Weighted averaging, logistic regression and the Gaussian response model*

Cajo J. F. ter Braak[1] & Caspar W. N. Looman[2] **
[1] *Institute TNO for Mathematics, Information Processing and Statistics, P.O. Box 100, 6700 AC Wageningen, The Netherlands;* [2] *Research Institute for Nature Management, P.O. Box 46, 3956 ZR Leersum, The Netherlands*

Keywords: Amplitude, Direct gradient analysis, Gaussian response curve, Logistic regression, Indicator value, Optimum, Tolerance, Unimodal response curve, Weighted average

Abstract

The indicator value and ecological amplitude of a species with respect to a quantitative environmental variable can be estimated from data on species occurrence and environment. A simple weighted averaging (WA) method for estimating these parameters is compared by simulation with the more elaborate method of Gaussian logistic regression (GLR), a form of the generalized linear model which fits a Gaussian-like species response curve to presence-absence data. The indicator value and the ecological amplitude are expressed by two parameters of this curve, termed the optimum and the tolerance, respectively. When a species is rare and has a narrow ecological amplitude – or when the distribution of quadrats along the environmental variable is reasonably even over the species' range, and the number of quadrats is small – then WA is shown to approach GLR in efficiency. Otherwise WA may give misleading results. GLR is therefore preferred as a practical method for summarizing species' distributions along environmental gradients. Formulas are given to calculate species optima and tolerances (with their standard errors), and a confidence interval for the optimum from the GLR output of standard statistical packages.

Introduction

If the relationships between species occurrences and values of a quantitative environmental variable conform to bell-shaped curves, then each species' curve can conveniently be summarized by an *indicator value* and an *ecological amplitude* (Ellenberg, 1979, 1982). The indicator values can subsequently be used to predict values of an environmental variable from species composition, simply by averaging the indicator values of species that are present (Ellenberg, 1979). The average indicator value can be weighted, to take account of differences in species abundance and in ecological amplitude (Goff & Cottam, 1967; Ter Braak & Barendregt, in press). Weighted averaging can also be used to estimate the indicator values themselves (de Lange, 1972; Salden, 1978). Values of the environmental variable are averaged over the samples in which a species occurs. (The average can be weighted by species abundance, but we consider only presence-absence data.) Weighted averaging is the basis of the ordination technique known as reciprocal averaging (Hill, 1973) and is implicit in Gasse & Tekaia's (1983) algorithm to establish a transfer function for estimating paleo-environmental conditions (pH) from fossil diatom assemblages. Hörnström (1981) used medians, instead of averages, in a similar context. But there is a problem with averaging, or taking medians: namely that the result can depend on the distribution of the quadrats along the environmental variable. When the distri-

* Nomenclature follows Heukels-van der Meijden (1983).

** We would like to thank Drs I. C. Prentice, N. J. M. Gremmen and J. A. Hoekstra for comments on the paper. We are grateful to Ir. Th. A. de Boer (CABO, Wageningen) for permission to use the data of the first example.

bution is uneven, all weighted averaging methods may potentially give misleading results (Greig-Smith, 1983, p. 130).

The estimation of indicator values is fundamentally a regression problem. Indicator values and ecological amplitudes can be estimated from presence-absence data by logistic regression, with a second-order polynomial in the environmental variable as linear predictor. This procedure, termed Gaussian logistic regression (GLR), fits a curve related to the Gaussian species response curve (Austin, 1980) but adapted for presence-absence data. The indicator value is then the 'optimum' (mode) of the curve. Logistic regression is a Generalized Linear Modelling technique (GLIM), and is the equivalent for presence-absence data of ordinary multiple and polynomial regression (Dobson, 1983; McCullagh & Nelder, 1983). Austin, Cunningham & Fleming (1984) showed the usefulness of GLM and GLR in their study of the occurrence of a range of eucalypt species in relation to temperature, rainfall, radiation and geology. There is no good evidence for the exact shape of a species response curve; we shall show that GLR is a practical method.

We compare the performance of weighted averaging and logistic regression, using stimulation and practical examples. We know from theory that logistic regression must give more accurate estimates of species' optima in *large* datasets in which the number of presences is not too small and for which the logistic model holds. But is logistic regression also worthwhile when the number of presences is small, say 10 or 20? There is no advantage in using an elaborate technique where a much simpler one would be equally good. Our simulations give some idea about the conditions under which weighted averaging compares reasonably well with logistic regression; but they also show that GLR is more generally applicable. Our results are also relevant in choosing between reciprocal averaging and Gaussian ordination (Ter Braak, in press).

Logistic regression

The 'presence-absence response curve' of a species describes the probability, $p(x)$, that the species occurs (in a quadrat of fixed size) as a function of an environmental variable x. Whittaker (1956), and others since, have observed that species typically show unimodal (bell-shaped) response curves. The 'Gaussian response curve' (Austin, 1980) is a simple bell-shaped curve in which the logarithm of abundance is a quadratic function of the environmental variable. Presence-absence data are more conveniently modelled with the *Gaussian logit curve,* in which the logit-transform of probability (Cox, 1970) is a quadratic function, (Fig. 1):

$$\log\left[\frac{p(x)}{1-p(x)}\right] = b_0 + b_1 x + b_2 x^2 = a - \tfrac{1}{2}\,(x-u)^2/t^2 \quad (1)$$

where u is the species *optimum* or indicator value (the value of x with highest probability of occurrence) and t is its *tolerance* (a measure of ecological amplitude). The parameter a is related to the *maximum* value of $p(x)$, which we shall call p_{max}. When p_{max} is small the shape of $p(x)$ is almost identical to that of a Gaussian curve; when p_{max} is close to 1 the Gaussian logit curve is flatter near the optimum (Fig. 1). The parameters b_0, b_1 and b_2 do not have a natural ecological meaning, but they can easily be estimated using logistic regression which is available in standard statistical packages including GENSTAT (Alvey *et al.,* 1977), GLIM (Baker & Nelder, 1978), BMDP (Nixon, 1981) and SAS (Barr *et al.,* 1982), and interpretable parameters can be obtained from them as follows:

$$\begin{aligned} &\text{optimum } u = -b_1/(2b_2) \\ &\text{tolerance } t = 1/\sqrt{(-2b_2)} \\ &\text{maximum probability } p_{max} = p(u) = \\ &\quad 1/\{1+\exp(-b_0 - b_1 u - b_2 u^2)\} \end{aligned} \quad (2)$$

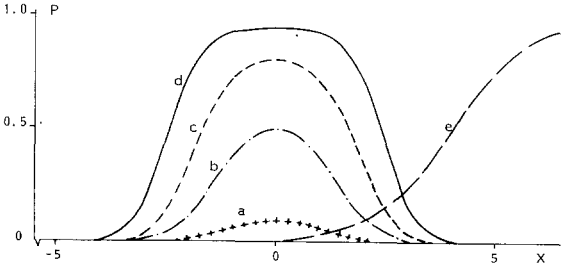

Fig. 1. Gaussian logit curves with $u=0$, $t=1$ and $p_{max}=0.1$ (a), 0.5 (b), 0.8 (c) and 0.95 (d) and a linear logit curve (e) (x: value for the environmental variable, $p(x)$: probability of finding this species at a value x).

(These formulas assume $b_2 < 0$. If $b_2 > 0$ the curve has a minimum instead of a maximum). Table 1 gives a sample program in GLIM (Baker & Nelder, 1978) for artificial data and Figure 2 shows the fitted curve. The sample program shows that this procedure of GLR is a special case of the Generalized Linear Model (see Dobson, 1983 for an introduction): (1) *response variable* is a *1/0*-variable, *y*, containing the presences and absences of the species in the quadrats; (2) *error distribution* is the binomial distribution with total *1*, also termed the Bernoulli distribution; (3) *link function* is the logit-transform, which links the expected value of *y* (*i.e.* the probability of occurrence) to (4) the *linear predictor* specified in the FIT-statement. In GLR the linear predictor is a quadratic polynomial in *x*. The user does not need to provide initial values for the parameters. The approximate standard errors of the estimated optimum and tolerance can be derived from the variances and covariances of b_1 and b_2 that are provided as options by the statistical packages. A confidence interval for the optimum can also be calculated. Details of these additional calculations are given in the Appendix.

Table 1. Sample program for Gaussian logistic regression in GLIM, with output for artificial data (S.E.: standard error of estimate). The program does not provide the estimates for p_{max}, u and t automatically; these estimates were computed by use of Eqs. (2), (A.1) and (A.2).

PROGRAM

```
$UNIT                                  16[1]
$DATA                                  X  Y[2]
$READ
 20   0   23   0   26    0   30   0
 33   0   36   0   40    0   43   0
 46   0   50   1   53    1   56   0
 60   1   70   1   80    0   90   0
$CALCULATE                             TOTAL = 1
$CALCULATE                             XQUAD = X*X
$YVARIATE                              Y[3]
$ERROR                                 BINOMIAL TOTAL
$LINK                                  LOGIT[4]
$FIT                                   X + XQUAD[5]
$DISPLAY                               E  $[6]
```

		ESTIMATE	S.E.
CONSTANT	(b_0)	−55.5	34.5
X	(b_1)	1.86	1.15
XQUAD	(b_2)	−0.015	0.009
p_{max}		0.90	-
u		62	3.3
t		5.8	1.8

Comments
[1] 16 data values.
[2] (x_i, y_i) being read.
[3] The response variable is *y* containing independent *1/0* data.
[4] Link function is the logit-transform.
[5] *x* and x^2 are the explanatory variables to be fitted.
[6] Displays the parameter estimates b_0, b_1, b_2 with standard error.

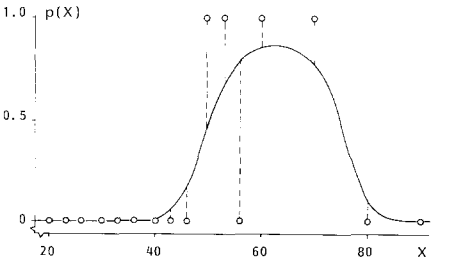

Fig. 2. Gaussian logit curve fitted by logistic regression to the artificial data (o) of Table 1.

The optimum cannot be estimated well if it lies outside or near the edge of the sampled range. In such cases the response curve is said to be truncated and b_2 in Eq. (1) could be set to zero; the effect is to fit a sigmoid curve, termed the linear logit curve (Fig. 1). Whether this simplification is acceptable statistically can be seen by a one-sided significance test on the value of b_2, in which b_2 divided by its standard error is compared with the Student *t*-distribution with $n-3$ degrees of freedom (n is the number of quadrats). If the null hypothesis ($b_2 \geq 0$) is rejected in favour of the alternative hypothesis ($b_2 < 0$), then the optimum is said to be significant.

A more general approach to statistical testing in GLIM is to compare the residual deviance of a model with that of an extended model (Austin *et al.*, 1984; Dobson, 1983). The additional terms in the model are significant when the difference in residual deviance is larger than the critical value of a chi-square distribution with k degrees of freedom, k being the number of additional parameters. (The residual deviance is defined by −2 log-likelihood and takes a similar role as the residual sum of squares in ordinary multiple regression). For example, to test the overall significance of GLR we also fit the model with both b_1 and b_2 in Eq. (1) set to zero and we compare the difference in residual devi-

ance with a chi-square with *2* degrees of freedom. The tests described in this paper are approximate; they are valid when the number of quadrats is large.

Weighted averaging

The weighted average for presence-absence data is simply the mean of the *x*-values over those quadrats in which the species occurs. Figure 3 shows how the weighted average depends on the distribution of sampled quadrats. Highly uneven distributions can even scramble the order of the weighted averages for different species (Fig. 3c). Truncation is an extreme form of uneven distribution, because the response curve is then not sampled over the whole range where the species can occur. Only in the special case of an even or uniform distribution over the whole range does the weighted average reliably estimate the optimum. The sample standard deviation (*SD*) of the *x*-values of those quadrats in which the species occurs is a simplistic estimate of ecological amplitude. Assuming the Gaussian logit response curve (1) and an even distribution of the quadrats, *SD* overestimates the tolerance *t*; the difference between the expected *SD* and *t* depends on the value of p_{max}, but is less than 12% when p_{max} is less than 0.5 (Looman, unpublished manuscript).

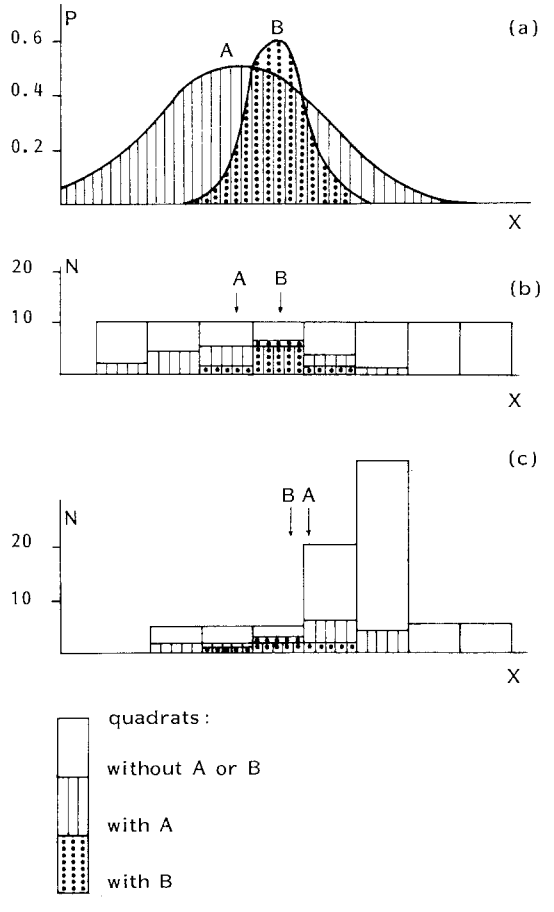

Fig. 3. The response curves of imaginary species A and B (a), the occurrence of these species in 80 samples, distributed evenly (b) or unevenly (c) along the environmental gradient. The weighted averages are indicated with arrows. The two sampling designs yield weighted averages that are in reversed order (*p*: probability of occurrence, *N*: number of quadrats, *x*: environmental variable).

Design of simulations

Presence-absence data were generated using a Gaussian logit response curve with *u* and *t* arbitrarily set to *0* and *1*, respectively. We further need to specify p_{max}, the number of quadrats per dataset and the distribution of the quadrats along the gradient. Table 2 shows the tested combinations and, for each combination, the expected number of presences per dataset. In case 1 of the distributions the *x*-values of the quadrats are equispaced on the interval from -5 to 5. In all the other cases the *x*-values are random. In cases $2-5$ their distribution is uniform with different degrees of truncation, negligible in case 2, asymmetric in cases 3 and 4 and symmetric in case 5. Another six cases were run with $p_{max}=0.5$ and 125 quadrats only (Table 3). In case 6 (Table 3) the curve is unevenly sampled with on average three times more quadrats in the interval [1, 5] than in the interval $[-5, 1]$, but

Table 2. Expected number of occurrences per dataset in the simulations specified by maximum probability of occurrence (p_{max}), number of quadrats and distribution of quadrats (case). (*U[a, b]*: uniform distribution of quadrats on the interval *a* to *b*).

	p_{max}	0.1	0.5	0.9	0.5	0.9
	no. of QUADRATS	375	65	25	125	50
C	1 EQUAL SPACING	10	10	10	19	19
A	2 *U[−5, 5]*	10	10	10	19	19
S	3 *U[−1, 5]*	13	13	12	25	23
E	4 *U[0, 5]*	10	10	10	19	19
	5 *U[−1, 1]*	32	30	22	57	44

with quadrats uniform within both intervals. Case 7 consists of quadrats uniformly distributed in the interval [−2, 5] but with quadrats from the interval [−1.5, 0.5] removed, giving a case with moderate truncation and an internal gap. For the remaining cases (8−11) we used normal (Gaussian) distributions of quadrats with different means and standard deviations; case 8 gives symmetric and cases 9 and 10 asymmetric truncation. In case 11 the curve is sampled over a short range with 95% of the quadrats in the interval [−0.5, 1.5].

Weighted averaging (WA) and Gaussian logistic regression (GLR) were obtained for each dataset using GENSTAT (Alvey *et al.*, 1977). For each combination in Tables 2 and 3 we simulated 100 datasets and summarized the results as means, medians and standard deviations of the weighted average and GLR-estimates calculated for each dataset. In cases where no optimum could be calculated ($b_2 \geq 0$), we treated the regression estimates as missing values. Estimated optima are also unreliable when b_2 is negative but close to zero; we therefore discarded simulations in which the estimated optimum lay more than ten times the tolerance outside the sampled interval. We also calculated means and standard deviations of the regression estimates over the cases in which the optimum was significant at the 10%-level. This selection summarizes the significantly non-monotone curves. No such selection was applied to weighted averaging, because in practice the weighted average is calculated irrespective of such evidence for unimodality. The efficiency of the weighted average with respect to the regression estimate for the optimum was then expressed as *MSE(GLR)/MSE(WA)* where *MSE* is the mean squared error, *i.e.* variance plus squared bias.

Comparison of WA and GLR

Equal spacing and uniform distribution without truncation

WA is as efficient as GLR when the x-values are equispaced (case 1). However, when the x-values are randomly distributed on a large interval (case 2), the efficiency of the weighted average is less. The efficiencies calculated from the runs of case 2 with, on average, 10 occurrences per simulated dataset (Table 2) were 1.0, 0.84 and 0.54 for $p_{max}=0.1$, 0.5 and 0.9, respectively, in agreement with theoretical values (Fig. 4) derived by Ter Braak & Barendregt (in press). The variance of the regression estimate in the simulation was slightly (<10%) larger than its theoretical value of t^2/(no. of occurrences) (cf. Ter Braak & Barendregt, in press), with the exception of the runs with only 25 quadrats (Table 2) where the difference was 50%.

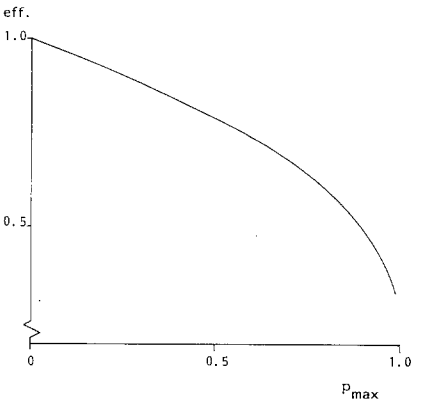

Fig. 4. The efficiency (ordinate) of weighted averaging with respect to Gaussian logistic regression to estimate the optimum for uniformly distributed quadrats without truncation (case 2, Table 2) decreases with increasing maximum probability of occurrence (abscissa).

Effect of distribution of quadrats

Table 3 summarizes the results of cases with 125 quadrats and $p_{max}=0.5$ and confirms that WA is sensitive to the distribution of the quadrats along the gradient, showing significant bias (t-test, $P<0.05$) in 7 cases. The optimum could not be estimated by GLR in 1% of the simulated datasets of Table 3, except in the cases 4 and 11 where this percentage was about 15%. GLR removes the bias of WA when the truncation is not too severe (cases 6−10). When it is severe (cases 3, 4 and 11) the regression estimate of the optimum shows a large bias in the opposite direction, but this bias is small in a statistical sense, as the standard error is high. The medians of the estimates show small bias in the same direction as WA. When the estimated curves are first tested for unimodality against monotonicity at the 10%-level, the remaining optima (*u-sig*) show selection bias; they are biased because an optimum is more likely to be significant

Table 3. Weighted averaging and Gaussian logistic regression compared on simulated datasets with eleven distributions of 125 quadrats along the environmental variable. Shown are means ± standard deviations and medians (*md*), multiplied by 100. The entries in the table must be compared with the true values: 0 for *u*, 100 for *t*, 50 for p_{max}, 112 for *SD*. The cases are explained in the text. *m*: average number of occurrences; *N-sig*: number out of 100 datasets showing a significant optimum and summarized under the headings *u-sig* and *t-sig*; *N a ± b*: normal distribution of quadrats with mean *a* and standard deviation *b*). For further symbols see text and Table 2.

CASE	*m*	*WA*	*u*	*md-u*	*u-sig*	*SD*	*t*	*md-t*	*t-sig*	p_{max}	*N-sig*
1 EQUAL SPACING	19	2±21	2± 21	0	2±21	108±16	94± 16	91	94±16	52±10	100
2 *U[−5, 5]*	19	3±28	3± 25	2	3±25	111±16	99± 16	98	99±16	51±10	100
3 *U[−1, 5]*	25	39±14	−22± 76	0	3±31	86±11	104± 31	98	96±19	53± 8	84
4 *U[0, 5]*	19	91±14	−88±403	33	60±21	63±11	104± 67	80	71±16	57±19	52
5 *U[−1, 1]*	58	1± 8	−3±116	2	2±11	55± 3	120± 80	89	67± 8	54± 7	30
6 UNEVEN	15	51±33	6± 30	7	6±30	114±21	94± 17	95	94±17	54±13	100
7 GAP	15	80±29	3± 35	1	2±35	106±29	93± 22	93	93±22	55±13	98
8 *N 0 ±2*	33	1±19	0± 22	1	0±22	98±11	99± 15	100	99±15	51± 7	100
9 *N 2 ±2*	22	50±19	−2± 37	4	0±32	97±14	99± 21	96	99±20	51± 8	99
10 *N 3 ±2*	14	72±24	0± 62	9	11±40	91±18	94± 28	91	91±21	54±12	94
11 *N 0.5±0.5*	55	44± 6	−70±488	14	27±18	45± 4	133±154	90	66±11	55±13	34

when it lies inside than when it lies outside the sampled interval. This bias is less than with WA. The efficiency of WA compared to GLR after the significance test lies between 0.2 and 0.6 except in the cases 1 and 2 and the unnatural cases 5 and 8 in which the quadrats lie symmetrically with respect to the true optimum.

The sampled *SD* underestimated the true *SD* in cases 3, 4, 5 and 11 with severe truncation (Table 3). Overestimation was never pronounced. GLR estimated the tolerance well; the bias shown in Table 3 is not significant ($P>0.05$). The median of the estimated tolerance is slightly biased downwards. After the significance test for unimodality the bias is downwards, but less than with the sample *SD*. GLR slightly overestimates the maximum probability with and without selection, the mean and median of the estimates being close together. WA provides no estimate for this probability. The remaining simulations of the cases 1−5 (Table 2) showed qualitatively similar features as reported here for $p_{max}=0.5$ and 125 quadrats.

The effect of number of quadrats

The efficiency of WA can be expected to decrease to zero with increasing numbers of quadrats in those cases in which WA is biased. This is because estimates by GLR are consistent, *i.e.* the bias in the estimates becomes smaller as the number of quadrats increases, and the variances become negligible with respect to the bias in WA. However, in our simulations with only 10−13 occurrences per dataset (Table 2) the variances are appreciable; consequently the efficiencies for estimating the optimum, after the significance test, were high (>0.9 in 10 out of the 12 simulations). Even 375 samples are not enough to get markedly better estimates with GLR than with WA, when $p_{max}=0.1$!

Standard errors and confidence interval

First, the standard errors found in the simulations are compared with the approximate standard errors provided by GLR for each estimated optimum and tolerance (see Appendix for the formulas used). The latter standard errors showed often a skew distribution with large outliers. As a result the average and the median of the estimated standard errors differed enormously, the average being much higher and the median slightly lower than the standard error found by simulation. Clearly the estimated optimum or tolerance is unreliable when the estimated standard error is huge, but when it is low, it may be over optimistic about the precision achieved. Secondly, in 1 085 (≈40%) of all simulations a 95%-confidence interval could be calculated (see Appendix). The true optimum lay outside the 95%-confidence interval in 3.9% of these 1 085 simulations, hence the interval gives higher confidence than its nominal value of 95%.

Examples with real data

The first real dataset concerns soil acidity (pH) and the occurrences of 15 species in 100 meadow samples, selected at random from the study of Kruijne *et al.* (1967). Figure 5 shows the fitted Gaussian logit curves for seven contrasting species. The Spearman rank correlation between the optima as estimated by GLR and the weighted averages was 0.93. (The optima for two species for which $\hat{b}_2$ was positive, but non-significantly different from zero, were set to $+\infty$ or $-\infty$, depending on whether the value of b_1 in the fit or the linear logit curve was positive or negative, respectively). However, the range of the weighted averages was much smaller than the range of the estimated optima (1.0 against more than 4.0 pH-units). A 90%-confidence interval for the optimum could be calculated for five species. For one of these species (*Bellis perennis*) the weighted average lies outside this confidence interval.

In the second example we used a much larger set of data, taken from Reijnen *et al.* (1981) and Gremmen *et al.* (1983). This dataset concerns the relation between species occurrence and soil moisture supply capacity in the Pleistocene part of West-Brabant (The Netherlands) with sandy to loamy soils. The distribution of soil moisture supply capacity in the 994 samples was markedly skewed, with many more 'wet' than 'dry' samples. For 121 of the 221 species that occurred in more than five samples, a 90%-confidence interval for the optimum could be calculated. The weighted average lies outside this interval for about half (65) of these species, always being on the wetter side of the confidence interval. Although p_{max} was less than 0.1 for about 75% of the species, WA is unreliable for estimating indicator values in this large dataset.

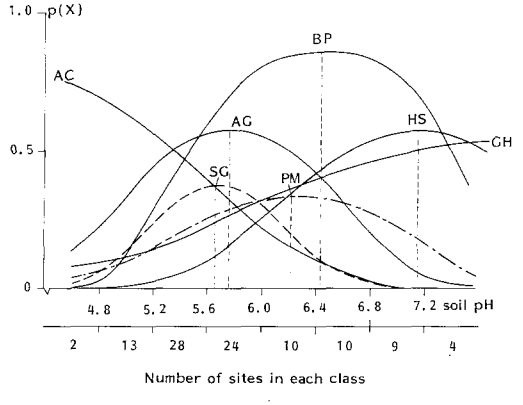

Fig. 5. Probability of occurrence of seven contrasting species in relation to soil acidity (pH) in meadows, as fitted with logistic regression. The curves can be identified by the code near their optimum indicated by dotted lines. The species arranged in order of their optima are: *Agrostis canina* (AC); *Stellaria graminea* (SG); *Alopecurus geniculatus* (AG); *Plantago major* (PM); *Bellis perennis* (BP); *Hordeum secalinum* (HS); *Glechoma hederacea* (GH).

Discussion

WA disregards species absences. Ashby (1936) pointed out that disregarding species absences may lead to erroneous conclusions, for instance that telegraph poles show an optimal pH-value (see Greig-Smith, 1983, p. 130). This effect is due to the distribution of quadrats. Nevertheless, WA is still being used (see Introduction), perhaps because of its simplicity. Our simulations provide a better reason; they suggest that WA performs reasonably well when the distribution of the quadrats along the environmental variable is not too uneven and when the response curve is not severely truncated. For rare species (species with low maximum probability of occurrence and/or narrow tolerance) WA is nearly as efficient as GLR in most situations. This result is irrespective of the distribution of the quadrats, provided the variance of the estimated optimum is large compared to the potential bias of the weighted average. In other cases WA can give misleading results. It is therefore safest always to use GLR.

To estimate optima and tolerances of species, the optima should ideally lie well within the range of environmental values of the samples. Further sampling considerations are provided by Mohler (1983). Attention should also be paid to confounding variables, *i.e.* variables that are influential and show a relation with the variable under consideration (see e.g. Breslow & Day, 1980). Ignoring confounding variables may give, for example, spuriously bimodal response curves (Austin *et al.*, 1984). The real power of logistic regression lies in the simultaneous analysis of the effect of several environmental variables, including potentially confounding variables (see Appendix). The Gaussian logit response curve is then just a convenient starting point in the process of model building.

Appendix

Standard errors for estimated u and t; confidence interval for u.

Denote the variance of the estimates of b_1 and b_2 in model (1) by v_{11} and v_{22} and their covariance by v_{12}. Using Taylor expansion we obtain that the variance of the estimated optimum and tolerance are approximately

$$\text{var}(\hat{u}) = (v_{11} + 4uv_{12} + 4u^2 v_{22})/(4b_2^2) \quad \text{(A.1)}$$

$$\text{var}(\hat{t}) = v_{22}/(-8b_2^3) \quad \text{(A.2)}$$

An approximate $100(1-\alpha)$%-confidence interval for the optimum is derived from Fiellers theorem (see Finney, 1964, p. 27–29). Let t_α be the ordinary Student t-deviate at chosen probability level α and with $n-3$ degrees of freedom (n is the number of quadrats). For example, $t_\alpha = 2.00$ for a 95%-confidence interval and 63 quadrats. Calculate $g = (t_\alpha^2 v_{22})/b_2^2$ and

$$D = 4b_2^2 \text{ var}(\hat{u}) - g(v_{11} - v_{12}^2/v_{22}) \quad \text{(A.3)}$$

$$u_{\text{lower}},\ u_{\text{upper}} = [\hat{u} + \tfrac{1}{2}\ g\ v_{12}/v_{22} \pm \tfrac{1}{2}\ t_\alpha(\sqrt{D})/b_2]/(1-g) \quad \text{(A.4)}$$

where the symbol $\pm$ is used to indicate addition and substraction in order to obtain the lower and upper limits of the confidence interval, respectively. If b_2 is not significantly different from zero ($g > 1$), then the confidence interval is of infinite length and, taken alone, the data must be regarded as valueless for estimating the optimum.

If model (1) is extended with another explanatory variable z to, for example (Austin *et al.*, 1984: Table 2)

$$\log [p/(1-p)] = b_0 + b_1 x + b_2 x^2 + c_1 z + c_2 z^2 \quad \text{(A.5)}$$

then the coefficients b_0, b_1, b_2, c_1 and c_2 can, again, be estimated with the mentioned statistical packages, together with variances and covariances. This model can easily be summarized by optima and tolerances with respect to x and z, because there is no interaction term, like $x.z$, in the model. To calculate the confidence interval for the optimum of respect to x (or z) from this model, the given formulas are still valid, apart from the number of degrees of freedom in t_α which must now be $n-5$.

References

Alvey, N. G., *et al.*, 1977. GENSTAT: a general statistical program. Rothamsted Experimental Station, Harpenden, England.

Ashby, E., 1936. Statistical ecology. Bot. Rev. 2: 221–235.

Austin, M. P., 1980. Searching for a model for use in vegetation analysis. Vegetatio 42: 11–21.

Austin, M. P., Cunningham, R. B. & Fleming P. M., 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio 55: 11–27.

Baker, R. J. & Nelder, J. A., 1978. The GLIM System, Release 3. Numerical Algorithms Groups, Oxford.

Barr, A. J., *et al.*, 1982. SAS User's Guide: Statistics, 1982 edition. SAS Institute Inc., Cary, 584 pp.

Breslow, N. E. & Day, N. E., 1980. Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case-Control Studies. IARC Scientific Publication, nr. 32, Lyon, 338 pp.

Cox, D. R., 1970. The Analysis of Binary Data. Methuen, London, 142 pp.

Dixon, W. J., 1981. BMDP Statistical Software, University of California Press, Berkeley, 726 pp.

Dobson, A. J., 1983. An Introduction to Statistical Modelling. Chapman & Hall, London, 125 pp.

Ellenberg, H., 1979. Zeigerwerte der Gefässpflanzen Mitteleuropas. 2nd ed. Scripta Geobotanica 9, Göttingen, 122 pp.

Ellenberg, H., 1982. Vegetation Mitteleuropas mit den Alpen in ökologischer Sicht. 3rd ed. Ulmer Verlag, Stuttgart, 989 pp.

Finney, D. J., 1964. Statistical Methods in Biological Assay. Griffin, London, 668 pp.

Gasse, F. & Tekaia, F., 1983. Transfer functions for estimating paleoecological conditions (pH) from East African diatoms. Hydrobiologia 103: 85–90.

Goff, F. G. & Cottam, G., 1967. Gradient analysis: the use of species and synthetic indices. Ecology 48: 783–806.

Greig-Smith, P., 1983. Quantitative Plant Ecology, 3rd ed. Butterworths, London, 359 pp.

Gremmen, N. J. M., Vreugdenhil, A. & Hermelink, P., 1983. Vegetatiekartering West-Brabant: de methodiek. Report 83/21 of the Research Institute for Nature Management, Leersum, The Netherlands, 58 pp.

Heukels, H. & Meijden, R. van der, 1983. Flora van Nederland. 20th ed. Wolters-Noordhoff, Groningen, 583 pp.

Hill, M. O., 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237–249.

Hörnström, E., 1981. Trophic characterization of lakes by means of qualitative phytoplankton analysis. Limnologica (Berlin) 13: 249–261.

Kruijne, A. A., Vries, D. M. de & Mooi, H., 1967. Bijdrage tot de oecologie van de Nederlandse graslandplanten (with english summary). Versl. Landbouwk. Onderz. 696. Pudoc, Wageningen, 65 pp.

Lange, L. de, 1972. An ecological study of ditch vegetation in the Netherlands. Ph.D. thesis, University of Amsterdam, Amsterdam, 112 pp.

McCullagh, P. & Nelder, J. A., 1983. Generalized Linear Models. Chapman & Hall, London, 260 pp.

Mohler, C. L., 1981. Effect of sampling pattern on estimation of species distributions along gradients. Vegetatio 54: 97–102.

Reijnen, M. J. S. M., Vreugdenhil, A. & Beije, H. M., 1981. Vegetatie en grondwaterwinning in het gebied ten zuiden van Breda. Report 81/24 of the Research Institute for Nature Management, Leersum, The Netherlands, 140 pp.

Salden, N., 1978. Beiträge zur Ökologie der Diatomeen (Bacillariophyceae) des Süsswassers. Decheniana, Beiheft 22: 1–238.

Ter Braak, C. J. F., in press. Correspondence analysis of incidence and abundance data, properties in terms of a unimodal response model. Biometrics 41.

Ter Braak, C. J. F. & Barendregt, L. G., in press. Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Biosci.

Whittaker, R. H., 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26: 1–80.

Accepted 15.3.1985.