Weighted Averaging of Species Indicator Values: Its Efficiency in Environmental Calibration

CAJO J. F. TER BRAAK AND LEO G. BARENDREGT
Institute TNO for Mathematics, Information Processing and Statistics,
P.O. Box 100, 6700 AC Wageningen, The Netherlands

Received 11 March 1985; revised 30 July 1985

ABSTRACT

A common bioassay problem in applied ecology is to estimate values of an environmental variable from species incidence or abundance data. An example is the problem of reconstructing past changes in acidity (pH) in lakes from diatom assemblages found in successive strata of the bottom sediment. The method of weighted averaging is based on indicator values, the indicator value of a species being, intuitively, the value of the environmental variable most preferred by that species. Indicator values of all species present in a site are averaged to give an estimate of the value of the environmental variable at the site. The average is weighted by species abundances, if known, with absent species having zero weight. Using field data, several authors have compiled lists of indicator values of species for various environmental variables for use in weighted averaging, e.g. pH indicator values of diatom species. In this paper the properties of the method of weighted averaging are studied, starting from the idea that indicator values are parameters of response curves that describe the expected abundance of each species in relation to the environmental variable. In practice the response curves must be estimated by regression methods, but here they are assumed to be known in advance. Conditions are derived under which the weighted average is a consistent and efficient estimator for the value of an environmental variable at a site. Because weighted averaging is central to the ordination technique known as reciprocal averaging or correspondence analysis, the conditions also define models that are implicitly invoked when reciprocal averaging is used in ecological ordination studies.

1. INTRODUCTION

Plant species need particular environmental conditions for regeneration, establishment, and growth. It should therefore be possible to infer the environmental conditions at a site from the species that occur there. This type of bioassay has become popular [3, 6, 9, 19] with the publication of lists of indicator values of species with respect to various environmental variables. For example, Ellenberg [8] has published indicator values of Central European

52 Vanderbilt Ave., New York, NY 10017 0025-5564/86/$03.50

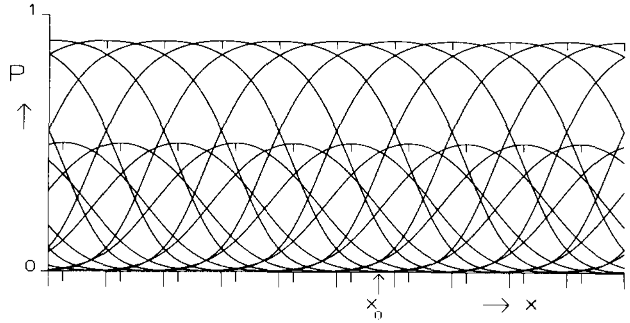

FIG. 1. Gaussian logit response curves of the probability $P = \mu_k(x)$ that a species (k) occurs at a site, against environmental variable x. Two sets of species are displayed, each with $t = 1$ and optima with spacing $d = 1$, having maximum probabilities of .5 and .9, respectively. x_0 is the value of x at a particular site.

plants with respect to site variables including soil moisture, pH, and nitrogen level. Ellenberg based the indicator values on his field observations of the conditions under which particular species occurred and, to a lesser extent, on laboratory experiments. For example, a plant species may prefer a particular soil moisture content, and not grow at all in places where the soil is either too dry and too wet. Intuitively, the indicator value is then the value most preferred by a species (cf. Figure 1). Ellenberg [8] did not give a precise definition of "indicator value." However, Ellenberg [7, 8] did describe a method to predict the value of an environmental variable: the method consists simply of averaging indicator values for the plant species that are present. For quantitative data, the average is weighted by species abundance, with absent species carrying zero weight. This method has been applied to vascular plants [12, 17, 21, 23, 25], to diatoms [20], and to aquatic organisms and the biological evaluation of water quality [19].

It might be thought easier to measure environmental variables at a site than to infer their values from the species that grow there. But often it is not. For example, total values over time may be required; repeated measurements are costly, while plants automatically integrate environmental conditions over time. This is one of the ideas behind biological evaluation of water quality and biomonitoring in general. There are also situations where it is impossible to measure environmental variables by direct means, whereas a biological record does exist. An example is the reconstruction of past changes in acidity (pH) in lakes, from diatom assemblages found in successive strata of the bottom sediment; this technique is an important tool in acid rain research. Most researchers in this area use the indicator values for acidity of diatom species as compiled by Hustedt in the 1930s [2]. A more sophisticated method, yet to be implemented, is to build firstly a (nonlinear) regression model from data on species occurrences and present pH in lakes, which yields for each species an estimated response curve for the probability of occurrence versus pH; and secondly to use these response curves for the calibration of pH from species data, for example by maximum likelihood estimation. Here the indicator value of a species is just a parameter of the response curve of that species, the mode of the curve being one possible definition of the indicator value.

In this paper we study the properties of weighted averaging of indicator values to estimate the value of a continuous environmental variable at a site. We do this by seeking conditions under which weighted averaging compares favorably with methods based on explicit response curves. We use assumptions (Section 2) that idealize the real world, among others that a single environmental variable determines the species composition at a site and that the response curves of the species with respect to this variable are already known. Certainly, weighted averaging is of little value if it has undesirable properties under ideal assumptions. On the other hand, there is no advantage in using an elaborate technique if a simpler one would be equally good. We answer two questions:

(1) How should indicator values of species be defined in terms of response curves to ensure that the weighted average is a consistent estimator? (The weighted average is called consistent if it converges in probability to the true value of the environmental variable as the number of species available increases.)

(2) What should the response curves look like to ensure that the weighted average is an efficient estimator? (An estimator is called efficient if its mean squared error is minimum.)

2. WEIGHTED AVERAGING AND RESPONSE CURVES: DEFINITIONS

Let x denote a quantitative environmental variable, and x_0 the value of this variable at a particular site. We want to estimate this value x_0 by checking which species (out of a large number) are present at that site or, more generally, the abundance of each species. Let Y_k be the abundance ($Y_k \geqslant 0$) of the kth species ($k = 1,2,3,\ldots$), and let u_k be its indicator value, usually taken from a published list of indicator values. To estimate x_0, ecologists commonly use the weighted average [7–9]

$$\hat{x}_{\mathrm{WA}} = \frac{\sum\limits_k Y_k u_k}{\sum\limits_k Y_k}, \tag{2.1}$$

where summations are over all species. To make sense, $\hat{x}_{\mathrm{WA}}$ and hence the values for u_k must have the same dimension as x. The indicator values are therefore location parameters on x.

To be a potential indicator, a species must show a distinct relation to the indicated environmental variable x. We define the relations between species and the environmental variable by a statistical response model with a response curve $\mu_k(x)$, a known function of x, for each species k. $\mu_k(x_0)$ specifies the expectation of the value Y_k observed at the site with value x_0 for x. The observational data will be assumed to be independent random variables with variances depending on the expectations only. The variance of Y_k is therefore a known function $v_k(x) = v^*(\mu_k(x))$. For presence-absence data Y_k is a Bernoulli variable and $\mu_k(x_0)$ is the probability that the kth species is present at a site with $x = x_0$. Then $v^*(\mu) = \mu(1-\mu)$. For counts, the data may be assumed to have a Poisson distribution so that $v^*(\mu) = \mu$, whereas for continuous quantitative data with constant coefficient of variation $[v^*(\mu) = c\mu^2]$ the data could have a Gamma distribution.

We consider response curves that form a location family, i.e. have identical (but arbitrary) shape and different positions along the real line. Formally, $\mu_k(x) = \mu(x - u_k)$ for some function $\mu(\cdot)$ that is almost everywhere continuous, and with location parameters for which we take the indicator values $\{u_k\}$. It follows that $v_k(x) = v(x - u_k)$, where $v(\cdot)$ is the variance function corresponding to $\mu(\cdot)$. We use asymptotics in which the number of species available for the estimation of x_0 increases indefinitely in such a way that the indicator values become increasingly densely spaced on every finite interval.

3. CONSISTENCY AND THE DEFINITION OF INDICATOR VALUE

Whether the weighted average is a "good" estimator depends on (1) the shape of the response curves, (2) the definition of indicator value, and (3) the distribution of the indicator values along the environmental variable. In this section we reverse the reasoning: we *require* that the weighted average be a consistent estimator of x_0, and from that requirement we derive conditions on the response curves, a definition of indicator value, and conditions on the distribution of the indicator values.

We express the number of indicator values at the point x by $\lambda[H_\lambda(x) - H_\lambda(x-0)]$, where λ is the average number of indicator values per unit length, $H_\lambda(x-0) = \lim_{y \uparrow x} H_\lambda(y)$, and $H_\lambda(\cdot)$ is a nondecreasing right-continuous stepfunction [in the terminology of measure theory, $H_\lambda(\cdot)$ is the distribution function of a discrete measure]. We suppose that for $\lambda \to \infty$ $H_\lambda(\cdot)$ converges to a distribution function with bounded and continuous derivative $h(\cdot)$. $h(\cdot)$ is the limiting density function of the indicator values. Now, $\hat{x}_{\mathrm{WA}} = T/R$, where $T = \lambda^{-1}\Sigma_k Y_k u_k$ and $R = \lambda^{-1}\Sigma_k Y_k$. It follows that T has expectation $\lambda^{-1}\Sigma_k u_k \mu(x_0 - u_k) = \int u\mu(x_0 - u)\, dH_\lambda(u)$, which for large λ approaches

$$\int u\mu(x_0 - u)h(u)\,du = x_0\int \mu(u)h(x_0 - u)\,du - \int u\mu(u)h(x_0 - u)\,du \tag{3.1}$$

Moreover, $\mathrm{var}(T) \to 0$ $(\lambda \to \infty)$ if and only if $\int x^2 v(x)\,dx$ exists; then T converges in probability to (3.1). Similarly, $R = \lambda^{-1}\Sigma_k Y_k$ converges in probability to $\int \mu(u)h(x_0 - u)\,du > 0$. Therefore T/R converges to x_0 if and only if $\int u\mu(u)h(x_0 - u)\,du = 0$. The latter condition should hold for every value of x_0; this condition may be fulfilled if the function $h(x)$ is constant, i.e. if the indicator values are evenly distributed. For particular $\mu(\cdot)$, certain almost periodic functions $h(\cdot)$ might do as well, but we believe these functions to be of no practical importance. For some $\mu(\cdot)$, e.g. the Gaussian curve [1, 9], constant $h(\cdot)$ is a necessary condition. If $h(x) = c$, we get $\int u\mu(u)\,du = 0$: the centroid of $\mu(\cdot)$ must be equal to zero. Consequently, the centroid of $\mu_k(x) = \mu(x - u_k)$ must be equal to u_k, or rephrasing, the indicator values must be the *centroids* of their response curves,

$$u_k = \frac{\int x\mu_k(x)\,dx}{\int \mu_k(x)\,dx}. \tag{3.2}$$

This definition of indicator value is necessary for the weighted average to be consistent. Note that defined in this way, the indicator value of a unimodal response curve is only equal to the most preferred value (mode or optimum) if the curve is symmetric. Note also that we had to assume in the derivation that both integrals in (3.2), and $\int x^2 v(x)\,dx$, exist. The weighted average is inconsistent for response curves that do not satisfy these conditions, e.g. monotone increasing or decreasing functions. The weighted average is also inconsistent for data with a constant variance function.

In conclusion, the weighted average is a consistent estimator of x_0 (for $\lambda \to \infty$) provided (1) the three aforementioned conditions on integrals of the response and variance curve hold, (2) the indicator values are centroids of the response curves, and (3) the indicator values are evenly distributed along the real line. Using central limit theorems and laws of large numbers valid for independent but nonidentically distributed random quantities [5], it follows that the weighted average is then asymptotically normal with variance [11, Equation (10.17), p. 247]

$$v_{\mathrm{WA}} = \frac{\sum_k (u_k - x_0)^2 v_k(x_0)}{\left[\sum_k \mu_k(x_0)\right]^2}. \tag{3.3}$$

4. THE MAXIMUM LIKELIHOOD APPROACH

When response curves can be expressed in parametric form, x_0 can be estimated by the method of maximum likelihood [4]. Maximum likelihood estimators are often good estimators in large samples: under mild conditions they are consistent and asymptotically normal with minimal variance [4, 5]. These assertions hold for our applications; the proof thereof goes along similar lines as in the standard case of independent and identically distributed random variables. Maximum likelihood is more widely applicable than weighted averaging.

For Bernoulli, Poisson, or Gamma random variables the maximum likelihood estimator is the solution for x_0 of the maximum likelihood equation [14]

$$\frac{\delta \log L}{\delta x_0} = \sum_k \frac{\mu_k'(x_0)[Y_k - \mu_k(x_0)]}{v_k(x_0)} = 0, \tag{4.1}$$

where $\mu_k'(x_0)$ denotes the derivative of $\mu_k(x)$ with respect to x, evaluated at x_0. Often the solution of (4.1) can only be obtained by numerical methods. The asymptotic variance of the maximum likelihood estimator is, as usual, the inverse of the information [4] and equals

$$v_{\mathrm{ML}} = \left[\sum_k \frac{\{\mu_k'(x_0)\}^2}{v_k(x_0)}\right]^{-1}. \tag{4.2}$$

When the distribution of Y_k is not fully specified, Equation (4.1) is a quasi-likelihood equation, which often gives estimators with good asymptotic properties [14]. This extension of (4.1) and (4.2) is important when count data are overdispersed with variance proportional to the mean.

5. EFFICIENCY AND SHAPE

For large numbers of species maximum likelihood will in general be more efficient than weighted averaging, but the latter method is much easier to use. It is therefore of interest to investigate whether there exists a shape of the response curves for which weighted averaging achieves, in terms of mean squared error, asymptotically unit efficiency with respect to maximum likelihood. With the species packing model [13, 22] in view, we adopt the location family of Section 2 with equispaced indicator values. In this situation both methods are consistent. It is therefore sufficient to compare the variances (3.3) and (4.2) for spacing $d \to 0$. It is proved in the Appendix that, asymptotically, $v_{\mathrm{ML}} \leqslant v_{\mathrm{WA}}$ with equality if and only if

$$\mu_k'(x) = -\frac{(x - u_k)v_k(x)}{t^2} \tag{5.1}$$

for t a nonzero constant. The differential equation (5.1) has a solution of the form

$$f(\mu_k(x)) = a - \frac{1}{2}\frac{(x-u_k)^2}{t^2}, \tag{5.2}$$

where the function $f(\cdot)$ depends on the variance function. The curves in (5.2) form a generalized linear model [14, 16], and the function $f(\cdot)$ is precisely the "natural" link function of such a model: the logistic function $f(\mu) = \log[\mu/(1-\mu)]$ for Bernoulli variables, the logarithmic function $f(\mu) = \log \mu$ for Poisson variables, and the inverse function $f(\mu) = -1/\mu$ (and $a < 0$) for Gamma variables. In (5.2) the parameter a is the maximum of $f(\cdot)$ attained at the indicator value, mode, or optimum u_k, and t, termed the tolerance, is a measure of curve width. For Poisson variables (5.2) is precisely the Gaussian response curve that is frequently invoked in plant ecological studies [1, 9].

For presence-absence data we propose to term (5.2) the Gaussian logit response curve (Figure 1). Its formula is

$$\mu_k(x) = \frac{\exp\left\{a - \frac{1}{2}(x-u_k)^2/t^2\right\}}{1+\exp\left\{a - \frac{1}{2}(x-u_k)^2/t^2\right\}}. \tag{5.3}$$

Instead of a we may use the parameter $p_{\max} = 1/(1+e^{-a})$, the maximum probability of occurrence. If $p_{\max} \to 0$, $\mu_k(x)$ approaches the Gaussian curve. Thus for many rare species, the two models are effectively the same. Using (3.3) and (4.2), we found numerically that for Bernoulli variables and Gaussian rather than Gaussian logit curves, the efficiency ($v_{\mathrm{ML}}/v_{\mathrm{WA}}$) of weighted averaging decreased from 1.0 to 0.8 when $p_{\max}$ was increased from near zero to 0.9.

The maximum likelihood variance (4.2) can be simplified by substitution of (5.1), which gives

$$v_{\mathrm{ML}} = t^4\left[\sum_k (u_k - x_0)^2 v_k(x_0)\right]^{-1}, \tag{5.4}$$

Because of the equal spacing of the indicator values,

$$\sum_k (u_k - x_0)^2 v_k(x_0) \approx t^2 \sum_k \mu_k(x_0). \tag{5.5}$$

For integrals the approximation (5.5) is an equality, as follows from (5.1) and integration by parts. Numerical calculations showed that the approximation in (5.5) is quite good, provided the indicator values are equispaced on a "large" interval I around x_0 with spacing less than t, where $I =$

$\{u \mid \mu(x_0 - u) > \delta,\ u \in \mathbb{R}\}$ for small δ. With (5.5) we obtain

$$v_{\mathrm{ML}} \approx \frac{t^2}{\sum_k \mu_k(x_0)}. \tag{5.6}$$

Substitution of (5.5) in (3.3) gives the same result for v_{WA}. A sample-based version of (5.6) is $t^2/\Sigma_k Y_k$.

We carried out a simulation study in which presence-absence data were generated according to the model (5.3) with $t = 1$, equispaced optima ($d \leqslant 1$: $d = 1$, 0.5, 0.25, 0.12, 0.06, or 0.03) on the interval $(-5,5)$ and maximum probability either .1 or .5 or .9. The minimum number of species was therefore 10. x_0 was always chosen close to the center of the interval, between 0 and $d/2$. The simulations were constrained to give at least two species occurrences per sample. In each case 1000 samples were generated. For each sample x_0 was estimated by weighted averaging and by maximum likelihood. All cases showed an efficiency in terms of mean squared error of 1.00, even when only 10 species were positioned on the interval. In most cases the mean squared error of both $\hat{x}_{\mathrm{WA}}$ and $\hat{x}_{\mathrm{ML}}$ exceeded the theoretical variance (5.6), but the excess was less than 12% when the average number of species occurrences per sample was larger than 5.

6. VARYING SPACING, MAXIMA, AND TOLERANCES

For the "optimal" response curves (5.2) the weighted average still has asymptotically unit efficiency when the species can be divided into sets such that within each set the species have equal maxima and equispaced optima with spacing less than t (Figure 1). An important example arises when the species are divided into sets on the basis of their response to another environmental variable. The result follows from (5.5): for each set of species (5.5) holds and can be substituted for each set in (3.3) and (5.4), which leads to (5.6) in both cases. However, this trick does not carry through when the tolerance varies between species, because substitution of (5.5) now involves different tolerances for different sets. As a result the efficiency can drop considerably when the tolerance varies. For example, with two tolerances differing by a factor of two, the efficiency drops to ca. 0.6 in the logistic model with maximum probability of occurrence .5. Full efficiency can then be retained by using a tolerance-weighted version of the weighted average,

$$\hat{x}_{\mathrm{WAT}} = \sum_k \frac{Y_k u_k}{t_k^2} \Bigg/ \sum_k \frac{Y_k}{t_k^2}. \tag{6.1}$$

In (6.1) good indicator species get more weight than bad ones, an intuitively reasonable idea used already by Zelinka and Marvan [24]. The results of this section suggest that equality of tolerances is a more critical assumption in the weighted average (2.1) than equality of maxima and equal spacing.

7. RANDOM INDICATOR VALUES AND RANDOM RESPONSE CURVES

The shapes of response curves may vary between species. In this section we mimic this variability by assuming that response curves arise from a "superpopulation" model consisting of three parts:

(1) A Poisson point process P that generates indicator values $\{u_k\}$ on the real line with intensity function $\lambda h(x)$ [$\lambda > 0$ and $h(x) > 0$ for every x].

(2) A stochastic process S that generates shapes $M(x)$ for response curves, independently for any indicator value u_k generated by P. Any realization of $M(x)$ is a bounded, nonnegative continuous function on the real line such that $x^2M(x)$ and $x^2V(x) \in L^1(-\infty,\infty)$, where $V(\cdot)$ is the variance function corresponding to $M(\cdot)$, and $\int xM(x)\,dx = 0$. Expectation and variance with respect to S are denoted by E_S and var_S.

(3) A translation of $M(x)$ over u_k: $M_k(x) = M(x - u_k)$.

The model will be termed the translation model. It is proved in the Appendix that the weighted average is consistent ($\lambda \to \infty$) if $h(x) = 1$. Then P is a homogeneous Poisson process, and the indicator values are said to be randomly spaced. The asymptotic variances are then

$$v_{\mathrm{WA}} = \frac{\int (u - x_0)^2 E_S\{V(u) + M^2(u)\}\,du}{\lambda\left[\int E_S M(u)\,du\right]^2} \tag{7.1}$$

and

$$v_{\mathrm{ML}} = \left[\lambda \int E_S\left\{\frac{[M'(u)]^2}{V(u)}\right\}du\right]^{-1} \tag{7.2}$$

respectively. v_{WA} is always strictly greater than v_{ML}. For the response curves (5.2) (process S degenerate) and random spacing, the efficiency of weighted averaging increases to unity when the maximum of $\mu(\cdot)$ decreases to 0, as shown in Figure 2 for logistic $f(\cdot)$. To obtain the variances in the case of equal instead of random spacing between the indicator values, $M^2(u)$ in (7.1) must be replaced by $\mathrm{var}_S\{M(u)\}$, whereas (7.2) remains the same. In this case $v_{\mathrm{ML}} \leqslant v_{\mathrm{WA}}$ with equality if and only if the response curves are nonrandom and satisfy (5.2).

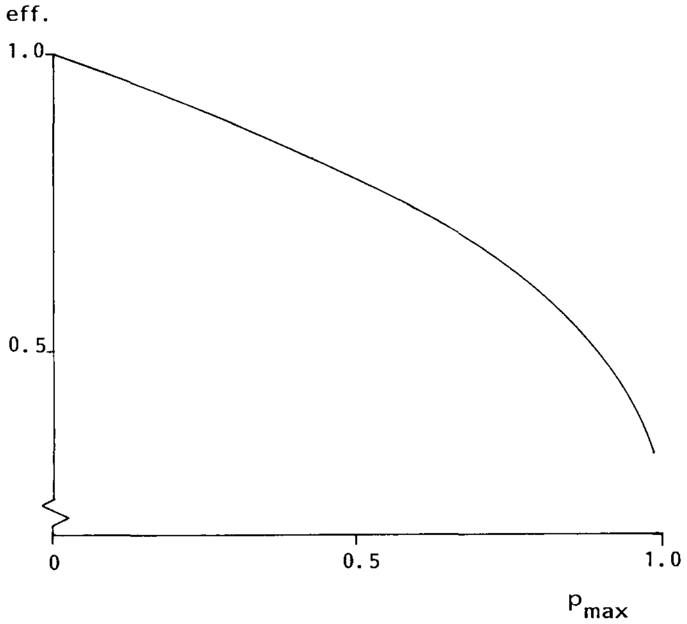

FIG. 2. The efficiency of weighted averaging with respect to maximum likelihood against the maximum probability of occurrence ($p_{\max}$) for Gaussian logit curves with randomly spaced optima and equal maxima and tolerances [eff $= v_{\mathrm{ML}}/v_{\mathrm{WA}} = (t/\tau)^2$].

To simplify (7.1) for Bernoulli variables we define the *commonness* α and the *standard deviation* τ of the expected response curve $\mu(x) = E_S\{M(x)\}$ by

$$\alpha = \int \mu(x)\,dx \quad \text{and} \quad \tau^2 = \frac{\int x^2\mu(x)\,dx}{\alpha} \tag{7.3}$$

From (7.1) we obtain [cf. (5.6)]

$$v_{\mathrm{WA}} = \frac{\tau^2}{\lambda\alpha}. \tag{7.4}$$

An unbiased estimator for (7.4) is the usual sample variance of the mean of the indicator values of the species present at the site. It is only in this special case that the indicator values might be considered as independent "samples" from a probability distribution.

Simulations, as in Section 5, with Gaussian logit curves (5.3), but with random, instead of equispaced, optima showed calculated efficiencies that agreed with the asymptotic efficiencies shown in Figure 2. The mean squared errors exceeded the theoretical variances (5.6) and (7.4), the convergence to the theoretical variances being slower than in Section 5. For random optima the excess was less than about 15% when the average number of species occurrences per sample was larger than 10.

8. DISCUSSION

This paper shows that a method proposed and used by community ecologists, namely weighted averaging, performs well under a model advocated by evolutionary ecologists, namely the species packing model [13]. This model is based on the idea that competing species evolve to occupy maximally separated niches with respect to a limiting resource. This idea applies as well to the occurrence of competing species along habitat variables [22]. Response curves should therefore have minimal overlap; hence, equally spaced indicator values. It should be noted that our asymptotic theory ignores another consequence of this model, namely that there exists a limiting similarity beyond which competing species cannot coexist. The minimal spacing derived by MacArthur and Levins [13] is about equal to the standard deviation of the response curves. But direct gradient analyses often show much closer spacings than that [9, 22]. Moreover, in lists of indicator values such as Ellenberg [8], the values coincide for many species. Of course, many species are coexisting without seriously competing.

Our results suggest that the distribution of the indicator values along the indicated variable should be even. But for Ellenberg's [8] list with about 2000 plant species the indicator values show uneven and markedly skew distributions [6, Figure 11]. A change of scale of the environmental variables could alleviate this problem. However, such a change modifies the response curves as well as their centroids. If the indicator values are centroids on the present scale, a nonlinear change of scale would destroy this desirable property. An alternative estimator is obtained by replacing Y_k with $Y_k/h(u_k)$ in (2.1). This estimator can be shown to be consistent under the model of Section 7. However, when the species packing model does hold in a part, say A, of a *multi*dimensional habitat space, possibly uneven *marginal* distributions of indicator values do not destroy the attractive properties of the usual weighted average (2.1). More specifically, when the indicator values are regularly spaced and the value $\mathbf{x}_0$ of the site lies well within A (i.e., there is a subset B of A such that $B = \{\mathbf{u} \mid \mu(\mathbf{x}_0 - \mathbf{u}) > \delta,\ \mathbf{x}_0 \in \mathbb{R}^n,\ \mathbf{u} \in \mathbb{R}^n\}$ for small δ), then for decreasing spacing along all n environmental variables:

(1) The weighted average is consistent if each indicator value is the centroid of the response curve that is obtained after integration of

the corresponding response surface over the remaining $n-1$ dimensions, and the integrals, defined in Section 3, of the "marginal" response curve exist.

(2) The weighted average has asymptotically unit efficiency with respect to maximum likelihood if the response surfaces are the multivariate extension of (5.2), namely

$$f(\mu_k(x_1, x_2, \ldots, x_n)) \\ = a - \frac{1}{2}\left\{\frac{(x_1-u_{k1})^2}{t_1^2} + \frac{(x_2-u_{k2})^2}{t_2^2} + \cdots + \frac{(x_n-u_{kn})^2}{t_n^2}\right\}, \quad (8.1)$$

where $x_1, x_2, \ldots, x_n$ are the variables of a n-dimensional habitat space, u_{kj} and t_j are the optimum and tolerance of the k-th species with respect to x_j and $f(\cdot)$ is as in Section 5. [With maximum likelihood based on (8.1) the values of $x_1, x_2, \ldots, x_n$ at the site are estimated jointly.]

The first assertion can easily be verified. The second assertion follows from Section 6: for fixed, but unknown values of $x_1, x_2, \ldots, x_n$ the species have different maxima with respect to x_1, but can be divided into sets of species with equal maxima because of the regular spacing in multidimensional habitat space.

Weighted averaging ignores species that are absent, whereas the maximum likelihood method uses the response curves of all species. In maximum likelihood, absent species do potentially provide information on the environment. This paper shows that this information is negligible under the (multidimensional) species packing model. Another, more informal model under which absent species do not add much information arises when the maximum probability of occurrence is close to zero. Then, the probability of absence is close to unity—irrespective of the value of the environmental variable—and hence cannot strongly influence the likelihood (see also Figure 2). The probability of occurrence of a species, given the value of a factor, will be small in practice for most species, just because in most sites with that value the species will be absent due to other, unfavorable factors (cf. the effect of neglecting other variables in a multidimensional species packing model). Absences therefore often indicate little.

Weighted averaging is central to the algorithm of the ordination technique known as reciprocal averaging or correspondence analysis. Reciprocal averaging is commonly used in ecological ordination studies to analyse data on the incidence or abundance of species in samples [9]. The first few ordination axes are often interpreted as latent variables and are presumed to relate to underlying habitat variables. The results of this paper can be extended to provide a theoretical basis of the model that is implicitly invoked when reciprocal averaging is used. Under the conditions of the species packing model it can be shown that reciprocal averaging approximates the maximum likelihood solution of Gaussian-like response models in one latent variable. The stochastic model of Section 7 is an explicit formulation of the model that is used by Hill and Gauch [10] to scale the axes of (detrended) correspondence analysis.

APPENDIX

Proof of (5.1). We prove that

$$\frac{\left[\int \mu(x)\,dx\right]^2}{\int x^2 v(x)\,dx \cdot \int \{[\mu'(x)]^2/v(x)\}\,dx} \leqslant 1 \tag{A1}$$

with equality iff $\mu'(x) = -xv(x)/t^2$. The left hand side in (A1) is the asymptotic ($d \to 0$) efficiency $v_{\mathrm{ML}}/v_{\mathrm{WA}}$, because summations in (3.3) and (4.2) approach integrals for $d \to 0$, and after translation, $x_0 = 0$. We use the Cauchy-Schwartz inequality

$$\left[\int p(x)q(x)\,dx\right]^2 \leqslant \int p^2(x)\,dx \int q^2(x)\,dx \tag{A2}$$

for arbitrary functions $p(x)$ and $q(x) \in L^2(-\infty,\infty)$. Equality in (A2) holds iff $p(x) = cq(x)$ with c a constant. By setting

$$p(x) = x\sqrt{v(x)} \quad \text{and} \quad q(x) = \frac{\mu'(x)}{\sqrt{v(x)}} \tag{A3}$$

and assuming that $x\mu(x) \to 0$ for $x \to \pm\infty$, so that

$$\int x\mu'(x)\,dx = -\int \mu(x)\,dx, \tag{A4}$$

we obtain (A1) with equality iff $xv(x) = c\mu'(x)$, from which (5.1) follows with $c = -t^2$. The condition $c < 0$ arises from the assumption above (A4).

Outline proof of (7.1). Expectations and (co)variances are required of $R = \Sigma_k Y_k$ and $T = \Sigma_k Y_k u_k$. These are calculated by dividing the real line into small intervals with midpoints $u_{(i)}$ ($i = \ldots, -2, -1, 0, 1, 2, \ldots$) and width Δ. The expectations correspond to the formulae in Section 3 with $\mu(u)$ replaced by $\lambda E_S M(u)$; hence $\hat{x}_{\mathrm{WA}}$ is consistent if $h(x)$ is constant. We show the derivation of the variances for $x_0 = 0$ and $h(x) = 1$. Repeated use is made of the decomposition of the variance as the sum of two components: (a) the average conditional variance, and (b) the variance of the conditional average [18, Equation (2b.3.6), p. 97]. Species with indicator values that lie in the ith interval contribute to var(R) an amount

$$c_i = \lambda\Delta\left[E_S\{V(u_{(i)})\} + \mathrm{var}_S\{M(u_{(i)})\} + E_S^2\{M(u_{(i)})\}\right], \qquad \text{(A5)}$$

and to var(T) an amount $u_{(i)}^2 c_i$. The last two terms in (A5) can be combined to give $E\{M^2(u_{(i)})\}$. The total variance can be obtained by summing over all intervals, because the data from different intervals are independent, due to the properties of the Poisson process. Replacing sums by integrals gives, with $g(u) = E_S\{V(u) + M^2(u)\}$,

$$\begin{aligned} \mathrm{var}(R) &= \lambda\int g(u)\,du, \\ \mathrm{var}(T) &= \lambda\int u^2 g(u)\,du, \\ \mathrm{cov}(R,T) &= \lambda\int u g(u)\,du. \end{aligned} \qquad \text{(A6)}$$

Because $u^2M(u)$ and $u^2V(u) \in L^1(-\infty,\infty)$, we have var($T/\lambda$), var($R/\lambda$), and cov($R/\lambda, T/\lambda$) $\to 0$ for $\lambda \to \infty$; this and Taylor expansion of T/R [11, Equation (10.17), p. 247] yield (7.1).

Outline proof of (7.2). Let $\hat{x}$ denote the maximum likelihood estimator, D_y the first x derivative of the log likelihood (4.1) evaluated at y, and I the total information evaluated at x_0. Without confusion, the symbol x will now be used for x_0. A first order Taylor expansion of $D_{\hat{x}}$ in x_0 gives [4, Chapter 9.2, Equation (19)]

$$D_{\hat{x}} = D_x - (\hat{x} - x)I. \qquad \text{(A7)}$$

Equating (A7) to zero, as in (4.1), and solving for $\hat{x} - x$ shows that, asymptotically ($\lambda \to \infty$),

$$\mathrm{var}(\hat{x}) = \frac{\mathrm{var}(D_x)}{I^2}. \qquad \text{(A8)}$$

Conditionally on S and P, the expectation of D_x is equal to zero and its variance is the inverse of (4.2). Unconditionally, the variance of D_x is therefore equal to the quantity between square brackets in (7.2). The total information is the expectation over S and P of the conditional information.

This expectation is equal to the variance of D_x; hence, from (A8) we obtain (7.2).

We would like to thank Drs. I. C. Prentice, M. O. Hill, and J. A. Hoekstra for valuable comments. Drs. T. A. B. Snijders and M. J. M. Jansen contributed to Section 3.

REFERENCES

1 M. P. Austin, On non-linear species response models in ordination, *Vegetatio* 33:33–41 (1976).

2 R. W. Batterbee, Diatom analysis and the acidification of lakes, *Philos. Trans. Roy. Soc. London Ser. B* 305:451–477 (1984).

3 R. Böcker, I. Kowarik, and R. Bornkamm, Untersuchungen zur Anwendung der Zeigerwerten nach Ellenberg, *Verh. Ges. Oekol.* 11:35–56 (1983).

4 D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, 1974.

5 H. Cramér, *Mathematical Methods of Statistics*, Princeton U.P., Princeton, N.J., 1946.

6 K.-J. Durwen, Zur Nutzung von Zeigerwerten und artspezifischen Merkmalen der Gefässpflanzen Mitteleuropas für Zwecke der Landschaftsökologie und -planung mit Hilfe der EDV-Voraussetzungen, Instrumentarien, Methoden und Möglichkeiten, *Arbeitsber. Lehrst. Landschaftsökologie Munster* 5:1–138 (1982).

7 H. Ellenberg, Unkrautgesellschaften als Mass für den Säuregrad, die Verdichtung und andere Eigenschaften des Ackerbodems, *Ber. Landtech.* 4:130–146 (1948).

8 H. Ellenberg, Zeigerwerten der Gefässpflanzen Mitteleuropas, *Scripta Geobotanica* 9:1–122 (1979).

9 H. G. Gauch, *Multivariate Analysis in Community Ecology*, Cambridge U.P., Cambridge, 1982.

10 M. O. Hill and H. G. Gauch, Detrended correspondence analysis: An improved ordination technique, *Vegetatio* 42:47–58 (1980).

11 M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, 4th ed., Griffin, London, 1977.

12 M. Kovâcs, Das Corno-quercetum des Mâtra-gebirges, *Vegetatio* 19:240–255 (1969).

13 R. H. MacArthur and R. Levins, The limiting similarity, convergence, and divergence of co-existing species, *Amer. Natur.* 101:377–385 (1967).

14 P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1983.

15 R. Mead and D. J. Pike, A review of response surface methodology from a biometric viewpoint, *Biometrics* 31:803–851 (1975).

16 J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, *J. Roy. Statist. Soc. Ser. A* 135:370–384 (1972).

17 S. Persson, Ecological indicator values as an aid in the interpretation of ordination diagrams, *J. Ecol.* 69:71–84 (1981).

18 C. R. Rao, *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.

19 V. Slâdecek, System of water quality from the biological point of view, *Arch. Hydrobiol. Beiheft* 7:1–218 (1973).

20 H. Van Dam, G. Suurmond, and C. J. F. Ter Braak, Impact of acidification on diatoms and chemistry of Dutch moorland pools, *Hydrobiologia* 83:425–459 (1981).

21 G. Van Wirdum, Linking up the natec subsystem in models for the water management, *Comm. Hydrol. Res. TNO (Centr. Organ. Appl. Sci. Res. Neth.) Proc. Inf.* 27:108–128 (1981).

22 R. H. Whittaker, S. A. Levin, and R. B. Root, Niche, habitat and ecotope, *Amer. Natur.* 107:321–338 (1973).

23 R. Wittig and K.-J. Durwen, Ecological indicator value spectra of spontaneous urban floras, in *Urban Ecology* (R. Bornkamm, J. A. Lee, and M. R. D. Seaward, Eds.), Blackwell, Oxford, 1982, pp. 23–32.

24 M. Zelinka and P. Marvan, Zur Präzisierung der biologischen Klassification der Reinheit fliessender Gewässer, *Arch. Hydrobiol.* 59:389–407.

25 L. Zhang, Vegetation ecology and population biology of *Fritillaria meleagris* L. at the Kungsängen Nature Reserve, Eastern Sweden, *Acta Phytogeogr. Suec.* 73:1–92 (1983).