The

Connecticut

Agricultural

Experiment

Station,

New Haven

Predicting Defoliation by the Gypsy Moth in Connecticut

BY RONALD M. WESELOH

Bulletin 938

October 1996

ABSTRACT

A procedure has been developed to predict forest defoliation by the gypsy moth, *Lymantria dispar* L., in Connecticut using past defoliation, egg mass counts, and physiographic features to fit logistic regression models. In a validation study, the models predicted the amount of defoliation quite accurately, and when defoliation was substantial, they also predicted the locations of defoliation accurately. However, when actual defoliation was very low, locations where defoliation had been predicted did not often correspond to actual areas of defoliation. The procedure should be useful in helping to make control decisions, and an automated implementation of the method has been developed and is described.

Predicting Defoliation by the Gypsy Moth in Connecticut

By Ronald M. Weseloh

For the last several years, gypsy moth, *Lymantria dispar* L, population numbers have been low in Connecticut. This appears to be due to the activity of a pathogenic fungus, *Entomophaga maimaiga* Humber, Shimazu & Soper, that was discovered in Connecticut in 1989 (Andreadis and Weseloh 1990). However, the ecology of this pathogen is not well enough known to predict with any certainty its impact on future gypsy moth numbers. Thus, the ability to predict forest defoliation by the gypsy moth in Connecticut is desirable to assist regulatory and management personnel. At present, winter egg mass counts on an 11.25 km grid system throughout Connecticut serve as rough guidelines for municipalities, state, and federal agencies. Towns can also request more detailed surveys. While the procedure provides valuable guidance to municipalities and citizens' groups that are concerned with gypsy moth control, interpretations are highly dependent on the experience of the predictor, and the detailed surveys can be time consuming. A more objective approach that relies exclusively on the statewide data would be desirable and less expensive.

Fortunately, there has been considerable progress in developing objective procedures for gypsy moth defoliation prediction, including the use of non-linear or linear regressions (Gage et al. 1990, Gansner et al. 1985, Williams et al. 1991, Montgomery 1990, and Liebhold et al. 1993). Perhaps the potentially most useful approach involves geographical information systems (GIS) and geostatistics (Liebhold et al. 1991, Liebhold et al. 1993a, Liebhold et al. 1993b, Hohn et al. 1993, Gribko et al. 1995) and cellular transition models (Zhou and Liebhold 1995), that explicitly take into account the spatial structure of the data.

Weseloh (in press) developed a procedure similar to a method described by Gribko et al. (1995) which included the use of physiographic features and estimates of previous defoliation and egg mass counts to predict future defoliation in Connecticut through application of a logistic regression model. This procedure was cumbersome, involving the use of kriging to interpolate egg mass counts, and having 10 independent variables and 12 interactions in the logistic regression. Kriging is a geostatistical technique that is useful for interpolating between sample points. Results have the lowest variance of all interpolation techniques, and kriging is especially useful when sample points are clumped. However, kriging requires knowledge of the spatial correlations in the data, and is computationally tedious (see Isaaks and Srivastava (1989) for further explanations of kriging and other geostatistical techniques). For routine use, a simpler procedure that can be easily automated would be desirable. This bulletin describes such a procedure and presents data on the method's ability to predict gypsy moth defoliation in Connecticut. An automated implementation of the process is described in the Appendix.

METHODS

In this study, extensive use was made of the Geographical Information System, IDRISI (Eastman 1993), for data and map analysis.

Sources of Data

U.S. Geological Survey Digital Elevation Model files obtained from the Map Library of the University of Connecticut were manipulated using IDRISI to produce a map of ground elevation above sea level at 2 km resolution for the State of Connecticut. Also, a Connecticut map at the same resolution giving percent of soil areas having poor drainage was constructed from data from the U.S. Department of Agriculture Soil Conservation Service, National Cooperative Soil Survey, Fort Worth, Texas.

Digitized maps of defoliation from 1969 to 1995 were obtained from aerial sketch maps of defoliation in Connecticut obtained from personnel in the State Entomologist's office. Details of the digitizing process are given in Weseloh (in press). Defoliation was coded at 5 levels: 0 = 0% to 9%, 1 = 10% to 25%, 2 = 26% to 50%, 3 = 51% to 75%, and 4 = 76% to 100%, and map resolution was 2 km. An average defoliation map was generated from the yearly maps from 1969 to 1994 by adding defoliation levels for each resolved point for all years and then dividing by the number of years.

From 1975 through 1988 in Connecticut, regulatory personnel carried out egg mass counts on 51 plots (0.023 ha) distributed on a square, 16 km grid in oak-dominated areas throughout the state. In these plots, all new egg masses seen on the trunks of trees were counted. From 1989 to the present, 102 plots distributed on a grid every 11.25 km have been sampled. The number of egg masses per acre (0.4 ha) was converted to the natural logarithm after 1 was added to each value. To convert the transformed egg mass counts to maps at 2 km resolution, values were interpolated using inverse-distance weighting (Isaaks and Srivastava 1989), in which the six closest sample points to the grid point for which the estimated value was desired were weighted by the inverse of the distance between the sample points and the estimated point. The weighted average of these sample points was used as the estimate. This method of interpolation was used instead of kriging because it is less computationally expensive and, because egg mass counts were distributed in a regular grid, should give comparable interpolations to kriging.

Defoliation Prediction

A separate logistic regression model (see Hosmer and Lemeshow 1989) was developed for each prediction year. (This procedure is different from that of Weseloh (in press), in which one model was used for prediction in all years.) The independent variables for each 2 km by 2 km map area were: transformed, interpolated egg mass counts for the 8 years previous to the prediction year; coded defoliation levels for the 2 years before each egg mass count; elevation; average defoliation; and percent of the soil region that had poor drainage. The last three variables did not change in the different years, and were simply repeated for each year. These variables had been found by Weseloh (in press) to be important in defoliation prediction. The dependent variable was the defoliation for the summer after each egg mass count was made, converted to 0 if < 25% defoliation occurred and 1 if >25% occurred. Four two-level interactions between independent variables were then constructed by multiplying: (1) average defoliation by interpolated egg mass counts, (2) average defoliation by coded defoliation for the year before egg mass counts, (3) elevation by interpolated egg mass counts, and (4) interpolated egg mass counts by coded defoliation for the year before egg mass counts were done.

To predict defoliation, parameters estimated from the fitting of the model were used with interpolated egg mass counts from the winter before the prediction year, coded defoliation values from the 2 years before the prediction year, elevation, average defoliation, % poor soil drainage, and the four interactions to obtain probability estimates at each 2 by 2 km area throughout Connecticut. Predictions were made only for 1985 and 1989-1995 because of deficiencies in egg mass data in other years.

Comparisons between predictions and actual defoliation were made by visual inspection and by aggregating values to 6 km resolution and calculating the correlation coefficient between actual and predicted defoliation pairs for each year.

RESULTS

Maps of elevation and average defoliation from 1969 to 1994 are presented in Fig. 1. Average defoliation is related to elevation over much of Connecticut except in the southeast, where persistent defoliation tends to occur. The characteristics leading to this persistence are not known.

Fig. 2 shows examples of interpolated egg mass numbers for 1990 as determined by kriging in Weseloh (in press) or inverse weighting as described in this paper. Results are very similar, confirming that when sample points are evenly spaced, inverse weighting is comparable to kriging, the generally preferred technique. The adoption of inverse weighting for interpolation greatly simplified the automation of the prediction procedure.

Maps showing predicted defoliation and actual defoliation for the different years are presented in Fig. 3. For the first 4 years (1985 and 1989-1991), patterns of expected and actual defoliation were remarkably similar, both in extent and location. For 1992-1995 this was not the case, probably because so little defoliation actually occurred in the later years. However, the amount of defoliation, if not the location, appeared to be adequately predicted in all years. This is shown in Fig. 4 (top), in which the number of 2 x 2 km cells of predicted defoliation that were above 0.10 probability are compared to the number of cells of actual defoliation for each year. There is good agreement between the amount of predicted and actual defoliation. However, when the correlation coefficients are compared (Fig. 4, bottom), only for the years in which defoliation was relatively extensive were correlations high, as is also evident from the maps of Fig. 3.

This procedure should be useful in providing quantitative, spatially-explicit information about defoliation in Connecticut. The predictions may not be more accurate than those made by experienced regulatory personnel, but the result in the form of a map of defoliation probabilities should be usable by more people. To aid in making the procedure more available, the process has been automated through use of the GIS "IDRISI" (Eastman 1993), and custom programs written by the author. Details are given in the Appendix.

REFERENCES CITED

Andreadis, T.G., and R.M. Weseloh. 1990. Discovery of *Entomophaga maimaiga* in North American gypsy moth, *Lymantria dispar*. Proc. Natl. Acad. Sci. USA 87: 2461-2465.

Eastman, J.R. 1993. IDRISI Version 4.1, Clark University Graduate School of Geography, Worcester, Massachusetts.

Gage, S.H., T.M. Wirth, and G.A. Simmons. 1990. Predicting regional gypsy moth (Lymantriidae) population trends in an expanding population using pheromone trap catch and spatial analysis. Environ. Entomol. 19: 370-377.

Gansner, D.A., O.W. Herrick, and M. Ticehurst. 1985. A method for predicting gypsy moth defoliation from egg mass counts. North. J. Appl. For. 2: 78-79.

Gribko, L.S., A.M. Liebhold and M.E. Hohn. 1995. Model to predict gypsy moth (Lepidoptera: Lymantriidae) defoliation using kriging and logistic regression. Environ. Entomol. 24: 529-537.

Hohn, M.E., A.M. Liebhold and L.S. Gribko. 1993. Geostatistical model for forecasting spatial dynamics of defoliation caused by the gypsy moth (Lepidoptera: Lymantriidae). Environ. Entomol. 22: 1066-1075.

Hosmer, D.W., and S. Lemeshow. 1989. Applied Logistic Regression. John Wiley & Sons. New York, 307 pp.

Isaaks, E.H. and R.M. Srivastava. 1989. Applied Geostatistics. Oxford University Press, New York, Oxford. 561 pp.

Liebhold, A.M., X. Xhang, M.E. Hohn, J.S. Elkinton, M. Ticehurst, G.L. Benzon and R.W. Campbell. 1991. Statistical analysis of gypsy moth (Lepidoptera: Lymantriidae) egg mass populations. Environ. Entomol. 20: 1407-1417.

Liebhold, A.M., R.E. Rossi, and W.P. Kemp. 1993a. Geostatistics and geographic information systems in applied insect ecology. Ann. Rev. Entomol. 38: 303-327.

Liebhold, A.M., E.E. Simons, A. Sior, and J.D. Unger. 1993b. Forecasting defoliation caused by the gypsy moth from field measurements. Environ. Entomol. 22: 26-32.

Montgomery, M.E. 1990. Role of site and insect variables in forecasting defoliation by the gypsy moth. pp. 73-84 in Watt, A.D., S.R. Leather, M.D. Hunter, and N.A.C. Kidd [eds.]. Population Dynamics of Forest Insects. Intercept, Andover, Hampshire, England. 408 pp.

Weseloh, R.M., 1996. Developing and validating a model for predicting gypsy moth (Lepidoptera: Lymantriidae), defoliation in Connecticut. J. Econ. Entomol. (in press).

Williams, D.W., R.W. Fuester, W.W. Metterhouse, R.J. Balamm, R.H. Bullock and R.J. Chianese. 1991. Oak defoliation and population density relationships for the gypsy moth (Lepidoptera: Lymantriidae). J. Econ. Entomol. 84: 1508-1514.

Zhou, G., & A.M. Liebhold. 1995. Forecasting the spatial dynamics of gypsy moth outbreaks using cellular transition models. Landscape Ecol. 10: 177-189.

ACKNOWLEDGMENTS

I wish to thank Morgan Lowry for her help in completing this project; Douglas Dingman for the use of a digitizing camera; and regulatory personnel in the office of Carol Lemmon, Deputy Connecticut State Entomologist, for providing the defoliation and egg mass data (all at The Connecticut Agricultural Experiment Station). I also thank Patrick McGlamery, Map Librarian (University of Connecticut) for making available Digital Elevation Model files. This research was supported by a U.S. Department of Agriculture Forest Service Cooperative Forest Health Program Focus Fund.

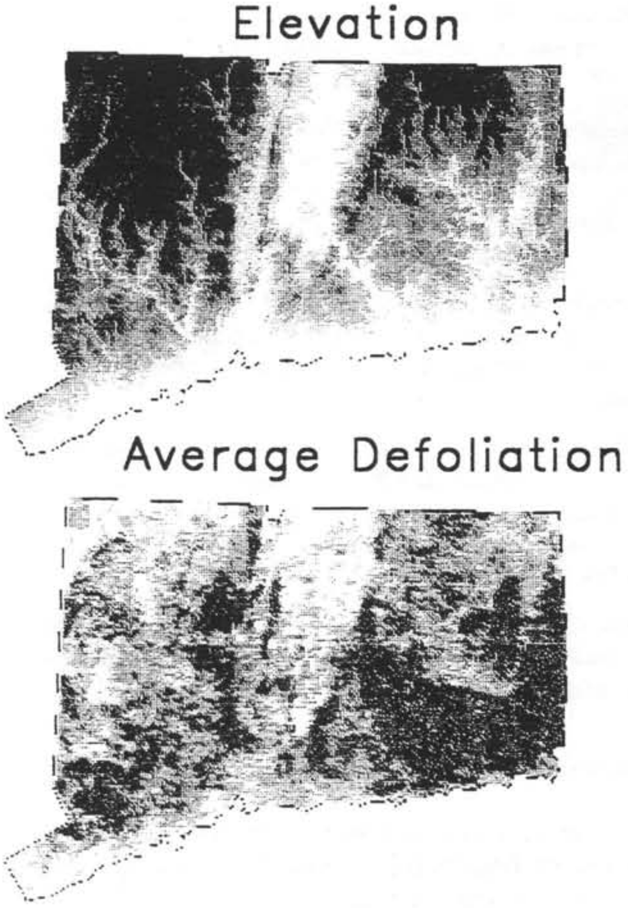

Fig. 1. Maps of Connecticut elevation in meters and average defoliation from 1969-1994. In this and subsequent maps, darker pixels have the higher values.

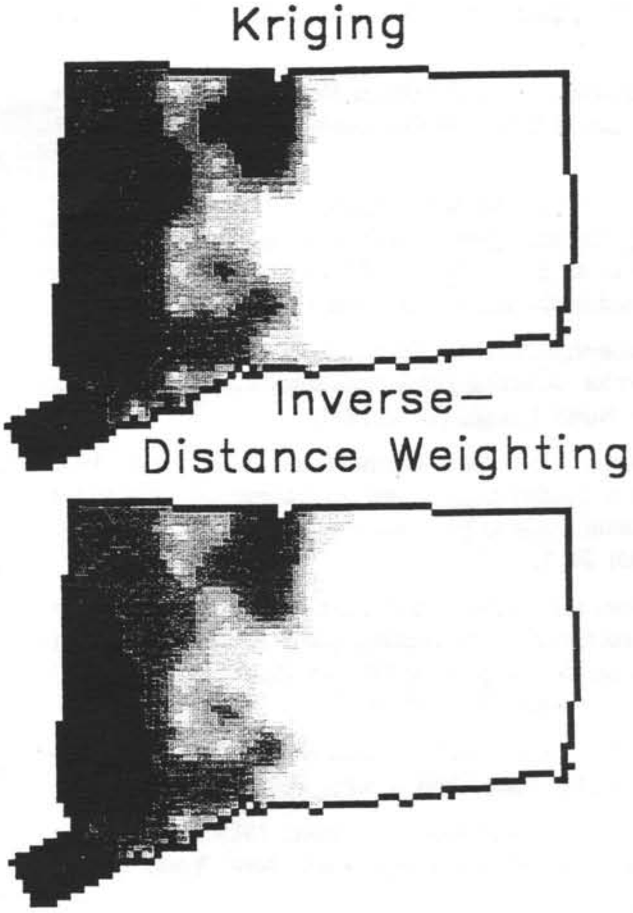

Fig. 2. Comparison of interpolated egg mass counts in Connecticut for 1990 when the interpolation procedure was kriging (top) or inverse distance weighting (bottom).

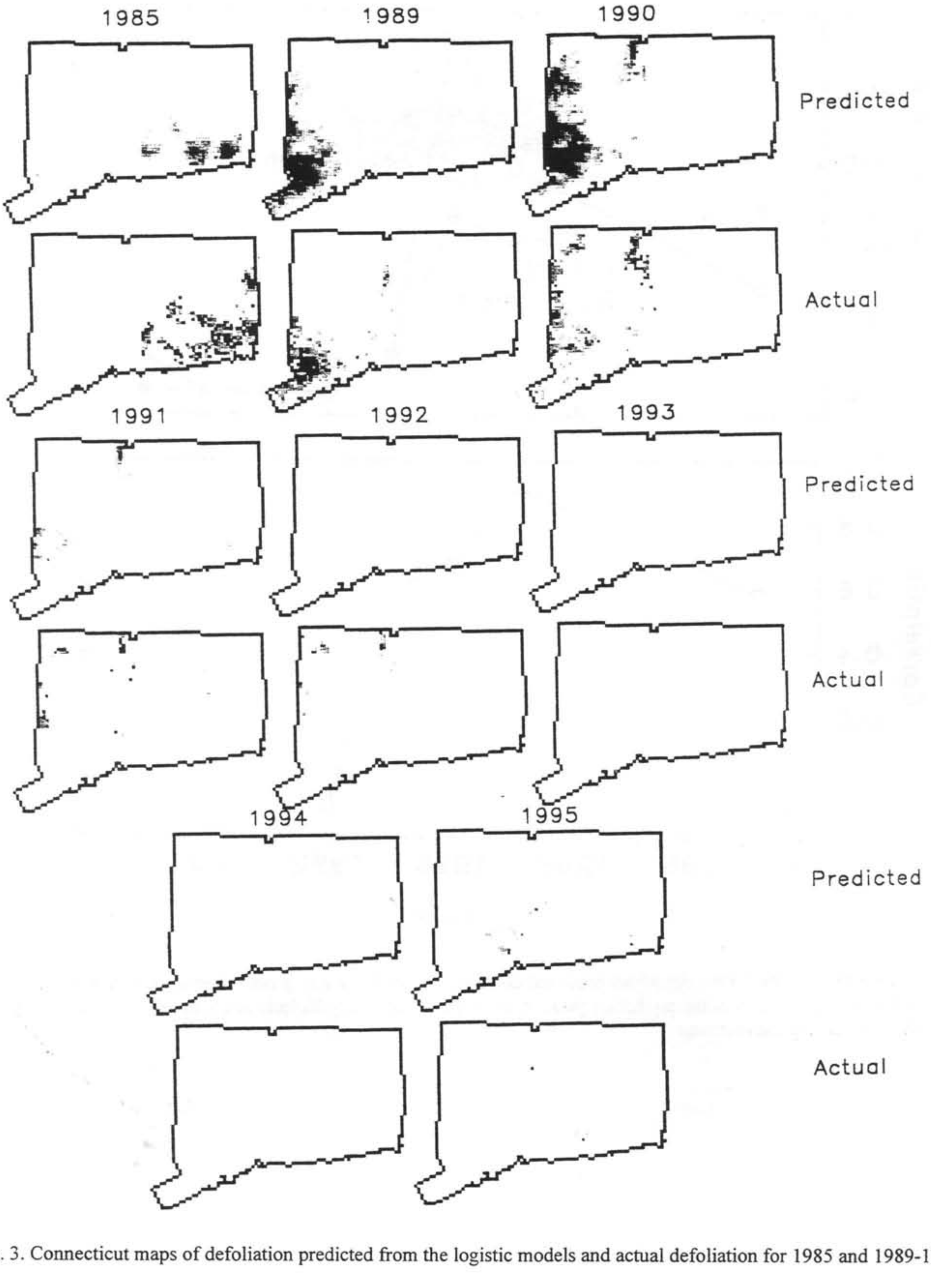

Fig. 3. Connecticut maps of defoliation predicted from the logistic models and actual defoliation for 1985 and 1989-1995.

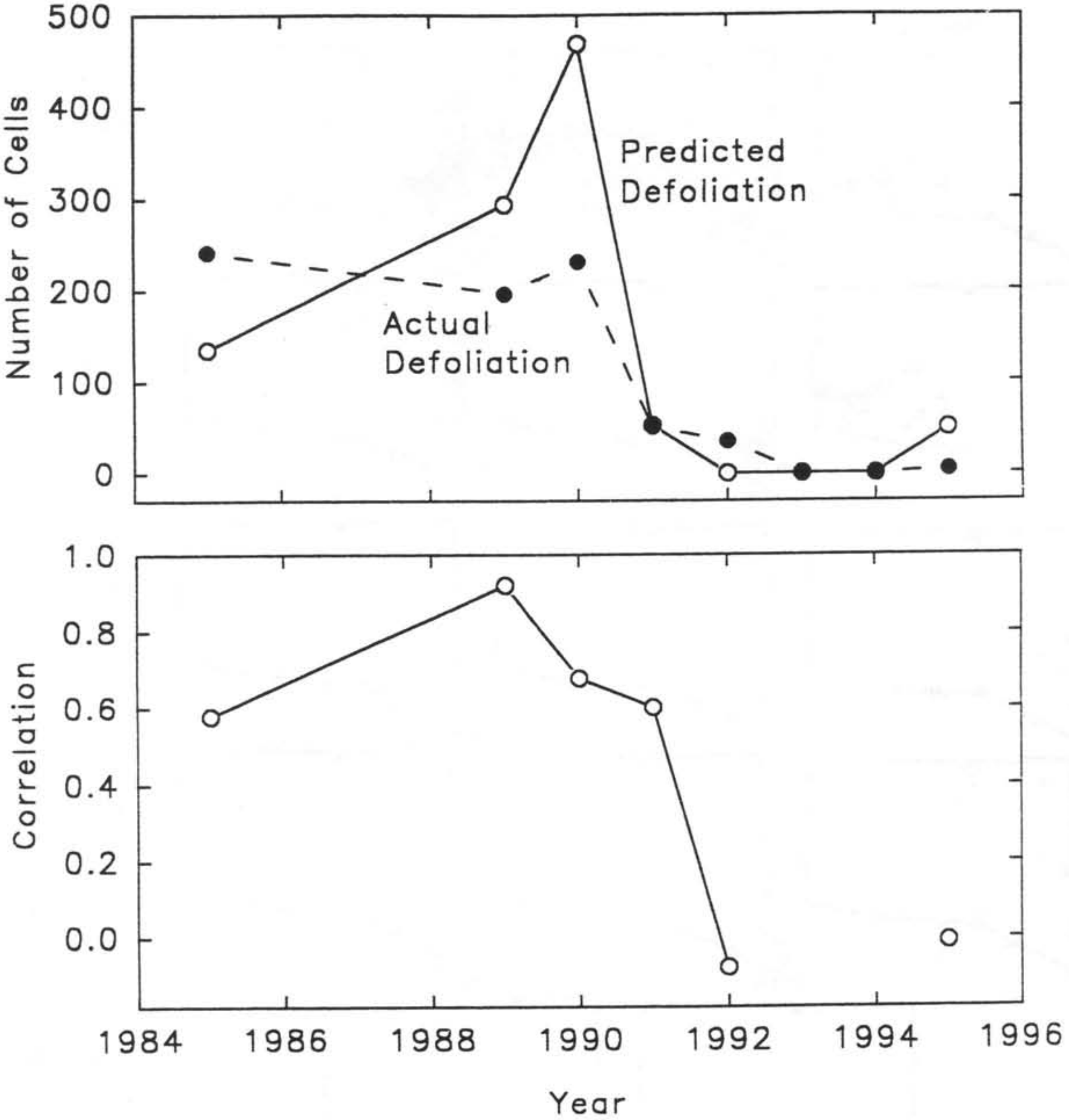

Fig. 4. Number of 2 km by 2 km cells where predicted defoliation probability was greater than 0.1 and number of cells where actual defoliation occurred in the prediction years (top), and correlation coefficients in the prediction years between predicted and actual defoliation (bottom).

APPENDIX

DESCRIPTION OF THE AUTOMATED PREDICTION PROCESS

This is a description of an automated procedure used to predict defoliation for a given year in Connecticut using spatial information about egg mass counts, defoliation data, elevation, soil drainage, and a 25-year average of defoliation, all mapped at a 2 km resolution throughout Connecticut. The procedure makes extensive use of and is integrated with the raster-based Geographic Information System Software called IDRISI (Eastman 1993). The software runs on an IBM compatible PC. The purpose of this description is to document the data files and software involved with each step of the process.

INPUT DATA

Before running the program, several data files are needed. A description of these files follows:

BASEDATA.DAT: This file contains most variables needed for prediction. It is a simple ASCII file that can be generated by any word processing program that will save the results in "Text Only" format. It was originally generated from IDRISI maps having the following characteristics: 84 columns by 64 rows, minimum longitude -73.75, maximum longitude -71.75, minimum latitude 40.92, maximum latitude 42.08. A binary "mask" was used to extract the data (3253 cases) from the areas of these maps that were within the boundaries of the State of Connecticut, as indexed by latitude and longitude. Each data case represents a particular point on the map. There is no header, and the variables are in adjacent columns in the order given:

Column: The number of the column of the original map at which the point is located.

Row: The number of the row of the original map at which the point is located.

Average Defoliation: The average defoliation at the point from 1969 to 1994, when defoliation has been coded as 0 (0%-9%), 1 (10-25%), 2 (26-50%), 3 (51-75%), or 4 (76-100% defoliation).

Elevation in Meters: of each point.

Percent Poor Drainage: The percent of the soil region within which the point is located that has poor drainage as determined in the State Soil Survey Geographic Data Base, USDA Soil Conservation Service.

Previous Defoliation: Thirteen successive variables of coded defoliation for the 13 years before the prediction year.

Previous Egg Mass Counts: Eight successive variables giving the natural logarithm of (egg mass counts/acre (0.4 ha) + 1) from the grid survey in Connecticut for the 8 winters before the winter of the year for which prediction is desired. The IDRISI module INTERPOL was used to interpolate values from the grid survey to map points using distance-weighted averaging, with the weights being the reciprocal of the distance of the grid survey counts to the map point.

NEWMASS.DAT: A file of the egg mass counts in the winter before the year for which prediction is desired. It is a non-headed ASCII file in which the first column has the grid point identification (in the form of "M-1" [Note: the letter MUST be capitalized]) and the second column has the number of egg masses per acre (0.4 ha).

NEWDEF.IMG and NEWDEF.DOC: These are an IDRISI image file and its corresponding document file, respectively, for the defoliation in the year before prediction is desired. The image is coded and has the same ranges of latitude and longitude as already described, but with 504 columns and 387 rows.

SMALLBOR.IMG and SMALLBOR.DOC: These are an IDRISI image file and its corresponding document file. The image file has "1" at points within Connecticut and "0" otherwise, as indexed by latitude and longitude. The bounds of the image and the number of rows and columns are the same as described for the data in the file "BASEDATA.DAT".

TOWNS.VEC and TOWNS.DVC: These are an IDRISI vector file and its corresponding document file of town boundaries in Connecticut, as indexed by latitude and longitude.

DESCRIPTION OF PREDICTION PROCEDURE

The prediction of defoliation is done by execution of the batch file called "DEFPRED.BAT". This file is reproduced below:

```
echo off
eggvec x \defmap\newmass.dat \defmap\t1
interpol x 1 t1 t2 1 1 1.0 Y 1 2 84 64
scalar x t2 t3 4 1000
overlay x 3 t3 smallbor newmass
dotitle \defmap\newmass.doc \defmap\newmass.doc
color x a newmass grey y 0 0 0 -1 0 -1 towns 14 y
scalar x newmass t1 1 1
overlay x 3 t1 smallbor t2
pointvec x t2 t3
convec \defmap\t3.vec \defmap\datamass.dat
contract x newdef t1 1 6 6
scalar x t1 t2 1 1
overlay x 3 t2 smallbor t3
pointvec x t3 t4
convec \defmap\t4.vec \defmap\datadef.dat
dofiles
cls
echo FITTING DATA TO LOGISTIC MODEL WILL TAKE ABOUT AN HOUR
deflogic x 10 2 3 4 5 6 7 8 9 10 11 1
forcast
color x a lastyear ibm y 0 0 0 -1 0 -1 towns 14 y
color x a nextyear ibm y 0 0 0 -1 0 -1 towns 14 y
```

When started, this batch file controls the execution of a series of programs that are described, in the order in which they are executed, below:

EGGVEC A program that converts the winter's egg mass file, "NEWMASS.DAT", into an IDRISI vector file ("T1") whose x and y values are specified as the latitude and longitude of the grid survey points. Also, 1 is added to the egg mass counts per acre (0.4 ha), converted to the natural logarithm, multiplied by 1000, and saved as an integer. (This last step is required because the next program will only execute successfully with vector files that have attributes as integers.)

INTERPOL An IDRISI program that is used to interpolate values (using distance-weighted averaging as described above) at the grid counts in the vector file "T1" and save results to a map called "T2".

SCALAR An IDRISI program that divides every value of "T2" by 1000 and saves the results in "T3". This step restores the egg mass data to its log values.

OVERLAY An IDRISI program that multiplies each point in "T3" by the corresponding value in the IDRISI map "SMALLBOR" and stores the results in the IDRISI map, "NEWMASS". Because the points in SMALLBOR have the value "1" only in the portions of the map that include Connecticut and "0" otherwise, this operation serves to mask out areas of "T3" that are not included in Connecticut.

DOTITLE A program than changes the Title of "NEWMASS" from the default.

COLOR An IDRISI program that displays the interpolated egg mass map "NEWMASS" so it can be checked for accuracy.

SCALAR Used again, this time to add the value "1" to each data point in "NEWMASS". Results are saved to a file named "T1".

OVERLAY Used to mask out the added values of "1" in all areas not covered by the State of Connecticut in "T1". The new map is called "T2". The result of the last two operations is to ensure that every part of the map that is covered by Connecticut has a value larger than 0.

POINTVEC An IDRISI program that converts the image map, "T2", to an IDRISI vector file of points ("T3") for all parts of the map in Connecticut. This program only saves points where the value is greater than 0, which is the rational for the last 2 steps.

CONVEC Converts the vector file "T3" into an ASCII file that has the natural logarithm of (egg mass counts +1) in the first column, the longitude of the point in the second column, and the latitude of the point in the third column. The resulting file is called "DATAMASS.DAT".

CONTRACT An IDRISI program that takes the map, "NEWDEF", and shrinks the number of rows and columns by 6, from 504 columns by 387 rows to 84 columns by 64 rows in the file called "T1". The contraction is done by removing values rather than averaging. This makes the defoliation map comparable to the interpolated egg mass map.

SCALAR Used to add 1 to every value in "T1" and save the results to "T2".

OVERLAY Masks out the added values of 1 in all areas not covered by the State of Connecticut in "T2". The resulting map is called "T3".

POINTVEC Converts the map, "T3", to an IDRISI vector file of points, "T4".

CONVEC Converts the vector file, "T4", into an ASCII file comparable to "DATAMASS.DAT", except that the first column contains the defoliation coding (1-4) from the defoliation maps. The resulting file is called "DATADEF.DAT".

DOFILES Combines the files "BASEDATA.DAT", "DATADEF.DAT" and "DATAMASS.DAT" for use in the logistic regression program. Data for the oldest defoliation year and egg mass year in "BASEDATA.DAT" are dropped, defoliation from the last year contained in "DATADEF.DAT" and the egg mass data contained in "DATAMASS.DAT" are added, and the resulting file is saved as "FULLDATA.DAT". This file will later be used for prediction once the logistic model has been fitted. The program also constructs the file that will be used to develop the logistic regression model by building "TEMP.DAT". This file has in successive columns: (1) a binomial variable of defoliation for a particular year ("0" if <26% defoliation, "1" if >25% defoliation) that serves as the dependent variable, (2) the average coded defoliation from 1969 to 1994, (3) the elevation in meters, (4) the percent of each soil area that has poor drainage (as given by the State Soil Survey Geographic Data Base, USDA Soil Conservation Service), (5) interpolated natural logarithm of (egg mass numbers + 1) for the winter before the defoliation represented by the dependent variable occurred, (6) coded defoliation for the year before the winter of the egg mass counts and (7) for the year before that, and 4 interactions produced by multiplying: (8) average 25-year defoliation by interpolated natural logarithm of (egg mass numbers + 1) (Interaction 1), (9) average defoliation by coded defoliation the year before (Interaction 2), (10) elevation by interpolated natural logarithm of (egg mass numbers + 1) (Interaction 3), and (11) interpolated natural logarithm of (egg mass numbers + 1) by coded defoliation the year before (Interaction 4). For example, if defoliation is to be predicted for 1995, "TEMP.DAT" would include binarized defoliation for 1987 (dependent variable) for each point in Connecticut. This would be matched with egg mass data from 1987, with coded defoliation from 1985 and 1986, as well as the data on average defoliation, elevation, poor soil drainage, and interactions. Appended to this would be similar binarized defoliation for 1988 and so on until all 8 previous years up to and including 1994 had been added. Thus, the data on average defoliation, elevation, and poor soil drainage would be repeated 8 times, once for each set of defoliation and egg mass data. When this data file has been assembled, it is scaled by dividing each variable by the average value of that variable. The results are stored in a file called "LOGIT.DAT", and the averages of the variables are stored in a file called "SCALEPAR.DAT". The scaling is done to stabilize calculations in the next step.

DEFLOGIC Fits a logistic regression model to the data in the file "LOGIT.DAT". The dependent variable is the binarized defoliation data for each year with the independent variables being average defoliation from 1969 to 1994, elevation in meters, percent of poor soil drainage, interpolated natural logarithm of egg mass numbers for the winter before the year of the dependent variable, coded defoliation data from 1 year and 2 years before the egg mass data was obtained, and the interaction terms described above. The program produces an output listing the maximum likelihood loss function of the fitted model and parameter values, standard errors, and Wald statistic associated with each independent variable. The parameter values are saved to a file called "LOGITPAR.DAT".

FORCAST uses the variable averages stored in "SCALEPAR.DAT" to rescale the parameter values stored in "LOGITPAR.DAT", and then uses these rescaled parameters and the data stored in "FULLDATA.DAT" and the logistic regression model to predict defoliation for the year before and the year for which prediction is desired. It then saves the predicted values into IDRISI image files called "LASTYEAR" and "NEXTYEAR", and also constructs the associated IDRISI document files.

COLOR An IDRISI program that displays the predicted defoliation from the last year as a check to determine how well the model fits to known data, and then displays the predicted defoliation for the year of interest.

OTHER CONSIDERATIONS

Every time the program is run, it constructs a new "LASTYEAR" and "NEXTYEAR" image. Thus, to save a particular prediction map, it should be renamed. Future predictions can easily be made as long as the "FULLDATA.DAT" file constructed for the predictions of the year before is available. Just rename this file as "BASEDATA.DAT". The "NEWMASS.DAT" and "NEWDEF" files should be prepared for the new year, and the programs executed by typing "defpred".

The Connecticut Agricultural Experiment Station, founded in 1875, is the first experiment station in America. It is chartered by the General Assembly to make scientific inquiries and experiments regarding plants and their pests, insects, soil and water, and to perform analyses for State agencies. The laboratories of the Station are in New Haven and Windsor; its Lockwood Farm is in Hamden. Single copies of bulletins are available free upon request to Publications; Box 1106; New Haven, Connecticut 06504. ISSN 0097-0905