypothesized that the rate of change of the weather variables, rather than the values of the variables themselves, are effective for predicting aphid abundance.

The work reported in this paper compares two ANN models, the standard multi-layer perceptron (MLP) and the evolving connectionist system (ECoS). Rather than predicting the absolute abundance of aphids for a given time period, the ANN predict when an outbreak is about to occur. Predictions were made using both a sliding time-window of weather variables, and the rate of change of each of those variables. Three research questions were investigated in this work:

1) Can evolving connectionist systems out-perform classical ANN models when predicting outbreaks?
2) Can the rate of change of weather variables be used to predict outbreaks of aphids?
3) Does the size of the temporal window of weather variables affect the prediction accuracy?

There are two main original contributions in this paper: Firstly, I re-cast the aphid prediction problem as a classification problem, rather than a time-series prediction problem; Secondly, I show that the rate of change of weather variables are sufficient to predict outbreaks.

*Information Technology Programme, Auckland Institute of Studies, P.O. Box 2995, Auckland, New Zealand (email: mjwatts@ieee.org).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

II. Methods

A. Data

The initial source of the data was Crop and Food of Lincoln, Canterbury, New Zealand. Crop and Food is a government research agency responsible for developing better methods of agriculture. Aphids were caught using a suction trap at the Lincoln Crop and Food research station. A suction trap is a large, vertically mounted pipe that has a fan at the bottom that generates a suction. Small flying insects are drawn into the pipe and trapped in a filter before reaching the fan. The filter was removed from the trap at the end of each week, and the insects inside were classified and counted. Only counts of *R. padi* were included in this data set. Twenty two years of aphid counts were available, for the period mid-1981 to mid-2004. These counts are shown in figure 1. From this plot it is clear that the aphids undergo periodic outbreaks. The period between outbreaks is irregular, as are the magnitudes of the outbreaks.

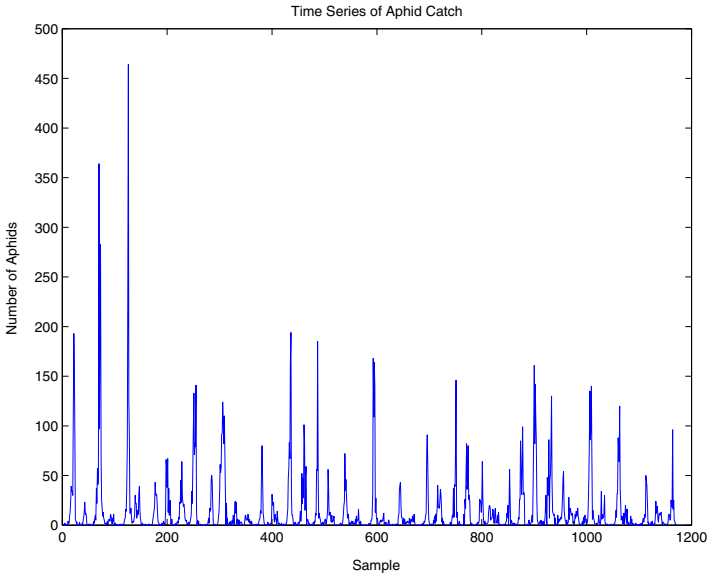

Fig. 1. Aphid abundances

Thirteen weather variables were also recorded continuously over this time period, as listed in table I. As the potential rainfall deficit variable ranged quite highly, the log of this variable was also included, as previous work [12] had shown that this improved performance.

The abundances were converted into outbreak classes by applying a threshold to the aphid counts. Any values greater than or equal to the threshold were converted to unity, while any values less than the threshold were converted to zero. The threshold used here was seventeen, which meant that 20% of the weekly counts were considered to be outbreaks. It is important to bear in mind that the number seventeen was the number of aphids caught at a single point (the location of the suction trap). There would therefore be many, many more aphids in the wild attacking the nearby crops. Predictions were made one week ahead, that is, the goal was to predict whether an outbreak would occur in the following week.

TABLE I
Weather variables used.

Average rainfall (mm)
Cumulative rainfall (mm)
Wind run (km/day)
Maximum air temperature (°C)
Minimum air temperature (°C)
Mean air temperature (°C)
D-days, cumulative temperature for the week
Grass temperature (°C)
Soil temperature at 100 centimetres below ground (°C)
Penman potential evaporation (mm)
Potential deficit of rainfall, accumulated excess of Penman over rainfall
Vapour pressure (hecto Pascals)
Solar radiation (MJ/m^2)

All variables were linearly normalized to the range zero to unity. As *R. padi* is known to have a three week life cycle, a temporal element must be included in the input variables. Two approaches were used for this. In the first, a sliding window was passed over each of the variables. Window sizes of three, four, five and six were tested. The number of aphids observed per week was included as an input variable that was also windowed. This gave an input vector of size 45, 60, 75 or 90 elements. The second approach sprang from the observation in previous work [12] that the changes in the weather variables were just as useful for predicting abundance as the absolute values of the variables. In this paper, rather than using the changes in values as predictors, the rate of change of each variable over a particular time window was calculated . The time windows were the same size as the sliding windows (three, four, five or six weeks). The advantage of this approach is that the input vector always has fifteen elements in it, no matter the size of the time window. The rate of change was calculated by fitting a linear regression line over the values in that window, and taking the gradient.

B. Cross Validation

Data was divided into years, and the years grouped together into ten subsets so that ten-fold cross-validation could be carried out. For each fold, one of the subsets was held out as a test data subset and the ANN trained on the remaining nine subsets. The network was then recalled over the held-out test subset to test the networks generalization ability. At the end of the ten folds, the generated predictions from the test subsets were combined into a single data set and compared with the known outputs. Accuracy was measured as a simple percentage of true positive, true negative, and overall examples correctly classified. Since only 20% of the examples were positive, it was possible to have a high overall percentage while only classifying the negative examples correctly. In other words, an ANN could score an overall accuracy of 80%, and a true negative accuracy of 100%, simply by classifying every example as negative. To deal with this bias, accuracies were also measured using Cohen's

kappa statistic [1]. The advantage of the kappa statistic is that it is a single metric that is not biased by an unbalanced number of examples from the two classes. Thus, while the overall percentage accuracy will be biased by the fact that only 20% of the weeks represent outbreaks (making it easy to achieve at least 80% accuracy), the kappa statistic requires accuracy over both classes in order to achieve a high score. Accuracies were also measured over each held-out subset so that ten accuracies were available for each cross-validation run. These sets of accuracies were needed to perform a statistically valid comparison between modeling approaches.

C. Modeling with MLP

This work used standard MLP trained with back-propagation with momentum. The networks had two hidden neuron layers and each layer, apart from the input layer, used logistic activation functions. A variety of hidden neuron layer sizes and back-propagation parameters were investigated via a trial-end-error process, where selection of the better parameters was based on the kappa statistic over the combined generalization data set.

D. Evolving Connectionist Systems

Evolving Connectionist Systems (ECoS) are a family of constructive ANN that expand and adapt their internal structure during learning. ECoS were first proposed in 1998 [2], and there are many models within the ECoS family, including the Evolving Fuzzy Neural Network EFuNN [3] and the Simple Evolving Connectionist System SECoS [7]. For a review of ECoS networks, see [13]. ECoS networks are fast learning models, and have been shown to have comparable or superior performance to conventional models like MLP [8].

The experiments in this work used the SECoS model. This model has three layers of neurons: the input layer, which has the same function as the input neuron layer of MLP; the evolving layer, which is where neurons are added during training; and the output layer, which again has the same function as the output neuron layer of MLP. SECoS has four training parameters: the first two are the sensitivity threshold and the error threshold, which control the addition of neurons to the evolving layer. The other two are the two learning rates, which control the adaptation performed in the input to evolving layer connection weights, and the evolving layer to output layer connection weights. The software used is available from [15]. As SECoS are able to train quickly compared with MLP, the optimal training parameters were approximated by performing an exhaustive combinatorial search over the parameter space. Selection of the training parameters was made by inspection of the kappa statistic over the combined generalization data set and of the number of neurons added to the evolving layer. Some parameter combinations will cause an evolving layer neuron to be added to the SECoS for every training example, which is not optimal as it leads to over-fitting and inefficient use of computing resources. Therefore, parameter combinations that caused neurons to be added for every training example were rejected and the selection made from the remaining results.

III. Results

After extensive experimentation with the architecture and training parameters of the MLP, no MLP was able to produce an acceptable accuracy. The kappa statistic for each MLP over both the training and testing data sets was consistently zero. This was caused by the inability of the MLP to achieve a true positive accuracy greater than zero. In other words, the MLP were unable to predict outbreaks at all. This was the case for both the windowed inputs, for all window sizes, and for the gradient inputs, for gradients calculated over all window sizes.

Conversely, the SECoS networks were able to learn to predict outbreaks to an acceptable level of accuracy. The accuracies of SECoS trained using moving windows of variables are presented in table II, while the results for the gradients are in table III. These results show that SECoS are able to predict outbreaks no matter the input representation scheme. There is no variation in accuracy or size of the SECoS for the gradient representation scheme, while the amount of variation in accuracy for the window scheme is small. There is a substantial variation in the size of the SECoS for the window representation scheme. This is understandable, as different numbers of input variables will require different numbers of neurons to model. There were no significant differences (two-tailed t-test, p=0.1) between the kappa statistics for either representation scheme, while the SECoS were significantly larger for the windowing representation, with the exception of window size of three, which was extremely small.

TABLE II

Accuracies of SECoS trained over windows of weather variables. The column "Window" is the number of weeks in the window. Column "Kappa" is the Cohen's kappa statistic. "TN %" is the percentage of true negative examples correctly classified, "TP %" is the percentage of true positive examples correctly classified, and "Overall" is the percentage of all examples correctly classified, whether negative or positive. "Neurons" is the number of neurons in the evolving layer of the SECoS after training. Accuracies are generalization accuracies over the entire data set

Window	Kappa	TN %	TP %	Overall %	Neurons
3	0.414	87.4	54.8	80.9	28.4
4	0.436	81	70.2	78.8	208
5	0.423	84.7	61.1	80	210.7
6	0.414	89.4	51	81.7	187.4

IV. Discussion

Three research questions were posed in the Introduction to this paper. The results allow them to be answered as follows:

1) Can evolving connectionist systems out-perform classical ANN models when predicting outbreaks?

TABLE III

ACCURACIES OF SECOS OVER THE GRADIENTS OF WEATHER VARIABLES. THE COLUMN "WINDOW" IS THE NUMBER OF WEEKS OVER WHICH THE GRADIENT WAS MEASURED. OTHER COLUMN LABELS ARE AS IN TABLE II

Window	Kappa	TN %	TP %	Overall %	Neurons
3	0.379	80.4	63.5	77	101
4	0.379	80.4	63.5	77	101
5	0.379	80.4	63.5	77	101
6	0.379	80.4	63.5	77	101

None of the MLP were able to predict the outbreaks at all. SECoS networks were able to predict outbreaks at least 54.8 % of the time for a window size of three weeks, and up to 70.2 % of the time for a window size of four weeks. From these results, it can be concluded that the particular evolving connectionist system used here, SECOS, out-perform MLP for this problem.

2) Can the rate of change of weather variables be used to predict outbreaks of aphids?

The answer to this is an unqualified *yes*. Firstly, the kappa statistics for the SECoS trained over the gradient of the input variables were all above zero, which corresponds to an accuracy better than chance. The true-positive percentages were all above 50 %, again indicating a better than chance performance. Secondly, there were no significant differences in the kappa statistics between the SECoS trained on the gradients of the input variables and the SECoS trained on the windows of input variables.

3) Does the size of the temporal window of weather variables affect the prediction accuracy?

The answer to this question is less clear. Varying the size of the input window altered the best accuracy of the SECoS trained on windowed weather variables, and led to a wide variation in the number of neurons added to the SECoS evolving layer during training. However, the performance of the SECoS trained on the gradients of the weather variables did not alter with the size of the window over which the gradients were measured. While the data definitely varied with the different window sizes, it seems that the variation was not enough to cause any changes in accuracy.

It is not currently known why MLP were unable to predict outbreaks, when the previous research has shown that they are able to model the abundance [9], [10]. However, when modeling abundance a common problem was that the MLP badly under-estimated the magnitude of the outbreaks. Thus, while they were able to predict an upward swing in aphid numbers, they were not able to detect a large increase. It is possible that the predicted upward swings in the abundance were below the threshold value chosen here, but that by itself does not entirely explain the inability to predict outbreaks.

While the aphid numbers and weather variables used in this study are considered to be data of high confidence, there are several other factors that complicate the results. Firstly, data is not available on the use of control measures such as pesticides during the study period. The use of pesticides would reduce the number of aphids present and potentially prevent further outbreaks. Secondly, data is not available on the amount of crops in existence at each time period. As *R. padi* is a phytophagous insect, their numbers are more likely to increase when a larger amount of food is available for them. Finally, while aphids have several natural predators, the abundance of these predators is not known. Clearly, an increase in the abundance of predators would lead to a decrease in the abundance of aphids, at least until an equilibrium was reached.

Overall, this work has shown that prediction of outbreaks of *R. padi* is possible. Future work will focus on improving the prediction accuracy. Ultimately, an outbreak prediction system could assist with the effective management of these crop pests.

REFERENCES

[1] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement* vol. 20, pp. 3746, 1960.

[2] N. K. Kasabov, "The ECOS framework and the ECO learning method for evolving connectionist systems. Journal of Advanced Computational Intelligence, 2(6):195202, 1998.

[3] N. K. Kasabov, "Evolving Fuzzy Neural Networks - Algorithms, Applications and Biological Motivations" *Methodologies for the Conception, Design and Application of Soft Computing*, 1998.

[4] G. Lankin, S. P. Worner, S. Samarasinghe and D. A. J. Teulon, "Can artificial neural network systems be used for forecasting aphid flight patterns?", *New Zealand Journal of Plant Protection* vol. 54, pp. 188-192, 2001.

[5] G. Lankin, "Neural network models for the prediction of autumn migration of the cereal aphid *Rhopalosiphum padi* at Lincoln, Canterbury, New Zealand", M.Sc. Thesis, Lincoln University, Canterbury, New Zealand, 2002.

[6] G. G. Thomas, G. K. Goldwin and G. M. Tatchell, "Associations between weather factors and the spring migration of the damson-hop aphid, *Phorodon humuli*, *Annals of Applied Biology*, vol. 102, pp. 7-17, 1983.

[7] M. J. Watts and N. K. Kasabov, "Simple evolving connectionist systems and experiments on isolated phoneme recognition", in *Proceedings of the first IEEE conference on evolutionary computation and neural networks*, San Antonio, IEEE Press, pp. 232-239, 2000.

[8] M. J. Watts. "Evolving connectionist systems: characterisation, simplification, formalisation, explanation and optimisation". PhD thesis, University of Otago, New Zealand, 2004.

[9] M. J. Watts, S. P. Worner, G. O. Lankin, and D. A. Teulon, "Comparison of MLP and ECoS Networks for the Prediction of the Flight of Aphids in Autumn Sown Wheat Crops". In: *Proceedings Conference on Neuro-Computing and Evolving Intelligence 2004, NCEI'04, 13-15 December, 2004, AUT Technology Park, 581 Great South Road, Auckland, New Zealand*. Editors: Nik Kasabov, Zeke S. H. Chan, 2004.

[10] M. J. Watts and S. P. Worner, "Using Multi-Layer Perceptrons to Model the Lincoln Aphid Data Set". Technical Report, Bio-Protection and Ecology, Lincoln University. ISBN 978-0-86476-176-7. February, 2007.

[11] M. J. Watts and S. P. Worner, "Comparison of Multi-Layer Perceptrons and Simple Evolving Connectionist Systems over the Lincoln Aphid Data Set". Technical Report, Bio-Protection and Ecology, Lincoln University. ISBN 978-0-86476-175-9. February, 2007.

[12] M. J. Watts and S. P. Worner, "Further Sensitivity Analysis of Simple Evolving Connectionist Systems Applied to the Lincoln Aphid Data Set". Technical Report, Bio-Protection and Ecology, Lincoln University. ISBN 978-0-86476-177-5. February, 2007.

[13] M. J. Watts, "A decade of Kasabovs Evolving Connectionist Systems: A Review". *IEEE Transactions on Systems, Man and Cybernetics Part C - Applications and Reviews*, vol. 39 no. 3, pp. 253-269, 2009.

[14] S. P. Worner, G. O. Lankin, S. Samarasinghe, and D. A. J. Teulon, "Improving Prediction of Aphid Flights by Temporal Analysis of Input Data for an Artificial Neural Network ", *New Zealand Plant Protection* vol. 55, pp. 321-316, 2002.

[15] The ECoS Toolbox. Retrieved from http://ecos.watts.net.nz/Software/Toolbox.html.