The use of multiple regression via principal components in forecasting early season aphid (Homoptera: Aphididae) flight

G.G. Howling

School of Biological Sciences, University of Birmingham, UK

R. Harrington

Department of Entomology and Nematology, AFRC Institute of Arable Crops Research, Rothamsted Experimental Station, Harpenden, Herts, UK

S.J. Clark

Department of Statistics, AFRC Institute of Arable Crops Research, Rothamsted Experimental Station, Harpenden, Herts, UK

J.S. Bale

School of Biological Sciences, University of Birmingham, UK

Abstract

The Rothamsted Insect Survey has kept records of aerial aphid activity using suction traps since 1965. Previous work has shown that, for certain species, there is a linear relationship between the date of first record in the trap each year and the mean temperature during the preceding winter. This paper describes and evaluates a more complex technique which relates the date of first record to a range of weather variables over 28 winter time periods. The technique uses principle components analysis to remove correlations between weather variables before these are regressed on the date of first record. Data from 1966 to 1988 inclusive are used to generate models to predict the date of first record of *Myzus persicae* (Sulzer) at Rothamsted, and the predictive values of models using both simple and multiple regression are assessed using data from 1989 to 1992. The accuracy of the multiple regression models was no greater than that of the simple regression models during these years. However, the multiple regression approach identified relationships with other variables for time periods when the correlation with mean temperature was weaker and may therefore be more widely applicable.

Introduction

Aphids cause damage to crops both directly, by the removal of phloem, and indirectly, by the transmission of virus diseases. Crop plants such as sugarbeet (*Beta vulgaris*) and potatoes (*Solanum tuberosum*) are particularly susceptible to damage if infected with virus during the early stages of their development (Watson *et al.*, 1946), and therefore an aphid infestation tends to be of greater economic importance the earlier in the growing season it occurs. Predictions of the timing of aphid migration into crops can enable growers to make rational decisions on the use of insecticides. For example, granular insecticides would normally be applied to

Correspondence: R. Harrington, Department of Entomology and Nematology, AFRC Institute of Arable Crops Research, Rothamsted Experimental Station, Harpenden, Herts. AL5 2JQ, UK.

potato and sugarbeet crops at the time of planting, and can be expected to remain active for six to eight weeks after this (Dewar, 1986). A system which could provide an early indication of the likelihood of aphid activity whilst the insecticide is effective would allow growers to avoid unnecessary applications at the time of planting. For these predictions to be of maximum value in rationalizing use of granular insecticides, they would need to be issued by the beginning of March, although a later prediction may be of value in warning growers of the need for aphicidal sprays later in the growing season.

The Rothamsted Insect Survey has used 12.2 m suction traps (Macaulay *et al.*, 1988) to monitor aerial aphid activity in Great Britain since 1965, and since then a network of such traps has been established throughout much of Europe (Tatchell, 1991). For some species at some sites, the date of first record and the total number of aphids caught up to 1 July are linearly related to mean temperature during the preceding winter, and simple linear regression (Turl, 1980; Walters & Dewar, 1986; Harrington *et al.*, 1990) or multiple linear regression (A'Brook, 1983; Harrington *et al.*, 1989) techniques have been used to develop predictive models. Simple linear regression models do not always account for a sufficient proportion of the total variance to make a reliable prediction, probably because the model is attempting to describe a complex relationship using only one predictor variable. Previous multiple regression studies have used weather variables which are likely to have been correlated with one another, and this makes interpretation of the models difficult. Harrington *et al.* (1991) investigated the value of an improved multiple regression technique in forecasting the date of the first record of 49 aphid species at six sites in 1989. The models used were particularly successful in predicting the date of the first record of four species (*Myzus persicae* (Sulzer), *Macrosiphum euphorbiae* (Thomas), *Sitobion avenae* (Fabricius) and *Phorodon humuli* (Schrank)) at a range of sites, and the methodology warrants further testing and comparison with the more straightforward simple linear models. This paper describes the multiple regression technique in greater detail and assesses the abilities of the models using this technique, and using simple regression with mean temperature, to predict the date of the first record of *Myzus persicae* in the Rothamsted suction trap in the years from 1989 to 1992.

Methods

Sources of data

Aphid data were obtained from the Rothamsted Insect Survey database. The Julian date of the first record of *M. persicae* at Rothamsted in each year was used in the analyses. This was assumed to be an indication of the time at which the aerial aphid population reached the threshold of detection using the suction traps, and therefore an index of the time of the start of the migration from winter hosts, potentially to crop plants. The use of the first aphid trapped is likely to result in stochastic sampling errors, although Harrington *et al.* (1990) found that the first record was more closely related to winter mean temperature than either the third or the fifth record. Data from 1966–1988 were used for the construction of the prediction models.

Table 1. Variables used in the analysis

1. Mean screen temperature
2. Mean screen minimum temperature
3. Mean grass minimum temperature
4. Mean accumulated day degrees below zero
5. Mean accumulated day degrees above zero
6. Mean 10 cm soil temperature
7. Mean 30 cm soil temperature
8. Mean rainfall
9. Mean sunshine duration
10. Minimum screen minimum temperature
11. Minimum grass minimum temperature

Weather data were obtained for each site from the Agricultural and Food Research Council's ARCMET database. To assess the relative importance of weather at different times during the winter, separate analyses were done for each of the months from November to May, and each consecutive combination of these months, giving 28 analyses in total. Nine weather variables were selected for inclusion in the model according to their availability and perceived biological significance, and these were averaged over each of the 28 winter periods. The absolute minimum screen and grass temperatures for each time period were also included, making 11 variables in total (table 1).

Statistical analysis

Multiple regression can be used to find the most important weather variables affecting date of first catch. However, high correlations are often found between weather variables. This can cause difficulties in assessing the relative importance of each of the regressor variables as well as limiting the usefulness of the model when predictions are made from data which are outside the range of those used to construct the model (Abraham & Ledolter, 1983). It is therefore desirable to remove these correlations before continuing with the construction of potential forecasting models. A principal components analysis applied to the 11 original variables generates a set of scores which retains the total variation in, and has the same dimensionality as, the original dataset. Moreover, each score is a linear combination of all the variables such that the scores are all uncorrelated, and are therefore more suitable than the original variables for use in multiple regression analysis. A drawback is that if an individual year has a missing value in any one variable then the values of all the other variables for that year are ignored in the principal components analysis. This was an important consideration in the selection of the variables to be used in the analysis, although a limited number of missing values could be estimated using the MULTMISS procedure in Genstat 5 (Payne *et al.*, 1987). A principal components analysis was performed after the 11 variables had been put on a common scale of 0–100. This was achieved by assigning the smallest datum in each variable a value of zero, the

https://doi.org/10.1017/S0007485300029278

largest a value of 100, and placing the remaining data on a linear scale between.

Multiple regression models were produced by regressing the aphid data on the principal components scores derived from the weather data for each time period, using the stepwise multiple linear regression methods of Genstat 5. The F-values for the inclusion and exclusion of variables were both set to 4.0, allowing terms to be added to, or deleted from, the model only if this resulted in a significant improvement at approximately the 5% level (Montgomery & Peck, 1982, p. 278). Models were backtransformed to obtain coefficients for the original variables.

Simple regression models were produced by regressing the aphid data on the mean screen temperature for each time period.

Model evaluation

The coefficient of multiple determination, R^2, can be interpreted as the 'percentage variance accounted for' and is often used as an indication of the adequacy of a linear model. However, R^2 can be influenced by factors other than the fit of the data, such as the steepness of the regression line (Barrett, 1974), and can at best only indicate the strength of the relationship within the existing data. Thus a model with a large value of R^2 may not perform well in forecasting future values. It is also inappropriate to use R^2 for comparisons of models in different situations, in this case different winter time periods for example.

To obtain a measure of model prediction accuracy the models generated using both techniques were assessed using the prediction error sum of squares (PRESS) technique (Allen, 1974). The PRESS value is the sum of the squared differences between the observed data points and their predicted values based on the remaining n-1 points. By substituting PRESS for the error sum of squares in the equation used for R^2, a 'prediction R^2' can be calculated (Montgomery & Peck, 1982, pp. 430–434). This indicates how much variability in predicting new observations the model is expected to explain.

Models using data up to the end of March and having a prediction R^2 greater than an arbitrary value of 40% for both techniques were further evaluated using test data from 1989 to 1992. These data were not used in the construction of the models. Predictions based on new data were calculated by substituting new values into the regression equation and confidence intervals at the 95% level were calculated for the predictions using the method suggested by Montgomery & Peck (1982), p. 125.

A regression model may provide accurate predictions when the new data lie within the range of those used in its construction, but be inaccurate when used outside that range, as the model contains no information about the relationships between the variables under the new conditions (Weisberg, 1985). It is therefore important to know when a prediction is based on an extrapolation of the model. In simple linear regression this is easily determined but, with multiple regression models, where the interpolation zone is represented by an ellipsoid in multidimensional space, it is possible for a new value to be outside this zone even if all of the regressor variables are within the range of those used to construct the model. The multiple regression models were tested for extrapolation using a method described by Montgomery & Peck (1982), p. 142.

Results and discussion

Of the 28 simple linear regression models with mean temperature alone, February was the most significant single month in determining the timing of the spring migration of *M. persicae* at Rothamsted (table 2, last column). The relationship with mean January temperature was very weak although models using time periods which contained both of these months almost always had a larger prediction R^2 than February alone. Similarly, the relationship with mean temperature in March was very weak but the inclusion of this month in a model often resulted in an increase in the prediction R^2. Models constructed using multiple regression appeared to follow a similar pattern (table 2) but comparisons between time periods are more difficult as the model selected for each uses different combinations of the weather variables. In general, mean screen tempera-

Table 2. Comparison of prediction R^2 of the models generated for each of the time periods using simple and multiple regression techniques.

Time period	Type of model		
	multiple		simple
	prediction R^2	influential weather variables*	prediction R^2
Nov	0	–	0
Nov–Dec	0	–	0
Nov–Jan	0	–	0
Nov–Feb	51	9,5,3	40
Nov–Mar	55	11,2,4,7,1	53
Nov–Apr	75	4,11,7,2,6	53
Nov–May	74	7,3,6,11,9,5	64
Dec	7	9,8	0
Dec–Jan	11	8,3,11,5	4
Dec–Feb	60	7,5,1,3,10,6	51
Dec–Mar	52	4,1,2,11,7,5	58
Dec–Apr	59	4,2,8,7,1	62
Dec–May	73	2,7,11,8,6	70
Jan	19	1,5	15
Jan–Feb	68	5,2,1,7,11	73
Jan–Mar	83	1,5	69
Jan–Apr	80	1,5	77
Jan–May	79	5,6,7,1,3	78
Feb	54	7,5,6,1,2,10,4	51
Feb–Mar	63	11,10,3,8,5	48
Feb–Apr	61	8,7,9,11,5	48
Feb–May	72	8,9,7,6,11	54
Mar	56	5,1,2,7,4,8	0
Mar–Apr	48	11,8,10,4,7	9
Mar–May	50	8,9,1,5	19
Apr	17	7,6,2,1,5	0
Apr–May	34	8,3,10	0
May	0	–	0

*Variables are as in table 1.

https://doi.org/10.1017/S0007485300029278

Table 3. Predicted and observed Julian date of first catch of *Myzus persicae* at Rothamsted, 1989–92 ± 95% confidence limits. Predictions in parentheses are based on extrapolations of the models producing them.

Model period	1989		1990		1991		1992	
	simple	multiple	simple	multiple	simple	multiple	simple	multiple
Nov–Feb	119 ± 34	(104 ± 31)	111 ± 35	(106 ± 39)	159 ± 33	(115 ± 47)	135 ± 33	(81 ± 49)
Nov–Mar	(110 ± 32)	(98 ± 34)	(101 ± 34)	108 ± 30	146 ± 30	(102 ± 49)	126 ± 31	(71 ± 58)
Dec–Feb	112 ± 31	118 ± 27	108 ± 32	(142 ± 34)	161 ± 30	149 ± 31	138 ± 29	(126 ± 33)
Dec–Mar	(106 ± 31)	117 ± 32	(101 ± 32)	(112 ± 32)	149 ± 28	161 ± 30	130 ± 29	145 ± 30
Jan–Feb	117 ± 23	117 ± 25	139 ± 22	(94 ± 29)	163 ± 23	160 ± 26	133 ± 23	133 ± 25
Jan–Mar	(109 ±2 5)	(90 ± 21)	125 ± 24	(111 ± 27)	146 ± 23	(108 ± 26)	123 ± 24	(111 ± 30)
Feb	126 ± 32	129 ± 31	(111 ± 34)	(112 ± 33)	168 ± 32	171 ± 31	132 ± 31	137 ± 30
Feb–Mar	(116 ± 34)	123 ± 25	(103 ± 36)	107 ± 26	145 ± 32	(140 ± 31)	(122 ± 33)	(107 ± 29)
Mar	131 ± 48	132 ± 31	(129 ± 49)	111 ± 31	131 ± 48	(129 ± 34)	133 ± 47	113 ± 31
Actual	92		112		174		99	

ture, accumulated day degrees above zero and soil temperature were influential variables in models based on data from the months of January, February and March. In models for the later periods examined, where the mean screen temperature was less important, minimum temperature, rainfall and sunshine duration tended to be more important. However, the model for March highlighted accumulated day degrees above zero and mean screen temperature as important variables, whereas simple regression showed the relationship with mean screen temperature to be very weak.

Models having a prediction R^2 greater than 40% were obtained from both techniques in 16 of the 28 time periods, and the eight of these 16 which used data up to the end of March were evaluated using new test data. Models for the month of March alone were also tested on new data in an attempt to resolve the above anomaly.

Despite having, on average, a prediction R^2 11 percentage points larger than the simple regression model for the same time period, the multiple regression models did not produce consistently more accurate predictions in 1989 to 1992 (table 3). The winter and spring of 1989 and 1990 were exceptionally mild in Britain and approximately half of the predictions for these years were based on extrapolations from the models. The relatively small size of the original dataset used in the models in this paper (23 years) means that predictions based on extrapolations could be a common occurrence with both techniques. This happened more frequently with the multiple regression technique (19 out of 36 cases) than with the simple regression technique (10 out of 36 cases). If no long-term changes in weather patterns were occurring, the likelihood of new data being outside the range of those used to construct the models would decrease as the dataset got larger. In a scenario of global climate change, however, this would not be the case.

The confidence limits associated with the predictions in table 3 were similar using both methods. The magnitude of the confidence limits reflects the variation in the date of first record which is not accounted for by the models and hence is usually greater for models having a small prediction R^2. From 1966 to 1988 the range of dates of first record of *M. persicae* at Rothamsted was between Julian days 102 and 205. Although the confidence limits of most predictions occupied approximately half of the year to year range, the mean error of interpolated simple and multiple models was much smaller than this (24 days and 23 days, respectively).

It is interesting to note that in 1989 and 1992, predictions based on extrapolations of both types of model performed better than those based on interpolations. The first *M. persicae* caught at Rothamsted in 1989 and 1992 were both earlier than the earliest aphid caught during 1966–1988 and therefore outside the range of the data on which the models were built. Since an interpolation from a regression model is unlikely to be outside the range of previous experience, predictions resulting from an extrapolation of a model would be expected to be closer to the actual date for these two years than those based on interpolations of the model producing them.

The relationship between mean temperature in the periods including January, February and March, and the date of the start of the spring migration of *M. persicae* at Rothamsted is particulary strong. As a result of this, the multiple regression method did not appear to increase significantly the accuracy of the predictions although, as discussed above, the addition of more years of data to the models may alter the relative performance of the two techniques. The multiple regression technique was able to identify relationships with other weather variables for time periods where the relationship with mean temperature was weak and, although these did not result in more accurate predictions in the present analysis, it is possible that the multiple regression models may be of more use for modelling other species of aphid, possibly at different sites, and other groups of insects where the relationship with mean temperature is weaker.

Empirical models such as the ones presented in this paper are potentially useful but their value may be increased by the incorporation of the results of labora-

https://doi.org/10.1017/S0007485300029278

tory and field experimentation. For example, a detailed knowledge of the cold-hardiness of each aphid species may suggest different thresholds for the accumulated temperature variables. A sound understanding of the factors which influence the overwintering survival of aphids can be of use in the rejection of models where the observed relationship comes about by chance alone. The multiple regression model for March appeared, from its prediction R^2, to be much more robust than the corresponding simple regression model but only resulted in a better prediction in 1990 and 1992. This highlights the importance of establishing a logical cause-effect relationship before a prediction model is used.

There is little point in using a complex analysis unless the models formed are more robust than those obtained from a simpler approach. Although the parameters included in the multiple regression models can be resolved into the original weather variables, the mechanism of the model is not immediately obvious and so cause-effect relationships can be less easy to establish than with the simple linear regression models. In order to make useful comparisons between these two methods it will be necessary to test both techniques on additional data from other years, and also to perform detailed comparative analyses on other species of aphid.

Acknowledgements

The authors are grateful to the staff of the Rothamsted Insect Survey and to Dr J.N. Perry for his constructive criticism of the manuscript. This research was funded by the Agricultural and Food Research Council (grant PG24/524).

References

Abraham, B. & Ledolter, J. (1983) *Statistical methods for forecasting*. 445 pp. New York (US), Wiley.

A'Brook, J. (1983) Forecasting the incidence of aphids using weather data. *Bulletin OEPP/EPPO Bulletin* **13,** 229–233.

Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16,** 125–127.

Barrett, J.P. (1974) The coefficient of determination—some limitations. *American Statistician* **28,** 19–20.

Dewar, A.M. (1986) Forecasting and control of virus yellows and aphids in sugar beet. *British Sugar Beet Review* **52,** 49–52.

Harrington, R., Dewar, A.M. & George, B. (1989) Forecasting the incidence of virus yellows in sugar beet in England. *Annals of Applied Biology* **114,** 459–469.

Harrington, R., Tatchell, G.M. & Bale, J.S. (1990) Weather, life cycle strategy and spring populations of aphids. *Acta Phytopathologica et Entomologica Hungarica* **25,** 423–432.

Harrington, R., Howling, G.G., Bale, J.S. & Clark, S.J. (1991) A new approach to the use of meteorological and suction trap data in predicting aphid problems. *Bulletin OEPP/EPPO Bulletin* **21,** 499–505.

Macaulay, E.D.M., Tatchell, G.M. & Taylor, L.R. (1988) The Rothamsted Insect Survey 12-metre suction trap. *Bulletin of Entomological Research* **78,** 121–129.

Montgomery, D.C. & Peck, E.A. (1982) *Linear regression analysis*. 504 pp. New York (US), Wiley.

Payne, R.W., Lane, P.W., Ainsley, A.E., Bicknell, K.E., Digby, P.G.N., Harding, S.A., Leech, P.K., Simpson, H.R., Todd, A.D., Verrier, P.J., White, R.P., Gower, J.C., Tunnicliffe Wilson, G. & Patterson, L. J. (1987) *GENSTAT 5 Reference Manual*. Oxford, Clarendon Press.

Tatchell, G.M. (1991) Monitoring and forecasting aphid problems. pp. 215–230 *in* Peters, D.C., Webster, J.A. & Chlouber, C.S. (Eds) *Aphid-plant interactions: populations to molecules Miscellaneous Publications. Oklahoma Agricultural Experiment Station* no. 132.

Turl, L.A.D. (1980) An approach to forecasting the incidence of potato and cereal aphids in Scotland. *Bulletin OEPP/EPPO Bulletin* **10,** 135–141.

Walters, K.F.A. & Dewar, A.M. (1986) Overwintering strategy and the timing of the spring migration of the cereal aphids *Sitobion avenae* and *Sitobion fragariae*. *Journal of Applied Ecology* **23,** 905–915.

Watson, M.A., Watson, D.J. & Hull, R. (1946) Factors affecting the loss of yield of sugar beet caused by beet yellows virus, I. Rate and date of infection, date of sowing and harvesting. *Journal of Agricultural Science* **36,** 151–166.

Weisberg, S. (1985) *Applied linear regression*. 324 pp. New York (US), Wiley.

(Accepted 8 March 1993)