IMPROVING PREDICTION OF APHID FLIGHTS BY TEMPORAL ANALYSIS OF INPUT DATA FOR AN ARTIFICIAL NEURAL NETWORK

S.P. WORNER¹, G.O. LANKIN¹, S. SAMARASINGHE² and
D.A.J. TEULON³

¹*Soil, Plant and Ecological Sciences Division, P.O. Box 84, Lincoln University,
Canterbury*

²*Environmental Management and Design Division, P.O. Box 84, Lincoln
University, Canterbury*

³*New Zealand Institute for Crop & Food Research Limited, Private Bag 4704,
Christchurch*

Corresponding author: worner@lincoln.ac.nz

ABSTRACT

Weather data in its raw form frequently contains irrelevant and noisy information. Often the hardest task in model development, regardless of the technique used, is translating independent variables from their raw form into data relevant to a particular model. A sequential or cascading temporal correlation analysis was used to identify weather sequences that were strongly correlated with aphid trap catches recorded at Lincoln, Canterbury, New Zealand, over 1982-2000. Trap catches in the previous year and 13 weather sequences associated with eight climate variables were identified as significant predictors of aphid trap catch during the autumn flight period. The variables were used to train artificial neural network (ANN) models to predict the size of autumn aphid migrations into cereal crops in Canterbury. Such models would assist cereal growers to make better informed and more timely pest management decisions. ANN predictive performance was compared with multiple regression predictions using jackknifed data. The ANN gave superior prediction compared with multiple regression over 13 jackknifed years.

Keywords: neural networks, prediction, aphid flights, sequential temporal cascading correlation.

INTRODUCTION

Weather conditions influence aphid population dynamics either directly, by affecting development and survival, or indirectly by affecting natural enemies or the aphid host plant. Several studies have shown aphid migrations to be correlated with the weather conditions prevailing prior to migratory flights and have used these relationships to predict such flights (Carter et al. 1982; Thomas et al. 1983; Worner et al. 1995). Several aphid species are vectors of barley yellow dwarf virus (BYDV), an important cause of yield loss in wheat crops in New Zealand. In Canterbury, the most important aphid vector is *Rhopalosiphum padi*. Any prior information concerning the expected size and timing of the flight patterns of this species would benefit growers by better informing their pest management decisions.

Artificial neural networks (ANN) are nonlinear mapping structures that attempt to mimic the computational processes of the brain to learn to identify important relationships and key variables in complex data sets (Lankin et al. 2001; Smith 1996). Once the neural network has been trained to recognise patterns, the resulting model can make predictions by detecting similar patterns in future data. Lankin et al. (2001) used twenty years of climatic and aphid flight data to train a neural network to predict the date and size of *Rhopalosiphum padi* migrations at Lincoln, Canterbury, New Zealand. Despite the resulting

New Zealand Plant Protection 55:312-316 (2002)

ANN model fitting the training data set extremely well, its prediction of new data was poor. The most likely explanation was that the model overfitted the data, describing the noise so well that it was unable to generalise. In its raw form the climatic data used to train the artificial neural network clearly included much irrelevant information.

The objective of this study was to develop an improved ANN trained with input data from which much irrelevant information was removed. This required a sequential temporal cascading correlation analysis to identify weather sequences prior to the beginning of the flight, that were highly correlated with aphid numbers caught during the flight period. Selected sequences were used as input variables in multiple regression and neural network models to predict autumn *R. padi* flight in Canterbury.

MATERIALS AND METHODS

Nineteen years of aphid trap catches recorded by Crop & Food Research, Lincoln, Canterbury, and weekly weather data consisting of 12 weather variables recorded at the Canterbury Agricultural Research Centre, Lincoln, were used. The data series for all variables started from week 27 of 1981 and extended to week 26 of 2000. The autumn flight period for each year was identified according to the methods of Lankin (2002). The number of aphids caught over the autumn flight period was correlated with weather conditions prior to migration using a sequential temporal cascading correlation algorithm written in MATLAB 4.2c.1 (The Mathworks 1993).

The number of aphids caught during the flight period for each year was correlated with the mean value of each weather variable averaged over x consecutive weeks. The duration of x varied sequentially from 1 week to a maximum of 10 weeks, and the starting date used for x was displaced sequentially (cascaded) by 1 week throughout the time series for each correlation.

The following independent variables were analysed: cumulative weekly rainfall (mm); maximum, minimum and mean air temperature ($^{\circ}\text{C}$); grass minimum temperature ($^{\circ}\text{C}$); solar radiation (MJ/m^2); wind run (km/day); vapour pressure (hPa); Penman potential evaporation (mm); potential deficit (accumulated excess of Penman over rainfall); average weekly degree days (above a threshold of 0°C); and soil temperature at 100 cm below ground ($^{\circ}\text{C}$). The number of aphids caught per week in the previous spring were also included in the correlations.

The weather sequences with the highest correlation coefficients were selected. A further selection was made based on a best subsets multiple regression (Minitab 11.12) that indicated the most significant independent variables and therefore the most parsimonious model.

Using the selected input variables, several different neural network models were trained using Neuroshell II (release 4.0) and the best seven were chosen according to the minimum error and highest R^2 values for the training set. Using the bootstrap method, 5 years were randomly chosen to be left out of training as validation data for each neural network model. The predicted values from the trained network and the observed values in the validation data set could then be compared.

Using the same input variables, a single year's data at a time was removed from the best subsets multiple regression model. A new model based on the remaining years was re-formulated. This model was then used to predict aphid flights for the missing year. This is a statistical technique that is known as "jackknifing" (Efron & Gong 1983). For both neural network and regression models, the dependent variable was $\log_{10}(\text{aphids})$. For each jackknifed year the predicted number of aphids migrating and the observed number were compared. Predictions made by the neural networks and multiple regression models were also compared with each other.

RESULTS

The sequential temporal cascading correlation method resulted in a large number of linear correlations. Those that satisfied two conditions were selected for further analysis. First, weather sequences that had a correlation coefficient (r) higher than +0.65 or lower

than -0.65 were chosen. Second, because a predictive model was needed, only those sequences occurring prior to Julian week 6 were considered. Julian week 6 was the earliest week of the beginning of migration over the 19 year period.

The model that combined the lowest number of significant variables, a C_p (Mallow's index) similar to the number of variables in the model and a very high coefficient of determination ($R^2 = 98.2$) was chosen from the best subsets regression. Trap catches in the previous year and 13 weather sequences associated with eight climate variables were chosen as input for the models (Table 1).

TABLE 1: Coefficients, standard deviations, T and P values for the best subset multiple regression.

Variable names	Coefficient	SD	T	P	Variable definition
$Cr_{(40-41)}$	9.57	1.72	5.56	0.005	Cumulative rain, weeks 40, 41 previous year
$Cr_{(48)}$	48.12	10.24	4.70	0.009	Cumulative rain, week 48 previous year
$Dd_{(41)}$	-57.62	9.19	-6.27	0.003	DDs above 0°C, week 41 previous year
$Dd_{(3-6)}$	-23.12	3.46	-6.69	0.003	DDs above 0°C, weeks 3, 4, 5, 6 current year
$MxT_{(6)}$	-368.40	67.23	-5.48	0.005	Max. air temperature, week 6 current year
$MnT_{(51)}$	-309.90	67.37	-4.60	0.01	Min. air temperature, week 51 previous year
$MnT_{(2-5)}$	1034.50	166.5	6.21	0.003	Min. air temp., weeks 2, 3, 4, 5 current year
$Nap_{(51)}$	-10.67	3.41	-3.12	0.035	Number of aphids, week 51 previous year
$Ppe_{(48)}$	285.88	79.52	3.60	0.023	Penman pot. evap., week 48 previous year
$Sr_{(48)}$	-91.26	19.19	-4.76	0.009	Solar radiation, week 48 previous year
$Sr_{(1-2)}$	37.98	7.37	5.15	0.007	Solar radiation, weeks 1, 2 current year
$Vp_{(48)}$	-169.20	27.86	-6.07	0.004	Vapour pressure, week 48 previous year
$Vp_{(1-2)}$	155.91	27.51	5.67	0.005	Vapour pressure, weeks 1, 2 current year
$W_{(41)}$	3.95	1.03	3.85	0.018	Wind run, week 41 previous year

Seven different neural networks were trained. Each had a different combination of yearly input data randomly selected as training data set, test data set and validation data set. The validation set comprised data left aside from the training process (a similar technique to jackknifing) to be used as independent data to test prediction accuracy. If a year was included in more than one validation set, the ANN predictions for that year were averaged. Because of the way validation was handled by the neural network software, only 13 of the 19 years were available to test the ANN predictions.

The neural network models performed better than the standard regression models (Table 2) with an absolute mean error of 122 aphids and a high correlation ($r = 0.96$) between observed and predicted aphids. This compares with an absolute mean error of

TABLE 2: Actual and predicted $\log_{10}$ (aphids) and backtransformed values, calculated from neural network and multiple regression models for new (jackknifed) data. Correlation coefficients (r values) are given for the correlation between the predicted values and the corresponding actual values in column 2 (observed $\log_{10}$ (aphids)) and column 3 (observed number of aphids).

Year	Observed		Predicted					
	$\log_{10}$ (aphids)	Number aphids	Neural networks			Multiple regression		
			$\log_{10}$ (aphids)	Number aphids	Actual - predicted	$\log_{10}$ (aphids)	Number aphids	Actual - predicted
82	1.9190	83	2.0501	112	-29	1.5993	40	43
83	1.7242	53	1.8901	78	-25	2.1716	148	-95
85	2.5428	349	2.5028	318	31	2.0530	113	236
86	2.8836	765	2.7266	533	232	3.0734	1184	-419
87	2.9907	979	2.7558	570	409	3.2994	1993	-1014
88	1.7708	59	2.0837	121	-62	1.6580	46	14
89	2.0128	103	2.0421	110	-7	2.2307	170	-67
90	2.7015	503	2.3508	224	279	1.9802	96	408
91	2.1959	157	2.1884	154	3	2.0798	120	37
94	1.7160	52	1.9780	95	-43	1.3598	23	29
96	2.6374	434	2.4399	275	159	3.0250	1059	-625
98	2.7041	506	2.4285	268	238	3.1716	1485	-979
2000	2.3201	209	2.1618	145	64	1.9923	98	111
Correlation coefficient r =			0.95	0.96		0.80	0.86	
Absolute mean error =					122			314

314 aphids and a correlation between observed and predicted number of aphids of $r = 0.86$ for the multiple regression models.

DISCUSSION

Other authors have investigated using weather sequences prior to insect population events to predict population dynamics. Often the periods used conform to official seasons, the beginning and end of which are not necessarily relevant to the species (Worner et al. 1995). Thomas et al. (1983) were the first to use what the authors call sequential temporal cascading correlation to identify weather sequences of relevance to the species of interest. Thomas et al. (1983) used temperature and rainfall, either individually or combined, to predict the timing of spring migration of the damson-hop aphid (*Phorodon humuli*) into hop-gardens in England. They did not attempt to predict the number of migrating aphids.

Ding et al. (1996) suggested that while neural networks are useful in real-world applications involving complex systems, they are most useful when trained on good quality data. In this study, pre-processing the input data greatly improved the performance of both the ANN and multiple regression models. While the ANN gave more accurate predictions, the multiple regression models also performed well. However, it is important to point out that some of the 14 independent variables chosen for model input were correlated with each other, thus violating the assumptions of regression analysis. In contrast, input variables that are correlated do not affect the predictive performance of ANN.

The ANN in this study would provide very good prediction of the expected size of aphid populations for growers but clearly could be further refined. ANN predictive ability will be increased by adding to the time series as data becomes available. As well, the addition

of daily weather data and variables such as plant phenology and sowing dates and the use of a form of calibration model (Smith 1996) to improve prediction could be investigated.

REFERENCES

- Carter, N.; Dixon, A.F.G.; Rabbinge, R. 1982: Cereal aphid populations: biology, simulation and prediction. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands. 91 p.
- Ding, X.; Canu, S.; Denoeux, T. 1996: Neural network based models for forecasting. *In*: Taylor, J.G. *ed.* Neural networks and their applications. John Wiley and Sons Ltd. Pp. 153-167.
- Efron, B.; Gong, G. 1983: A leisurely look at the bootstrap, the jackknife, and cross validation. *Am. Stat.* 37: 36-48.
- Lankin, G. 2002: 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.
- Lankin, G.; Worner, S.P. Samarasinghe, S.; Teulon, D.A.J. 2001: Can artificial neural network systems be used for forecasting aphid flight patterns. *N.Z. Plant Prot.* 54: 188-192.
- Smith, M. 1996: Neural networks for statistical modelling. Thomson Computer Press, USA. 235 p.
- The MathWorks, Inc. 1993: MATLAB® High-performance numeric computation and visualisation software. User's Guide. The MathWorks, Inc., Massachusetts.
- Thomas, G.G.; Goldwin, G.K.; Tatchell, G.M. 1983: Association between weather factors and the spring migration of the damson-hop aphid, *Phorodon humuli*. *Ann. Appl. Biol.* 102: 7-17.
- Worner, S.P.; Tatchell, G.M.; Woiwod, I.P. 1995: Predicting spring migration of the damson-hop aphid *Phorodon humuli* (Homoptera: Aphididae) from historical records of host-plant flowering phenology and weather. *J. Appl. Ecol.* 32: 17-28.