A method for predicting the probability of species occurrence using data from systematic surveys

M. G. LE DUC, M. O. HILL and T. H. SPARKS

Institute of Terrestrial Ecology, Monks Wood Experimental Station, Abbots Ripton, Huntingdon, Cambs., PE17 2LS

ABSTRACT

Presence data for species in 2-km squares, recorded systematically during the B.S.B.I. Monitoring Scheme, were smoothed to derive probability response surfaces for *Euonymus europaeus* L., *Hyacinthoides non-scripta* (L.) Chouard ex Rothm., *Trientalis europaea* L. and *Veronica montana* L. Logistic regression was used to predict species frequencies from the response surfaces together with information on species occurrence in 10-km squares. Predicted frequencies were compared with those reported in some recent county floras. Agreement was generally good, but county differences in recording intensity were apparent.

INTRODUCTION

Accurate information on the spatial distribution of plants is now needed more than ever as human impacts on the environment intensify. Agricultural expansion and intensification (Green 1989), atmospheric pollutants (e.g. nitrogen compounds – see Tamm 1991) and climate change (Huntley *et al*. 1989) – thought to be a consequence of the increasing release of 'greenhouse' gases – are all seen to result in habitat change and species loss. Some gains are also to be expected as governments try to reduce agricultural surpluses by extensification and habitat creation, for example planting new woods on farms (Insley 1988).

Currently, and perhaps foreseeably, it is not possible to predict the presence or absence of a species from a knowledge of environmental factors and autecological characteristics alone. Prediction is dependent on good floristic survey data. Plant distribution maps with 10-km square resolution (Perring & Walters 1976), and local floras with tetrad (2-km square) resolution, are examples of such data for Britain. Dony (1963) described how floristic surveys can be used to predict the numbers of species occurring in tetrads. Hill (1991) demonstrated a method for using environmental data to estimate the probability of finding bird and plant species in 10-km squares. He concluded that the quality of the estimates varied with habitat preference, and that those species with strong edaphic requirements (e.g. *Helianthemum nummularium* (L.) Miller) were only poorly predicted in a broad-scale analysis.

For many species, the frequency of occurrence in tetrads provides a better indication of local abundance than a map of distribution at the 10-km square scale. However, a complete survey of vascular plants in Britain and Ireland at the tetrad scale would hardly be feasible, even if it were desirable. Fortunately, a systematic survey of a selected subset of tetrads not only is feasible but was accomplished by the B.S.B.I. Monitoring Scheme (Rich & Woodruff 1990). Data from this survey can be used to estimate the probability of finding species in tetrads that were not surveyed, and hence give an indication of local frequency.

The main purpose of this paper is to develop and compare methods for estimating such probabilities, using data from the Monitoring Scheme and other systematic surveys. In addition, we show how probability estimates can be used to generate species frequency maps at the national scale.

MATERIALS AND METHODS

The B.S.B.I. Monitoring Scheme (funded by the Nature Conservancy Council and the Department of the Environment, Northern Ireland) was a survey carried out in 1987 and 1988 and administered

TABLE 1. SAMPLING STATISTICS FOR THE B.S.B.I. MONITORING SCHEME
The British subset used in this work excludes data from Ireland and the Channel Islands.

Sample units	Sample size	Actual number surveyed	Number in British subset
10-km squares	429	425	298
Tetrads (A, J & W only)	1114	1080	796
Mean number of tetrads per 10-km square	2·60	2·54	2·67

through *B.S.B.I. News* (see Ellis 1986; Rich 1986, 1987, 1988, 1989). For the survey, one in nine of the 10-km squares were systematically selected from the British and Irish National Grids. Within selected 10-km squares, presence records for plant species were recorded in each of three systematically positioned tetrads (designated A, J and W). Some tetrads did not contain land, so that, on average, slightly fewer than three tetrads were sampled per 10-km square (Table 1). The Monitoring Scheme data are held by the Biological Records Centre (B.R.C.) at the Institute of Terrestrial Ecology (I.T.E.), Monks Wood. They are in ORACLE database format on a VAX computer cluster running under VMS (Rich & Woodruff 1990).

Records of species presence or absence in tetrads were smoothed to a response surface whose z-axis value is the probability of finding that species in the local tetrad. Each smoothed value is a weighted average of the neighbouring values, with weights specified by the bivariate Gaussian function with a root-mean-square deviation 30 km (Fig. 1). This smoothing radius was chosen because 30 km is the spacing of the Monitoring Scheme 10-km squares. A smaller radius would result in a response surface that showed marked local variation, reflecting frequencies in individual 10-km squares.

The smoothed value is

$$p_i = \frac{\sum_{k=1}^{n} w_k \alpha_{ik}}{\sum_{k=1}^{n} w_k}$$

where $w_k = \exp(-(x_k^2 + y_k^2)/r^2)$, p_i = estimated probability of finding the i^{th} species in the target tetrad, w_k = weight assigned to the k^{th} tetrad in the sample area, α_{ik} = value (1 or 0) specifying presence or absence of the i^{th} species in the k^{th} tetrad, r = smoothing radius (30 km), and x_k and y_k are the easting and northing distances of the k^{th} tetrad from the target tetrad. A smoothing radius of 30 km ensures that 98% of the weight comes from within a 60 km radius. Note that the summation is taken over tetrads surveyed for the Monitoring Scheme. A tetrad near the coast is given a smoothed value by averaging over nearby tetrads inland. This average is taken over a smaller number of points than for a non-coastal position, but is not otherwise affected by proximity to the sea.

Since presence and absence data are not normally distributed, the method of logistic regression analysis (cf. Jongman *et al.* 1987) was used to estimate species frequency in 10-km squares. Each Monitoring Scheme 10-km square was allocated a species frequency value which was calculated as the ratio of the number of occupied tetrads to the number of recorded (maximum three) tetrads. These values were regressed against the mean of the expected probabilities, estimated from the response surface, averaging probabilities over all the tetrads (25 maximum) within that square. Two models were considered: firstly, a model using only the spatially smoothed probability as independent variable (Model 1 below); secondly, a model (Model 2) using the spatially smoothed probability together with 10-km presence and absence data. For this purpose, 10-km data were obtained from the records held by B.R.C. at I.T.E., Monks Wood. These data comprise validated plant records from a variety of sources and were the records used to plot the *Atlas of the British flora* (Perring & Walters 1976).

The regression models, fitted by means of generalized linear modelling using the GENSTAT computer package, were

$$\log_e\left(\frac{q_i}{1-q_i}\right) = a_i + b_i\bar{p}_i \qquad \text{Model 1}$$

$$\log_e \left(\frac{q_i}{1-q_i}\right) = a_i + b_i\bar{p}_i + c_iB_i \qquad \text{Model 2}$$

where q_i = probability of finding the i^{th} species in a given tetrad of a Monitoring Scheme 10-km square, $\bar{p}_i$ = mean estimated probability of occurrence smoothed over the tetrads in the 10-km square, B_i = presence or absence (one or zero) of the i^{th} species in the 10-km square, and a_i, b_i & c_i are constants.

The accuracy of the smoothed probability surface was further investigated using a validation set of data from independent surveys obtained from a selection of those English county Floras meeting three criteria. Firstly, publication had to be relatively recent; secondly, records had to be available in atlas form for ease of data extraction; thirdly, mapping had to be at tetrad or 1-km square resolution. Those selected were for Bedfordshire (Dony 1976), Devon (Ivimey-Cook 1984), Durham (Graham 1988), north-east Essex (Tarpey & Heath 1990), Hertfordshire (Dony 1967), Kent (Philp 1982), Leicestershire (Primavesi & Evans 1988) and Sussex (Hall 1980). None of the available atlases from Wales or Scotland met the criteria (McCosh 1988). Only those 10-km squares falling wholly within the county (or vice-county) boundaries were considered. For each species and each 10-km square a table of presences out of the number of tetrads per 10-km square (25) was produced. For the north-east Essex Flora the published data are for 1-km squares and were summarized for each tetrad prior to processing.

Data from the county atlases were compared with both point estimates from simple Gaussian smoothing and predicted values from each of the logistic regression models. The basis for the comparison was the average number of presences in tetrads per 10-km square, county by county. Analysis of variance was used to test the significance of differences. Accuracy of predictions was measured by the root-mean-square difference between predicted and observed values.

To illustrate the technique we have selected four species, namely *Euonymus europaeus* L., *Hyacinthoides non-scripta* (L.) Chouard ex Rothm., *Trientalis europaea* L. and *Veronica montana* L. *E. europaeus* is a southern species of calcareous soils. *T. europaea* is a boreal species having a requirement for cooler northern winters. The other two species are generally distributed in older woodlands, but *H. non-scripta* is much the commoner of the two. Tetrad presences and absences (obtained from the B.S.B.I. Monitoring Scheme database) for each species have been plotted in Fig. 2. Version 6 of the UNIRAS computer package (I.U.C.C. Information Services Group 1989) was used for this and subsequent distribution maps and figures. Orkney and Shetland have been omitted. For them, as for the Isle of Man (which was included, but which had only three tetrads), a larger smoothing radius than 30 km might be desirable.

RESULTS

The response surfaces obtained by Gaussian smoothing are illustrated in Fig. 3. Regression coefficients and significance levels for Models 1 and 2 are shown in Table 2. Highly significant results can be expected because the independent regression variables were derived from the observed values (dependent variables) by smoothing. Both Models 1 and 2 contain the derived variable $\bar{p}_i$.

The comparisons between county atlas records and the estimated values from Gaussian-smoothed and regression models are shown in Table 3. The Gaussian-smoothed values were obtained by summing p_i for each tetrad in the 10-km square; predicted values from Models 1 and 2 were obtained by inserting appropriate $\bar{p}_i$ values into the regression equations to obtain values of q_i. Although many of the estimated values were close to those expected from the county Floras there were some substantial differences (Table 3).

The mean deviation (bias) was smallest for the prediction method using Model 2, but the bias of all three methods was small and not statistically significant (Table 4). The root-mean-square error for Model 2 was less than for Model 1 and approached that of the Gaussian-smoothed probabilities. The analysis of variance shows no effect due to species but a highly significant county effect. The bulk of the county effect can be attributed to underestimation by the three methods of tetrad frequencies in Kent and possible over-estimation in Bedfordshire.

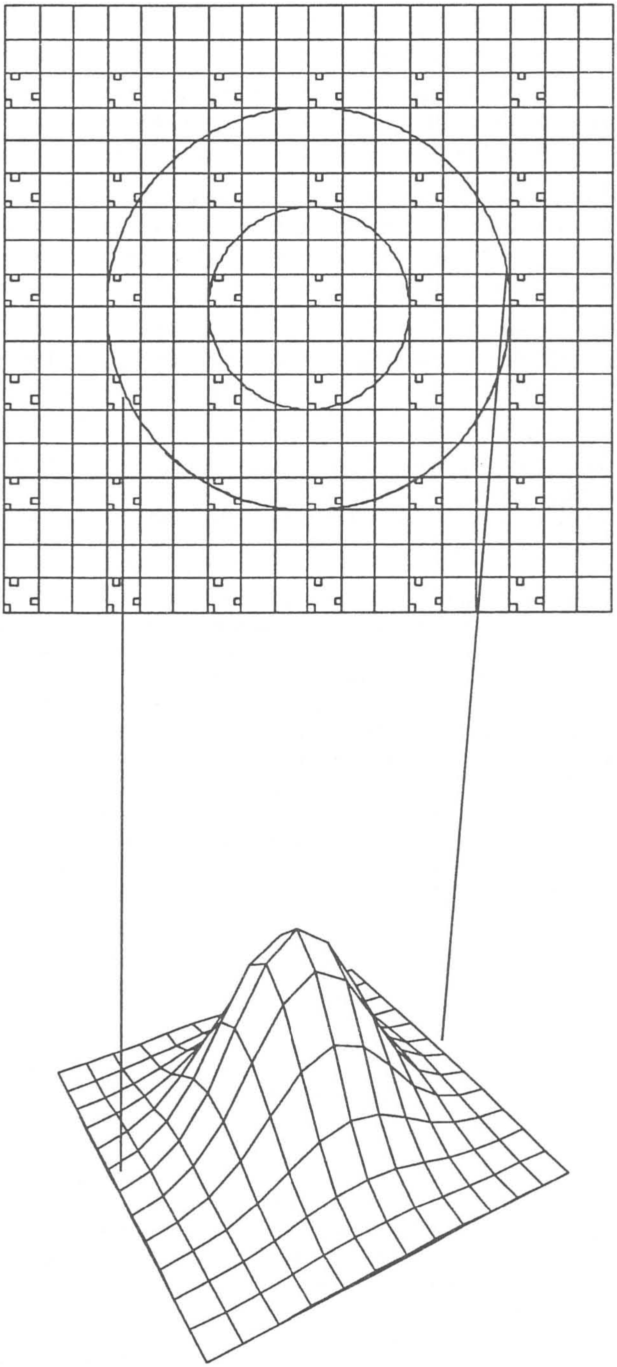


FIGURE 1. Gaussian smoothing of occurrence in tetrads (2-km squares). At any point in Britain the probability of a species being found in that tetrad is estimated as a weighted mean local frequency. Weights are defined by a Gaussian function with root-mean-square deviation 30 km. The diagram shows the weight function projected on to a 10-km square grid with the A, J and W tetrads for the one-in-nine sample indicated.

TABLE 2. LOGISTIC REGRESSION OF OBSERVED AGAINST PREDICTED FREQUENCIES IN 10-KM SQUARES OF THE B.S.B.I. MONITORING SCHEME

Coefficients a_i, b_i and c_i are defined in the text for Models 1 and 2. Degrees of freedom were (1, 296) for Model 1 and (2, 295) for Model 2.

Species	a_i	b_i	c_i	Deviance explained (%)	Significance
		Model 1			
Euonymus europaeus	−4·27	8·50	—	74·3	p<0·001
Hyacinthoides non-scripta	−3·86	7·40	—	69·2	p<0·001
Trientalis europaea	−5·51	11·71	—	83·3	p<0·001
Veronica montana	−3·90	9·29	—	62·7	p<0·001
		Model 2			
Euonymus europaeus	−10·35	6·35	7·34	80·8	p<0·001
Hyacinthoides non-scripta	−10·28	6·86	6·80	71·9	p<0·001
Trientalis europaea	−10·20	7·68	6·73	87·9	p<0·001
Veronica montana	−10·81	7·65	7·67	72·1	p<0·001

TABLE 3. OBSERVED (1) AND PREDICTED (2–4) NUMBERS OF TETRADS OCCUPIED BY SPECIES PER 10-KM SQUARE IN SELECTED COUNTIES

1 – average number of tetrads occupied according to the county atlases; 2 – expected value using the Gaussian-smoothed Monitoring Scheme data; 3 and 4 – expected values using regression models 1 and 2 respectively.

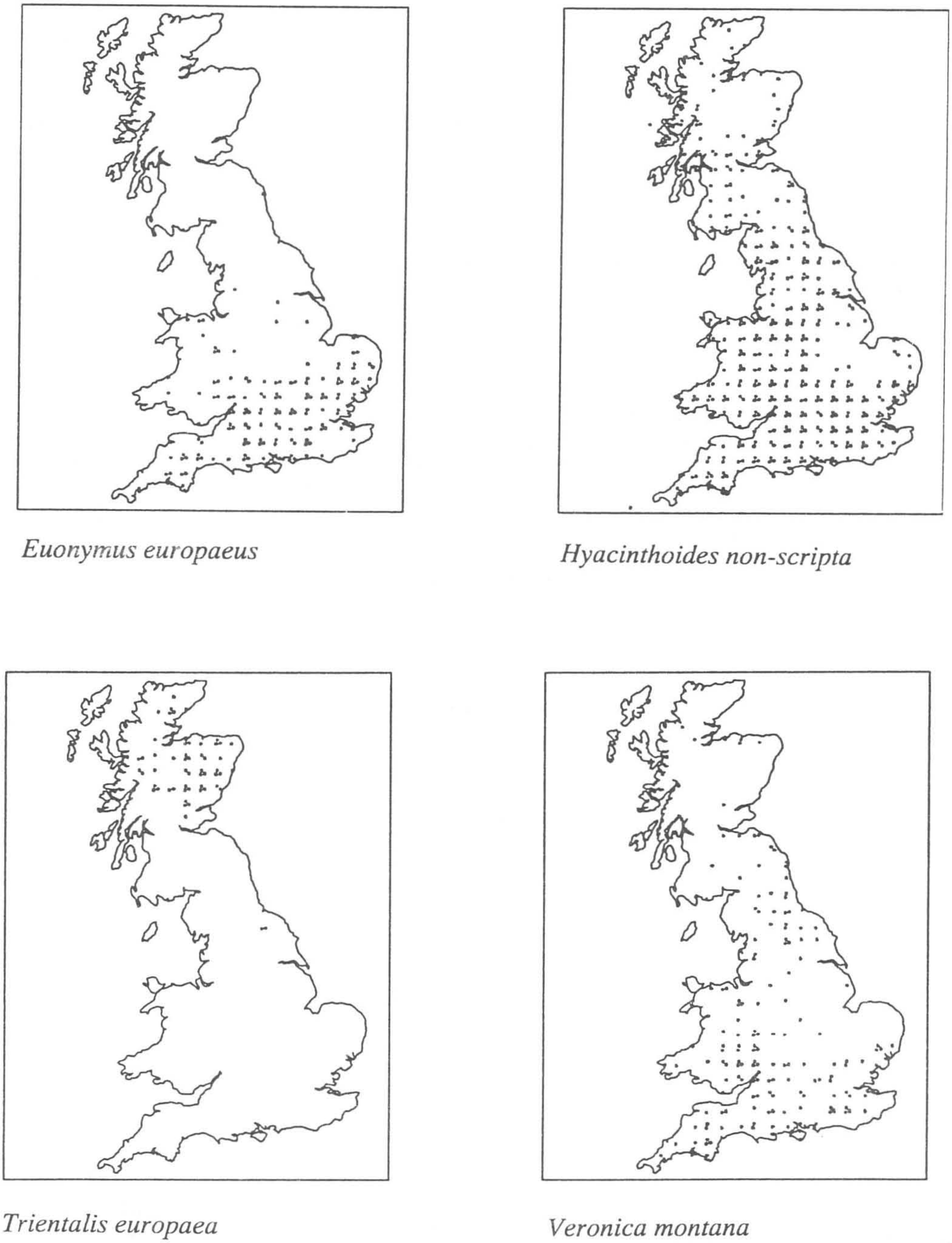
	Beds.	Devon	Durham	Essex	Herts.	Kent	Leics.	Sussex	Total
n	5	49	15	7	5	22	10	24	137
				Euonymus europaeus					
1	11·4	9·8	0·0	13·0	15·2	19·7	0·4	15·4	11·0
2	15·7	14·1	0·0	15·1	19·6	8·0	0·5	11·4	10·4
3	18·4	15·4	0·4	17·5	22·9	4·8	0·4	11·3	10·7
4	18·0	11·4	0·3	13·0	21·9	7·0	0·7	11·3	9·4
				Hyacinthoides non-scripta					
1	14·8	20·1	13·1	15·6	20·2	23·2	14·4	23·8	19·6
2	20·1	18·8	17·0	19·4	22·8	13·2	8·3	18·4	17·1
3	21·8	20·4	18·8	21·6	23·6	12·8	5·8	20·4	18·2
4	21·8	20·4	16·7	21·4	23·4	13·2	5·7	20·4	18·0
				Trientalis europaea					
1	0	0	0	0	0	0	0	0	0
2	0	0	0·2	0	0	0	0	0	0
3	0·2	0·1	0·1	0·1	0·2	0·1	0·1	0·1	0·1
4	0	0	0·1	0	0	0	0	0	0
				Veronica montana					
1	0·8	12·0	6·9	5·0	11·0	14·2	3·6	16·0	11·1
2	4·1	13·2	11·4	7·2	10·0	6·1	3·1	10·5	9·9
3	2·4	17·0	14·5	6·0	11·5	5·9	1·5	12·4	11·7
4	2·7	15·2	11·8	6·7	12·1	6·7	1·4	11·9	10·9

n = number of 10-km squares.

DISCUSSION

Smoothed distribution maps (Fig. 3) demonstrate the potential of the Monitoring Scheme data for depicting probabilities of occurrence in tetrads. Similar smoothed maps could be used in future to compare survey and re-survey results given a common survey protocol.

The ability of all the methods, including simple Gaussian smoothing and regression, to predict the



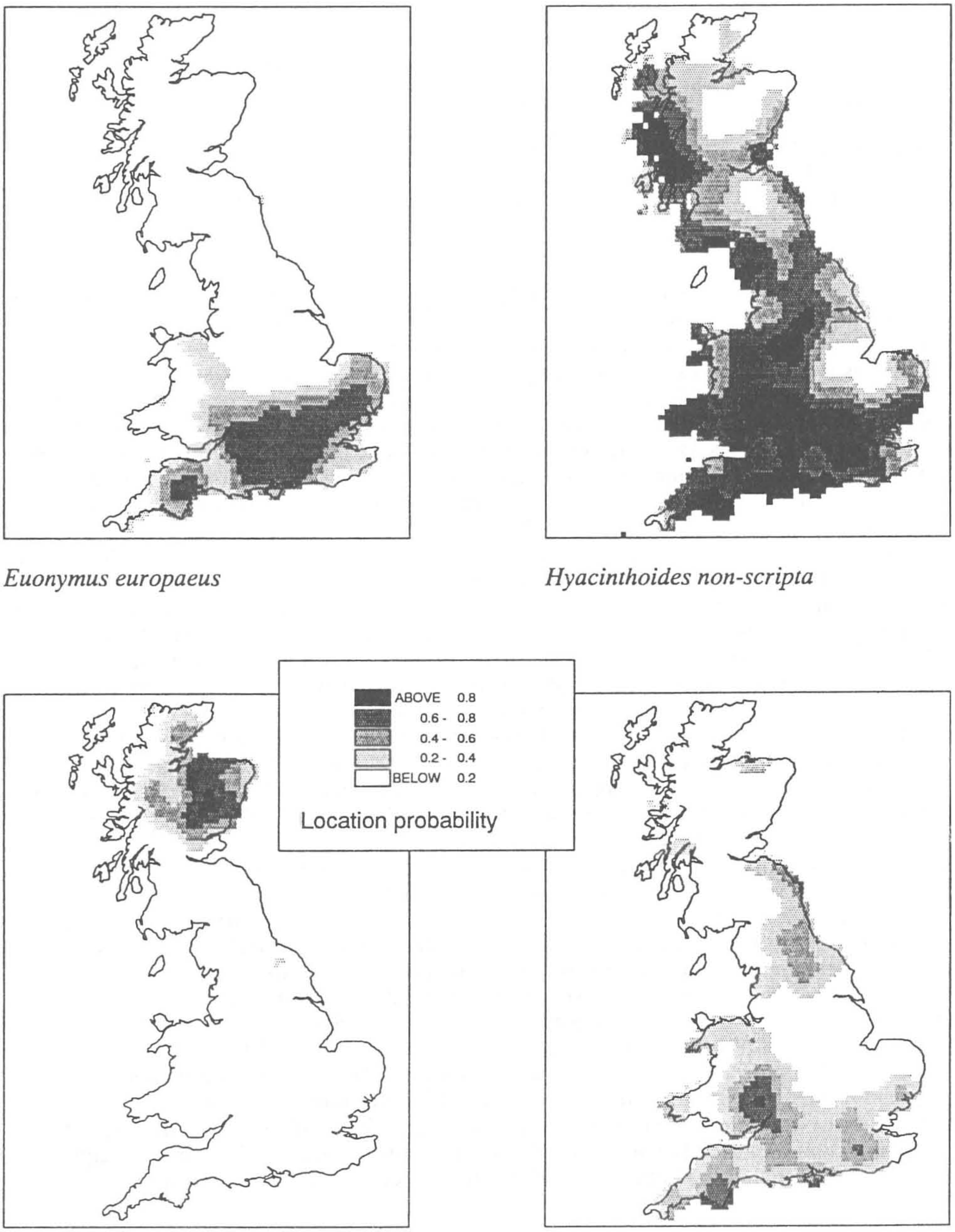
Euonymus europaeus

Hyacinthoides non-scripta

Trientalis europaea

Veronica montana

FIGURE 2. Species occurrence in tetrads (2-km squares) recorded during the B.S.B.I. Monitoring Scheme.



Euonymus europaeus

Hyacinthoides non-scripta

Trientalis europaea

Veronica montana

FIGURE 3. Probabilities of species occurrence in tetrads (2-km squares), estimated by smoothing the data in Fig. 2 with a Gaussian function.

TABLE 4. ANALYSIS OF THE DIFFERENCES BETWEEN TETRAD FREQUENCIES FOR 10-KM SQUARES, COMPARING FREQUENCIES OF *EUONYMUS EUROPAEUS*, *HYACINTHOIDES NON-SCRIPTA* AND *VERONICA MONTANA* PREDICTED FROM THE MONITORING SCHEME WITH THOSE OBSERVED IN COUNTY FLORAS

Models 1 and 2 are defined in the text. Method effect refers to tests of the null hypothesis that the mean deviation is zero.

Observed mean county density	Mean deviation (bias)	RMSE	Method effect t_{14}	Species effect $F_{2,14}$	County effect $F_{7,14}$	Kent vs others $F_{1,14}$
Gaussian smoothed	−0·50	4·91	−1·05 ns	0·25 ns	13·11 ***	56·6 ***
Regression Model 1	0·32	6·10	0·46 ns	0·10 ns	8·88 ***	39·0 ***
Regression Model 2	−0·19	5·34	−0·36 ns	0·00 ns	8·97 ***	37·6 ***

RMSE = root-mean-square error; ns = not significant; *** = p<0·001.

B.S.B.I. Monitoring Scheme data was generally quite good. The mean deviation (bias) was smallest for the prediction method using Model 2, the root-mean-square error indicating its advantage over Model 1. However the error was least for simple Gaussian smoothing.

There was a notable and statistically very significant difference between counties (Table 4). In terms of effort per tetrad, Kent was more intensively surveyed for the county Flora than for the Monitoring Scheme, whilst Bedfordshire was less so. In any survey the uniformity of sampling effort is of great importance. The B.S.B.I. Monitoring Scheme was carefully controlled with this objective (Rich & Woodruff 1990), but differences must inevitably have occurred. Variation also exists between the county Floras, some being over-sampled in comparison with the Monitoring Scheme, whilst others were relatively under-sampled.

For validation we have selected county Floras with a high and fairly uniform sampling coverage. Even though the per-tetrad effort may sometimes have been less than that achieved by the Monitoring Scheme, overall they will all have had more intensive sampling. Thus the resolution of the response surfaces produced from the Monitoring Scheme will be poorer than those which could be obtained from the county Floras. In general we would expect those species with a fairly general but patchy distribution, such as those requiring habitats in old woods, to be less easy to predict than those species with distributions depending on some more widespread factor of the physical environment such as climate or soil type. This seems to be the case when comparing the deviances explained for *E. europaeus* and *T. europaea* on the one hand, with *H. non-scripta* and *V. montana* on the other (Table 2). It is also supported by the closer agreement between overall county atlas data and the Gaussian-smoothed response surface (rows 1 and 2 in Table 3) for *E. europaeus* than for *H. non-scripta* and *V. montana*.

The ability to predict species presence or absence using regression methods also seems to be somewhat species-specific (Table 2). Those whose distribution is strongly restricted by specific environmental factors such as climate (*E. europaeus* and *T. europaea*) are seen to be better predicted than the others. Predictions were substantially improved by including information on 10-km square occurrence (Model 2). It is interesting that the coefficients a_i, b_i, c_i in Model 2 were so close in value that a single regression would have sufficed for all four species.

One of the main advantages of the logistic regression approach is that it can readily be extended to include other information (Le Duc *et al.* 1992). Such information might include, for instance, soil type (Avery 1973) and local climate (Bendelow & Hartnup 1980). Perhaps more important for many widespread species would be inclusion of additional habitat information such as the presence of woods, rivers, or a coastline. Such information is now becoming available in, for instance, the I.T.E. land classification database (Bunce *et al.* 1981). The more accurately the present frequency of a species can be estimated the better we shall be able to detect change in the future.

CONCLUSIONS

In Great Britain, sufficiently good survey data are now available to derive reliable national estimates of the probability of species occurrence in tetrads. Such estimates can be validated using independent data from county Floras. Using Gaussian-smoothed data from the Monitoring Scheme, combined with additional information about each tetrad, regression models can be developed which would improve the accuracy of estimates. These estimates can be used in future to detect the effects of major disturbances such as climate change or large-scale shifts in land use.

ACKNOWLEDGMENTS

This work was partly supported by a grant from the Forestry Research Co-ordination Committee. We thank Chris Preston for his assistance in selecting the county atlases, and Tim Rich for suggestions on use of the Monitoring Scheme database.

REFERENCES

Avery, B. W. (1973). Soil classification in the Soil Survey of England and Wales. *J. Soil Sci.* **24**: 324–338.

Bendelow, V. C. & Hartnup, R. (1980). *Climatic classification of England and Wales*. Soil Survey Technical Monograph No. 15. Harpenden.

Bunce, R. G. H., Barr, C. J. & Whittaker, H. A. (1981). *Land classes in Great Britain: Preliminary descriptions for users of the Merlewood method of land classification*. Merlewood Research and Development Paper No. 86. Grange-over-Sands.

Dony, J. G. (1963). The expectation of plant records from prescribed areas. *Watsonia* **5**: 377–385.

Dony, J. G. (1967). *Flora of Hertfordshire*. Hitchin.

Dony, J. G. (1976). *Bedfordshire plant atlas*. Luton.

Ellis, G. (1986). The new mapping scheme. *B.S.B.I. News* **43**: 7–9.

Graham, G. G. (1988). *The flora and vegetation of County Durham*. Durham.

Green, B. H. (1989). Agricultural impacts on the rural environment. *J. Appld Ecol.* **26**: 793–802.

Hall, P. C. (1980). *Sussex plant atlas*. Brighton.

Hill, M. O. (1991). Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *J. Biogeog.* **18**: 247–255.

Huntley, B., Bartlein, P. J. & Prentice, I. C. (1989). Climatic control of the distribution and abundance of beech (*Fagus* L.) in Europe and North America. *J. Biogeog.* **16**: 551–560.

Insley, H. (1988). *Farm woodland planning*. Forestry Commission Bulletin No. 80. London.

I.U.C.C. Information Services Group (1989). *UNIRAS reference guide, version 6*. Inter University Committee on Computing, University of Manchester, Manchester.

Ivimey-Cook, R. B. (1984). *Atlas of the Devon flora*. Exeter.

Jongman, R. H. G., Ter Braak, C. J. F. & Van Tongeren, O. F. R. (1987). *Data analysis in community and landscape ecology*. Wageningen.

Le Duc, M. G., Sparks, T. H. & Hill, M. O. (1992). Predicting potential colonisers of new woodland plantations. *Aspects Appld Biol.* **29** (In press).

McCosh, D. J. (1988). Local Floras – a progress report. *Watsonia* **17**: 81–89.

Perring, F. H. & Walters, S. M., eds (1976). *Atlas of British flora*, 2nd ed. Wakefield.

Philp, E. G. (1982). *Atlas of the Kent flora*. Maidstone.

Primavesi, A. L. & Evans, P. A. (1988). *Flora of Leicestershire*. Leicester.

Rich, T. C. G. (1986). The B.S.B.I. monitoring scheme. *B.S.B.I. News* **44**: 11.

Rich, T. C. G. (1987). B.S.B.I. monitoring scheme. *B.S.B.I. News* **45**: 9–12, **46**: 7, **47**: 8–12.

Rich, T. C. G. (1988). B.S.B.I. monitoring scheme. *B.S.B.I. News* **48**: 8–10, **49**: 16–17, **50**: 16–17.

Rich, T. C. G. (1989). B.S.B.I. monitoring scheme. *B.S.B.I. News* **51**: 17, **52**: 19.

Rich, T. C. G. & Woodruff, E. R. (1990). *B.S.B.I. Monitoring Scheme 1987–1988*. Unpublished report to the Nature Conservancy Council.

Tamm, C. O. (1991). *Nitrogen in terrestrial ecosystems: questions of productivity, vegetational changes and ecosystem stability*. Ecological Studies 81. Berlin.

Tarpey, T. & Heath, J. (1990). *Wild flowers of north-east Essex*. Colchester.

(*Accepted February 1992*)