Unimodal models to relate species to environment

Promotor: dr. ir. L. C. A. Corsten
hoogleraar in de wiskundige statistiek

Co-promotor: dr. I. C. Prentice
wetenschappelijk medewerker,
Institute of Ecological Botany, University of Uppsala

Cajo J. F. ter Braak

Unimodal models to relate species to environment

Proefschrift
ter verkrijging van de graad van
doctor in de landbouwwetenschappen,
op gezag van de rector magnificus,
dr. C. C. Oosterlee,
in het openbaar te verdedigen
op maandag 16 november 1987
des namiddags te vier uur in de aula
van de Landbouwuniversiteit te Wageningen

ISN 29630

Aan mijn moeder

Deze uitgave is voor f 25,- te bestellen bij de Groep Landbouwwiskunde, postbus 100, 6700 AC Wageningen, onder vermelding van 'Unimodal models'.

This publication can be ordered from the Agricultural Mathematics Group, Box 100, NL-6700 AC Wageningen, The Netherlands. The price is Dfl 25.

Stellingen

1. Gewoonlijk leiden statistici vanuit een model en een optimaliteitscriterium de optimale techniek af. In technieken die niet op die manier tot stand gekomen zijn, wordt het inzicht vergroot door te zoeken naar een bijbehorend optimaal model.
   Dit proefschrift.

2. Een benadering van een statistische techniek is soms redelijker dan de statistische techniek zelf.
   Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). J. R. Statist. Soc. B 48: 259-302.
   Dit proefschrift.

3. Hoofdcomponentenanalyse en correspondentie-analyse verschillen in metriek. Achter dit verschil gaat een verschil in model schuil.
   Dit proefschrift.

4. Partiële kleinste-kwadratenregressie en Procrustes-analyse benadrukken respectievelijk de variabelen en de objecten in één singuliere-waardenontbinding van de matrix van covarianties tussen de variabelen in de ene configuratie van objecten en die in de andere.
   Aastveit, A. H. & Martens, H. (1986). ANOVA interactions interpreted by Partial Least Squares regression. Biometrics 42: 829-844.
   Sibson, R. (1978). Studies in the robustness of multidimensional scaling: procrustes statistics. J. R. Statist. Soc. B 40: 234-238.

5. Expertsystemen kunnen een kader bieden voor groei van kennis over levensgemeenschappen.

6. Net als variantie is de diversiteit van een levensgemeenschap een eigenschap van de tweede orde en dus moeilijker te schatten dan dichtheden van aparte soorten.

7. Het promotiereglement van de Landbouwuniversiteit sluit met de eis dat stellingen vatbaar moeten zijn voor bestrijding wiskundige stellingen uit.

8. Modelbouwers zijn optimisten, statistici pessimisten.

9. „Was sind das für Zeiten, wo
   Ein gespräch über Bäume fast ein Verbrechen ist
   Weil es ein Schweigen über so viele Untaten einschliesst."

   Brecht's dichtregels zijn ook van toepassing op wetenschappelijke kontakten met Zuidafrikanen.
   Brecht, B. (1973). An die Nachgeborenen (1938). In: Svendborger Gedichte, Suhrkamp.

10. Sport is betaalde arbeid of het afreageren daarvan.

Cajo J. F. ter Braak
„Unimodal models to relate species to environment"
Wageningen, 16 november 1987

Cajo J. F. ter Braak

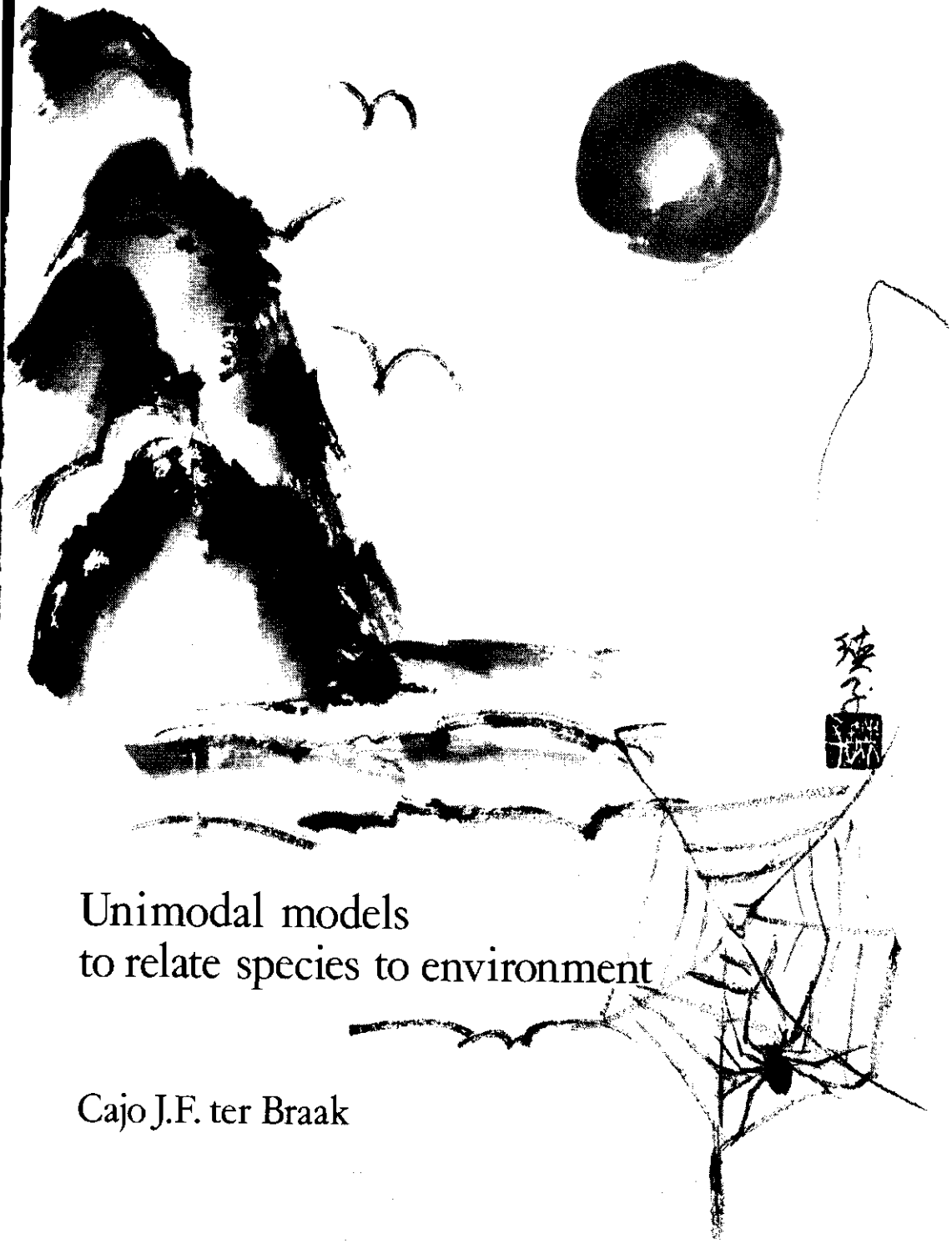
Unimodal models to relate species to environment

Proefschrift
ter verkrijging van de graad van
doctor in de landbouwwetenschappen,
op gezag van de rector magnificus,
dr. C. C. Oosterlee,
in het openbaar te verdedigen
op maandag 16 november 1987
des namiddags te vier uur in de aula van de
Landbouwuniversiteit, Generaal Foulkesweg 1A
te Wageningen

Promotor: **dr. ir. L. C. A. Corsten**
hoogleraar in de wiskundige statistiek

Co-promotor: **dr. I. C. Prentice**
wetenschappelijk medewerker, Institute of Ecological Botany, University of Uppsala

Na afloop borrel in café Troost tegenover de aula



Unimodal models to relate species to environment

Cajo J.F. ter Braak

Samenvatting

Bij de theoretische onderbouwing van natuurbeheer en milieu-effect-rapportage moeten de gevolgen worden getaxeerd van milieu-ingrepen op levensgemeenschappen. Kennis over de relatie tussen milieuvariabelen en het voorkomen van soorten is daarbij onontbeerlijk. Ecologen proberen die relaties te achterhalen door op verschillende monsterplekken soorten te inventariseren (op aan/afwezigheid of abundantie) en tevens huns inziens relevante milieuvariabelen te meten. Het onderzoek, dat tot dit proefschrift heeft geleid, richtte zich op het ontrafelen van de vereiste veronderstellingen van statistische methoden, die vaak door ecologen worden toegepast en op het ontwikkelen van een nieuwe techniek.

Vanuit klassiek statistisch oogpunt zijn soortgegevens moeilijk te verwerken:
- er zijn veel soorten bij betrokken (10-500);
- heel wat soorten komen maar op weinig plekken voor, dus de gegevens zitten vol nullen;
- verbanden tussen soorten en milieuvariabelen zijn vaak niet rechtlijnig, maar ééntoppig: een plant bijvoorbeeld groeit bij voorkeur onder een voor die soort optimale vochtconditie en wordt zowel op drogere als op nattere monsterplekken minder aangetroffen. Een wiskundig model voor een eentoppig verband is het Gaussische responsiemodel.

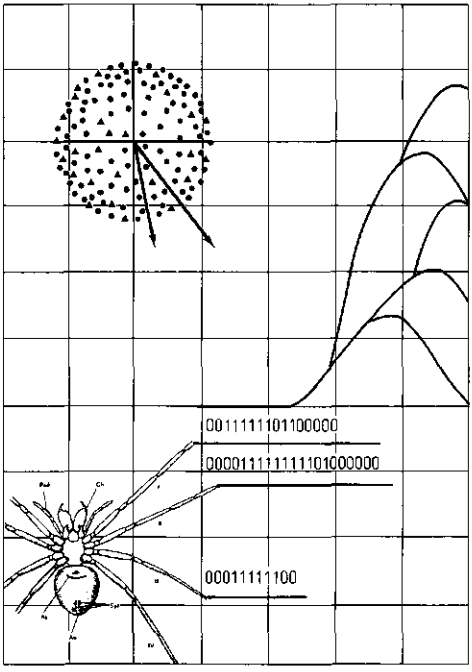
Klassieke methoden als lineaire regressie, hoofdcomponentenanalyse en canonische correlatie-analyse kunnen niet zinnig worden gebruikt, omdat ze van rechtlijnige verbanden uitgaan. Eén van de methoden, waar ecologen wel mee werken, is correspondentie-analyse. Het inzicht in het achterliggende responsiemodel hiervan liet tot voor kort te wensen over. Via correspondentie-analyse wordt een ordening in soorten en monsterplekken aangebracht (ordinatie) om de structuur in de gegevens te laten zien. De ordening wordt vervolgens aan de milieuvariabelen gekoppeld. Het is een indirecte methode om relaties op te sporen, ofwel een methode voor indirecte gradiënten-analyse.

Correspondentie-analyse werd omstreeks 1935 ontwikkeld, maar staat bij ecologen pas in de belangstelling sinds 1973. Toen leidde M. O. Hill de techniek opnieuw af als het herhaald toepassen van gewogen middelen - een methode waar ecologen al sinds 1930 mee vertrouwd zijn. Gewogen middelen heeft het voordeel van de eenvoud bij toepassing op ecologische gegevens. Deze techniek kan voor twee verschillende doelstellingen worden gebruikt. Ten eerste kan het optimum van een soort voor een milieuvariabele ermee geschat worden. Ten tweede kan bij bekende optima de waarde van een milieuvariabele op een monsterplek worden geschat (calibratie) aan de hand van de soortensamenstelling (dit is ook de methode die Ellenberg aanbeveelt voor gebruik van zijn milieu-indicatiegetallen).

In hoofdstuk 2 wordt het schatten van optima met gewogen middelen vergeleken met de resultaten van niet-lineaire regressie op basis van het Gaussische responsiemodel. Onder bepaalde voorwaarden blijken deze twee methoden precies overeen te komen. In andere gevallen schat men door gewogen middelen het optimum onzuiver en verdient niet-lineaire regressie de voorkeur. Bovendien kunnen met niet-lineaire regressie responsiemodellen met meer dan één milieuvariabele worden aangepast. In hoofdstuk 3 wordt het schatten van de waarde van een milieuvariabele via gewogen middelen afgezet tegen calibratie op basis van het Gaussische responsiemodel. Ook hier blijken de technieken soms equivalent te zijn. Hoofdstuk 4 gaat in op correspondentie-analyse. Er wordt aangetoond, dat correspondentie-analyse onder bepaalde voorwaarden een benadering geeft van ordinatie op basis van het Gaussische responsiemodel, wat qua rekentechniek veel ingewikkelder is.

Indirecte methoden voor het opsporen van relaties hebben een belangrijk nadeel. Een aantal milieuvariabelen kan de soortensamenstelling zo sterk beïnvloeden, dat het effect van andere interessante milieuvariabelen niet meer te achterhalen is. Alleen directe methoden als niet-lineaire regressie ondervangen dit probleem, maar niet-lineaire regressie met veel soorten en milieuvariabelen is zeer bewerkelijk. In hoofdstuk 5 wordt een veel eenvoudiger directe methode voorgesteld, canonische correspondentie-analyse. In hoofdstuk 6 blijkt canonische correspondentie-analyse een multivariate uitbreiding van gewogen middelen te zijn. De resultaten kunnen grafisch weergegeven worden. In hoofdstuk 7 wordt een uitbreiding met covariabelen besproken, wat leidt tot partiële canonische correspondentie-analyse. Er wordt tevens op gewezen dat Gaussische modellen en canonische correspondentie-analyse kunnen worden toegepast op afhankelijkheidstabellen.

Hoofdstuk 8 beschrijft onderzoek om ecologische amplitudes van planten ten opzichte van de vochtschaal van Ellenberg te bepalen op basis van alleen soortgegevens. Hoe consequent de vochtindicatiegetallen zijn is ook onderzocht. Hoofdstuk 9 tenslotte geeft een overzicht van gradiënten-analyse. Er is een computerprogramma ontwikkeld, CANOCO, waarmee het merendeel van de behandelde technieken kan worden uitgevoerd.


0011111101100000
00001111111101000000
00011111100

Voorwoord

Dit proefschrift is voortgekomen uit mijn werkzaamheden als consulterend statisticus voor het Rijksinstituut voor Natuurbeheer (RIN). Herman van Dam en Paul Opdam waren daar de eersten die mij advies vroegen over ordinatie en cluster-analyse. Via het WAFLO-project en de SWNBL-studies brachten Rien Reijnen, Jaap Wiertz, Niek Gremmen, Geert van Wirdum en Douwe van Dam me in contact met milieu-indicatie-getallen van hogere planten. Hun vragen en opmerkingen, en ook die van Hans van Biezen, hebben me bijzonder geïnspireerd. Later vormde ook het EKOO-project van Piet Verdonschot een stimulans. De directies van het RIN, het voormalige IWIS-TNO en het ITI-TNO ben ik zeer erkentelijk voor de ruimte en vrijheid die ik heb gekregen om dit onderzoek vorm te geven. Ik wil hiervoor met name danken de heren J.C.A. Zaat (IWIS-TNO) en ir. A.A.M. Jansen (Groep Landbouwwiskunde) en dr. A.J. Wiggers, dr. R.A. Prins, dr. A.B.J. Sepers en dr. C.H. Gast (allen van het RIN).

Tijdens een conferentie over statistische ecologie in 1978 te Parma kwam ik in contact met Rob Hengeveld en Bas Kooijman. Mede van hen heb ik geleerd hoe belangrijk unimodale modellen zijn voor de ecologie en hoe moeilijk ordinatie dan is. Tijdens mijn studiejaar (1979/1980) in Newcastle upon Tyne leerde ik Colin Prentice kennen. Hij werd mijn goeroe zonder wie ik dit onderzoek niet tot een goed einde had kunnen brengen. Mijn bezoek in 1980 aan Mark Hill in Bangor heeft grote invloed gehad. Ik was toen, mede door het contact met professor Corsten, zeer gecharmeerd van de elegantie van de hoofdcomponenten-analyse-biplot. Mark sprak zijn misprijzen uit over de toepassing hiervan in de ecologie, maar kon mij niet duidelijk maken wat het model was achter zijn "detrended correspondence analysis". In 1981, terug in Nederland, nam ik deel aan de PAO-cursus "Niet-lineaire multivariate analyse" te Leiden waarbij ik kennis maakte met het werk van Albert Gifi. Hoewel "niet-lineair" veelal "monotoon" betekende, heb ik veel aan de cursus gehad. Het werk van Willem Heiser daarin over ontvouwing ging wel uit van unimodale modellen. Pas later heb ik ingezien hoe nauw mijn eigen werk aansluit bij de hoofdstukken 6 en 8 van zijn proefschrift. Willem merkte ook de grote overeenkomst op tussen canonische correspondentie-analyse en Abby Israëls' redundantie-analyse voor nominale variabelen. Willem en Abby, hartelijk dank voor de vele zinnige discussies!

Een bijzonder stimulerende invloed hebben ook Onno van Tongeren, Rob Jongman en Caspar Looman gehad. Bedankt voor de goede samenwerking tijdens en na de PAO-cursussen "Numerieke methoden voor de verwerking van ecologische gegevens".

Ik dank ook mijn collega's op het Staringgebouw voor de prettige contacten. Zonder de secretariële ondersteuning door Mary Mijling en Joke van de Peppel en de technische ondersteuning door Martha de Vries zou dit onderzoek alleen maar bij een idee gebleven zijn. De mensen van de tekenafdeling en de fotoafdeling van het ICW wil ik hartelijk danken voor het teken- en fotowerk dat ze tussen de bedrijven door voor me hebben gedaan. De bibliotheek en het rekencentrum van het Staringgebouw verleenden uitstekende service!

De afbeelding op het voorkaft van dit proefschrift is gemaakt door Eiko Kondo met de Sumi-è schildertechniek en die van het achterkaft door Frank Arnoldussen. Hiervoor mijn hartelijke dank.

Een proefschrift is pas een proefschrift als het onderworpen is aan het kritische oog van een promotor. Professor Corsten wil ik bijzonder bedanken voor alle aandacht die hij aan dit proefschrift heeft besteed.

Tenslotte wil ik iedereen bedanken die aan de totstandkoming van dit proefschrift heeft bijgedragen, maar niet met name is genoemd.

UNIMODAL MODELS TO RELATE SPECIES TO ENVIRONMENT

CONTENTS

[1]) Published in Vegetatio 65: 3-11, 1986 (with C.W.N. Looman). Reproduced with permission of Dr. W. Junk Publishers.

[2]) Published in Mathematical Biosciences 78: 57-72, 1986 (with L.G. Barendregt). Reproduced with permission of Elsevier Science Publishing Co.

[3]) Published in Biometrics 41: 859-873, 1985. Reproduced with permission of the Biometric Society.

[4]) Published in Ecology 67: 1167-1179, 1986. Reproduced with permission of the Ecological Society of America.

[5]) Published in Vegetatio 69: 69-77, 1987. Reproduced with permission of Dr. W. Junk Publishers.

[6]) To appear in: Classification and related methods of data analysis. H.H. Bock, ed., North-Holland, Amsterdam, 1988.

[7]) Published in Vegetatio 69: 79-87, 1987 (with N.J.M. Gremmen). Reproduced with permission of Dr. W. Junk Publishers.

[8]) To appear with minor modifications in Advances of Ecological Research, 1988 (with I.C. Prentice).

Chapter 1. GENERAL INTRODUCTION

Introduction

In the last decades, many people have become aware of the human potential to cause environmental change both on a local scale (e.g., a temperature increase in a river by a power-plant) and on a global scale (e.g., acid rain, CO_2 increase by burning fossil fuels). To assess the impact of environmental change on biological communities, one needs to know the relations between environmental variables and the occurrence of species. Such knowledge is indispensable also for nature management.

Ecologists attempt to acquire knowledge about species-environment relations from data on biological communities and their environment. Typically, several sites are selected and at each site the occurrence or abundance of each species of a taxonomic group is recorded and environmental variables that ecologists believe to be important, are measured. So the data consist of two sets: data on the occurrence or abundance of several species at sites and data on several environmental variables measured at the same sites. A "site" is the basic sampling unit separated in space or time from other sites, e.g. a quadrat, a woodlot or a light trap. The design of ecological field studies is discussed by Green (1979) and Jager and Looman (1987). The design of impact studies in the strict sense is reviewed by Stewart-Oaten et al (1986).

This thesis deals with methods for the statistical analysis of ecological data on species and environmental variables. Such data have several features that make them special in a statistical sense:

1. the number of species is large (10 - 500),
2. the data are either binary (presence/absence of a species at a site) or, if they are quantitative, they contain many zero values for sites at which a species is absent. Measures of abundance, like density of animals or relative cover of plants, are highly variable and always show a skew distribution.
3. Relationships between species and quantitative environmental variables are generally nonlinear. Species abundance or probability of occurrence is often a unimodal function of the environmental variables.

The importance of unimodal relationships between species and environment has been realized since the beginning of this century. For example, Shelford's law of tolerance (1919: in Odum, 1971) states that a species not only requires a certain minimum amount of a resource (as in Liebig's law) but also that species do not tolerate more than a certain maximum amount of the resource. Hesse (1924: in Thienemann, 1950) stated a more general law: each species thrives best at a particular optimum value of an environmental variable and cannot survive when the value is either too low or too high. In the introduction to a study of the relationship between some Orthoptera species and moisture, Gause (1930: p. 307) stated that "the law of Gauss is the basis of ecological curves", but also that "we must not forget that factors exist (such as competition, for instance) which produce changes in different sections of the curve of distribution (Du Rietz, '21)." This warning still holds (Austin, 1980). In later work, Gause became more interested in competition and developed his principle of competitive exclusion (Gause, 1934).

Whittaker (1956, 1967) also stressed that species generally show unimodal relationships with environmental variables. Gauch and Whittaker

(1972) popularized the Gaussian curve as an attractively simple model for unimodal relationships. The formula of the Gaussian curve (Fig. 1) is

$$Ey_{ik} = c_k \, e^{-\frac{1}{2}(x_i - u_k)^2 / t_k^2} \qquad (1)$$

with y_{ik} the abundance of species k at site i (i = 1, ..., n; k = 1, ..., m) and Ey_{ik} is the expected abundance,
x_i the value of environmental variable x at site i,
c_k the maximum of the curve for species k,
u_k the optimum of species k, i.e. the value of x for which the maximum is attained,
t_k the tolerance of species, which is a measure of curve breadth or ecological amplitude.

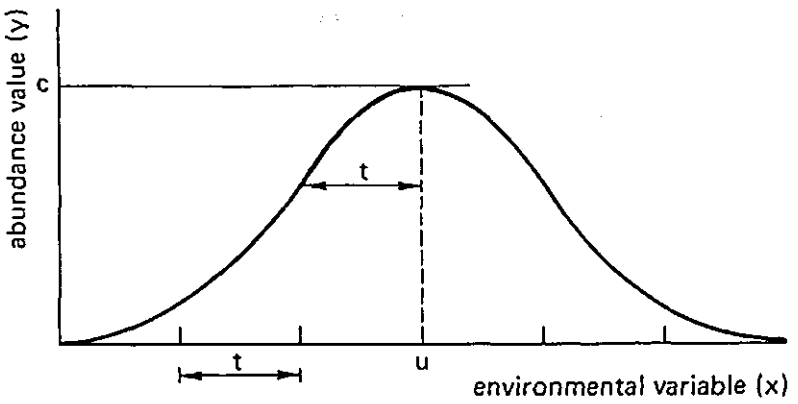


Fig. 1 The Gaussian response curve for the abundance value (y) of a species against an environmental variable (x). (u = optimum or mode; t = tolerance; c = maximum.)

Gauch and Chase (1974) developed an algorithm to estimate the species parameters (c_k, u_k, t_k) by nonlinear least-squares regression. By doing so, they made explicit that the Gaussian curve of Eq. (1) represents a response function, not a probability distribution. The species is considered to respond to the environmental variable: in the terminology of regression, the abundance of a species is the response variable and the environmental variable is the explanatory variable. An example of "Gaussian regression" is given by Westman (1980).

It should be noted that a unimodal curve may appear monotonic if only a limited range of the environmental variable is sampled. In such cases, the estimates of the parameters of Eq. (1) are ill-determined; in particular, the optimum cannot be estimated well, and a monotonic statistical model (e.g. fitting a straight line) is more appropriate. Unimodal relationships become visible when a sufficient range of the environmental variable(s) is considered. However, if the data are collected over a sufficient range of environments for species to show unimodal (or more complex) relationships with environmental variables, it is clearly inappropriate to analyse these relationships by standard statistical methods that assume linear relationships such as multiple linear regression (without squared terms in

the environmental variables) (Montgomery and Peck, 1982), principal components analysis (Jolliffe, 1986), factor analysis (Lawley and Maxwell, 1971), redundancy analysis (van den Wollenberg, 1977), canonical correlation analysis (Gittins, 1985) and LISREL models (Jöreskog and Sörbom, 1981). With multiple regression, unimodal models can be fitted by including squared terms in the environmental variables in the regression equation (e.g. Alderdice, 1972; Forsythe and Loucks, 1972), but multiple regression is unattractive in this context because the response variable (the abundance of a species) often has a skew distribution which cannot be transformed to symmetry because of the many zero values.

Ecologists have therefore used and adapted non-standard techniques to analyse their data (see e.g. Greig-Smith, 1983). Most conspicuously, ordination and cluster analysis have become very popular as reflected in the recent text books by Green (1979), Gauch (1982), Greig-Smith (1983), Legendre and Legendre (1983), Pielou (1984), Kershaw and Looney (1985), Digby and Kempton (1987) and Jongman et al (1987). These techniques are commonly used to reduce the multi- species data to a few ordination axes or a few relatively homogenous clusters. The ordination axes or clusters are then interpreted in the light of whatever is known about the species and the environment. This interpretation arises in an informal way, if explicit environmental data are lacking, or in a formal statistical way, if environmental data were collected. If many environmental variables were measured, ordination or cluster analysis are sometimes applied to the environmental data as well and the results are compared with the ordination or cluster analysis of the species data (see e.g. Wiens and Rotenberry, 1981; Bates and Brown, 1981; Holder-Franklin and Wuest, 1983; Earle et al, 1986). In this way the whole analysis becomes rather complicated. Species are related to environment in an indirect manner, hence Whittaker's (1967) term "indirect gradient analysis". Whittaker contrasted this with direct gradient analysis, which is similar to what statisticians call regression - i.e., the abundance of each species is described in relation to environmental variables.

Among the possible ordination techniques, ecologists most often use either principal components analysis, with various forms of prior transformation of the species data (Noy-Meir et al, 1975), or reciprocal averaging (alias correspondence analysis). Multidimensional scaling has also received attention, mainly in comparative studies of ordination techniques. Principal components analysis was the earlier technique to be used in ecology, with an application by Goodall (1954) but since Hill (1973) introduced reciprocal averaging to ecologists, reciprocal averaging has gained markedly in popularity over principal components analysis. Hill and Gauch (1980) later introduced detrended correspondence analysis as an improved form of reciprocal averaging, and this method has in recent years become possibly the most popular technique of all. This may be so partly because an efficient computer program (DECORANA) became available (Hill, 1979), but also because the new technique proved exceptionally effective for simulated data generated with the Gaussian model (Hill and Gauch, 1980).

In their 1980 paper, Hill and Gauch based the improvements made in detrended correspondence analysis on a "species packing model", that is a model in which the species have Gaussian curves which equispaced optima, equal maxima and equal tolerances (Fig. 2). But the rationale for this model is difficult to follow -- partly because mathematics is avoided -- and the Gaussian model appears to come out of thin air. Neither the 1980 paper, nor Hill's other papers (Hill, 1973, 1974), explain why correspondence analysis is suited for the analysis of data that follow the Gaussian model. The same is true of other rationales for correspondence analysis, most of which

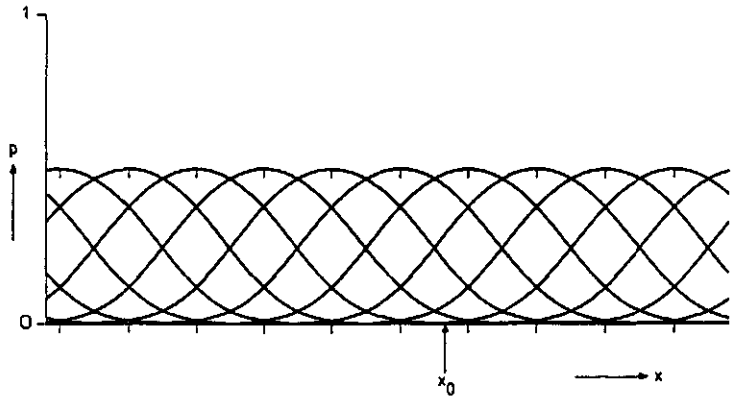


Fig. 2 Species packing model: Gaussian logit curves of the probability (p) that a species occurs at a site, against environmental variable x. The curves shown have equispaced optima, equal tolerances and equal maximum probabilities of occurrence (p_{max} = 0.5). x_0 is the value of x at a particular site.

concern categorical data (Nishisato, 1980; Gifi, 1981; Greenacre, 1984; Tenenhaus and Young, 1985). This thesis resulted from an attempt to understand the properties of correspondence analysis in terms of a unimodal model since this would provide a rationale for ecologists' use of correspondence analysis in indirect gradient analysis. I then began to explore methods that relate species directly to environment - methods like linear regression or canonical correlation analysis, but then in a form appropriate for the analysis of unimodal relationships.

Structure of the thesis

Four main types of statistical problems are dealt with in this thesis. Each type is specified for the Gaussian curve of Eq. (1), as follows (Table 1):

1. Regression - where parameters of a species are estimated from data of the corresponding species and of the environmental variable; so for species k, c_k, u_k, t_k are estimated from (y_{ik}) and (x_i) $[i = 1, \ldots, n]$.
2. Calibration - where the value of an environmental variable at a site is estimated from data of species and parameters of species; so for site i, x_i is estimated from (y_{ik}) and (c_k, u_k, t_k) $[k = 1, \ldots, m]$. Calibration is here a type of multi-species bio-assay. An example is the calibration of pH to reconstruct past changes in pH in lakes from fossil diatoms found in successive strata of the bottom sediment (Battarbee, 1984). [The way in which the term calibration is used in this thesis is somewhat narrow; more usually, the estimation of the species parameters from a training set is included.]
3. Ordination - where the parameters of species and the values of sites are estimated from data of the species; so, for all sites and species, x_i, c_k, u_k and t_k are estimated from (y_{ik}) $[i = 1, \ldots, n; k = 1, \ldots, m]$.

4. Constrained ordination = in which the values of the sites are not free parameters as in ordination, but are constrained to be a linear combination of environmental variables. Here, the parameters of species and the coefficients of the linear combination are estimated from the data of the species and the environmental variables.

Ecologists have developed much simpler methods than nonlinear regression and calibration. For both problems they invented heuristically the method of weighted averaging. It is shown in this thesis that, under simplifying circumstances, the method of weighted averaging gives efficient estimates of the optimum (u_k) of a Gaussian curve, in the regression context (Chapter 2), and of x_i in the calibration context (Chapter 3). The later chapters build further on these results. By applying the method of weighted averaging both ways and in an iterative fashion, Hill (1973) derived "reciprocal averaging", alias correspondence analysis. When Hill invented reciprocal averaging, correspondence analysis was already in existence, but was seldomly applied to ecological data. In chapter 4, correspondence analysis is shown to give an approximate solution to ordination on the basis of the Gaussian model. In the same way, canonical correspondence analysis is derived as an approximate solution to constrained ordination (Chapter 5). Canonical correspondence analysis satisfies ecologists' desire for a simple, robust method to relate species to environmental variables, if the relationships are assumed to be unimodal. In Chapter 6, canonical correspondence analysis is shown to be a multivariate extension of weighted averaging. In Chapter 7, the case is considered where the environmental variables are divided in a set of variables-of-interest and a set of covariables, leading to partial canonical correspondence analysis. It is also shown that constrained ordination can be seen as a form of constrained regression. Chapter 8 is a case study of a rather special estimation problem (Table 1). The concluding chapter 9 gives a synthesis of linear and unimodal methods to relate species to environment.

The remainder of this GENERAL INTRODUCTION gives a sketch of the context in which the chapters of this thesis were written. This is done for each of the main types of statistical problems just distinguished.

Table 1: Types of problems studied in the chapters of this thesis and the unknown parameters that are to be estimated, with special reference to the parameters of the Gaussian curve (1).

Type of problem	site values $\{x_i\}$	species parameters $\{c_k, u_k, t_k\}$	heuristic method	Chapter
regression	known	unknown	weighted averaging	2
calibration	unknown	known	weighted averaging	3
ordination	unknown	unknown	correspondence analysis	4
constrained ordination	linear combination of environmental variables	unknown	canonical correspondence analysis	5,6,7
unnamed	unknown	u_k known; c_k, t_k unknown	weighted averaging	8

Regression

Suppose a researcher wants to investigate whether diatoms are good indicators of the acidity (pH) of lakes, with the aim to reconstruct, subsequently, pH from fossil diatoms found in successive strata of the bottom sediment. A sample of n lakes is selected. For each lake, some material is taken from the upper layer of the sediment and pH is measured.
In the laboratory, a slide for use under the microscope is made from the material sampled and the species (or taxa) that are present in the slide are identified. For simplicity, suppose that only presence/absence of species is recorded. The survey so results in the presences and absences of, say, m species in the n lakes ("sites"). Let $y_{ik} = 1$ or 0 depending on whether species k is present or absent in lake i, respectively ($i = 1,\ldots,n$; $k = 1,\ldots,m$). For typical data, most of the species will have a relative frequency in the sample below 0.05, and only very few species will reach 0.3.

The first step is to describe the relationship of the probability of occurrence (p) of each species against pH. What comes to mind is to carry out logit regression of the data of each particular species on pH, for example by the model

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x + b_2 x^2 \qquad (2)$$

where p is shorthand for Ey_{ik}, x is pH and b_0, b_1 and b_2 are regression coefficients, a triple for each species. The quadratic term is included because the relationship can be non-monotonic. By deviance tests, it can be tested whether $b_2 = 0$, or whether $b_1 = b_2 = 0$. If $b_1 = b_2 = 0$, then the species is not an indicator for pH. If $b_2 < 0$, then the curve has an optimum; if the maximum of the curve is small, the curve resembles the Gaussian curve and, therefore, is termed the Gaussian logit curve, in Chapter 2.

Logit regression is a recent development (Cox, 1970). It was not widely available before the introduction of the generalized linear model (Nelder and Wedderburn, 1972). Ecologists have used and developed other methods. One such method is to divide pH in K classes, to crosstabulate the species presence/absence and pH-classes in a 2 × K table, and to calculate a chi-squared statistic, or an "information" statistic (Guillerm, 1971; Kwakernaak, 1984) which is related to the G-test (a deviance test). I will not discuss this method further. In this thesis, I am interested in variation along continuous variables, termed gradients by ecologists. Another simple method is at the center of this thesis. From the time of Gause (1930) till today (Charles, 1985), many ecologists have analysed their data by the method of weighted averaging. In this method, the relationship of species with an environmental variable is characterized by the weighted average

$$\bar{u}_k^{WA} = \frac{\sum_{i=1}^{n} y_{ik} x_i}{\sum_{i=1}^{n} y_{ik}} \qquad (3)$$

and the weighted standard deviation

$$\bar{t}_k^{WA} = \left[\frac{\sum_{i=1}^{n} y_{ik}(x_i - \bar{u}_k^{WA})^2}{\sum_{i=1}^{n} y_{ik}}\right]^{1/2} \qquad (4)$$

In this thesis, Eqs. (3) and (4) are considered as "simple-minded" estimates of the optimum and tolerance of the Gaussian (logit) curve, and their statistical properties are studied. Weighted averaging is used both for presence/absence data and for abundance data. For presence-absence data the method reduces to the calculation of the mean and standard deviation of the environmental variable for those sites in which the species is present. An intuitive rationale is as follows. With pH as the environmental variable, a species with a particular optimum for pH will be present most frequently at sites with pH close to its optimum. So an intuitively reasonable estimate of the optimum is to take the average of pH of sites in which the species is present.

In statistics, means and standard deviations estimate the expectation and standard deviation of probability distributions. With some imagination, the values of x where the species is present can be considered to derive from a distribution. The distribution concerned can be obtained by factoring its density, f(.), by

$$f(\text{species is present at } x) = g(x)\, p(\text{species is present} \mid x) \qquad (5)$$

where g(x) represents the probability density function of the environmental variable x in the population sampled and $p(. \mid x)$ is a conditional density. Because the response, y, is binary (1/0),

$$p(\text{species is present} \mid x) = E(y \mid x) \qquad (6)$$

which shows that $p(. \mid x)$ is a response function, denoted by $\mu_k(x)$ for species k in Chapter 3. The weighted average (Eq. (3)) is an unbiased estimator of the expectation of the distribution with density f(.). But, what is of interest is a parameter of the response function $\mu_k(x)$, for example, the centroid of $\mu_k(x)$, $\int x\, \mu_k(x)dx / \int \mu_k(x)dx$, or the optimum of $\mu_k(x)$. If g(x) is constant (x has a uniform distribution), the centroid of $\mu_k(x)$ coincides with the expectation of the distribution with density f(.). If $\mu_k(x)$ is symmetric, for example, the Gaussian logit curve, then the centroid coincides with the optimum. So, the weighted average is an unbiased estimator of the optimum, if x has a uniform distribution and the response function is symmetric.

In Chapter 2, the weighted average is compared by simulation and real data with the estimator of the optimum obtained from logit regression. In the simulations, the data were generated from Eq. (2), the Gaussian logit curve. The distribution of the environmental variable, the number of sites sampled and the maximum probability of occurrence were varied in the simulations. For equispaced values of x_i, the weighted average and the regression estimator for the optimum resulted in almost identical values and are therefore equally efficient. The results also showed that the weighted average is a reasonable efficient estimator of the optimum, if the distribution of the environmental variable is uniform, or if the species has few occurrences and a small tolerance. The simulations thus confirmed for small samples what was expected from the asymptotic theory given in Chapter 3 and Chapter 4. In large samples in which the distribution of the environmental variables is not uniform, weighted averaging may however give estimates with nonnegligible bias.

Logit regession has several advantages over weighted averaging by allowing

- approximate statistical tests to be carried out,
- approximate confidence intervals for the optimum to be constructed,
- quantitative predictions,
- other shapes of curve to be fitted, e.g. by fitting splines,
- joint analysis of the effects of several environmental variables.

This research was a stimulus for Barendregt et al (1985) to develop their ICHORS model. This model is a set of logit regression equations relating the probability of occurrence of water plants to water chemistry variables, fitted to data from 800 samples from polders in the Vecht-region. The equations are used to evaluate the possible effects of changes in water management for these polders (see also Barendregt et al, 1986). The equations were fitted by a step-wise regression procedure in which the square of each variable considered was added to the model.

However, relating species to environment by multiple logit regression is not without problems. Outliers form a serious problem (Looman, 1985). If interaction effects of environmental variables are to be considered, the number of parameters in the models becomes large. The parameters are likely to become ill-determined. The number of parameters can be reduced by fitting a hierarchy of models and by deciding by statistical tests whether a simpler model is still acceptable. This is however a rather complicated procedure, often leading to qualitatively different models for different species (Looman, 1985). It will depend on the context whether such a complex procedure is worthwile. The experiences of Looman (1985) with multiple logit regression were an important stimulus to me to search for a simpler direct method to relate species to environment (Chapters 5-7).

Calibration

The example of the previous section is continued. After having described the relationship of diatom species with pH, the researcher wants to produce estimates of the pH in the past from fossil diatom remains. He/she takes a core from the sediment, splits the core into thin sections and identifies which species are present in each section. In addition, the sections are dated by methods analogous to the ^{14}C-method. The only problem considered here is how to estimate pH from the presences and absences of the species. It is a nonlinear multivariate calibration problem. The notation used is the same as in the previous section, but it should be noted that the sites now refer to thin sections of a core and that the values $\{x_i\}$ are unkowns.

Nonlinear multivariate calibration has not received much attention in the statistical literature. The approach proposed in Chapter 3 is based on extra -admittedly unrealistic- assumptions.

1. The parameters of the response curve of each of the species are determined with great precision, so that they can be considered as known constants,
2. the responses of the species, given pH, are independent.

With these assumptions, the pH can be estimated from the presences and absences of the species by the maximum likelihood method. Here, the likelihood is maximized numerically.

In vegetation science, Ellenberg (1948) developed a much simpler method to estimate the value of an environmental variable at a site from the plant species that grow there. The method is based on "indicator values" of species with respect to the environmental variable. Ellenberg (1948) did not give a precise definition of "indicator value", but, intuitively, it is the optimum (= the value most preferred by the species). So, the weighted average in Eq. (2) can be considered as an estimator of the indicator value. In Ellenberg's method, the value of an environmental variable is estimated by the weighted average of indicator values of species growing at the site; in our notation,

$$\tilde{x}_i^{WA} = \frac{\sum_{k=1}^{m} y_{ik} u_k}{\sum_{k=1}^{m} y_{ik}} \qquad (7)$$

So it is a weighted averaging method, but "the other way round" compared to Eq. (3). For presence-absence data, the method reduces to averaging of optima of species that are present. An intuitive rationale is as follows.
In a site with a particular pH, species with an optimum close to that pH will be present most frequently. So, an intuitively reasonable estimate of pH is to take the average of optima for pH of the species present. Ellenberg's method was proposed independently by Whittaker (1948: in Gauch, 1982), Pantle and Buck (1955) and continues to receive interest (e.g. von Tumpling, 1966; Durwen, 1982, Gauch, 1982, Böcker et al, 1983, Melman et al, 1985, Sladeçek, 1986).

Chapter 3 is a bold attempt to reconstruct the model that Ellenberg (1948) may have had in mind when he proposed weighted averaging of indicator values as a calibration method. This is done by investigating with which model the method has attractive statistical properties, namely consistency and efficiency. It turned out that, for presence-absence data, the Gaussian logit curve is the only response model under which the weighted average can achieve asymptotically an efficiency of 1 compared to the maximum likelihood estimator. Unit efficiency is actually achieved with a species packing model (Fig. 2), in which the Gaussian logit curves of the species have equispaced optima, equal maxima and equal tolerances. For abundances that are Poissonian, the Gaussian curve has this property. So chapter 3 shows that the Gaussian logit model has a more than casual relation to the method of weighted averaging. In the context of regression, weighted averaging can also achieve unit efficiency (Chapter 2), but the theoretical analysis is carried out for calibration because then only a single parameter is involved.

In the example, a simple method to infer pH from diatoms is thus to estimate the optima from a training set by Eq. (3) and to use Eq. (7) to produce estimates of pH for thin sections of the core. (In this approach, averages are taken twice, so that the range of pH is shrunken. This defect can be repaired by linear rescaling on the basis of a simple linear regression of pH on $\tilde{x}_i$ in the training set.) Using counts of diatoms, Ter Braak and Van Dam (in prep.) compared this method with the maximum likelihood method. They found that the maximum likelihood method performed only slightly better than weighted averaging as judged by the mean squared prediction error in a test set.

Calibration by weighted averaging-applied-twice is the natural end-point of a historical development that started with Imbrie and Kipp (1971). To reconstruct past sea-surface temperature from Foraminifera, Imbrie and Kipp (1971) considered applying inverse regression to a training data set, i.e. regression of temperature on the abundances of the species. But this method was considered inappropriate as the abundances of species showed multicollinearity. So, they decided to reduce the abundances of the species to a few axes by principal components analysis and to regress temperature on these axes (this is termed principal components regression; Jolliffe, 1986). The resulting equation was used for reconstruction. Roux (1979) produced better estimates of temperature, at least in the training set, by replacing principal components analysis by correspondence analysis.
By rearranging species and sites in the data matrix in order of their scores on the first axis of correspondence analysis, he obtained a matrix with large

abundance values near the principal "diagonal" of the matrix and small values elsewhere. Such matrices arise when relationships are unimodal.

Gasse and Tekaia (1983) were concerned about the fact that only part of the information on the relationship of species to x is retained in the first few axes of the correspondence analysis. They suggested the following improvement in their attempt to estimate pH from diatoms. They divided pH into four classes and, next, applied correspondence analysis to a species-by-class data matrix, each entry of which contains the total abundance of a species in sites with a pH of the corresponding class. The final calibration equation was obtained by a multiple regression of pH on the axes of the correspondence analysis. Despite its complexity, the method is closely related to weighted averaging-applied-twice. Both methods are special cases of canonical correspondence analysis (Chapter 5). The main difference is that, in the method of Gasse and Tekaia (1983), pH is divided in classes whereas pH is treated as a quantitative variable in weighted averaging-applied-twice.

Ordination

With ordination, one enters the realm of explorative data analysis. If one has not measured any environmental variable, one can still attempt to construct a latent variable that explains the abundances of the species observed at the sites by way of the Gaussian model. Ordination is then a method to detect a simple structure in the data, or a method to reduce the dimensionality of the data (from m to 1 or 2).

Gauch et al (1974) fitted Gaussian curves to vegetation data by the least-squares method. However, the least-squares method is not very attractive because abundances tend to have a very skew distribution. In a paper that remained largely unnoticed, Kooijman (1977a) fitted Gaussian curves by the maximum likelihood method under the assumption that the abundances were independent Poissonian counts. Kooijman (1977a) was the first to fit the two-dimensional Gaussian model in which species have Gaussian response surfaces against two latent variables. The computer programs developed by Kooijman (1976b) were written in APL, which limited their use. An application is described in Kooijman and Hengeveld (1979). A recent overview of one-dimensional Gaussian ordination, including algorithms, is given by Ihm and van Groenewoud (1984).

Gaussian ordination has not become popular among ecologists because of its computational complexity and its strong and explicit assumptions. Hill (1973) developed a simpler method with the same aim: reciprocal averaging, alias correpondence analysis. Hill (1973) is one of the many independent inventors and reinventors of correspondence analysis (Tenenhaus and Young, 1985). Hill suggested the technique as a natural extension of the method of weighted averaging, known to him via Whittaker's (1956) paper. If Eqs. (3) and (7) are applied alternately to a data matrix $\{y_{ik}\}$, the values of (u_k) and (x_i) converge to the first nontrivial axis of correpondence analysis (Hill, 1973; Chapter 4 and Chapter 9). Under simplifying conditions, this first axis is an approximation to the latent variable of Gaussian ordination as estimated by maximum likelihood (Chapter 4). The conditions needed are a combination of those needed in Chapter 2 and 3 for the weighted average to be an efficient estimator of u_k and of x_i, respectively. This results holds true for presence/absence data and abundance data that follow the Poisson distribution.

Independently, Ihm and van Groenewoud (1984) compared correspondence analysis and Gaussian ordination. They defined a variant of the Gaussian

model that is attractive if sites vary in "size", so that only relative abundance values are meaningful. I shall discuss this variant in some detail as it provides an interesting link with the analysis of contingency tables by correspondence analysis. Their model (Equation 3.2.1 of the paper) is (with $t_k = t$)

$$Ey_{ik} = r_i c_k e^{-\frac{1}{2}(x_i - u_k)^2/t_k^2} \tag{8}$$

Compared with Eq. (2), r_i is an extra parameter, which accounts for the size of site i. The model is useful for compositional data also; r_i then accounts for the constant-sum constraint (Dawid, 1982; ter Braak, 1987). By expanding the quadratic term in Eq. (8) and assuming $t_k = t$, Ihm and van Groenewoud (1984: section 5.1) obtain

$$Ey_{ik} = r_i^* c_k^* e^{u_k x_i/t^2} \tag{9}$$

with $r_i^* = r_i \exp(-\frac{1}{2} x_i^2/t^2)$ and $c_k^* = c_k \exp(-\frac{1}{2} u_k^2/t^2)$, and by using a first order Taylor expansion,

$$Ey_{ik} = r_i^* c_k^*(1 + u_k x_i/t^2) \tag{10}$$

A simple estimate of $r_i^* c_k^*$ is $y_{i+}y_{+k}/y_{++}$, so that, with t=1 and y_{ik} replacing Ey_{ik}, we obtain

$$y_{ik} = \frac{y_{i+}y_{+k}}{y_{++}}(1 + u_k x_i) \tag{11}$$

This is the reconstitution formula (of order 1) of correspondence analysis (Chapter 4: Eq. (2.4)). So the model of Eq. (8) is shown to resemble the "model" of correspondence analysis. The estimation equations are similar too, as shown by Goodman (1981); Eq. (9) is Goodman's RC-model for two-way contingency tables. The similarity can also be shown by extending the analysis of Chapter 4. Eq. (8) can be rewritten in a form similar to Eq. (3.1) of Chapter 4, namely

$$\log Ey_{ik} = \phi_i + a_k - \frac{1}{2}(x_i - u_k)^2/t_k^2 \tag{12}$$

where $\phi_i = \log r_i$ and $a_k = \log c_k$. Under Poisson sampling, Eqs. (3.2) and (3.3) of Chapter 4 are then the maximum likelihood equations for u_k and x_i (with $u_{ik} = Ey_{ik}$). The appoximations made in Chapter 4 are valid for this model too and lead to the transition formulae of correspondence analysis. The equality of Eqs. (8) and (9), for $t_k = t$, is the solution of the apparent paradox, noted in Chapter 4, that both a unimodal model and a (generalized) bilinear model stand at the basis of correspondence analysis. In chapter 7, a multidimensional form of Eqs. (9) and (12) are considered, which - when approximated - reduces to multiple correspondence analysis. Chapter 7 so provides a link between multiple correspondence analysis and a loglinear model for contingency tables. The loglinear model contains main effects and multiplicative terms. Van der Heijden and de Leeuw (1985) and van der Heijden (1987) use correspondence analysis to analyse the residuals of an additive loglinear model. Such an analysis is an approximation to a loglinear model with both additive and multiplicative terms (van der Heijden and Worsley,

1986). Gabriel (1978) considered a linear (not loglinear) model with both additive and multiplicative terms.

The possibility of analysing unimodal relationships with correspondence analysis was first noted by Mosteller (1948 in Torgerson 1958: p. 338). Mosteller showed that Guttman's principal components analysis of categorical data (Torgerson, 1958) could also be used to analyse point items (binary observations with unimodal "trace lines" with respect to the latent variable). Heiser (1981) proved that correspondence analysis has interesting properties for ordering sites when relationships are unimodal (see also Heiser, 1986).

Since the introduction of correspondence analysis, ecologists have been concerned about the arch effect. This is the phenomenon that the second axis of correspondence analysis is a quadratic function of the first axis (Hill, 1974; Gauch, 1982). By careful mathematical analysis, Schriever (1983) established when the arch occurs. A qualitative explanation that can be understood by ecologists is given in Chapter 8 (section IV C); see also Jongman et al (1987: section 5.2.3) and the discussion of Chapter 7. Although the explanation makes clear that the arch is sometimes an artifact of the method, the debate will continue whether it is always an artifact (Pielou 1984; Heiser, 1986, 1987; van Rijckevorsel, 1987). In detrended correspondence analysis (Hill and Gauch, 1980) the arch is removed by a modification of the reciprocal averaging algorithm. In simulations (Chapter 4), this modification was shown to improve the approximation to two-dimensional Gaussian ordination. The modification may occasionally lead to new artifacts (Minchin, 1987), which led me to develop a simpler alternative method of detrending (Chapter 9). The new method of detrending by polynomials is incorporated in the computer program CANOCO (ter Braak, 1987).

Rival approaches to ordination on the basis of a unimodal model are maximum likelihood Gaussian ordination (Ihm and van Groenewoud, 1984), unfolding (Heiser, 1987) and multidimensional scaling (Prentice, 1977; Faith et al, 1987; Minchin, 1987). In nonmetric unfolding, the model does not need to be Gaussian, but must still be symmetric (Heiser, 1987). The multidimensional scaling approach appears to allow even more complex models when used with an appropriate measure of similarity (Faith et al, 1987). These rival approaches are computationally far more demanding than detrended correspondence analysis, and require good starting values. Such values can be derived from detrended correspondence analysis (Chapter 4).

Constrained ordination

Ordination is also popular among ecologists even when environmental variables have been measured. The approach is then to interpret the ordination axes (estimates of latent variables) in terms of the environmental variables - an indirect way of relating species to environment.

There is a problem with this indirect approach. Ordination of species data is not designed to detect the effect on the species of any environmental variable at all. So the effect of a variable one is particularly interested in can be poorly represented in the ordination or even be missed completely. This problem can be overcome by using regression instead of ordination. Building non-linear models by regression is demanding in time and computation, when the effects of several environmental variables on a set of species are of interest (see the section on regression). A considerable simplification is possible if species react to the same linear combination of environmental variables, according to a common response model. Such a model is the Gaussian ordination model in which the latent variable is constrained

to be a linear combination of environmental variables,

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij} \qquad (13)$$

where z_{ij} is the value of environmental variable j at site i and b_0, $b_1 \ldots b_q$ are parameters. By inserting Eq. (13) in Eq. (1), we obtain the model of canonical Gaussian ordination (Chapter 5)

$$Ey_{ik} = c_k \, e^{-\frac{1}{2}(b_0 + \sum_j b_j z_{ij} - u_k)^2 / t_k^2} \qquad (14)$$

which is, of course, just a particular non-linear regression model. Under the same simplifying conditions as in the precious section, the model reduces to canonical correspondence analysis (Chapter 5), a constrained form of correspondence analysis.

When I wrote chapter 5, I chose the adjective "canonical" because of the relation of the technique with canonical correlation analysis, which is the standard linear method of relating two sets of variables (here, species and environmental variables). It turns out that the linear method of redundancy analysis (van den Wollenberg, 1977) is even more closely related (Chapter 9). Fortunately, "canonical" is still an apt adjective for another reason. It is shown in Chapter 7 that Eq. (14) is the (one-dimensional) canonical form of a particular nonlinear regression model.

The idea of constrained ordination may be new to ecology, but has already been around for some time in psychometry (see de Leeuw and Heiser, 1980). Heiser (1981: sections 8.3 and 8.4) proposed a constrained unfolding model closely related to the model of canonical correspondence analysis. Imposing constraints on the solution of correspondence analysis is not new either as it is the basis of the Gifi system of multivariate analysis of nominal and ordinal variables (Gifi, 1981; de Leeuw, 1984). Even the type of equations for solving canonical correspondence analysis are not new; Israëls (1984) derived the same eigenvalue equations in his redundancy analysis of qualitative variables (see also Israëls, 1987 and Lauro and d'Ambra, 1984). Yet canonical correspondence analysis is new, because it was not clear in advance that these developments were useful in relating species to environmental variables according to a unimodal model. In chapter 7 a Gaussian model is proposed that takes into account the effects of covariables; this is the natural endpoint of the general approach in this thesis, i.e. the approximation of complicated Gaussian models by correspondence analysis techniques.

I hope this thesis will encourage ecologists to go beyond exploratory ordination, data analysts to understand the limitations of correspondence analysis techniques, and statisticians to bridge the gap between correspondence analysis techniques and nonlinear regression models.

References

Alderdice, D.F., 1972. Factor combinations responses of marine poikilotherms to environmental factors acting in concert. Pages 1659-1722 in: O. Kinne (editor): Marine ecology, vol 1, part 3., Wiley, New York.

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

Barendregt, A., J.T. de Smidt & M.J. Wassen, 1985. Relaties tussen milieufaktoren en water- en moerasplanten in de Vechtstreek en de omgeving van Groet. Interfacultaire Vakgroep Milieukunde, RU Utrecht.

Barendregt, A., M.J. Wassen, J.T. de Smidt & E. Lippe, 1986. Ingreep-effect voorspelling voor waterbeheer. Landschap 1: 41-55.

Bates, J.W. & D.H. Brown, 1981. Epiphyte differentiation between Quercus petraea and Fraxinus excelsior trees in a maritime area of South-West England. Vegetatio 48: 61-70.

Battarbee, R.W., 1984. Diatom analysis and the acidification of lakes. Philosophical Transactions of the Royal Society of London 305: 451-477.

Böcker, R., I. Kowarik & R. Bornkamm, 1983. Untersuchungen zur Anordnung der Zeigerwerte nach Ellenberg. Verhandlungen der Gesellschaft für Ökologie 11: 35-56.

Charles, D.F., 1985. Relationships between surface sediment diatom assemblages and lakewater characteristics in Adirondack lakes. Ecology 66: 994-1011.

Cox, D.R., 1970. The analysis of binary data. Chapman and Hall, London.

Dawid, A.P., 1982. Discussion to "The statistical analysis of compositional data" by J. Aitchison. Journal of the Royal Statistical Society series B 44: 162-163.

de Leeuw, J., 1984. The GIFI system of nonlinear multivariate analysis. Pages 415-424 in: E. Diday et al eds.). Data Analysis and Informatics 3, North-Holland, Amsterdam.

Digby, P.G.N. & R.A. Kempton, 1987. Multivariate analysis of ecological communities. Chapman and Hall, London.

Durwen, K.-J., 1982. Zur Nutzung von Zeigerwerten und artspezifischen Merkmalen der Gefaszpflanzen Mitteleuropas für Zwecke der Landschaftsoekologie und -planung mit Hilfe der EDV - Voraussetzungen, Instrumentarien, Methoden und Möglichkeiten. Arbeitsberichte des Lehrstuhls Landschaftsoekologie, Münster 5: 1-138.

Earle, J.C., H.C. Dutchie & D.A. Scruton, 1986. Analysis of the phytoplankton composition of 95 Labrador lakes, with special reference to natural and anthropogenic acidification. Canadian Journal of Fisheries and Aquatic Science 43: 1804-1811.

Ellenberg, H., 1948. Unkrautgesellschaften als Mass für den Säuregrad, die Verdichtung und andere Eigenschaften des Ackerbodens. Berichte über Landtechnik, Kuratorium für Technik und Bauwesen in der Landwirtschaft 4: 130-146.

Faith, D.P., P.R. Minchin & L. Belbin, 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69: 57-68.

Forsythe, W.L. & O.L. Loucks, 1972. A transformation for species response to habitat factors. Ecology 53: 1112-1119.

Gabriel, K.R., 1978. Least squares approximation of matrices by additive and multiplicative models. Journal of the Royal Statistical Society, Series B 40: 186-196.

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

Gauch, H.G., 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Gauch, H.G., & R.H. Whittaker, 1972. Coenocline simulation. Ecology 53, 446-451.

Gauch, H.G., G.B. Chase & R.H. Whittaker, 1974. Ordinations of vegetation samples by Gaussian species distributions. Ecology 55: 1382-1390.

Gause, G.F., 1930. Studies on the ecology of the Orthoptera. Ecology 11: 307-325.

Gause, G.F., 1934. The struggle for existence. Williams and Wilkin, Baltimore.

Gifi, A., 1981. Nonlinear multivariate analysis. DSWO-press, Leiden.
Gittins, R, 1985. Canonical analysis. A review with applications in ecology. Springer-Verlag, Berlin.
Goodall, D.W., 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Australian Journal of Botany 1: 39-63.
Goodman, L.A., 1981. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. Journal of the American Statistical Association 76: 320-334.
Green, R.H., 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York.
Greenacre, M.J., 1984. Theory and applications of correspondence analysis. Academic Press, London.
Greig-Smith, P., 1983. Quantitative Plant Ecology. 3rd edition. Blackwell Scientific Publications, Oxford.
Guillerm, J.L., 1971. Calcul de l'information fournie par un profil écologique et valeur indicatrice des espèces. Oecologia Plantarum 6: 209-225.
Heiser, W.J., 1981. Unfolding analysis of proximity data. Thesis. University of Leiden, Leiden.
Heiser, W.J., 1986. Undesired nonlinearities in nonlinear multivariate analysis. Pages 455-469 in: E. Diday et al. (editors): Data analysis and Informatics 4. North Holland, Amsterdam.
Heiser, W.J., 1987. Joint ordination of species and sites: the unfolding technique. In: New developments in numerical ecology. (P. Legendre and L. Legendre, eds.), Springer-Verlag, Berlin, in press.
Hill, M.O., 1973. Reciprocal averaging: an eigenvector method of ordination. Journal of Ecology 61: 237-249.
Hill, M.O., 1974. Correspondence analysis: a neglected multivariate method. Applied Statistics 23: 340-354.
Hill, M.O., 1979. DECORANA - A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics. Cornell University, Ithaca, New York.
Hill, M.O. & H.G. Gauch, 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42: 47-58.
Holder-Franklin, M.A. & L.J. Wuest, 1983. Population dynamics of aquatic bacteria in relation to environmental change as measured by factor analysis. Journal of Microbiological Methods 1: 209-227.
Ihm, P. & H. van Groenewoud, 1984. Correspondence analysis and Gaussian ordination. COMPSTAT lectures 3: 5-60.
Imbrie, J. & N.G. Kipp, 1971. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. Pages 71-181 in: K.K. Thurekian (ed.): The late Cenozoic glacial ages. Yale University Press, New Haven.
Israëls, A.Z., 1984. Redundancy analysis for qualitative variables. Psychometrika 49: 331-346.
Israëls, A.Z., 1987. Eigenvalue techniques for qualitative data. Thesis. University of Leiden, Leiden.
Jager, J.C. & C.W.N. Looman, 1987. Data Collection. Chapter 2 in: R.H.G. Jongman, C.J.F. ter Braak & O.F.R. van Tongeren (eds.): Data Analysis in Community and Landscape Ecology, Pudoc, Wageningen.
Jongman, R.H.G., C.J.F. ter Braak & O.F.R. van Tongeren, 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen.
Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, Berlin.

Jöreskog, K.G. & D. Sörbom, 1981. LISREL: Analysis of linear structural relationships by the method of maximum likelihood. International Educational Services, Chicago.
Kershaw, K.A. & J.H.H. Looney, 1985. Quantitative and dynamic plant ecology. 3rd edition. Edward Arnold, London.
Kooijman, S.A.L.M., 1977a. Species abundance with optimum relations to environmental factors. Annals of System Research 6: 123-138.
Kooijman, S.A.L.M., 1977b. Inference about dispersal patterns. Thesis. University of Leiden, Leiden.
Kooijman, S.A.L.M. & R. Hengeveld, 1979. The description of a non-linear relationship between some carbid beetles and environmental factors. Pages 635-647 in: "Contemporary Quantitative Ecology and Related Econometrics." (G.P. Patil and M.L. Rossenzweig, eds.): Intern. Co-operative Publ. House, Fairland, Maryland.
Kwakernaak, C., 1984. Information applied in ecological land classification. Pages 59-66 in: J. Brandt & P. Agger (eds.): Methodology in landscape ecological research and planning. Vol. III; theme III: Methodology of Data Analysis. Roskilde Universitetsforlag GeoRue, Roskilde.
Lauro, N. & L. D'Ambra, 1984. L'analyse non symetrique des correspondances. Pages 433-446 in: E. Diday et al (eds.). Data Analysis and Informatics 3, North-Holland, Amsterdam.
Lawley, D.M. & A.E. Maxwell, 1971. Factor Analysis as a Statistical Method. 2nd edition, Butterworth, London.
Legendre, L. & P. Legendre, 1983. Numerical Ecology. Elsevier Scientific Publishing Company, Amsterdam.
Looman, C.W.N., 1985. Responsies van slootplanten op standplaats factoren: uitwerking van een methode. Rapport Studiecommissie Waterbeheer Natuur, Bos en Landschap, Postbus 20020, 3502 LA Utrecht.
Melman, Th.C.P., P.H.M.A. Clausman, H.A.U. de Haes, 1985. Voedselrijkdom-indicatie van graslanden. Vergelijking en toetsing van drie methoden voor het bepalen van de voedselrijkdom-indicatie van graslandvegetaties. Centrum voor Milieukunde - Mededeling 19, Leiden.
Minchin, P.R., 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69: 89-107.
Montgomery, D.C. & E.A. Peck, 1982. Introduction to linear regression analysis. Wiley, New York.
Nelder, J.A. & R.W.M. Wedderburn, 1972. Generalized linear models. Journal of the Royal Statistical Society, Series A 135: 370-384.
Nishisato, S., 1980. Analysis of categorical data: dual scaling and its applications. Toronto University Press, Toronto.
Noy-Meir, I., D. Walker & W.T. Williams, 1975. Data transformations in ecological ordination. II. On the meaning of data standardization. Journal of Ecology 63: 779-800.
Odum, E.P., 1971. Fundamentals of Ecology. 3rd edition. W.B. Saunders Company, Philadelphia.
Pantle, R. & H. Buck, 1955. Die biologische Ueberwachung der Gewässer und die Darstellung der Ergebnisse. Gas- und Wasserfach 96: 604.
Pielou, E.C., 1984. The interpretation of ecological data. A primer on classification and ordination. Wiley, New York.
Prentice, I.C., 1977. Non-metric ordination methods in ecology. Journal of Ecology 65: 85-94.
Roux, M., 1979. Estimation des paléoclimats d'après l'ecologie des foraminifères. Les Cahiers de l'Analyse des Données 4: 61-79.

Schriever, B.F., 1983. Scaling of order-dependent categorical variables with correspondence analysis. International Statistical Review 51: 225-238.
Sládecek, V., 1986. Diatoms as indicators of organic pollution. Acta hydrochim. hydrobiol. 14: 555-566.
Stewart-Oaten, A., W.W. Murdoch & K.P. Parker, 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.
Tenenhaus, M. & F.W. Young, 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50: 91-119.
ter Braak, C.J.F., 1987. CANOCO - a FORTRAN program for canonical community ordination by [partial][detrended][canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). TNO Institute of Applied Computer Science, Wageningen.
Thienemann, A., 1950. Verbreitungsgeschichte der Süszwassertierwelt Europas. E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart.
Torgerson, W.S., 1958. Theory and methods of scaling. Wiley, New York, 460 pp.
van der Heijden, P.G.M. & J. de Leeuw, 1985. Correspondence analysis used complementary to loglinear analysis. Psychometrika 50: 429-447.
van der Heijden, P.G.M. & K.J. Worsley, 1986. Comment on " Correspondence analysis used complementary to loglinear analysis. PRM 86-01, Dept. of Psychology, Leiden, Psychometrika, to appear.
van Rijckevorsel, J., 1987. The application of fuzzy coding and horseshoes in multiple correspondence analysis. DWSO-press, Leiden.
van den Wollenberg, A.L., 1977. Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42: 207-219.
von Tümpling, W., 1966. Ueber die statistische Sicherheit soziologischer Methoden in der biologischen Gewässeranalyse. Limnologica (Berlin) 4: 235-244.
Westman, W.E., 1980. Gaussian analysis: identifying environmental factors influencing bell-shaped species distributions. Ecology 61: 733-739.
Whittaker, R.H., 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs 26: 1-80.
Whittaker, R.H., 1967. Gradient analysis of vegetation. Biological Reviews of the Cambridge Philosophical Society 49: 207-264.
Wiens, J.A. & J.T. Rotenberry, 1981. Habitat associations and community structure of birds in shrubsteppe environments. Ecological Monographs 51: 21-41.

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. Hockstra 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.

Vegetatio 65, 3–11 (1986).

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:

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

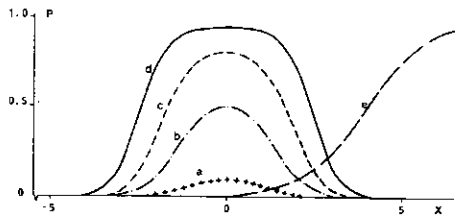


Fig. 1. Gaussian logit curves with $u=0$, $t=1$ and $p_{max}=0.1$ (a), 0.5 (b), 0.8 (c) and 0.95 (d) and a linear logit curve (e) (x: value for the environmental variable, $p(x)$: probability of finding this species at a value x).

(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.

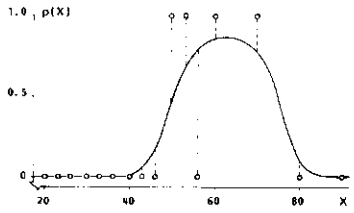


Fig. 2. Gaussian logit curve fitted by logistic regression to the artificial data (○) of Table 1.

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.

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).

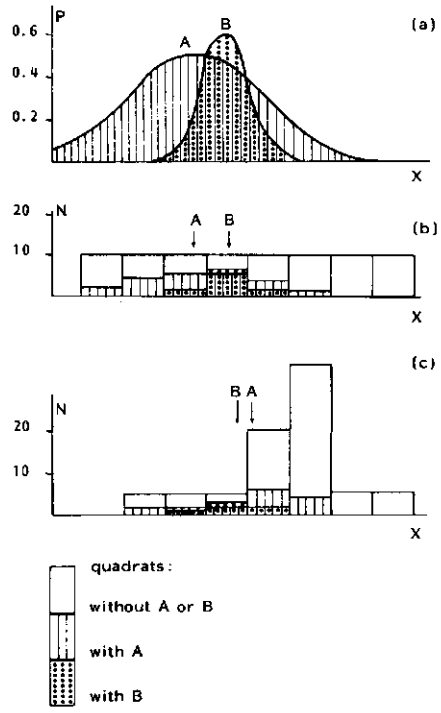


Fig. 3. The response curves of imaginary species A and B (a), the occurrence of these species in 80 samples, distributed evenly (b) or unevenly (c) along the environmental gradient. The weighted averages are indicated with arrows. The two sampling designs yield weighted averages that are in reversed 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%.

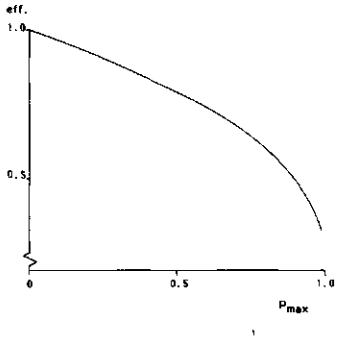


Fig. 4. The efficiency (ordinate) of weighted averaging with respect to Gaussian logistic regression to estimate the optimum for uniformly distributed quadrats without truncation (case 2, Table 2) decreases with increasing maximum probability of occurrence (abscissa).

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 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.

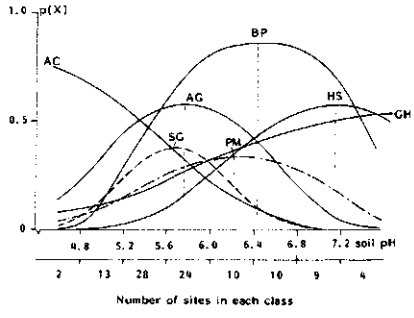


Fig. 5. Probability of occurrence of seven contrasting species in relation to soil acidity (pH) in meadows, as fitted with logistic regression. The curves can be identified by the code near their optimum indicated by dotted lines. The species arranged in order of their optima are: *Agrostis canina* (AC); *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

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

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

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)

$$\log\,[p/(1-p)] = b_0 + b_1x + b_2x^2 + c_1z + c_2z^2 \quad \text{(A.5)}$$

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.

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

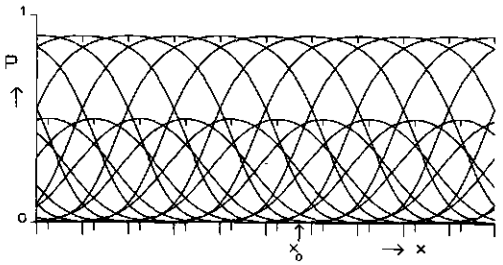


FIG. 1. Gaussian logit response curves of the probability $P = \mu_k(x)$ that a species (k) occurs at a site, against environmental variable x. Two sets of species are displayed, each with $t = 1$ and optima with spacing $d = 1$, having maximum probabilities of .5 and .9, respectively. x_0 is the value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

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

where summations are over all species. To make sense, $\hat{x}_{WA}$ 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, $\hat{x}_{WA} = T/R$, 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, $\operatorname{var}(T)\to 0$ $(\lambda\to\infty)$ 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]

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

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 $I =$

$\{u \mid \mu(x_0 - u) > \delta,\ u \in \mathbf{R}\}$ 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,

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

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

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

and

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

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 $v_{ML} \leqslant v_{WA}$ with equality if and only if the response curves are nonrandom and satisfy (5.2).

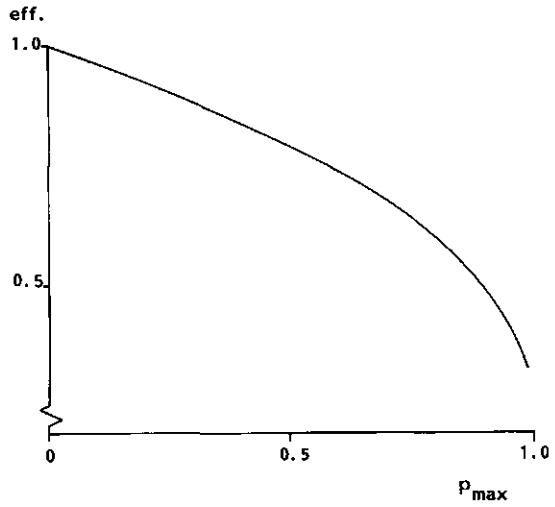


FIG. 2. The efficiency of weighted averaging with respect to maximum likelihood against the maximum probability of occurrence (p_{max}) for Gaussian logit curves with randomly spaced optima and equal maxima and tolerances [eff $= v_{ML}/v_{WA} = (t/\tau)^2$].

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)]

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

An unbiased estimator for τ^2 is the usual sample variance 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 $B = \{\mathbf{u} \mid \mu(\mathbf{x}_0 - \mathbf{u}) > \delta,\ \mathbf{x}_0 \in \mathbf{R}^n,\ \mathbf{u} \in \mathbf{R}^n\}$ 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

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

with equality iff $\mu'(x) = -xv(x)/t^2$. The left hand side in (A1) is the asymptotic ($d \to 0$) efficiency v_{ML}/v_{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 $R = \sum_k Y_k$ and $T = \sum_k Y_k u_k$. 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 $\hat{x}_{WA}$ 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

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

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 $g(u) = E_S\{V(u) + M^2(u)\}$,

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

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)]

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

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

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

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. Kovács, Das Corno-quercetum des Mátra-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. Sládecek, 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).

BIOMETRICS 41, 859–873
December 1985

Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model

Cajo J. F. ter Braak

TNO Institute of Mathematics, Information Processing and Statistics,
P. O. Box 100, 6700 AC Wageningen, The Netherlands

SUMMARY

Correspondence analysis is commonly used by ecologists to analyze data on the incidence or abundance of species in samples. The first few axes are interpreted as latent variables and are presumed to relate to underlying environmental variables. In this paper correspondence analysis is shown to approximate the maximum likelihood solution of explicit unimodal response models in one latent variable. These models are logistic–linear for presence/absence data and loglinear for Poisson counts, with predictors that are quadratic in the latent variable. The approximation is best when the maxima and tolerances (widths) of the response curves are equal and the species' optima and the sample values of the latent variable are equally spaced. It is still fairly good for uniformly distributed optima and sample values, as shown by simulation. For the models extended to two latent variables, the approximation is often bad because of the horseshoe effect in correspondence analysis, but improves considerably in the simulations when this effect is removed as it is in detrended correspondence analysis.

1. Introduction

Correspondence analysis is a multivariate technique primarily developed for the analysis of contingency table data (Nishisato, 1980; Greenacre, 1984). However, in ecology and archaeology, correspondence analysis is commonly applied to incidence or abundance matrices (Gauch, 1982). In ecology these matrices typically record the presence/absence or abundance of species in samples, e.g., plant species in quadrats or animal species in areas. Such matrices are not transformed to m-way contingency tables "on the grounds that the data are essentially asymmetric and the absences indicate little" (Hill, 1974). Clearly a different rationale is needed for the application of correspondence analysis to incidence or abundance data. A pertinent result concerns so-called Petrie matrices (a Petrie matrix is an incidence matrix which has a block of consecutive 1's in every row and in every column, the block of the first row starting in the first column and the block of the last row ending in the last column). The result says that if a matrix can be rearranged to a Petrie matrix by a permutation of rows and columns, then this permutation is generated by the first nontrivial solution of correspondence analysis (see Hill, 1974).

Hill (1973) introduced correspondence analysis to ecology, under the name of "reciprocal averaging." He suggested the technique as a natural extension of the method of weighted averaging used in Whittaker's (1956) "direct gradient analysis." Whittaker, among others, observed that species typically show unimodal (bell-shaped) response curves with respect to environmental gradients. 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 or too wet.

Key words: Correspondence analysis; Detrended correspondence analysis; Dual scaling; Ecology; Generalized linear models; Joint plot; Reciprocal averaging; Species packing model; Unfolding; Unimodal response model.

Each species is therefore largely confined to a specific interval along an environmental variable. The value most preferred by a species was termed its "indicator value" or optimum. In Whittaker's method, the indicator value of a species is estimated by taking the average of the values of the environmental variable in those samples in which the species occurs. (For quantitative data, the average is weighted by species abundance.) Conversely, with known indicator values of species, weighted averaging is used to estimate the value of an environmental variable in a sample from the species that it contained [see e.g., Kovács (1969) for an application]. Hill (1973) showed that if iterated, this process of "reciprocal averaging" converges to a solution independent of initial indicator values, namely the first nontrivial axis of correspondence analysis (see also Greenacre, 1984, §4.2). Hill's method therefore amounts to arranging samples and species along a latent variable, an activity Whittaker (1967) termed "indirect gradient analysis." After such analysis, attempts are made to identify the latent variable by comparison with known variation in the environment (Gauch, 1982). The Petrie matrix provides a deterministic example of a response model wherein the response curves are (weakly) unimodal "block functions." Unimodal models also play an important role in unfolding theory (Coombs, 1964).

In this paper, correspondence analysis is regarded as an estimation method for latent variable models and is compared with maximum likelihood under parametric unimodal response models with respect to one or two latent variables. The models considered are loglinear and logistic–linear models with predictors that are quadratic in the latent variable(s). Ter Braak and Barendregt (in press) showed that these are the only models with Poisson and binomial error, respectively, for which the weighted average of indicator values can achieve unit asymptotic efficiency with respect to maximum likelihood. The comparison gives some idea about the model that is implicitly invoked when correspondence analysis is applied to incidence or abundance data. This comparison is important because the maximum likelihood approach may be computationally too demanding for the numbers of species and samples commonly encountered in ecological research. Moreover, when the maximum likelihood approach is considered worthwhile, the results suggest that good initial estimates can be derived from correspondence analysis or, for two latent variables, from detrended correspondence analysis (Hill and Gauch, 1980).

2. Correspondence Analysis

Nishisato (1980) takes the view that correspondence analysis, alias dual scaling, assigns real numbers or "scores" to rows and columns of a table so as to optimize a particular criterion. Consider a species-by-sample matrix $\mathbf{Y} = [y_{ki}]$ $(k = 1, \ldots, m;\ i = 1, \ldots, n)$ of nonnegative real numbers, denoting the presence/absence ($y_{ki} = 1$ or 0) or count of individuals of each of m species in n samples. Let $\mathbf{u} = [u_k]$ $(k = 1, \ldots, m)$ and $\mathbf{x} = [x_i]$ $(i = 1, \ldots, n)$ contain the scores for species (rows) and samples (columns), respectively. In correspondence analysis these scores are chosen so that the weighted sum of squares of the sample scores is maximum with respect to the weighted sum of squares of the sample scores within species, i.e., the criterion maximized is

$$D^2 = \sum_i y_{+i}(x_i - z)^2 / \sum_k \sum_i y_{ki}(x_i - u_k)^2, \tag{2.1}$$

where $z = \sum_i y_{+i}x_i/y_{++}$ and the subscript + denotes summation over that subscript. Maximization of D^2 will give each species a score close to the scores of those samples in which it is abundant. (An alternative interpretation of this criterion is given in Section 4.3.) With the Lagrange method of multipliers and the sample scores centred so that $z = 0$, we obtain after some rearrangement the *transition formulae* of correspondence analysis (with

$\alpha = 0$):

$$\lambda^{1-\alpha} x_i = \sum_k y_{ki} u_k / y_{+i} \quad (i = 1, \ldots, n), \tag{2.2}$$

$$\lambda^{\alpha} u_k = \sum_i y_{ki} x_i / y_{k+} \quad (k = 1, \ldots, m), \tag{2.3}$$

where λ is a real number ($0 \leqslant \lambda \leqslant 1$). The extra parameter α governs the scaling of the species scores and the sample scores with respect to one another. There are three choices of α in common usage, namely $\alpha = 0$, 1, or $\frac{1}{2}$. Criterion (2.1) leads to $\alpha = 0$. With $\alpha = 0$, the species scores u_k are weighted averages of the sample scores x_i [equation (2.3)] and the sample scores are proportional to the weighted averages of the species scores [equation (2.2)]. With $\alpha = 1$, the role of species and samples is interchanged, also in the criterion being maximized. The third choice, $\alpha = \frac{1}{2}$, is a compromise in that it treats species and sample scores in a symmetric way.

The transition formulae have more than one solution. All solutions can be obtained from the singular value decomposition of $\mathbf{R}^{-1/2}\,\mathbf{Y}\mathbf{C}^{-1/2}$ (see Hill, 1974) with $\mathbf{R} = \text{diag}(y_{k+})$ and $\mathbf{C} = \text{diag}(y_{+i})$. When the left and right normalized singular vectors in this decomposition are denoted by $\mathbf{q}_s$ and $\mathbf{r}_s$, corresponding to singular value $\rho_s = \sqrt{\lambda_s}$ ($s = 0, 1, 2, \ldots$), then the solutions are $\mathbf{u}_s = \rho_s \mathbf{R}^{-1/2}\mathbf{q}_s y_{++}^{1/2}$ and $\mathbf{x}_s = \mathbf{C}^{-1/2}\mathbf{r}_s y_{++}^{1/2}$. The solutions are the "axes" of correspondence analysis and λ_s is termed the eigenvalue of the sth axis. The maximum singular value is always 1, corresponding to the trivial solution in which all sample and species scores equal 1. The first nontrivial solution ($s = 1$) is orthogonal to the trivial solution, hence satisfies the previously applied centering $z = 0$, and maximizes the criterion D^2 with $\mathbf{u} = \mathbf{u}_1$, $\mathbf{x} = \mathbf{x}_1$, and $D^2 = 1/(1 - \lambda_1)$. Moreover, the singular value decomposition implies that the species and sample scores, $\mathbf{u}$ and $\mathbf{x}$, approximate the data in a weighted least squares sense by the bilinear model (see Nishisato, 1980)

$$\frac{y_{ki} - e_{ki}}{e_{ki}} \approx u_k x_i \tag{2.4}$$

with $e_{ki} = y_{k+} y_{+i} / y_{++}$, the expectation under the assumption of row/column independence in contingency tables.

3. A Unimodal Response Model

From now on the species-by-sample matrix $\mathbf{Y}$ will be assumed to consist either of counts y_{ki} that are independent Poisson variables with expected value μ_{ki}, or of presence/absence (1/0) data that are independent Bernoulli variables with probability μ_{ki} that the kth species is present in the ith sample. The models assumed for μ_{ki} are loglinear and logistic-linear models (Nelder and Wedderburn, 1972) in which the linear predictor is a quadratic polynomial in the latent variable x. It is convenient to write these models in the form

$$\text{link}(\mu_{ki}) = a_k - \tfrac{1}{2}(x_i - u_k)^2 / t_k^2, \tag{3.1}$$

where link is the logarithmic function for counts and the logistic function for the 1/0 data. In (3.1) the parameters for the kth species are a_k, the *maximum* on log or logit scale; u_k, the *mode* or *optimum* (i.e., the value of x for which the maximum is attained); and t_k, the *tolerance*, a measure of ecological amplitude. The value of the latent variable in the ith sample is x_i, which is treated as a fixed incidental parameter. Figure 1 displays an example for 1/0 data. The loglinear model is precisely the "Gaussian" response curve that is put forward by ecologists as an ideal for species responses along a gradient [see Austin (1976) and Gauch (1982) for reviews].

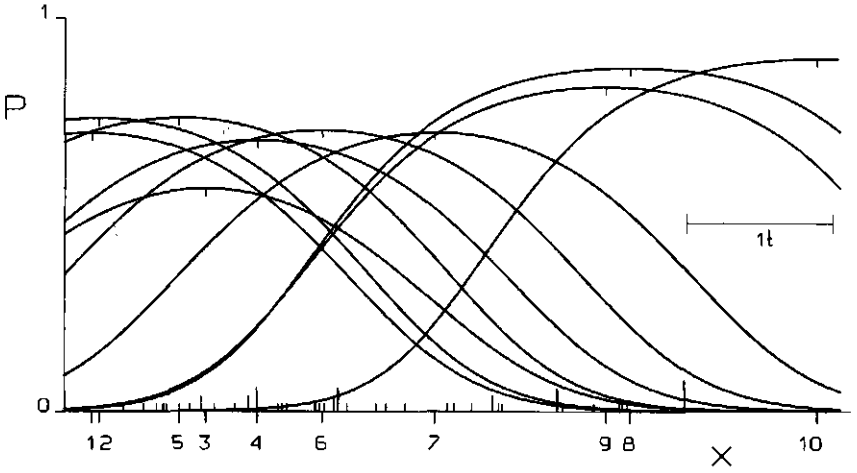


Figure 1. Unimodal response curves (3.1) for the probability (P) of occurrence along a latent variable (x), fitted by correspondence analysis to Table 2. The species optima and sample points are indicated by ticks below and above the abscissa. The length of a tick is proportional to the number of sample points. The numbers below the optima correspond to row numbers in Table 2. The horizontal bar is 1 tolerance unit.

The arbitrariness in the scale of the latent variable can be resolved, for example by centering as in correspondence analysis ($\sum_i y_{+i}x_i = 0$) and by setting the mean square of the tolerances to unity ($\sum_k t_k^2/m = 1$), so that the latent variable can be measured in (mean) tolerance units. Then, the maximum likelihood equations for the parameters $\mathbf{x} = [x_i]$ ($i = 1, \ldots, n$) and $\mathbf{u} = [u_k]$ ($k = 1, \ldots, m$) become, after some rearrangement,

$$x_i = \sum_k \frac{y_{ki}u_k}{t_k^2} \Big/ \sum_k \frac{y_{ki}}{t_k^2} - \left[\sum_k \frac{(x_i - u_k)\mu_{ki}}{t_k^2} \Big/ \sum_k \frac{y_{ki}}{t_k^2}\right], \tag{3.2}$$

$$u_k = \sum_i y_{ki}x_i/y_{k+} - \left[\sum_i (x_i - u_k)\mu_{ki}/y_{k+}\right]. \tag{3.3}$$

These (implicit) equations could be simplified further by using the maximum likelihood equations for the parameters $\mathbf{a} = [a_k]$ ($k = 1, \ldots, m$), but for the comparison with correspondence analysis, (3.2) and (3.3) are sufficient.

4. Theoretical Comparisons

Hill's approach to correspondence analysis makes plausible that the species scores and sample scores in Section 2 play a role similar to the species optima and sample values in Section 3; that is why similar symbols are used in Sections 2 and 3. Our aim is to show that the terms between square brackets in (3.2) and (3.3) are negligible in certain cases, so that the maximum likelihood equations reduce effectively to the transitional formulae (2.2) and (2.3) of correspondence analysis. These cases are as follows: either μ_{ki} is small or μ_{ki} is symmetric around x_i and around u_k.

4.1 *Equations for the Sample Scores*

For the comparison of the estimation equations (2.2) and (3.2), let us first assume that x is a manifest environmental variable, and that the species' tolerances are equal ($t_k = t = 1$). With known species' optima and maxima, a missing value of the environmental variable in a sample can be estimated by using (3.1) as calibration relation. The naive estimator is the weighted average (2.2) with $\alpha = 1$. The maximum likelihood equation (3.2) would give the same result when the term between square brackets is negligible, e.g., if for all species the maximum of μ_{ki} as a function of x is close to 0 ($a_k \to -\infty$). This case may have some practical relevance, as it implies very sparse matrices, which are not uncommon in ecology.

A more interesting case arises when μ_{ki} is symmetric around x_i. This happens under the species packing model (MacArthur and Levins, 1967). This is an ecological model based on the idea that during evolution species evolve to occupy maximally separated niches with respect to a limiting resource. Christiansen and Fenchel (1977, Chap. 3) provide a lucid introduction. With x the resource, maximally separated niches mean minimal overlap between the response curves and thus, for a given number of species on a fixed-length interval and equal maxima, equal spacing between the optima (apart from edge effects). If in this situation (i) the interval is longer than, say, 10 tolerance units, (ii) the spacing between the optima on this interval is closer than ca. 1 and (iii) the sample value x_i is well within this interval, then the term between square brackets is negligible because of the symmetry in the model (3.1). Simulations showed that under the stated conditions the weighted average has, in terms of mean squared error, an efficiency of 1.00 with respect to the maximum likelihood estimator (with an uninformative prior for x_i). Moreover, Ter Braak and Barendregt (in press) showed that the asymptotic efficiency is unity when the spacing decreases to 0 on an interval of increasing length and that in the class of response curves that form a location family on x, the models considered here are the only models with this property.

The weighted average still has approximately unit efficiency when the species maxima and optima vary in a cyclic pattern along the environmental variable, i.e., when the species can be divided into sets so that within each set the species have equal maxima and equally-spaced optima with spacing less than 1 tolerance unit. However, the efficiency may drop considerably when the tolerance varies. For example, with two tolerances differing by a factor 2, the efficiency drops to ca. .6 in the logistic model with maximum probability of occurrence .5. In that case the term between square brackets still vanishes, but what remains is not a simple weighted average. If the tolerances are known a priori, then the weighted average should be applied to y_{ki}/t_k^2, instead of to y_{ki}, in order to retain high efficiency.

More realistically, let us assume a superpopulation of response curves in which (i) the optima are independently and uniformly distributed on an interval (cf. Whittaker, Levin, and Root, 1973), (ii) the species maxima are either constant or random variables independent of the species optima, and (iii) the tolerances are equal. In this superpopulation the numerator of the term in square brackets in (3.2) vanishes in expected value, provided the sample value x_i is, again, well within the interval on which the optima are uniformly distributed. Because expectation is involved now, neglecting the term in square brackets makes weighted averaging less efficient with respect to maximum likelihood. In the logistic model with equal maxima, the asymptotic efficiencies are .96, .79, and .50 when the maximum probability of occurrence is .1, .5, and .9, respectively (Ter Braak and Barendregt, in press).

With $\alpha = 1$, the difference between the correspondence analysis equation (2.2) and the maximum likelihood equation (3.2) for latent x is the term between square brackets. The above comparisons for manifest x indicate in which situations neglecting this term does

not affect the solution too much. Note that equation (2.2) does not involve the species maxima and, further, that for equation (2.2) to be efficient for all samples, the sampled interval should be amply contained in the interval of the optima. With the choice $\alpha = 1$ the latter condition is pre-assumed.

4.2 *Equations for the Species Optima*

When the sample values are known a priori, estimation of the optima is a regression problem. From the symmetry between sample values and species optima in model (3.1) when the maxima and tolerances are equal, we deduce that the results of the previous section carry over to those species whose optima lie well within the sampled interval. For those species the weighted average is therefore asymptotically fully efficient with respect to the maximum likelihood estimator of the optimum, when the sample points are equally spaced with spacing less than 1 tolerance unit, and has a somewhat lower efficiency when the sample points are independently and uniformly distributed over the sampled interval (Ter Braak and Looman, in press). (That the maximum and the tolerance are to be estimated as well does not matter, because for these species the estimator for the optimum has under the stated conditions negligible correlation with the estimators for the maximum and the tolerance.) However, for species whose optima lie near the edge of, or even outside, the sampled interval, the weighted average is biased toward the center of the sampled interval, because these species' response curves are truncated. For example, the weighted average always gives a value inside the sampled interval, whereas the true optimum may lie outside this interval. This is where the eigenvalue λ of correspondence analysis comes in. With $\alpha = 1$ as in the previous section, equation (2.3) can be rewritten as

$$u_k = \sum_i y_{ki}x_i/y_{k+} - (\lambda - 1)u_k. \tag{4.1}$$

The term $(\lambda - 1)u_k$ can be considered as an overall correction term for the bias, or, alternatively, as a crude approximation to the term between square brackets in the maximum likelihood equation (3.3). The first nontrivial solution to the transition formulae has an eigenvalue λ closest to 1 and is therefore the solution where the least correction is required. This must be the solution with the longest underlying gradient, because the edge effects that cause the bias decrease with increasing length of the sampled interval. Although the correction term acts in the right direction, it overcorrects for optima well within the sampled interval and still undercorrects for optima on the edge of or outside the sampled interval. This observation explains the "compression of the first axis' ends relative to the axis middle" (Gauch, 1982) in correspondence analysis.

4.3 *Scaling of the Correspondence Analysis Solution*

The choice of α in the transition formulae (2.2) and (2.3) affects the scaling of the species scores with respect to the sample scores. If the sampling interval is contained well within the interval of the species optima, then α should naturally be 1 (§4.1). If the converse applies, then α should be 0. In practice, the intervals may coincide or may only partly overlap. The choice of α is then arbitrary and should be decided upon by other means (see §6.2).

The standardization of the sample scores also requires attention. Commonly the dispersion s^2 of the sample scores, $s^2 = \sum_i y_{+i}x_i^2/y_{++}$, is set equal to the eigenvalue λ, so that differences between sample scores approximate "chi-squared distances" between samples (see, e.g., Greenacre, 1984, p. 82). In the maximum likelihood approach (§3), the mean squared tolerance is set to unity. Assuming the loglinear model and the species packing

model, Hill (1979) estimated the mean squared tolerance by $\sum_k \sum_i y_{ki}(x_i - u_k)^2/y_{++}$ and standardized the correspondence analysis solution so that this estimator becomes 1. Hill's standardization gives as dispersion of the sample scores $1/(1 - \lambda)$ for $\alpha = 0$ (see §2) and $\lambda/(1 - \lambda)$ for $\alpha = 1$. Under the species packing model an alternative interpretation of criterion (2.1) is therefore that correspondence analysis maximizes the dispersion of the sample scores, subject to maintaining species response curves with unit mean squared tolerances. (By contrast, principal component analysis maximizes the variance of the sample scores subject to the condition that the sample scores are a normalized linear combination of the species' abundances.)

4.4 *Conclusion*

In conclusion, the transition formulae of correspondence analysis approximate the maximum likelihood equations for model (3.1). For equally-spaced optima and sample points, and equal maxima and tolerances, correspondence analysis uses a rough approximation to correct for edge effects. For uniformly distributed optima and sample points a second kind of approximation is involved, namely that the expectation is taken with respect to these uniform distributions over these parts of the maximum likelihood equations that do not depend on the data y_{ki}. The equality of the species maxima does not appear to be a crucial assumption. For unequal and unknown tolerances the approximation is worse, because the transition formulae then need to be weighted as well by the tolerances, which is not done in correspondence analysis.

5. Two Latent Variables

5.1 *A Unimodal Model*

The obvious extension of model (3.1) with equal tolerances to two latent variables is

$$\text{link}(\mu_{ki}) = a_k - \tfrac{1}{2}(x_{i1} - u_{k1})^2 - \tfrac{1}{2}(x_{i2} - u_{k2})^2. \tag{5.1}$$

The maximum likelihood equations for $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{u}_1$, $\mathbf{u}_2$ are analogous to (3.2) and (3.3) and nothing new arises in the comparison with the transition formulae. However, the edge effects due to truncation are likely to be more severe in two dimensions. First, there is more edge; second, the bias of the weighted average for, say, u_{k1} will in general depend not only on u_{k1} but also, through μ_{ki}, on u_{k2}. Approximating this bias by $(\lambda_1 - 1)u_{k1}$ is thus dubious; yet only with such approximations do the maximum likelihood equations reduce to the transition formulae of correspondence analysis.

5.2 *Detrended Correspondence Analysis*

Hill and Gauch (1980) developed detrended correspondence analysis as a heuristic modification of correspondence analysis, designed to correct two major "faults": (i) that the ends of the first axis are often compressed relative to the axis middle (see §4.2); (ii) that the scores of the second axis frequently show a systematic, often quadratic relation with those of the first axis. The latter fault, known as the horseshoe or arch effect, can be proven to occur for certain matrices (Hill, 1974, Proposition 8; Schriever, 1983).

Hill and Gauch (1980) adopt the species packing model to remedy the compression problem. The "species turnover rate" (assumed constant) can be estimated at a point along the gradient by the dispersion of the scores of the species present in a sample at that point. Hill and Gauch therefore try to equalize the mean within-sample dispersion of the species scores at all points along the axis by rescaling the species scores [see Hill (1979) for the details]. Thereafter the sample scores are simply derived by weighted averaging.

The horseshoe effect is considered by Hill and Gauch (1980) as "a mathematical artifact, corresponding to no real structure in the data." They eliminate the horseshoe by "detrending." Detrending intends to assure that, at any point along the first axis, the mean value of the sample scores on the subsequent axes is approximately 0. To this end the first axis is divided into a number of segments and within each segment the sample scores on axis 2 are adjusted by centering them to zero mean. The program by Hill (1979) uses running segments for this purpose. This process of detrending is built into the reciprocal averaging algorithm that generates the normal correspondence analysis solution, and replaces the usual orthogonalization procedure. Subsequent axes are derived similarly by detrending with respect to each of the existing axes.

Detrended correspondence analysis has been tested on data sets simulated under the Gaussian response model in one to four dimensions and was found to recover the structure of the data well (Hill and Gauch, 1980; Gauch, Whittaker, and Singer, 1981).

6. Numerical Comparisons

6.1 *Introduction*

The theoretical comparisons described so far are approximate and are supplemented in this section by numerical comparisons, using simulated data sets and one real data set. The performance of correspondence analysis is judged by correlations of the sample scores with the real values and by log-likelihood.

6.2 *Methods*

Data were simulated under the response models (3.1) and (5.1) in one and two dimensions, respectively, using unit tolerance and equal maxima. The optima and sample points were drawn in each simulation independently from a uniform distribution on an interval and rectangle with prechosen length and sides, respectively. Ecologists refer to such simulations as coenocline and coenoplane simulations [see Gauch (1982)]. The simulations were constrained to give at least three occurrences in each sample and at least three occurrences per species, to ensure that all parameters could be estimated.

Subroutines from Hill (1979) were used to calculate the (detrended) correspondence analysis solution for the species optima and sample scores with $\alpha = 1$ and Hill's (1979) standardization (§4.3). With these scores and $t = 1$ the species maxima were estimated by maximum likelihood, analytically in case of Poisson counts (Kooijman, 1977), and numerically in case of 1/0 data. For this solution the likelihood was calculated. In this simple approach the choice of α is arbitrary, but influences the likelihood. In a second approach this problem was solved by calculating for each species the regression of the species' responses on the sample scores. This is easy because models (3.1) and (5.1) are generalized linear models (Nelder and Wedderburn, 1972). The tolerances were kept fixed to 1 in the regressions.

The maximum likelihood solution was derived by alternating "regressions" to estimate the species parameters and "calibrations" to estimate the sample parameters, the latter being centred and, in two dimensions, rotated to principal axes in each iteration (Kooijman, 1977). Thus, regression and calibration replace the simple weighted averages in the two-way averaging algorithm to derive the correspondence analysis solution. In each regression step and each calibration step the Gauss–Newton method was used with Gallant's (1975) chopping rule for stepshortening, and a primitive method that prevented parameters from iterating to infinity. As usual, it cannot be guaranteed that the overall maximum of the likelihood is found, but the algorithm is at least hill climbing. This optimization method is akin to the EM algorithm (Dempster, Laird, and Rubin, 1977), the difference being that

with the EM algorithm it is assumed that the incidental parameters are random, whereas in this paper they are treated as fixed parameters. EM maximizes therefore a marginal likelihood (Bock and Aitkin, 1981), whereas here the joint likelihood is maximized. The (detrended) correspondence analysis solutions and also, when available, the true parameter values provided the initial parameter values.

6.3 *Simulation Results*

Table 1 summarizes simulations of incidence matrices (A–E) and matrices with counts (F–I), the former simulated from the logistic response curves (3.1), the latter from the loglinear response surfaces (5.1), all with unit tolerance. The maximum probability of occurrence is .7 in A, B, and C, and .5 in D and E. The maximum count is either 5 (F, G, H) or 1 (I).

Table 2 shows an example of B in which the length of the sampled interval is 5 tolerance units and Figure 1 displays its correspondence analysis solution. Although some of the species scores are out of order, the correlation of the scores of samples and of species with the true values is over .9 and the deviance is even lower than under the true parameter values. Table 1 shows that in all simulations correspondence analysis performed well for the first dimension, but in simulations F–I, badly for the second dimension. Detrended correspondence analysis is comparable to correspondence analysis in one dimension (A–E), but far superior in two dimensions (F–I).

Table 1

Results of simulations of the models (3.1) *and* (5.1) *with unit tolerance, for* 1/0 *data in one dimension* (A–E) *and for Poisson counts in two dimensions* (F–I). *Shown are average values of at least four simulations (first axis* 1, *then axis* 2, *if appropriate).*

Simulation	A	B	C	D	E	F	G	H	I
No. of species	30	10	30	30	30	40	40	40	40
No. of samples	20	50	50	50	50	50	50	50	50
Range of u	12	6	5	5	3	10; 5	5; 5	7; 4	7; 4
Range of x	10	5	4	4	2	8; 4	4; 4	6; 3	6; 3
Value of a	1	1	1	0	0	1.6	1.6	1.6	0
No. of par.	79	69	109	109	109	218	218	218	218
df	521	431	1391	1391	1391	1782	1782	1782	1782
Eigenvalues (× 100)									
CA	90	50	38	52	18	88; 63	61; 49	77; 44	81; 57
DCA	90	50	38	52	18	88; 45	61; 39	77; 34	81; 44
Deviances									
Null model	634	654	1941	1641	1936	3448	4316	4000	1477
True par.	327	483	1556	1396	1883	836	1377	1225	856
CA	308	458	1506	1289	1778	1696	1708	1958	907
DCA	292	445	1533	1324	1789	1010	1433	1194	681
CA + REGR	264	441	1475	1280	1758	1167	1320	1374	754
DCA + REGR	279	423	1495	1309	1781	775	1255	1070	642
ML	217	417	1440	1259	1739	648	1170	994	598
Correlation with true sample scores (× 100)									
CA	98	90	95	95	67	97; 57	—	98; 64	96; 53
DCA	98	90	96	91	51	98; 83	—	99; 91	96; 77
ML	99	86	94	92	67	99; 95	—	99; 93	96; 77

No. = number; u = species optima; x = sample scores; par. = parameters; df = degrees of freedom; CA = correspondence analysis; DCA = detrended correspondence analysis; (D)CA + REGR = (D)CA followed by regression on (D)CA sample scores; ML = maximum likelihood.

—: Meaningless.

Table 2

Incidence matrix simulated from unimodal response curves (3.1) *under condition* B *in Table* 1, *The species* (*rows*) *and samples* (*columns*) *are arranged in increasing order of the true optima and sample values, respectively.*

```
111111111101100101010000010001000000000000000000
111011111001111111001100100100000000000000000000
110000011001011011111111100100000110000000000000
011100111110110010111011010101111101011000000000
111111100100111101111001110100101100000000100000
001100010111011010111101111011101111101100000000
000101010110000111110111110010001011111011001111000
000000000000000110100001001000111111011111111111111
000000000000101000001110010011011001101111011111111
000000000000000000000000000010001010010111010011111
```

In two dimensions each solution of correspondence analysis showed the horseshoe, most in F and H, least in G and I. The lower the maximum of the response curves, the better correspondence analysis (D vs C and I vs H), in accordance with the theory. The simulations also confirm the observation of Hill and Gauch (1980) that correspondence analysis works more satisfactorily with square sampling regions as compared to rectangular regions (G vs F, H). In order to determine whether the success of detrended correspondence analysis is due to the rescaling of the axes or to the detrending, some tests were done with rescaling, but without detrending. These tests showed a slight, but unimportant improvement over the results of correspondence analysis. The success of detrended correspondence analysis is therefore mainly due to the detrending.

The eigenvalues showed little variation between simulations of the same type; for example, in A and F the standard deviations were below 0.05.

The estimates of the species optima can be improved by regressing each species response on the sample scores, as can be seen from the drop in the deviance (Table 1) and the increase in correlation with the true optima (not shown). The deviance after regression on the sample scores from detrended correspondence analysis was in nearly all simulations less than the deviance under the true parameters.

The maximum likelihood solution has, by definition, the lowest deviance, but does not always give the highest correlation with the true sample scores. Of the three sets of initial values used to derive the maximum likelihood solution, the true values and the values from detrended correspondence analysis gave nearly identical solutions. Starting from the correspondence analysis solution, the maximization procedure frequently became trapped in a local maximum in simulations F–I.

For statistical tests and confidence regions it is tempting to assume that deviances are chi-squared distributed. This assumption is risky in this context because the number of parameters increases with the number of observations. Indeed, the true parameter values lie outside the usual 95% confidence region in 34% of the 29 simulations of the one-dimensional model and in 12% of the 24 simulations of the two-dimensional model.

6.4 *A Real Data Set*

The real data set, taken from Van der Aart and Smeenk-Enserink (1975), concerns the distribution of twelve wolfspiders (Lycosidae) in a dune area and consists of their accumulated catches in 100 samples. The maximum count in the data is 189, far higher than in the simulations, but zeroes are as equally abundant as in the simulations. Correspondence analysis was applied to these data, giving .65 and .42 for the first two eigenvalues. The

sample scores of the second axis showed a clear quadratic trend with respect to those of the first axis. Removing this trend, detrended correspondence analysis resulted in a second eigenvalue of .09. This small value indicates that the second axis is unimportant for these data, which agrees with the results of Kooijman (1977), who fitted one- and two-dimensional Gaussian response models to these data by maximum likelihood.

Table 3 shows the results of loglinear regressions of the catches of the wolfspiders on the sample scores of the first axis of detrended correspondence analysis. When a quadratic term was added to the model, the deviance decreased considerably for nine of the twelve spider species. Their fitted curves are all unimodal (see Figure 2). The rescaling of the axis in detrended correspondence analysis appears advantageous for these data, as the quadratic fit with respect to the first axis of the usual correspondence analysis resulted in a 50% higher deviance. The full maximum likelihood solution (with equal tolerances) gave a deviance of 4890, 30% lower than the deviance of the quadratic model in Table 3. Yet the sample scores as estimated by maximum likelihood showed a high correlation (.95) with those of detrended correspondence analysis.

Van der Aart and Smeenk-Enserink (1975) also characterized the vegetation and the soil around 28 of the 100 pitfall traps. They state, "The sites were selected in such a way that as many biotope types as possible were represented." Interpreting the first axis of detrended correspondence analysis as a latent variable, we can therefore attempt to relate this latent variable to the measured environmental variables. A multiple regression of the first axis' scores on the logarithms of the variables soil water content, percentage of bare sand, and percentage cover by mosses accounted for 90% of the variance. All three variables contributed to this regression, as judged by t tests on the regression coefficients. The first axis can therefore be interpreted as a composite gradient of soil moisture and openness of the habitat. A possible explanation for these results is that wolfspiders require an open habitat for hunting purposes but, on the other hand, require moisture to avoid desiccation. Each species balances these conflicting requirements in its own way and is therefore largely confined to a specific interval along the composite gradient of soil moisture and openness. Other factors related to soil moisture or openness cannot be excluded to be operational.

Table 3

Loglinear regressions of catches of wolfspiders (k) on the sample scores (x_i) of the first axis of detrended correspondence analysis. Given are the deviance of the null model and the decreases in deviance when the loglinear model is extended successively with a linear ($b_{k1}x$) and a quadratic term ($b_{k2}x^2$). Provided $b_{k2} < 0$, the quadratic model fits Gaussian response curves with unequal tolerances [equation (3.1)]. The spiders are arranged in order of the species score of the first axis.

			Successive decrease in deviance	
	Model: $\log \mu_{ki} =$	Deviance b_{k0}	$+ b_{k1}x_i$	$+ b_{k2}x_i^2$
k	**Wolfspider**			
1	*Pardosa lugubris*	1494	1159	11
2	*Zora spinimana*	935	245	341
3	*Pardosa nigriceps*	3109	388	1490
4	*Trochosa terricola*	3671	1033	1743
5	*Pardosa pullata*	4504	427	2570
6	*Arctosa lutetiana*	315	18	149
7	*Aulonia albimana*	958	93	488
8	*Alopecosa cuneata*	1396	57	696
9	*Pardosa monticola*	4103	130	3023
10	*Alopecosa accentuata*	856	329	202
11	*Alopecosa fabrilis*	864	693	24
12	*Arctosa perita*	340	254	3
	Total	22545	4826	10740

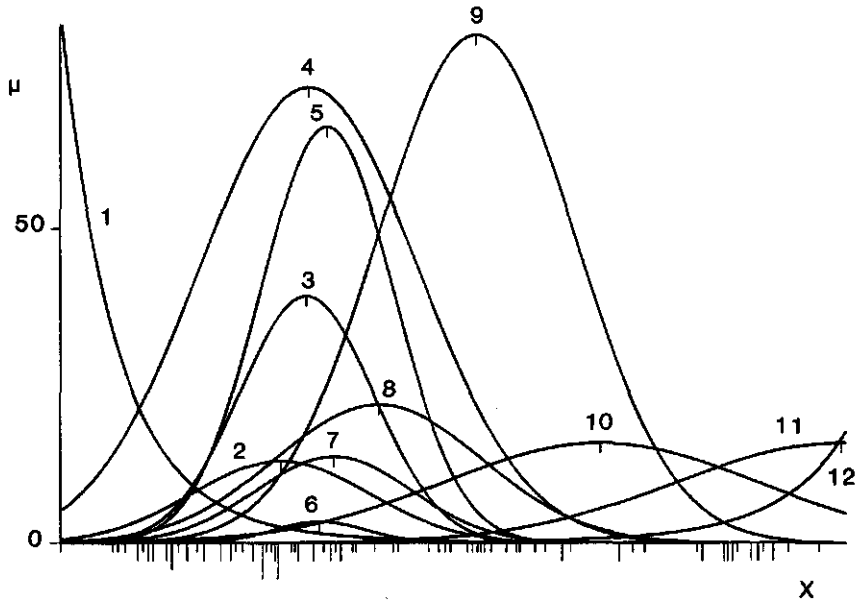


Figure 2. Unimodal response curves (3.1) for the expected number (μ) of wolfspiders along the first axis of detrended correspondence analysis (x), fitted by loglinear regression (Table 3, last column). The curves are labelled by the species identification numbers of Table 3. The sample points are indicated by ticks below the abscissa (length proportional to number). Data from Van der Aart and Smeenk-Enserink (1975).

7. Discussion

Both the unimodal model (3.1) with $t_k = t$ and the bilinear model (2.4) stand at the basis of correspondence analysis. The clue to this apparent paradox is data transformation. In linear regression, data transformation can be used to linearize *monotone* relationships. In multivariate analysis, data transformation can also be used to linearize nonmonotone relationships. Correspondence analysis is not the only example. Kooijman (1977) showed that principal component analysis recovers exactly the parameters of equal tolerance Gaussian curves and surfaces from error-free data when the data matrix is centered by rows and by columns after log transformation. Aitchison (1983) proposed this transformation to overcome the difficulty of the constant-sum constraint in principal component analysis of compositional data. He notices that "the nonlinearity of the logarithmic function opens up the possibility of coping with curvature in data sets ...," but does not refer to the Gaussian or unimodal response model. [His Figure 2(b) clearly shows the unimodal response of constituent F along the first principal component.] Ihm and Van Groenewoud (1975) used a different transformation to analyze Gaussian response curves by principal component analysis. Their method requires the same assumptions as correspondence analysis about the distribution of the optima and the sample points.

Four conditions (equal tolerances, equal or independent maxima, and equally-spaced or uniformly distributed optima and sample points) are needed to show that (detrended) correspondence analysis provides an *approximate* solution to the unimodal models (3.1) and (5.1). How realistic are these assumptions in practice and how robust is correspondence

analysis to violations of the assumptions? Some checks on the assumptions are possible, e.g., by regressing each species' responses on the derived sample scores, allowing the tolerances and maxima to vary among species, and I suggest that this should be done routinely, if only to determine the goodness-of-fit of the model for descriptive purposes. Ihm and Van Groenewoud (1975) and Kooijman (1977) reported that the optima and sample values as estimated by their methods are fairly robust against unequal tolerances, as did Hill and Gauch (1980) for detrended correspondence analysis. The four conditions are not needed in the maximum likelihood approach, taken by Gauch, Chase, and Whittaker (1974) for normal data, Kooijman (1977) for Poisson data, and Goodall and Johnson (1982) for presence/absence data. Yet, the maximum likelihood approach is applied seldom in ecological research because of its computational complexity and the lack of reliable and flexible software (Gauch, 1982). Another reason might be that correspondence analysis appears to be "nonparametric." However, this paper reveals its close connection with "Gaussian" response curves with equal tolerances.

Commonly high values in the data matrix are downweighted in correspondence analysis by, for example, a prior square root transformation. However, when the variance is proportional to the mean, transformation is not required (Wedderburn, 1974). Overdispersion then inflates the mean deviance, not necessarily implying lack of fit. When the type of dispersion or lack of fit is allowed to vary between species, all problems of common factor analysis are lurking in the way.

Principal component analysis and correspondence analysis are rival methods for dimensionality reduction for abundance data (Gauch, Whittaker, and Wentworth, 1977; Greig-Smith, 1983), both allowing "major features" of the data to be visualized in joint plots of species and sample scores. The geometrical interpretation of a principal component plot is based on the bilinear model, as stressed by Gabriel (1971), who termed the plot a *biplot*. The value of a variable, as approximated by the biplot, changes linearly across the plot. Correspondence analysis therefore gives a biplot of the transformed data values (2.4). However, in terms of the original data Y the joint plot of correspondence analysis is not a biplot, because the model *for the original data* is unimodal rather than bilinear. The original value of a variable, as approximated by a correspondence analysis plot, is maximum at this variable's point in the plot and decreases with distance from that point, disregarding for a moment the fact that (detrended) correspondence analysis provides only an approximate solution to the unimodal models (3.1) and (5.1). We may interpret the correspondence analysis plot more informally as Benzécri et al. (1973) do. Their centroid principle (*le principe barycentrique*) is simply the transition formulae interpreted geometrically. Multidimensional unfolding provides the same kind of plot (Carroll, 1972).

Although principal component analysis and correspondence analysis model and display multivariate data in different ways, the resulting plots of the sample scores are sometimes similar. This happens when all unimodal surfaces are truncated to monotone surfaces over the region actually sampled, the monotone surfaces being approximated by planes in principal component analysis. In such cases the correspondence analysis solution with $\alpha = 1$ shows some species points close to the centroid of the sample points, whereas the other species' points fall outside the region where the sample points lie.

ACKNOWLEDGEMENTS

I would like to thank Dr I. C. Prentice for valuable discussions and comments.

RÉSUMÉ

L'analyse des correspondances est couramment utilisée par les écologistes pour analyser des données de présence/absence ou d'abondance d'espèces. Les tout premiers axes sont interprétés en termes de variables sous-jacentes conditionnant la distribution des espèces. On fait l'hypothèse que ces variables

sont liées aux variables de milieu non explicitées. Dans cet article, on montre qu'en utilisant l'analyse des correspondances, on obtient une solution approchée de la solution donnée par la technique du maximum de vraisemblance dans le cas de modèles de réponse unimodale à une variable sous-jacente. Les modèles utilisés sont des modèles logistiques–linéaires en ce qui concerne les données de présence/absence et log-linéaires pour des abondances suivant des lois de Poisson, les estimateurs étant des fonctions quadratiques de la variable sous-jacente. On obtient une approximation de meilleure qualité lorsque, d'une part, les maximum et les amplitudes (tolérances des espèces aux conditions de milieu) des courbes de réponse des espèces ont mêmes valeurs et que, d'autre part, les valeurs de la variable sous-jacente correspondant aux optimum de chaque espèce et aux pointes d'échantillonnage sont régulièrement réparties. L'approximation demeure satisfaisante pour des optimum et des valeurs correspondant aux échantillons distribués uniformément, ainsi que le montre la simulation. Pour des modèles à 2 variables sous-jacentes, l'approximation est souvent mauvaise en raison de la présence d'un effet Guttman. L'approximation est de bien meilleure qualité lorsque l'on réalise des simulations après avoir retiré cet effet, ce qui se produit lorsqu'on utilise une technique d'analyse des correspondances qui efface la tendance centrale du phénomène étudié.

References

Aart, P. J. M. van der and Smeenk-Enserink, N. (1975). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology* **25,** 1–45.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* **70,** 57–65.

Austin, M. P. (1976). On nonlinear species response models in ordination. *Vegetatio* **33,** 33–41.

Benzécri, J. P. et al. (1973). *L'Analyse des Données*: II. *L'Analyse des Correspondances.* Paris: Dunod.

Bock, R. D. and Aitkin, M. A. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46,** 443–459.

Braak, C. J. F. ter and Barendregt, L. G. (in press). Weighted averaging of species indicator values: Its efficiency in environmental calibration. *Mathematical Biosciences.*

Braak, C. J. F. ter and Looman, C. W. N. (in press). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio.*

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In *Multidimensional Scaling. Theory and Applications in the Behavioral Sciences. Vol.* I: *Theory,* R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds), 105–155. New York: Seminar Press.

Christiansen, F. B. and Fenchel, T. M. (1977). *Theories of Populations in Biological Communities.* Berlin: Springer-Verlag.

Coombs, C. H. (1964). *A Theory of Data.* New York: Wiley.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58,** 453–467.

Gallant, A. R. (1975). Nonlinear regression. *The American Statistician* **29,** 73–81.

Gauch, H. G. (1982). *Multivariate Analysis in Community Ecology.* Cambridge: Cambridge University Press.

Gauch, H. G., Chase, G. B., and Whittaker, R. H. (1974). Ordination of vegetation samples by Gaussian species distributions. *Ecology* **55,** 1382–1390.

Gauch, H. G., Whittaker, R. H., and Singer, S. B. (1981). A comparative study of nonmetric ordinations. *Journal of Ecology* **69,** 135–152.

Gauch, H. G., Whittaker, R. H., and Wentworth, T. R. (1977). A comparative study of reciprocal averaging and other ordination techniques. *Journal of Ecology* **65,** 157–174.

Goodall, D. W. and Johnson, R. W. (1982). Nonlinear ordination in several dimensions. A maximum likelihood approach. *Vegetatio* **48,** 197–208.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.

Greig-Smith, P. (1983). *Quantitative Plant Ecology,* 3rd ed. Oxford: Blackwell.

Hill, M. O. (1973). Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology* **61,** 237–249.

Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Applied Statistics* **23,** 340–354.

Hill, M. O. (1979). DECORANA—*A* FORTRAN *Program for Detrended Correspondence Analysis and Reciprocal Averaging.* Ithaca, New York: Cornell University.

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

Ihm, P. and Groenewoud, H. van (1975). A multivariate ordering of vegetation data based on Gaussian-type gradient response curves. *Journal of Ecology* **63,** 767–777.

Kooijman, S. A. L. M. (1977). Species abundance with optimum relations to environmental factors. *Annals of Systems Research* **6,** 123–138.

Kovács, M. (1969). Das Corno-quercetum des Mátra-gebirges. *Vegetatio* **19,** 240–255.

MacArthur, R. H. and Levins, R. (1967). The limiting similarity, convergence, and divergence of coexisting species. *American Naturalist* **101,** 377–385.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications.* Toronto: University of Toronto Press.

Schriever, B. F. (1983). Scaling of order-dependent categorical variables with correspondence analysis. *International Statistical Review* **51,** 225–238.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61,** 439–447.

Whittaker, R. H. (1956). Vegetation of the Great Smoky Mountains. *Ecological Monographs* **26,** 1–80.

Whittaker, R. H. (1967). Gradient analysis of vegetation. *Biological Reviews* **42,** 207–264.

Whittaker, R. H., Levin, S. A., and Root, R. B. (1973). Niche, habitat and ecotope. *American Naturalist* **107,** 321–338.

Received June 1984; *revised June* 1985.

Ecology, 67(5), 1986, pp. 1167–1179

CANONICAL CORRESPONDENCE ANALYSIS: A NEW EIGENVECTOR TECHNIQUE FOR MULTIVARIATE DIRECT GRADIENT ANALYSIS[1]

CAJO J. F. TER BRAAK
TNO Institute of Applied Computer Science, P. O. Box 100, 6700 AC Wageningen, The Netherlands, and Research Institute for Nature Management, Leersum, The Netherlands

Abstract. A new multivariate analysis technique, developed to relate community composition to known variation in the environment, is described. The technique is an extension of correspondence analysis (reciprocal averaging), a popular ordination technique that extracts continuous axes of variation from species occurrence or abundance data. Such ordination axes are typically interpreted with the help of external knowledge and data on environmental variables; this two-step approach (ordination followed by environmental gradient identification) is termed indirect gradient analysis. In the new technique, called canonical correspondence analysis, ordination axes are chosen in the light of known environmental variables by imposing the extra restriction that the axes be linear combinations of environmental variables. In this way community variation can be directly related to environmental variation. The environmental variables may be quantitative or nominal. As many axes can be extracted as there are environmental variables. The method of detrending can be incorporated in the technique to remove arch effects.

(Detrended) canonical correspondence analysis is an efficient ordination technique when species have bell-shaped response curves or surfaces with respect to environmental gradients, and is therefore more appropriate for analyzing data on community composition and environmental variables than canonical correlation analysis. The new technique leads to an ordination diagram in which points represent species and sites, and vectors represent environmental variables. Such a diagram shows the patterns of variation in community composition that can be explained best by the environmental variables and also visualizes approximately the "centers" of the species distributions along each of the environmental variables. Such diagrams effectively summarized relationships between community and environment for data sets on hunting spiders, dyke vegetation, and algae along a pollution gradient.

Key words: biplot; canonical correlation analysis; canonical correspondence analysis; detrended correspondence analysis; Gaussian model; gradient analysis; ordination; reciprocal averaging; regression; species–environment relations; unfolding; weighted averaging.

INTRODUCTION

Problems in community ecology often require the inferring of species–environment relationships from community composition data and associated habitat measurements. Typical data for such problems consist of two sets: data on the occurrence or abundance of a number of species at a series of sites, and data on a number of environmental variables measured at the same sites. (A "site" is the basic sampling unit, separated in space or time from other sites, e.g., a quadrat, a woodlot, a light trap, or a plankton sample.) When the data are collected over a sufficient habitat range for species to show nonlinear, nonmonotonic relationships with environmental variables, it is inappropriate to summarize these relationships by correlation coefficients or to analyze the data by techniques that are based on correlation coefficients, such as canonical correlation analysis (Gauch and Wentworth 1976, Gittins 1985). An alternative, two-step approach has become popular: (1) extract from the species data the dominant pattern of variation in community composition by an ordination technique, such as (detrended) correspondence analysis, and (2) attempt to relate this pattern (i.e., the first few ordination axes) to the environmental variables (Gauch 1982*a*). The particular merit of detrended correspondence analysis in this context is that it removes nonlinear dependencies between axes (Hill and Gauch 1980) and has been shown to be an efficient technique to extract one or more ordination axes ("gradients") such that species show unimodal (bell-shaped) response curves or surfaces with respect to these axes (Ter Braak 1985*b*). The axes can be thought of as hypothetical environmental gradients, which are subsequently interpreted in terms of measured environmental variables in the second step of the analysis. This two-step approach is essentially Whittaker's (1967) indirect gradient analysis.

What can be inferred from indirect gradient analysis? If the measured environmental variables relate strongly to the first few ordination axes, they can "account for" (i.e., they are sufficient to predict) the main part of the variation in the species composition. If the environmental variables do not relate strongly to the first few axes, they cannot account for the main part of the variation, but they may still account for some of the remaining variation—which can be substantial. Further, it is nontrivial to detect by indirect gradient anal-

[1] Manuscript received 18 March 1985; revised 12 November 1985; accepted 22 January 1986.

ysis the effects on community composition of a subset of environmental variables in which one is particularly interested (Carleton 1984). These limitations can only be overcome by methods of direct gradient analysis, in which species occurrences are related directly to environmental variables (Gauch 1982*a*). Methods of direct gradient analysis in current use consider essentially one species at a time. Simple methods involve plotting species abundance against a single environmental variable, or isopleths in a space of two environmental variables (Whittaker 1967). More elaborate methods use (generalized linear) regression methods (Austin et al. 1984, Bartlein et al. 1986) and are useful in studying simultaneously the effect of more than one environmental variable. Regression methods allow fitted response surfaces to assume a wide variety of shapes. However, when the number of species is large, separate regression analysis for each species may be impractical. Moreover, separate analyses cannot be combined easily to get an overview of how community composition varies with the environment (in particular, when the number of environmental variables exceeds two or three), and a multivariate method (based on a common response model) is required.

In this paper a multivariate direct gradient analysis technique is developed, whereby a set of species is related directly to a set of environmental variables. The new technique identifies an environmental basis for community ordination by detecting the patterns of variation in community composition that can be explained best by the environmental variables. In the resulting ordination diagram, species and sites are represented by points and environmental variables are represented by arrows. Such a diagram shows the main pattern of variation in community composition as accounted for by the environmental variables, and also shows, in an approximate way, the distributions of the species along each environmental variable. The technique thus combines aspects of regular ordination with aspects of direct gradient analysis. The rationale of the technique is derived from a species packing model wherein species are assumed to have Gaussian (bell-shaped) response surfaces with respect to compound environmental gradients. These gradients are assumed to be linear combinations of the environmental variables. The new technique is called canonical correspondence analysis, because it is a correspondence analysis technique in which the axes are chosen in the light of the environmental variables. Examples demonstrate that canonical correspondence analysis allows a quick appraisal of how community composition varies with the environment.

THEORY

Data and model

Suppose a survey of n sites lists the abundances or occurrences (presence scored as 1, absence as 0) of m species and the values of q environmental variables ($q < n$). Let y_{ik} be the abundance or presence/absence (1/0) of species k ($y_{ik} \geq 0$), and z_{ij} the value of environmental variable j at site i.

The first step in indirect gradient analysis is to summarize the main variation in the species data by ordination. The method of Gaussian ordination (Gauch et al. 1974) does this by constructing an axis such that the species data optimally fit Gaussian response curves along this axis. Then the response model for the species is the bell-shaped function

$$E(y_{ik}) = c_k \exp[-\tfrac{1}{2}(x_i - u_k)^2/t_k^2], \qquad (1)$$

where $E(y_{ik})$ denotes the expected (average) value of y_{ik} at site i that has score x_i on the ordination axis. The parameters for species k are c_k, the maximum of that species' response curve; u_k, the mode or optimum (i.e., the value of x for which the maximum is attained); and t_k, the tolerance, a measure of ecological amplitude. Ter Braak (1985*b*) showed that correspondence analysis approximates the maximum likelihood solution of Gaussian ordination, if the sampling distribution of the species abundances is Poisson, and if:

- C1) the species' tolerances are equal ($t_k = t$, $k = 1, \ldots, m$),
- C2) the species' maxima are equal ($c_k = c$, $k = 1, \ldots, m$),
- C3) the species' optima $\{u_k\}$ are homogeneously distributed over an interval A that is large compared to t,
- C4) the site scores $\{x_i\}$ are homogeneously distributed over a large interval B that is contained in A.

(The wording "homogeneously distributed" is used to cover either of two cases, namely (1) that the scores are equispaced, with spacing small compared to t, or (2) that the scores are drawn randomly from a uniform distribution.) Conditions C1–C3 imply a species packing model (Whittaker et al. 1973) with respect to the ordination axis. The species scores resulting from a correspondence analysis actually estimate the optima of the species in this model. Ter Braak (1985*b*) provided a similar rationale for correspondence analysis of presence-absence data. Conditions C1 and C2 are not likely to hold in most natural communities, but the usefulness of correspondence analysis in practice relies on its robustness against violations of these conditions (Hill and Gauch 1980).

The second step of indirect gradient analysis is to relate the ordination axis to the environmental variables, for example graphically, or by calculating correlation coefficients, or by multiple regression (see Montgomery and Peck 1982) of the site scores on the environmental variables

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij}, \qquad (2)$$

where b_0 is the intercept and b_j is the regression coefficient for environmental variable j. Note that the species optima u_k and sites scores x_i are estimated from the species data first; the regression coefficients b_j are estimated next, keeping x_i (and u_k) fixed. The species data are thus indirectly related to the environmental variables, via the ordination axis.

The technique proposed in this paper simultaneously estimates the species optima, the regression coefficients and, hence, the site scores by using the model described by Eq. 1, in conjunction with Eq. 2. Simultaneous estimation turns the technique into a direct gradient analysis method. In principle the method of maximum likelihood could be used to obtain the estimates. This analysis could be called Gaussian canonical ordination. It requires excessively heavy computation. The computational task can, however, be alleviated considerably if conditions C1–C4 hold. The reasoning that led from Gaussian ordination to correspondence analysis, now leads to the transition formulae of canonical correspondence analysis (see Appendix):

$$\lambda u_k = \sum_{i=1}^{n} y_{ik} x_i / y_{+k} \tag{3}$$

$$x_i^* = \sum_{k=1}^{m} y_{ik} u_k / y_{i+} \tag{4}$$

$$\boldsymbol{b} = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\boldsymbol{x}^* \tag{5}$$

$$\boldsymbol{x} = \mathbf{Z}\boldsymbol{b}, \tag{6}$$

where y_{+k} and y_{i+} are species and site totals, respectively, $\mathbf{R}$ is a diagonal $n \times n$ matrix with y_{i+} as the (i, i)-th element; $\mathbf{Z} = \{z_{ij}\}$ is an $n \times (q + 1)$ matrix containing the environmental data and a column of ones; and $\boldsymbol{b}$, $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are column-vectors: $\boldsymbol{b} = (b_0, b_1, \ldots, b_q)'$, $\boldsymbol{x} = (x_1, \ldots, x_n)'$, and $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)'$. The transition formulae define an eigenvector problem (see Appendix) that is akin to the eigenvector problem posed by canonical correlation analysis, λ in Eq. 3 being the eigenvalue. As in correspondence analysis, the equations have a trivial solution in which all site and species scores are equal and $\lambda = 1$; this trivial solution can either be disregarded or be excluded by requiring that the site scores are centered to zero mean, i.e., $\Sigma_i y_{i+} x_i = 0$.

Algorithm: reciprocal averaging and regression

The transition formulae can be solved by the following iteration algorithm of reciprocal averaging and multiple regression.

S1) Start with arbitrary, but unequal, initial site scores.

S2) Calculate species scores by weighted averaging of the site scores (Eq. 3 with $\lambda = 1$).

S3) Calculate new site scores by weighted averaging of the species scores (Eq. 4).

S4) Obtain regression coefficients by weighted multiple regression of the site scores on the environmental variables (Eq. 5). The weights are the site totals (y_{i+}).

S5) Calculate new site scores by Eq. 6 or, equivalently, Eq. 2. The new site scores are in fact the fitted values of the regression of the previous step.

S6) Center and standardize the site scores such that

$$\Sigma_i y_{i+} x_i = 0 \quad \text{and} \quad \Sigma_i y_{i+} x_i^2 = 1. \tag{7}$$

S7) Stop on convergence, i.e., when the new site scores are sufficiently close to the site scores of the previous iteration; otherwise go to S2.

This procedure is akin to the reciprocal averaging algorithm of correspondence analysis, but steps S4 and S5 are additional. The new technique is a correspondence analysis technique with restrictions (S4 and S5) on the site scores (cf. De Leeuw 1984). The final regression coefficients will be called canonical coefficients, and the multiple correlation coefficient of the final regression will be called the species–environment correlation. The species–environment correlation is a measure of how well the extracted variation in community composition can be explained by the environmental variables and is equal to the correlation between the site scores $\{x_i^*\}$, which are weighted mean species scores (calculated by Eq. 4), and the site scores $\{x_i\}$, which are a linear combination of the environmental variables (calculated by Eq. 2 or Eq. 6). This equality requires the assumption that sites are weighted proportional to y_{i+}, as in steps S4 and S6, and this weighting of sites is assumed in the calculation of means, variances, and correlations throughout the paper.

The standardization of the site scores in S6 is convenient in the algorithm, but it has more meaning ecologically to rescale the solution according to Eq. A.8 of the Appendix, as proposed by Hill (1979). Then, the tolerance of the fitted Gaussian response curves is (on average) about 1 unit, and a species' response curve can be expected to rise and decline over an interval of about 4 units.

More than one dimension and detrending

Second and additional axes can be extracted as in correspondence analysis by adding to the algorithm, after S5, a step that makes the trial site scores uncorrelated with the previous axes. The two-dimensional solution is intended to fit bivariate Gaussian response surfaces to the species data (Ter Braak 1985*b*) but often gives a bad fit because of the arch effect, an approximately quadratic dependence between the scores of the first two axes. This effect crops up whenever a short gradient is dominated by a long gradient (Gauch 1982*a*). The modifications of correspondence analysis that led to detrended correspondence analysis (Hill and Gauch 1980) can also be incorporated in canonical correspondence analysis; the rationale for detrending is the same. Detrending removes the arch effect and im-

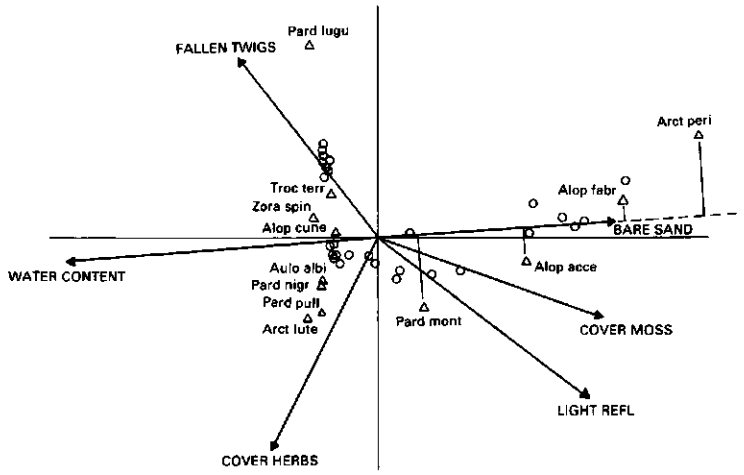


FIG. 1. The distribution of 12 species of hunting spiders caught in pitfall traps in a Dutch dune area. Canonical correspondence analysis (CCA) ordination diagram with pitfall traps (○), hunting spiders (△), and environmental variables (arrows); first axis is horizontal, second axis vertical. Shown also are the projections of the spider points labelled Arct peri, Alop fabr, Alop acce, and Pard mont onto the trajectory of the arrow of bare sand; the order of the projection points indicates the approximate ranking of the centers of the distributions of these spiders along the variable "percentage bare sand," *Arctosa perita* being found in habitats with the highest percentages of bare sand. The spider species are: Alop acce = *Alopecosa accentuata*, Alop cune = *Alopecosa cuneata*, Alop fabr = *Alopecosa fabrilis*, Arct lute = *Arctosa lutetiana*, Arct peri = *Arctosa perita*, Aulo albi = *Aulonia albimana*, Pard lugu = *Pardosa lugubris*, Pard mont = *Pardosa monticola*, Pard nigr = *Pardosa nigriceps*, Pard pull = *Pardosa pullata*, Troc terr = *Trochosa terricola*, Zora spin = *Zora spinimana*. The environmental variables are: Water Content = percentage of soil dry mass, Bare Sand = percentage cover of bare sand, Fallen Twigs = percentage cover of fallen leaves and twigs, Cover Moss = percentage cover of the moss layer, Cover Herbs = percentage cover of the herb layer, and Light Refl = reflection of the soil surface with cloudless sky.

proves the fit to the Gaussian model considerably in simulations where the true site and species scores are homogeneously distributed in a rectangle (the extension to two dimensions of conditions C3 and C4; Ter Braak 1985*b*). Detrending, however, also attempts to impose such a homogeneous distribution of scores on the data where none exists. The computer program CANOCO (Ter Braak 1985*a*) will also perform detrended canonical correspondence analysis. For a comparison of the detrended analysis with the non-detrended analysis, see Tests on Real Data.

Canonical coefficients and intraset correlations

For interpreting the ordination axes one can use the canonical coefficients and the intraset correlations. The canonical coefficients define the ordination axes as linear combinations of the environmental variables through Eq. 2, and the intraset correlations are the correlation coefficients between the environmental variables and these ordination axes. (The term intraset is used here to distinguish these correlations from the interset correlations between the environmental variables and the site scores $\{x_i^*\}$ that are derived from the species data.) For the rest of the analysis it is assumed that the environmental variables have been standardized to zero mean and unit variance prior to the analysis. This standardization removes arbitrariness in the units of measurement of the environmental variables and makes the canonical coefficients comparable to each other, but does not influence other aspects of the analysis.

By looking at the signs and relative magnitudes of the intraset correlations and of the canonical coefficients so standardized, we may infer the relative importance of each environmental variable for predicting the community composition. The canonical coefficients give the same information as the intraset correlations in the special case that the environmental variables are mutually uncorrelated, but may provide rather different information when the environmental variables are correlated with each other, as they usually are in field data. Both a canonical coefficient and an intraset correlation coefficient relate to the rate of change in community composition per unit change in the corresponding environmental variable, but in the former case it is assumed that other environmental variables are being held constant, whereas in the latter case the other environmental variables are assumed to covary with that one environmental variable in the particular way they do in the data set. When the environmental variables are strongly correlated with each other—for example, simply because the number of environmental variables approaches the number of sites—the effects

of different environmental variables on community composition cannot be separated out and, consequently, the canonical coefficients are unstable. This is the multicollinearity problem, well known to occur in multiple regression analysis (see Montgomery and Peck 1982). When this problem arises (the program CANOCO [Ter Braak 1985*a*] provides statistics to help detect it) one should abstain from attempts to interpret the canonical coefficients. Fortunately, the intraset correlations do not suffer from this problem and can still be used for interpretation purposes. One can also remove environmental variables from the analysis, keeping at least one variable per set of strongly correlated environmental variables; the eigenvalues and species–environment correlations will usually decrease only slightly. If the eigenvalues and species–environment correlations drop considerably, one has removed too many (or the wrong) variables.

In contrast to canonical correlation analysis, canonical correspondence analysis is not hampered by multicollinearity in the species data; the number of species is therefore allowed to exceed the number of sites.

Ordination diagram

The solution of canonical correspondence analysis can be displayed in an ordination diagram with sites and species represented by points, and environmental variables represented by arrows (see Fig. 1). The species and site points jointly represent the dominant patterns in community composition insofar as these can be explained by the environmental variables, and the species points and the arrows of the environmental variables jointly reflect the species' distributions along each of the environmental variables. For example, when an arrow refers to "water content," the diagram allows us to infer—by rules explained in the following paragraphs—which species largely occur in the wettest sites, which in the driest sites, and which in sites with intermediate moisture values. We shall limit the discussion to two-dimensional diagrams because these are the most convenient to visualize. The rules for construction and interpretation of higher-dimensional ordination diagrams are the same.

For the diagram to represent the approximate community composition at the sites, we must plot species scores and site scores that are weighted mean species scores, as in Hill's (1979) program DECORANA. Because each site point then lies at the centroid of the species points that occur at that site, one may infer from the diagram which species are likely to be present at a particular site. Also, insofar as canonical correspondence analysis is a good approximation to the fitting of Gaussian response surfaces, the species points are approximately the optima of these surfaces; hence the abundance or probability of occurrence of a species decreases with distance from its location in the diagram.

At which values of an environmental variable a species occurred in the data can conveniently be summarized by the weighted average. The weighted average of a species distribution (k) with respect to an environmental variable (j) is defined as the average of the values of that environmental variable at those sites at which that species occurs, the weighting of each site being proportional to species abundance, i.e.,

$$z_{kj} = \sum_{i=1}^{n} y_{ik} z_{ij} / y_{+k}. \qquad (8)$$

The weighted average indicates the "center" of a species' distribution along an environmental variable (Ter Braak and Looman 1986), and differences in weighted averages between species indicate differences in their distributions along that environmental variable. The ordination diagram of canonical correspondence analysis can be supplemented by arrows for the environmental variables to give a graphical summary of the weighted averages of all species with respect to all environmental variables.

The arrows for the environmental variables must be added in the following way. The position of the head of the arrow for an environmental variable depends on the eigenvalues of the axes and the intraset correlations of that environmental variable with the axes (see Appendix). The coordinate of the head of the arrow on axis s must be $[\lambda_s(1 - \lambda_s)]^{1/2}$ times the intraset correlation of the environmental variable with axis s, where λ_s is the eigenvalue of axis s and it is assumed that the species scores are standardized according to Appendix Eq. A.8, as before. By connecting the origin of the plot (the centroid of the site points) with each of the arrowheads, we obtain the arrows representing the variables (Fig. 1). How to construct such a diagram from a detrended canonical correspondence analysis is described in the Appendix. Only the directions and relative lengths convey information, so one can increase or reduce the lengths of all arrows to fit conveniently in the ordination diagram.

The ordination diagram so constructed allows the following interpretation. Each arrow determines a direction or axis in the diagram, obtained by extending the arrow in both directions (in your mind or on paper). From each species point we must drop a perpendicular to this axis. Fig. 1 shows an example. The arrow for water content has been extended (the axis happens to coincide with the arrow for bare sand) and perpendiculars have been dropped to this axis from four species points. The endpoints indicate the relative positions of the centers of the species distributions along the water content axis or, more precisely, they indicate in an approximate way the relative value of the weighted average of each species with respect to water content. From Fig. 1 we thus infer that *Arctosa perita* has the lowest weighted average with respect to water content (i.e., it largely occurs at the driest sites), *Alopecosa fabrilis* the second lowest value, and so on to *Arctosa lutetiana*, which is inferred to have the highest weight-

TABLE 1. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of hunting spider data (see Fig. 1): eigenvalues and species-environment correlation coefficients for the first three axes.

	Axis		
	1	2	3
	Eigenvalues		
DCA	0.58	0.16	0.02
CCA	0.53	0.21	0.06
DCCA	0.53	0.13	0.02
	Correlation coefficients		
DCA	0.96	0.92	0.88
CCA	0.96	0.93	0.64
DCCA	0.97	0.94	0.90

ed average (i.e., to occur largely at the wettest sites). In general, the approximate ranking of the weighted averages for a particular environmental variable can be seen easily from the order of the endpoints of the perpendiculars of the species along the axis for that variable. Further, the weighted averages are approximated in the diagram as deviations from the grand mean of each environmental variable, the grand mean being represented by the origin of the plot. A second useful rule for interpreting the diagram is therefore that the inferred weighted average is higher than average if the endpoint of a species lies on the same side of the origin as the head of an arrow does, and is lower than average if the origin lies between the endpoint and the head of the arrow.

These rules for interpreting the joint plot of species points and environmental arrows are identical to the rules for interpreting a biplot (Gabriel 1971). Biplots have been used so far primarily in connection with principal components analysis (Ter Braak 1983), but a biplot is essentially just a joint plot of two kinds of entities that allows a particular kind of quantitative interpretation (Gabriel 1981, Ter Braak 1983). The joint plot of species and environmental variables is, in fact, a biplot. This biplot provides a weighted least squares approximation of the weighted averages of the species with respect to the environmental variables (see Appendix). The measure of goodness of fit, $100 \times (\lambda_1 + \lambda_2)$/(sum of all eigenvalues), expresses the percentage variance of the weighted averages accounted for by the two-dimensional diagram. In interpreting percentages of variance accounted for, it must be kept in mind that the goal is not 100%, because part of the total variance is due to noise in the data (cf. Gauch 1982*b*). Even an ordination diagram that explains only a low percentage may be quite informative.

Finally, the length of an arrow representing an environmental variable is equal to the rate of change in the weighted average as inferred from the biplot, and is therefore a measure of how much the species distributions differ along that environmental variable. Important environmental variables therefore tend to be represented by longer arrows than less important environmental variables.

Relation of canonical correspondence analysis with weighted averaging ordination and discriminant analysis

Canonical correspondence analysis generalizes two existing techniques for direct gradient analysis. When a single quantitative environmental variable is considered, it reduces to weighted averaging ordination (Gauch 1982*a*), because x_i in Eq. 1 is then simply the value of this variable at site i, and fitting this model simplifies under condition C4 to weighted averaging (cf. Ter Braak and Looman 1986). With two quantitative environmental variables, the technique represents the same information in a two-dimensional diagram as weighted averaging ordination with respect to these variables, although the variables are not necessarily displayed as orthogonal directions in the ordination diagram. With a single nominal environmental variable, canonical correspondence analysis is a variant of discriminant analysis (canonical variate analysis) that is appropriate to a unimodal response model, and which can be obtained more simply from a correspondence analysis of a two-way table of species by (classes of) the nominal variable (Greenacre 1984: section 7.1). The cells of the table must contain the total abundances of each of the species in each of the classes. In the resulting ordination diagram the classes are represented by points. This equivalence suggests that it can be more natural to represent nominal environmental variables by points instead of arrows. The point for a class of a nominal environmental variable must be located at the centroid (the weighted average) of the sites belonging to that class. Classes consisting of sites with high values for a species will then tend to lie close to that species' point. Gasse and Tekaia (1983) applied this technique to establish a transfer function for estimating paleo-environmental conditions from diatom assemblages.

TABLE 2. Hunting spider abundance data from Fig. 1: canonical coefficients and the intraset correlations of environmental variables with the first two axes of canonical correspondence analysis (CCA). The environmental variables were standardized to unit variance after log-transformation. For a description of variables, see Fig. 1 legend.

Axis variable	Canonical coefficients		Correlation coefficients	
	1	2	1	2
Water Content	−0.51	−0.41	−0.93	−0.08
Bare Sand	0.33	−0.10	0.73	0.06
Fallen Twigs	−0.14	0.37	−0.43	0.78
Cover Moss	0.05	−0.27	0.69	−0.30
Cover Herbs	−0.28	−0.15	−0.32	−0.78
Light Refl	0.27	−0.03	0.64	−0.59

TABLE 3. Hunting spider abundance data, with species (rows) and sites (columns) arranged in order of the scores for the first axis of canonical correspondence analysis (CCA). Site numbers correspond to those of Van der Aart and Smeenk-Enserink (1975: Table 4). The species abundance data have been transformed by taking square roots; the integer part is shown, a blank denoting absence of the species and 9 denoting >80 individuals captured. For this table, the range of each environmental variable was divided into 10 equal-sized classes (denoted by 0–9) after the data were transformed. Abbreviations and a description of the biological system are given in legend of Fig. 1.

	Site numbers																										
	15	19	20	16	17	18	2	8	21	5	6	14	4	7	13	3	1	9	12	25	11	10	28	23	22	27	24	26
Species																												
Arct lute										1	2	1	1	3	1	1												
Pard lugu	2	3	3	2	1	2	1	7	4	1		1	1	1	1	1				1				1				
Zora spin	1	1	1	2	1		3	1	1	4	5	5	5	4	4	1	2			2								
Pard nigr		1		1			3	1		9	5	3	5	9	7	4	3	1	1	2								
Pard pull							6	1	1	8	4	8	9	9	8	6	6	1	2		1							
Aulo albi							5	2		3	2	2	4	4	4	3	2			1	1							
Troc terr	5	4	4	5	4	5	8	5	4	9	7	9	9	9	9	8	7	1	3	4	2	1	1	1	1			1
Alop cune		1	1	1		1	1	3	1	4	2	1	2	2	6	4	3	1	3	1	1							
Pard mont							1	1	1	1	3	3	2	5	4	5	7	5	9	3	9	4	2	2	1	1	1	
Alop acce										1			1	1	1	3	5	1	4	3	3	1	3	4	2	5	3	1
Alop fabr															1	1				3	1	1	3	3	4	3	4	2
Arct peri																							1	2	1	2	2	4
Environmental variable																												
Water Content	9	7	8	8	9	8	8	6	7	8	9	8	6	8	9	6	5	5	5	3	4	4	0	0	1	0	2	0
Bare Sand	0	0	0	0	0	0	0	0	0	0	5	0	0	0	3	0	0	0	0	7	0	8	7	6	7	5	7	9
Cover Moss	1	3	1	1	1	0	2	2	1	0	5	4	5	1	1	5	7	9	8	2	9	7	8	9	9	8	9	4
Light Refl	1	0	0	0	2	2	3	1	0	5	1	2	6	5	7	8	8	7	8	5	8	8	8	9	8	8	9	9
Fallen Twigs	9	9	9	9	9	9	3	9	9	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cover Herbs	5	2	0	0	5	5	9	6	2	9	6	9	9	9	9	9	9	6	8	8	7	5	6	6	0	6	5	2

TESTS ON REAL DATA

Hunting spider data

The first data set, taken from Van der Aart and Smeenk-Enserink (1975), concerns the distributions of 12 species of hunting spiders (Fig. 1) in a Dutch dune area, in relation to environmental data. The species data are the numbers of individuals of each species caught in pitfall traps over a period of 60 wk. Twenty-six environmental variables were measured at 28 of the pitfall traps. This number of variables is too large to sort out their independent effects on community composition. Eighteen variables were removed on a priori grounds, and two more variables were removed because they were strongly correlated with one of the remaining six variables (Fig. 1). The species data were transformed by taking square roots to down-weight high abundances; the environmental data were transformed by taking logarithms, as in the original paper.

The ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) are very similar for these data. The first eigenvalue of CCA is only slightly lower than the first eigenvalue of DCA, and the species–environment correlations of the first three axes are all high (Table 1). Apparently the measured environmental variables are sufficient to explain the major variation among the spider catches. From Table 2 we infer that the first axis is a moisture gradient, on which the drier sites have a high percentage of bare sand or of moss. The correlations of the second axis show a contrast between sites with a high cover of leaves and twigs and sites with a well-developed herb and moss layer.

From the species and site points in the CCA ordination diagram (Fig. 1) we infer, for example, that *Arctosa perita* and *Alopecosa fabrilis* reached their maximum abundance in the six pitfall traps represented on the right-hand side of the diagram, that *Pardosa monticola* had maximum abundance in the pitfall traps shown in the middle, and that *Pardosa lugubris* was most abundant in the cluster of pitfall traps represented in the top-left of the diagram. These inferences from the diagram largely agree with the data (cf. Table 3).

The arrows for environmental variables in Fig. 1 account, in conjunction with the species points, for 87% of the variance in the weighted averages of the 12 spiders with respect to the six environmental variables, the sum of all eigenvalues being 0.85. For example, projecting the spider points on the axis of percentage bare sand shows that *Arctosa perita* and *Alopecosa fabrilis* were mainly found in habitats with the highest percentages of bare sand, *Alopecosa accentuata* and *Pardosa monticola* in habitats with intermediate bare sand percentages, and the species on the left-hand side of the diagram in habitats with the lowest percentages of bare sand. For *Ar. perita, Al. fabrilis, Al. accentuata,* and *P. monticola,* the same ranking applies with respect to the cover of the moss layer. The ranking is more or less the reverse with respect to soil water content. *Arctosa lutetiana, Pardosa pullata, Pardosa nigriceps, Aulonia albimana,* and *Pardosa monticola* occurred in

TABLE 4. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of dyke vegetation data (see Fig. 2): eigenvalues and species-environment correlation coefficients for the first four axes.

	Axis			
	1	2	3	4
	Eigenvalues			
DCA	0.34	0.25	0.22	0.19
CCA	0.20	0.13	0.12	0.07
DCCA	0.20	0.12	0.09	0.05
	Correlation coefficients			
DCA	0.52	0.40	0.58	0.22
CCA	0.82	0.81	0.80	0.77
DCCA	0.83	0.81	0.76	0.66

habitats with a well-developed herb layer. *Pardosa lugubris* occupies an aberrant position in the diagram, being the single spider species that occurred mainly in habitats with a high cover of fallen leaves and twigs (i.e., in woods). *Trochosa terricola, Zora spinimana,* and *Alopecosa cuneata* occupy an intermediate position between the woody and grassier sites. Van der Aart and Smeenk-Enserink (1975) gave a similar description, but the CCA ordination diagram tells the main story at a glance. The DCCA ordination diagram provided essentially the same information. The main structure in the data is also clear from Table 3, where species and sites are reordered according to their scores on the first CCA axis. The species data show a diagonal band; soil water content decreases along the first axis, whereas percentage bare sand, cover of moss, and light reflection increase along this axis.

Dyke vegetation

De Lange (1972) studied the occurrences of macrophytes in dykes in the Netherlands in relation to electrical conductivity, phosphate and chloride concentration in the water, and soil type (clay, peaty soil, sand). A total of 125 fresh water dykes (conductivity <126 mS/m) were selected, with in total 133 plant species. Conductivity data were transformed by taking logarithms, because of a skewed distribution, and chloride concentration was transformed to chloride ratio (the share of chloride ions in the electrical conductivity; G. Van Wirdum, *personal communication*). The nominal variable "soil type" (with three classes) was dealt with, as in multiple regression (see Montgomery and Peck 1982: chapter 6), by defining two dummy environmental variables "peat" and "sand." (The variable "peat" takes the value 1 when a dyke has soil type "peat" and the value 0 otherwise. The variable "sand" is defined analogously. A dyke in clay thus scores the value 0 on each of the two variables. The canonical coefficient of "peat" then measures the difference in expected site scores between peaty and clay soils. Other choices of dummy variables could have been used equivalently, e.g., "clay" and "sand.")

Table 4 shows that the environmental variables are poorly related to the first four species axes of DCA. But by choosing the axes in the light of the environmental variables, by applying CCA or DCCA, the species–environment correlations increase considerably. The interpretation of the axes is unambiguous (Table 5): the first axis is defined by conductivity and phosphate, the second by the chloride ratio and soil type; the soil types further differentiate on the third and fourth axes. CCA and DCCA do not differ much for this data set. On the CCA ordination diagram (Fig. 2) the dykes are not displayed because the diagram would have been too crowded; the undisplayed dykes all lie in the open center region of Fig. 2. Fig. 2 accounts for 56% of the variance and shows that the weighted averages of the species with respect to conductivity and phosphate result in similar rankings; this similarity cannot be explained by the correlation between these variables in the data set, because this correlation is only 0.44. In contrast, the ranking with respect to chloride ratio is different. The soil types are also represented by arrows (Fig. 2). Species whose distribution is the most restricted to peaty soils lie somewhat to the top-left-hand corner of the diagram. Analogously, species with a distribution mainly on clay tend to lie somewhat to the bottom-right-hand corner of the diagram.

The eigenvalues (Table 4) show that the extracted gradients are quite short (cf. Gauch and Stone 1979). The scores (optima) of most species therefore lie outside the center region where the sites lie, and the probability of occurrence of such species simply increases

TABLE 5. Dyke vegetation data from Fig. 2: canonical coefficients and intraset correlations, as in Table 2. For a description of variables see Fig. 2 legend.

Axis	Canonical coefficients				Correlation coefficients			
variable	1	2	3	4	1	2	3	4
EC	0.27	0.03	−0.02	0.10	0.83	0.17	−0.25	0.20
Phosphate	0.30	0.01	0.16	−0.15	0.86	−0.08	0.30	−0.21
Chloride Ratio	0.01	0.30	−0.09	0.09	0.14	0.86	−0.30	0.29
Clay	0	0	0	0	0.27	−0.21	−0.89	−0.31
Peat*	−0.09	0.44	0.78	−0.03	−0.38	0.49	0.72	−0.17
Sand*	0.01	−0.30	0.58	0.99	0.13	−0.40	0.40	0.78

* Not standardized to unit variance.

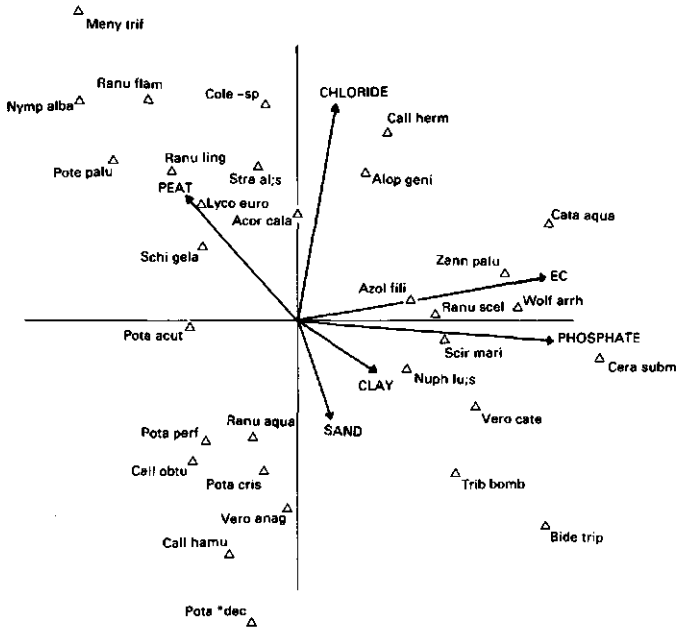


FIG. 2. Dyke vegetation data: CCA ordination diagram with plant species (△) and environmental variables (arrows); first axis is horizontal, second axis vertical. Species with positions near the center and some other species elsewhere are not shown because the diagram would have become too crowded. The plant species shown are: Acor cala = *Acorus calamus*, Alop geni = *Alopecurus geniculatus*, Azol fili = *Azolla filiculoides*, Bide trip = *Bidens tripartita*, Call hamu = *Callitriche hamulata*, Call herm = *Callitriche hermophroditica*, Call obtu = *Callitriche obtusangula*, Cata aqua = *Catabrosa aquatica*, Cera subm = *Ceratophyllum submersum*, Cole -sp = *Coleochaete* sp., Lyco euro = *Lycopus europaeus*, Meny trif = *Menyanthes trifoliata*, Nuph lu;s = *Nuphar lutea* (submerged form), Nymp alba = *Nymphaea alba*, Pota acut = *Potamogeton acutifolius*, Pota cris = *Potamogeton crispus*, Pota *dec = *Potamogeton decipiens*, Pota perf = *Potamogeton perfoliatus*, Pote palu = *Potentilla palustris*, Ranu aqua = *Ranunculus aquatilis s.l.*, Ranu flam = *Ranunculus flammula*, Ranu ling = *Ranunculus lingua*, Ranu scel = *Ranunculus sceleratus*, Schi gela = *Schizochlamys gelatinosa*, Scir mari = *Scirpus maritimus*, Stra al;s = *Stratiotes aloides* (submerged form), Trib bomb = *Tribonema bombycinum*, Vero anag = *Veronica anagallis-aquatica*, Vero cate = *Veronica catenata*, Wolf arrh = *Wolffia arrhiza*, Zann palu = *Zannichellia palustris*. The environmental variables are: EC = electrical conductivity, Phosphate = orthophosphate concentration, Chloride ratio = share of chloride ions in the electrical conductivity, and Clay, Peat, Sand (=type of soil surrounding the dyke).

or decreases monotonically along the gradients actually sampled, instead of being unimodal as required (see Theory). Condition C4 is clearly violated in this data set; nevertheless CCA worked well.

Algae along a pollution gradient

Fricke and Steubing (1984) sampled 25 sites in rivulets near the Ederstausee (Western Germany), recorded the abundances of 34 algae on a scale from 0 to 5, and measured seven environmental variables (Fig. 3), six of which (all but °D) were transformed by taking logarithms in our analysis because of skewed distributions. The first axis of DCA and that of CCA nearly coincided (Table 6), being a clear pollution gradient: positive correlations with ammonium, phosphate, biological oxygen demand (BOD5), and electrical conductivity, and a negative correlation with oxygen (Table 7). Although the ordination diagram of CCA (Fig.

TABLE 6. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of data on algae along a pollution gradient: eigenvalues and species–environment correlation coefficients for the first three axes.

	Axis		
	1	2	3
	Eigenvalues		
DCA	0.70	0.17	0.09
CCA	0.67	0.14	0.10
DCCA	0.67	0.08	0.05
	Correlation coefficients		
DCA	0.97	0.50	0.67
CCA	0.98	0.72	0.89
DCCA	0.98	0.80	0.79

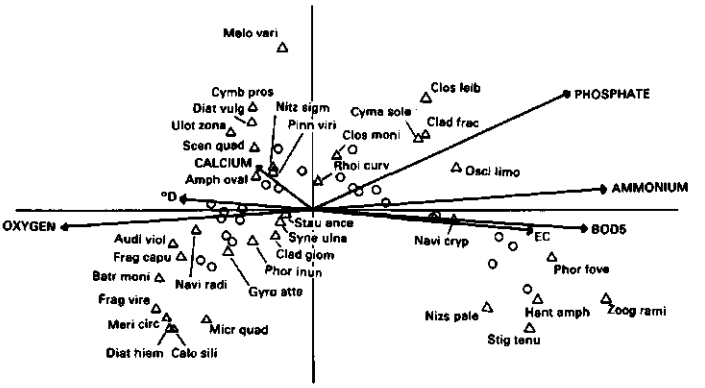


FIG. 3. Algae along a pollution gradient: CCA ordination diagram with algae (△), sites (○), and environmental variables (arrows); first axis is horizontal, second axis vertical. The algae are: Amph oval = *Amphora ovalis*, Audi viol = *Audionella violacea*, Batr moni = *Batrachospermum moniliforme*, Calo sili = *Caloneis silicula*, Clad frac = *Cladophora fracta*, Clad glom = *Cladophora glomerata*, Clos moni = *Closterium moniliferum*, Clos leib = *Closterium leibneinii*, Cyma sole = *Cymatopleura solea*, Cymb pros = *Cymbella prostata*, Diat hiem = *Diatoma hiemale mesodon*, Diat vulg = *Diatoma vulgare*, Frag capu = *Fragilaria capucina*, Frag vire = *Fragilaria virescens*, Gyro atte = *Gyrosigma attenuatum*, Hant amph = *Hantzschia amphioxis*, Melo vari = *Melosira varians*, Meri circ = *Meridion circulare*, Micr quad = *Microspora quadrata*, Navi cryp = *Navicula cryptocephala*, Navi radi = *Navicula radiosa*, Nizs pale = *Nitzschia palea*, Nitz sigm = *Nitzschia sigmoidea*, Osci limo = *Oscillatoria limosa*, Phor fove = *Phormidium foveolarum*, Phor inun = *Phormidium inundatum*, Pinn viri = *Pinnularia viridis*, Rhoi curv = *Rhoicophenia curvata*, Scen quad = *Scenedesmus quadricauda*, Stau ance = *Stauroneis anceps*, Stig tenu = *Stigeoclonium tenue*, Syne ulna = *Synedra ulna*, Ulot zona = *Ulotrix zonata*, Zoog rami = *Zoogloea ramigera*. The environmental variables are: Oxygen = oxygen concentration, BOD5 = biological oxygen demand, Ammonium = ammonium concentration, Phosphate = orthophosphate concentration, Calcium = calcium concentration, °D = German standard measure for the total concentration of calcium and magnesium, and EC = electrical conductivity.

3) explains most of the variance (73%), the diagram is unsatisfactory because of the arch effect (Gauch 1982*a*). The detrending in DCCA largely removes this effect (Fig. 4) and shows that the variation in species composition on the second axis is small ($\lambda_2 = 0.08$). This variation has surprisingly high correlation with the environmental variables (Table 6). The canonical coefficients of the second axis (Table 8) suggest that this minor component of the variation is related to the ratio of ammonium to phosphate.

In this example the interpretations of the CCA diagram and the DCCA diagram (Figs. 3 and 4) are not very different, but in more complicated data sets the difference can be large. As in regular ordination, detrending is a method to prevent the second axis from being obscured by dependence on the first.

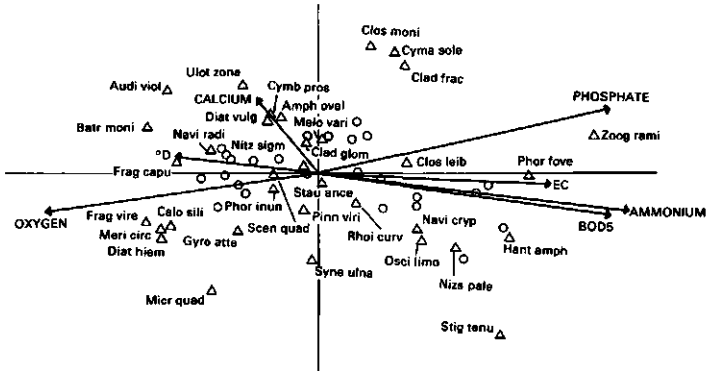


FIG. 4. Algae along a pollution gradient: DCCA ordination diagram. For an explanation of symbols see Fig. 3 legend.

TABLE 7. Data on algae along a pollution gradient, from Fig. 3: canonical coefficients and intraset correlations, as in Table 2. For a description of variables see Fig. 3 legend.

Axis variable	Canonical coefficients		Correlation coefficients	
	1	2	1	2
Oxygen	−0.47	0.20	−0.81	−0.06
BOD5	0.06	−0.11	0.88	−0.08
Ammonium	0.80	−0.07	0.94	0.09
Phosphate	−0.04	0.64	0.83	0.51
Calcium	−0.25	0.28	−0.19	0.19
°D	−0.07	−0.10	−0.44	0.05
EC	0.28	−0.27	0.71	−0.09

DISCUSSION

Canonical correspondence analysis provides an integrated description of species–environment relationships by assuming a response model that is common to all species, and the existence of a single set of underlying environmental gradients to which all the species respond. The same strong assumption is implicit in all ordination techniques. Canonical correspondence analysis has the advantage over other techniques in that it focuses on the relations between species and measured environmental variables and so provides an automated interpretation of the ordination axes.

Canonical correspondence analysis derives theoretical strength from its relation to maximum likelihood Gaussian canonical ordination under conditions C1–C4 and furthermore seems extremely robust in practice when these assumptions do not hold. The vital assumption is that the response surfaces of the species are unimodal, the Gaussian (bell-shaped) response model being the example for which the method's performance is particularly good. For the simpler case where species–environment relationships are monotone, the results can still be expected to be adequate in a qualitative sense (see Tests on Real Data: Dyke Vegetation). The method would not work if a large number of species were distributed in a more complex way, e.g., bimodally; the restriction to a unimodal model is necessary for practical solubility, but as Hill (1977) points out, a good choice of environmental variable should minimize the number of species with more complex distributions. Some care, however, is required with the interpretation of the ordination diagram when the additional assumptions (C1–C4) do not hold. Species in the center of the ordination diagram may then have their optima there, but may alternatively be unrelated to the axes. Which possibility is most likely can be decided upon by tabular rearrangement of the species data with respect to each axis, as is done in Table 3 for the first axis. Further work still needs to be done on the statistical significance of eigenvalues, species–environment correlations, and canonical coefficients.

As in correspondence analysis, any kind of transformation of the species abundance data may influence the results. When the abundance data have a very skewed distribution, it is recommended to transform them by taking square roots or logarithms. In this way we prevent a few high abundance values from unduly influencing the analysis. Because the compound environmental gradients constructed by canonical correspondence analysis are required to be linear combinations of environmental variables, nonlinear transformation of environmental variables can also be considered if there is some reason to do so. Prior knowledge about the possible impact of the environmental variables on community composition may suggest particular nonlinear transformations and particular nonlinear combinations, i.e., environmental scalars in the sense of Loucks (1962) and Austin et al. (1984). The use of environmental scalars can also circumvent the multicollinearity problem described in Theory: Canonical Coefficients. In contrast to the ordination techniques in common use, canonical correspondence analysis allows one to incorporate existing knowledge about species–environment relationships into the analysis and thus potentially is a more powerful tool to advance this knowledge.

Canonical correspondence analysis can be used fruitfully in combination with (detrended) correspondence analysis, as in the examples described. When the solutions do not differ much, we infer that the measured environmental variables can account for the main variation in the species data. When the solutions do differ, we infer either that the environmental variables account for less conspicuous directions of variation in the species data (when the correlations between species and environment axes are high) or that they cannot account for any of the variation (when the correlations are small). These possibilities considerably extend the analytical power of ordination by allowing comparison of results from indirect and direct gradient analysis techniques that have a common theoretical basis. Direct and indirect gradient analysis can also be combined in a single analysis to answer such questions as "Does the known environmental variation account for all the community variation, or is there a substantial residual variation?" Suppose we believe two environmental variables govern the species composition in a

TABLE 8. Data on algae along a pollution gradient, from Fig. 3: canonical coefficients and intraset correlations in DCCA. For a description of variables see Fig. 3 legend.

Axis variable	Canonical coefficients		Correlation coefficients	
	1	2	1	2
Oxygen	−0.37	0.05	−0.81	0.04
BOD5	0.07	0.21	0.88	−0.40
Ammonium	0.65	−0.60	0.95	−0.47
Phosphate	0.10	0.50	0.86	0.06
Calcium	−0.22	0.23	−0.19	0.37
°D	−0.06	−0.07	−0.43	0.18
EC	0.22	−0.17	0.70	−0.22

region. We may choose two ordination axes in the light of these variables, then extract further axes as in detrended correspondence analysis by reciprocal averaging and detrending with respect to all previous axes. The lengths of the extra axes measure the residual variation. The program CANOCO (Ter Braak 1985*a*) has an option to do such combined analyses. The same option allows analysis of nested data (subplots within plots, e.g., yearly vegetation records from several permanent plots, or bird records from woodlots in several regions). The first axes can be chosen to represent variation between plots, so that the further axes represent variation between subplots. Swaine and Greig-Smith (1980) used a variant of principal components analysis in this way to obtain an ordination of within-plot vegetation change in permanent plots; canonical correspondence analysis could be used for the same purpose but is not hampered by the unwarranted assumption of a linear relationship between species abundance and environment.

ACKNOWLEDGMENTS

I would like to thank Dr. I. C. Prentice for inspiring discussions. The comments of the reviewers helped to clarify a number of practical aspects of the technique.

LITERATURE CITED

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

Bartlein, P. J., I. C. Prentice, and T. Webb, III. 1986. Climatic response surfaces from pollen data for some eastern North American taxa. Journal of Biogeography **13**:35–57.

Carleton, T. J. 1984. Residual ordination analysis: a method for exploring vegetation-environment relationships. Ecology **65**:469–477.

De Lange, L. 1972. An ecological study of ditch vegetation in the Netherlands. Dissertation. University of Amsterdam, Amsterdam, The Netherlands.

De Leeuw, J. 1984. The GIFI system of nonlinear multivariate analysis. Pages 415–424 *in* E. Diday, M. Jambu, L. Lebart, J. Pagès, and R. Tomassone, editors. Data analysis and informatics III. North-Holland Publishing, Amsterdam, The Netherlands.

Fricke, G., and L. Steubing. 1984. Die Verbreitung von Makrophyten und Mikrophyten in Hartwasser-Zuflüsse des Ederstausees. Archiv für Hydrobiologie **101**:361–372.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**:453–467.

———. 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. Pages 147–173 *in* V. Barnett, editor. Interpreting multivariate data. J. Wiley and Sons, New York, New York, USA.

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

Gauch, H. G. 1982*a*. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, England.

———. 1982*b*. Noise reduction by eigenvector ordinations. Ecology **63**:1643–1649.

Gauch, H. G., G. B. Chase, and R. H. Whittaker. 1974. Ordination of vegetation samples by Gaussian species distributions. Ecology **55**:1382–1390.

Gauch, H. G., and E. L. Stone. 1979. Vegetation and soil pattern in a mesophytic forest at Ithaca, New York. American Midland Naturalist **102**:332–345.

Gauch, H. G., and T. R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio **33**: 17–22

Gittins, R. 1985. Canonical analysis. A review with applications in ecology. Springer-Verlag, Berlin, Germany.

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London, England.

Hill, M. O. 1977. Use of simple discriminant functions to classify quantitative phytosociological data. Pages 597–613 *in* E. Diday, L. Lebart, J. P. Pagès, and R. Tomassone, editors. Data analysis and informatics, I. Institut de Recherche d'Informatique et d'Automatique, Le Chesnay Cedex, France.

———. 1979. DECORANA: A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Section of Ecology and Systematics, Cornell University, Ithaca, New York, USA.

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

Loucks, O. L. 1962. Ordinating forest communities by means of environmental scalars and phytosociological indices. Ecological Monographs **32**:137–166.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. Multivariate analysis. Academic Press, London, England.

Montgomery, D. C., and E. A. Peck. 1982. Introduction to linear regession analysis. J. Wiley and Sons, New York, New York, USA.

Swaine, M. D., and P. Greig-Smith. 1980. An application of principal components analysis to vegetation change in permanent plots. Journal of Ecology **68**:33–41.

Ter Braak, C. J. F. 1983. Principal components biplots and alpha and beta diversity. Ecology **64**:454–462.

———. 1985*a*. CANOCO: A FORTRAN program for canonical correspondence analysis and detrended correspondence analysis. IWIS-TNO, Wageningen, The Netherlands.

———. 1985*b*. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics **41**:859–873.

Ter Braak, C. J. F., and C. W. N. Looman. In press 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio **65**:3–11.

Van der Aart, P. J. M., and N. Smeek-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology **25**:1–45.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews of the Cambridge Philosophical Society **42**:207–264.

Whittaker, R. H., S. A. Levin, and R. B. Root. 1973. Niche, habitat and ecotope. American Naturalist **107**:321–338.

APPENDIX

Here canonical correspondence analysis is shown to be (1) an approximation to Gaussian canonical ordination, (2) an eigenvector technique akin to canonical correlation analysis, and (3) a method for weighted least squares approximation of weighted averages of species with respect to environmental variables. For an explanation of the notation, see Theory.

The model of Gaussian canonical ordination is Eq. 1 in conjunction with Eq. 2 (see Theory). It is assumed that the species data are Poisson-distributed counts with $E(y_{ik}) = \mu_{ik}$ and that the species tolerances are all equal to 1. Then the maximum likelihood equations for u_k and b_j are, after some rearrangement, respectively:

$$u_k = \sum_i y_{ik}x_i/y_{+k} - \left[\sum_i (x_i - u_k)\mu_{ik}/y_{+k}\right] \quad (A.1)$$

$$\sum_i z_{ij}\left[\sum_k y_{ik}(x_i - u_k)\right] = \sum_i \left[\sum_k (x_i - u_k)\mu_{ik}\right]z_{ij}. \quad (A.2)$$

Under conditions C1–C4 and Eq. 7, we may use the approximations

$$\sum_k (x_i - u_k)\mu_{ik} \approx 0 \quad (A.3)$$

$$\sum_i (x_i - u_k)\mu_{ik} \approx -\lambda^* u_k y_{+k} \quad (A.4)$$

because μ_{ik} is symmetric about x_i and about u_k; the proportionality constant λ^* comes in because the species' curves are the more truncated the more their optima lie towards or beyond the edge of the sampling interval (Ter Braak 1985*b*). The transition formulae Eqs. 3–6 now follow from Eqs. A.1 and A.2 by using Approximations A.3 and A.4 and the equation $\lambda = 1 - \lambda^*$.

Starting from Eq. 5 we substitute for x^* (Eq. 4), u_k (Eq. 3), and finally x_i (Eq. 6) and obtain

$$(S_{21}S_{11}^{-1}S_{12} - \lambda S_{22})b = 0, \quad (A.5)$$

where $S_{21} = Z'Y$, $S_{12} = Y'Z$, $S_{11} = \text{diag}(y_{+1}, y_{+2}, \ldots, y_{+m})$, $S_{22} = Z'RZ$ and $Y = \{y_{ik}\}$. Similarly, successive substitutions in Eq. 3 lead to

$$(S_{12}S_{22}^{-1}S_{21} - \lambda S_{11})u = 0, \quad (A.6)$$

where $u = (u_1, \ldots, u_m)'$. Apart from the particular definitions of the matrices in Eqs. A.5 and A.6, these equations are the eigenvector equations of canonical correlation analysis, and the eigenvalue λ lies between 0 and 1 (Gittins 1985). The eigenvectors are all uncorrelated; using subscripts r and s for different axes we obtain that $u_r'S_{11}u_s = 0$, $b_r'S_{22}b_s = 0$ and $x_r'Rx_s = 0$. Algorithms based on Eq. A.5 or Eq. A.6 will in general be more efficient than the algorithm developed in Theory.

The first axis of canonical correspondence analysis does not maximize the species–environment correlation, i.e., the correlation between x and x^*. I have also developed an eigenvector technique that maximizes the species–environment correlation. This technique requires that the number of species is smaller than the number of sites. This requirement is often a nuisance in ecological research. As we have seen, the rationale for canonical correspondence analysis is different: it is, under conditions C1–C4, almost a maximum likelihood technique.

The weighted averages of the species with respect to the environmental variables in Eq. 8 are, in matrix notation, $W = S_{11}^{-1}Y'Z = S_{11}^{-1}S_{12}$, where $W = \{z_{kj}\}$. We want a least squares approximation of W in an ordination diagram. However, when a species total is low, the weighted average is imprecise (cf. Ter Braak and Looman 1986), so that it is not worthwhile to approximate that species' weighted averages very accurately in the diagram. This consideration suggests giving the species weights that are proportional to the species totals contained in S_{11}. The result would still depend on the scale of measurement of the environmental variables. To make the method scale-invariant we use S_{22}^{-1} as weights for the environmental variables. The desired weighted least squares approximation of W follows now from the singular value decomposition (see for example Greenacre 1984: Appendix A).

$$S_{11}^{1/2}WS_{22}^{-1/2} = S_{11}^{-1/2}S_{12}S_{22}^{-1/2} = P\Lambda^{1/2}Q', \quad (A.7)$$

where P and Q are orthonormal $m \times q$ and $q \times q$ matrices (respectively) and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q)$. For convenience of notation it is assumed here that $q \leq m$. This singular value decomposition is just another way to solve Eqs. A.5 and A.6 (see Mardia et al. 1979: chapter 10). With Hill's (1979) scaling of site and species scores, namely

$$\sum_{i,k} y_{ik}(x_i - u_k)^2 = y_{++}, \quad (A.8)$$

the coordinates of the species points are the first two columns of the matrix

$$U = y_{++}^{1/2}S_{11}^{-1/2}P(I - \Lambda)^{-1/2}, \quad (A.9)$$

and the coordinates of the points for the environmental variables are the first two columns of the matrix

$$B_e = y_{++}^{-1/2}S_{22}^{1/2}Q(I - \Lambda)^{1/2}\Lambda^{1/2} = y_{++}^{-1}Z'RX(I - \Lambda), \quad (A.10)$$

where the second equality follows after some algebra, with X the matrix whose s^{th} column is x_s. In this scaling $U'S_{11}U = y_{++}(I - \Lambda)^{-1}$ and $X'RX = y_{++}\Lambda(I - \Lambda)^{-1}$. It is easy to verify using Eqs. A.7, A.9, and A.10 that $W = UB_e'$. Therefore the points for species and environmental variables form a biplot (Gabriel 1971) in the sense that inner products approximate the elements of the matrix W, leading to a two-dimensional approximation W_2, say. A measure of goodness of fit is $(\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$, which is equal to $\text{trace}(S_{11}W_2S_{22}^{-1}W_2')/\text{trace}(S_{11}WS_{22}^{-1}W')$ and is, loosely speaking, the percentage variance in the weighted averages accounted for by the biplot. When the environmental variables are scaled to zero mean and unit variance (using y_{i+} as site weights), we obtain from Eq. A.10 that the coordinate of the point for environmental variable j on axis s must be $[\lambda_s(1 - \lambda_s)]^{1/2}$ times the correlation coefficient of the environmental variable with the site scores x_s. In detrended canonical correspondence analysis the coordinates of the points for the environmental variables are obtained from a multivariate regression of W on the first two columns of U, U_2 say:

$$B_e = W'S_{11}U_2(U_2'S_{11}U_2)^{-1} = Z'RX(U_2'S_{11}U_2)^{-1}, \quad (A.11)$$

which reduces to Eq. A.10 in canonical correspondence analysis.

Vegetatio 69: 69–77, 1987

The analysis of vegetation-environment relationships by canonical correspondence analysis*

Cajo J. F. Ter Braak[1,2]**
[1]*TNO Institute of Applied Computer Science, Statistics Department Wageningen, P.O. Box 100, 6700 AC Wageningen, The Netherlands, and* [2]*Research Institute for Nature Management, P.O. Box 46, 3956 ZR Leersum, The Netherlands*

Keywords: Canonical correspondence analysis, Correspondence analysis, Direct gradient analysis, Ordination, Species-environment relation, Trend surface analysis, Weighted averaging

Abstract

Canonical correspondence analysis (CCA) is introduced as a multivariate extension of *weighted averaging ordination*, which is a simple method for arranging species along environmental variables. CCA constructs those linear combinations of environmental variables, along which the distributions of the species are maximally separated. The eigenvalues produced by CCA measure this separation.

As its name suggests, CCA is also a correspondence analysis technique, but one in which the ordination axes are constrained to be linear combinations of environmental variables. The ordination diagram generated by CCA visualizes not only a pattern of community variation (as in standard ordination) but also the main features of the distributions of species along the environmental variables. Applications demonstrate that CCA can be used both for detecting species-environment relations, and for investigating specific questions about the response of species to environmental variables. Questions in community ecology that have typically been studied by 'indirect' gradient analysis (i.e. ordination followed by external interpretation of the axes) can now be answered more directly by CCA.

Introduction

Direct gradient analysis relates species presence or abundance to environmental variables on the basis of species and environment data from the same set of sample plots (Gauch, 1982). The simplest methods of direct gradient analysis involve plotting each species' abundance values against values of an environmental variable, or drawing isopleths for each species in a space of two environmental variables (Whittaker, 1967). With these simple methods one can easily visualize the relation between many species and one or two environmental variables.

Plant species experience the conditions provided by many environmental variables; therefore one might wish to analyse their joint effects. Multiple regression can be used for that purpose. However, despite some successful applications, e.g., Yarranton (1970), Austin (1971) and Forsythe & Loucks (1972), ordinary multiple regression has never become popular in vegetation science. Reasons for this include: (1) Each species requires separate analysis, so regression analysis may require an unreasonable amount of effort. (2) Vegetation data are often qualitative, or when they are quantitative the data contain many zero values for the plots at which a species is absent. In neither case do the data satisfy the assumption of a normal error distribution that is implicit in ordinary multiple regression. (3) Relationships between species and environmental variables are generally non-linear. Species abundance is often a single-peaked (bell-

* Nomenclature follows Heukels-Van der Meijden (1983). Flora van Nederland, 20th ed.

** I would like to thank the authors of the example data sets for permission to use their data, Drs M. O. Hill and H. G. Gauch for permission to use the code of the program DECORANA, and Drs I. C. Prentice, L. C. A. Corsten, P. F. M. Verdonschot, P. W. Goedhart and P. F. G. Vereijken for comments on the manuscript.

shaped) function of the environmental variables. (4) Environmental variables are often highly correlated, and so it can be impossible to separate their independent effects. Generalized Linear Modelling (Austin *et al.*, 1984; Ter Braak & Looman, 1986) provides a solution for (2) and (3), but (1) and (4) remain. Whenever the number of influential environmental variables is greater than two or three, it becomes difficult to put results for several species together so as to obtain an overall graphical summary of species-environment relationships.

A simple method is therefore needed to analyze and visualize the relationships between many species and many environmental variables. Canonical correspondence analysis (CCA) is designed to fulfil this need. CCA is an eigenvector ordination technique that also produces a multivariate direct gradient analysis (Ter Braak, 1986). CCA aims to visualize (1) a pattern of community variation, as in standard ordination, and also (2) the main features of species' distributions along the environmental variables.

Ter Braak (1986) derived CCA as a heuristic approximation to the statistically more rigorous (but computationally fraught) technique of Gaussian canonical ordination, and also showed CCA's relation to correspondence analysis (CA), alias reciprocal averaging (Hill, 1973). In this paper a simple, alternative derivation of CCA is given starting from the method of weighted averaging (WA).

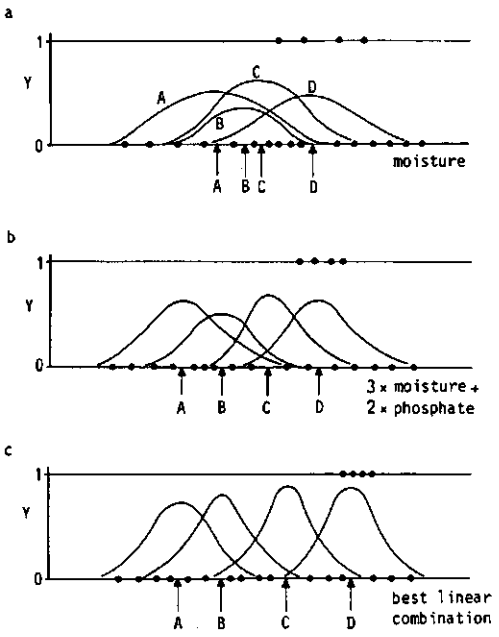


Fig. 1. Artificial example of single peaked response curves of four species (A–D) with respect to standardized environmental variables showing different degrees of separation of the species curves: (a) moisture; (b) a linear combination of moisture and phosphate, chosen apriori; (c) the best linear combination of environmental variables, chosen by CCA. Sites are shown by dots at $y = 1$ if species D is present and at $y = 0$ if species D is absent.

Theory

From weighted averaging to canonical correspondence analysis

Figure 1a shows an artificial example of single-peaked response curves for four species along an environmental variable (e.g. moisture). Species A occurs in drier conditions than species D. Fig. 1a shows presence-absence data for species D: the species is present at four of the sites.

How well does moisture explain the species' data? The fit could be formally measured by the deviance between the data and the curves, as in logistic regression (Ter Braak & Looman, 1986), but this idea will not be pursued here. Instead, a simple alternative based on the method of weighted averaging (WA) is used.

For each species a score can be calculated by taking the weighted average of the moisture values of the plots. For abundance data, this score is calculated as

$$u_k = \sum_{i=1}^{n} y_{ik} x_i / y_{+k} \qquad (1)$$

where u_k is the weighted average of the k-th (out of m) species, x_i is the (moisture) value of the i-th (out of n) site and y_{ik} is the abundance of species k at site i, and y_{+k} is the total abundance of species k. For presence-absence data the weighted average is simply the average of the moisture values of the plots in which the species is present. The weighted average

gives a first indication of where the species occurs along the moisture gradient (see the arrows in Fig. 1a). As a measure of how well moisture explains the species data, the *dispersion of the weighted averages* is used (see below). If the dispersion is large, moisture neatly separates the species curves, and moisture explains the species data well. If the dispersion is small, then moisture explains less.

To compare the explanatory power of different environmental variables, each environmental variable must first be standardized to mean 0 and variance 1. For technical reasons, weighted means and variances are used; each environmental variable is standardized such that

$$\sum_{i=1}^{n} y_{i+}x_i = 0 \text{ and } \sum_{i=1}^{n} y_{i+}x_i^2/y_{++} = 1 \qquad (2)$$

where y_{i+} is the total abundance at site i and y_{++} the overall total. The dispersion can now be written as

$$\delta = \sum_{k=1}^{m} y_{+k}u_k^2/y_{++} \qquad (3)$$

By calculating the dispersion for each environmental variable one can select the 'best' variable.

Now suppose that moisture is the 'best' single variable in the artificial example. However, someone might suggest a better variable, that is a combination of two others (see, e.g., Loucks, 1962). In the artificial example a combination of moisture and phosphate, namely (3 × moisture + 2 × phosphate), is shown to give a larger dispersion than moisture alone (Fig. 1b); and consequently the curves in Fig. 1b are narrower, and the presences of species D are closer together, than in Fig. 1a. So it can be worthwhile to consider not only the environmental variables separately but also all possible linear combinations of them, i.e. all 'weighted sums' of the form

$$x_i = b_1z_{i1} + b_2z_{i2} + \ldots + b_pz_{ip} \qquad (4)$$

where z_{ij} is the value of the j-th (out of p) environmental variable at site i, and b_j is the weight (not necessarily positive) belonging to that variable; x_i is the value of a compound environmental variable at site i. (It is assumed in equation (4) that each environmental variable is centered to a weighted mean of 0. Although not essential, it will also be convenient to standardize the environmental variables according to equation (2) so as to make the weights (b_j) comparable.)

CCA turns out to be *the technique that selects the linear combination of environmental variables that maximizes the dispersion of the species scores.* In other words, CCA chooses the optimal weights (b_j) for the environmental variables. In the Appendix it is shown that these optimal weights are the solution of the same eigenvalue equation as the one derived by another rationale in Ter Braak (1986), and that the first eigenvalue of CCA is actually equal to the (maximized) dispersion of species scores along the first CCA axis.

The second and further CCA axes also select linear combinations of environmental variables that maximize the dispersion of the species scores, but subject to the constraint of being uncorrelated with previous CCA axes. In principle, as many axes can be extracted as there are environmental variables.

From correspondence analysis to canonical correspondence analysis

CA also maximizes the dispersion δ in equation (3). But it does so irrespective of any environmental variable; that is, CA assigns scores (x_i) to sites such that the dispersion is absolutely maximum, the scores being standardized as in equation (2) (Nishisato, 1980). CCA is therefore 'restricted correspondence analysis' in the sense that the site scores are restricted to be linear combinations of supplied environmental variables.

A familiar algorithm to carry out CA is the reciprocal averaging algorithm (Hill, 1973). In Ter Braak (1986) this algorithm is extended with an additional multiple regression step so as to obtain the CCA solution. In each iteration cycle the trial site scores are regressed on the environmental variables (using y_{i+}/y_{++} as site weights) and the new trial scores are the fitted values of this regression. The FORTRAN program CANOCO (Ter Braak, 1985b) to carry out CCA is in fact just an extension of Hill's (1979) program DECORANA.*

CCA is restricted correspondence analysis, but the restrictions become less strict, the more environmental variables are included in the analysis. If $p \geq n-1$, then there are actually no restrictions any more; CCA is then simply CA. The arch effect may therefore crop up in CCA as it does in CA (Gauch, 1982). The method of detrending (Hill & Gauch, 1980) can be used to remove the arch and is available in the computer program

*The program is available from the author at cost price.

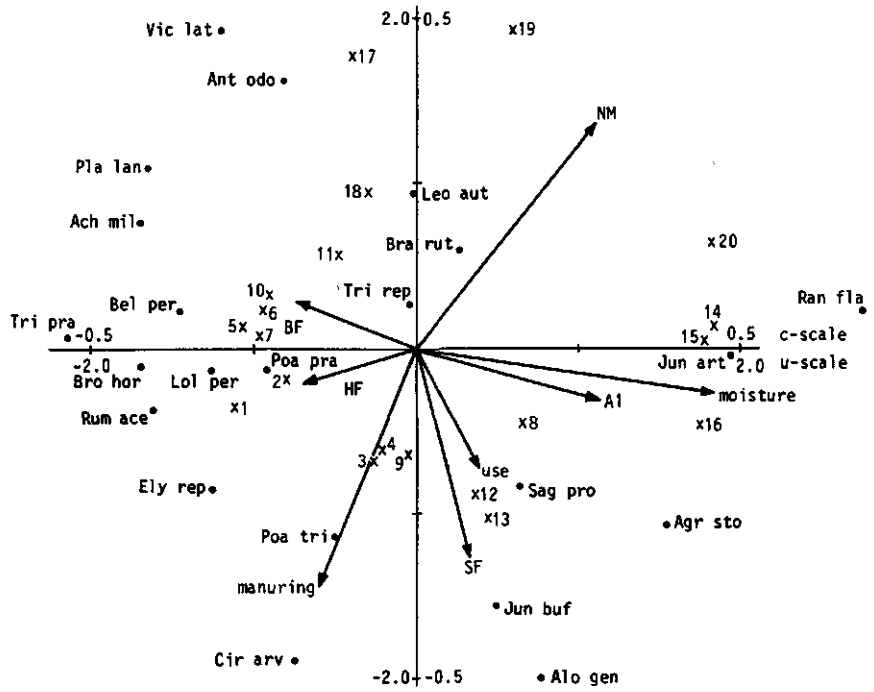


Fig. 2. Dune meadow data: CCA ordination diagram with relevés (x), plant species (•) and environmental variables (arrow); first axis horizontally, second axis vertically. For relevé numbers see Table 1. Abbreviations are given as *underlining* in full names in Table 1. The c-scale applies to the environmental arrows, the u-scale to species and sites points. Eight infrequent species are not shown because they lie outside the range of this diagram.

CANOCO (Ter Braak, 1985b). But in CCA the arch can be removed more elegantly by dropping superfluous environmental variables. Variables that are highly correlated with the 'arched' axis (often the second axis) are most likely to be superfluous.

CA is very susceptible to species-poor sites containing rare species in that it places such aberrant sites (and the rare species occurring there) at extreme ends of the first ordination axes (Gauch, 1982), relegating the major vegetation trends in the data to later axes. CCA does not show this 'fault' of CA, provided the sites that are aberrant in species composition are not so aberrant in terms of the environmental variables.

Ordination diagram

The ordination diagram of CCA displays sites, species and environmental variables (Fig. 2). The site and species points have the same interpretation as in CA. They display variation in species composition over the sites. The environmental variables are represented by arrows (Fig. 2). Loosely speaking, the arrow for an environmental variable points in the direction of maximum change of that environmental variable across the diagram, and its length is proportional to the rate of change in this direction. Environmental variables with long arrows are more strongly correlated with the ordination axes than those with short arrows, and so more closely related to the pattern of community variation shown in the ordination diagram.

Further insight into the ordination diagram of CCA can be obtained from yet another characterization of CCA. From equations (A.5) en (A.6) of the Appendix it follows that CCA is a

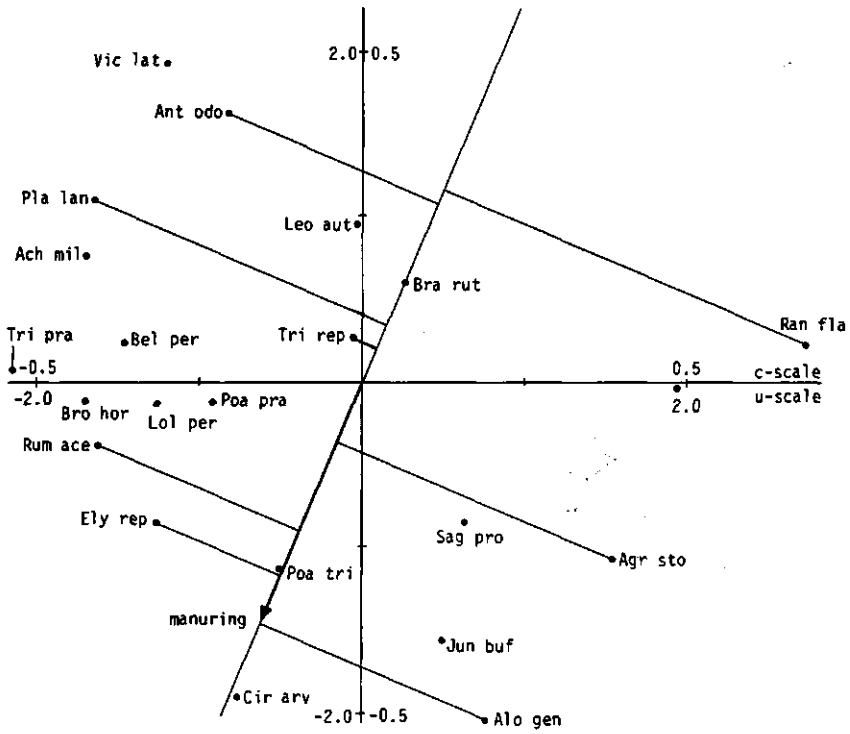


Fig. 3. Inferred ranking of the species along the variable quantity of manuring based on the biplot interpretation of Fig. 2. For explanation see the Ordination diagram section.

weighted principal components analysis applied to a matrix of species by environmental variables, the (k, j)-th element of which is the weighted average of species k with respect to environmental variable j (it is here assumed that each environmental variable is reduced to zero mean). CCA is a weighted analysis in the sense that species are given weights proportional to their total abundance (y_{+k}) and the environmental variables are weighted inversely with their covariance matrix. The intuitive advantage of the implicit species weights is that a weighted average for a species is imprecise when its total is low (Ter Braak & Looman, 1986) and is thus not worth much attention. Environmental variables are given equal weight irrespective of their variance or unit of measurement. (This type of weighting is also implicit in discriminant analysis (see Campbell & Atchley, 1981) and makes the analysis invariant to nonsingular linear transformations of the environmental variables). This characterization of CCA shows that the joint plot of species and environmental variables in the CCA ordination diagram can be interpreted similarly to a principal components biplot (Gabriel, 1971; Ter Braak, 1983), allowing inference of the approximate values of the weighted averages of each of the species with respect to each of the environmental variables.

The most convenient rule for quantitative interpretation of the CCA biplot (Ter Braak, 1986) is therefore as follows: each arrow representing an environmental variable determines a direction or 'axis' in the diagram; the species points can be projected on to this axis (see Fig. 3). The order of the projection points corresponds approximately to the ranking of the weighted averages of the species with respect to that environmental variable. The weighted average indicates the position of a species' distribution along an environmental variable (Fig. 1), and thus the projection point of a species also indicates this position, although approximately.

Table 1. Dune meadow data: data table with species (rows) and relevés (columns of one digit width) arranged in order of their scores on the first axis of CCA. Relevé numbers are printed vertically. The abundance values, as used in the analysis, are on a 1 – 9 scale to replace the Braun-Blanquet symbols r, +, 1, 2m, 2a, 2b, 3, 4, 5. Thickness of the A1 horizon is divided into ten equal-sized classes (denoted 0 – 9). The values 1, 2 and 3 for agricultural use refer to hayfield, haypasture and pasture, respectively. For further explanation of the environmental variables see text.

	relevés 1 111 11 11112 51670217834923894560
Trifolium pratense	2-52----------
Achillea millefolium	212243-2----------
Bromus hordeaceus	2--244----3----------
Plantago lanceolata	5-553-323----------
Rumex acetosa	5-63-------22-------
Bellis perennis	2---23--222----------
Elymus repens	44---4---446 ------
Lolium perenne	2766657-2652--4-----
Vicia lathyroides	----1-2-1----------
Poa pratensis	243444413544-24-----
Anthoxanthum odoratum	4-324--4-------4----
Cirsium arvense	----------2---------
Poa trivialis	624547---655494---2-
Trifolium repens	2-52653-2213322261--
Leontodon autumnalis	3-3335525222223622-2
Brachythecium rutabulum	2-622-4-62224-23-444
Juncus bufonius	---2--------443------
Sagina procumbens	------2---524223----
Alopecurus geniculatus	-----2---723855---4-
Hypochaeris radicata	------22-------5----
Aira praecox	-------2-------3----
Salix repens	--------2------3---5
Agrostis stolonifera	---------483454-4475
Juncus articulatus	----------4-4--334
Chenopodium album	----------1------
Empetrum nigrum	----------2----
Ranunculus flammula	----------22-2224
Eleocharis palustris	----------4-4584
Calliergonella cuspidata	----------4-33
Potentilla palustris	----------22--
thickness *A1*	401000012111331179 30
moisture	11112112122445555555
quantity of *manuring*	24231210044123311131
agricultural *use*	12231231122122313231
Standard Farming	01000000011011000010
Bio-dynamic Farming	00001110000000000000
Hobby Farming	10110000000100100000
Nature Management	00000001100000011101

The ordination diagrams of CCA and CA also share some of the shortcomings of WA (Ter Braak & Looman, 1986). The most important practical shortcoming is that species that are unrelated to the ordination axes tend to be placed in the center of the ordination diagram and are not distinguished from species that have true optima there. This problem can easily be circumvented by looking at a species-by-site data table in which species and sites are arranged in order of their scores on one of the ordination axes (cf. Table 1).

The CCA ordination diagram is not in any way hampered by high correlations between species, or between environmental variables.

Applications

Exploratory use of the ordination diagram

Batterink and Wijffels (report) studied the possible relation between vegetation and management of dune meadows on the island Terschelling (The Netherlands).

A subset of their data is analysed here to illustrate the ordination diagram of CCA. This subset consists of 20 standard plots recorded in 1982, and 30 plant species (Table 1).

Five environmental variables were recorded: (1) thickness of the A1 horizon, measured in millimeters; (2) moisture content of the soil, scored on a five-point scale in a semi-objective manner; (3) quantity of manuring, scored on a five-point scale on the basis of a questionnaire sent to the owners of the meadows; (4) agricultural use, a nominal variable with three classes – hayfield, haypasture and pasture; and (5) type of management, a nominal variable with four classes – standard farming, biodynamic farming, hobby farming and nature management.

CCA cannot directly cope with ordinal variables, like moisture and manuring here. Ordinal variables must either be treated as if they were quantitative, or as nominal variables. Here they were treated as quantitative. Nominal variables, like type of management, must be transformed to dummy variables as shown in Table 1. For instance, the dummy variable 'nature management' indicates which meadows received that type of management. Agricultural use was however treated as a quantitative variable (Table 1), because haypasture was considered as an intermediate between hayfield and pasture.

Two values were missing in the environment data. CCA cannot cope with missing values, so relevés with missing values in the environment data must be deleted. To avoid deletion, missing values were replaced here by the mean of the corresponding variable over the remaining plots.

Despite the crude measurement of the environmental variables, they nicely explain the major variation in the vegetation. The first two eigenvalues of CCA (λ_1 = 0.46 and λ_2 = 0.29) were not much reduced in comparison with those of standard CA (0.54 and 0.40), and the two-dimensional configurations of species and sites in the ordination diagrams

looked similar. The most conspicuous difference was that relevés 17 and 19 were outliers in CA and not so much in CCA (Fig. 2).

The configurations of species and sites in CCA (Fig. 2) must be interpreted as in CA (Ter Braak, 1985a). For instance, from Fig. 2 *Sagina procumbens* can be expected to have its maximum abundance in the relevés close to its point in Fig. 2 (relevés 8, 12 and 13) and to be absent in relevés far from that point.

Figure 2 accounts for 65% of the variance in the weighted averages of the species with respect to each of the environmental variables. This percentage is calculated as in principal components analysis by taking $100 \times (\lambda_1 + \lambda_2)/(\lambda_1 + \ldots + \lambda_p)$. It can be deduced from Fig. 2, for example, that *Cirsium arvense, Alopecurus geniculatus* and *Elymus repens* mainly occur in the highly manured meadows, *Agrostis stolonifera* and *Trifolium repens* in intermediately manured meadows, and *Ranunculus flammula* and *Anthoxanthum odoratum* in little manured meadows (see Fig. 3). The other arrows can be interpreted similarly. From Fig. 2 it can thus be seen at once which species occur mainly under wetter conditions (those on the right hand side of the diagram) and which ones prefer drier conditions (those on the left hand side of the diagram).

Multi-species trend surface analysis

CCA can be used to detect spatial gradients in vegetation data. A spatial gradient can be specified by a linear combination of two orthogonal coordinates, say, the x-coordinate (z_1) and y-coordinate (z_2) of the relevés, i.e. by $b_1z_1 + b_2z_2$. The strongest spatial gradient in vegetation data might be defined as that combination of z_1 and z_2 that maximally separates the spatial distributions of the species, and can thus be estimated by taking the x- and y-coordinates as environmental variables in a CCA. Put another way, CCA searches for the direction of the strongest vegetation zonation (cf. Fig. 1).

Such an analysis was applied to counts of 13 arable weeds in summer barley in May 1983 in 96 plots (0.5 ×0.5 m) in the experimental field 'Doeksen' (50 m × 100 m) (B. Post, unpubl).

The first CCA axis was defined by $b_1 = 0.0261$ and $b_2 = 0.0117$, so that the gradient was estimated to make $\tan^{-1}(b_2/b_1) = 24°$ with the x-coordinate axis. Further, the first eigenvalue was six times the second eigenvalue, which indicated that the gradient was a clear one. But, judged on the basis of the value of the first eigenvalue ($\lambda_1 = 0.09$), the amount of species turnover was quite small (cf. Gauch & Stone, 1979).

To verify the supposition that the gradient was related to moisture, percentage moisture was measured in the top soil (0–3 cm) in March 1985 (B. Post, unpubl). The strongest gradient in these moisture values had an angle of 34° with the x-coordinate axis and thus pointed approximately in the same direction as the gradient estimated by CCA from the 1983 weed data.

Vegetation succession

An example of application in a succession study on a rising sea-shore is found elsewhere in this volume (Cramer & Hytteborn, 1987). One of their questions was whether the vegetation succession tracks the land uplift (ca. 0.5 cm per year) or whether it lags behind.

This question was approached with detrended CCA with elevation and year as the 'environmental variables', through fitting the compound gradient $x = b_1 \times \text{elevation} + b_2 \times \text{year}$. The resulting weights were $b_1 = 0.054$ and $b_2 = 0.041$. Consequently, the equivalent change in vegetation per year is $b_2/b_1 = 0.76$ cm.

An approximate 95%-confidence interval for the change ranges from 0.4 cm to 1.1 cm and clearly includes the known land rise of ca 0.5 cm per year. The confidence interval was obtained from the standard errors of b_1 and b_2 in the final regression within the reciprocal averaging algorithm of CCA by using Fieller's theorem (see Finney, 1964, p. 27–29). The interval is presumably a little too short as it ignores that the CCA-axis is chosen optimally.

Discussion

CCA considerably extends the analytical power of ecological ordination. Questions like those tackled in the applications section above could formerly only be investigated by 'indirect gradient analysis', i.e. first extracting the ordination axes from the species data and subsequently interpreting the major axes in relation to environmental data – e.g. by regression analysis (Dargie, 1984), trend surface analysis (Gittins, 1968) or canonical correlation analysis (Carleton, 1984). Such two-step analyses ignore the minor axes of variation in community composition; yet 'minor' aspects of the variation

may still be substantial, especially in large data sets, and in some problems may be just the variation that one is actually interested in because of its relationship to particular external variables (see Jolliffe, 1982).

CCA works because species tend to have single-peaked response functions to environmental variables. When the response functions are simpler (e.g. approximately linear), the results can still be expected to be adequate in a qualitative sense, but it might then be advantageous to utilize instead the linear counterpart of CCA – redundancy analysis (Israëls, 1984). The weed data are a case in point. Because the number of species is quite small in that example, and the number of absences is small as well, these data could also be analysed from the beginning by canonical correlation analysis (Gittins, 1985). But canonical correlation analysis and redundancy analysis fail, when species do show single-peaked response functions (Gauch & Wentworth, 1976), i.e. in the case where CCA works best.

Appendix

Maximizing δ in Eq. (3) leads to CCA (Ter Braak, 1986) and CCA is a weighted principal components analysis applied to a matrix of weighted averages.

Let $\mathbf{Y} = \{y_{ik}\}$ and $\mathbf{Z} = \{z_{ij}\}$ be $n \times m$ and $n \times p$ matrices containing the species data and environmental data, respectively, and let $\mathbf{R} = \text{diag}(y_{1+}, y_{2+}, \ldots, y_{n+})$. Each environmental variable is centered to a weighted mean of 0, i.e. $\mathbf{Z}'\mathbf{R}\mathbf{1}_n = \mathbf{0}$, where $\mathbf{1}_n$ is an n-vector containing 1's. Further, let $\mathbf{S}_{11} = \text{diag}(y_{+1}, y_{+2}, \ldots, y_{+m})$, $\mathbf{S}_{12} = \mathbf{Y}'\mathbf{Z}$, $\mathbf{S}_{21} = \mathbf{Z}'\mathbf{Y}$, $\mathbf{S}_{22} = \mathbf{Z}'\mathbf{R}\mathbf{Z}$ and let $\mathbf{u}$ and $\mathbf{b}$ be vectors of order m and p, containing the species scores u_k and the weights b_j, respectively.

By inserting Eq. (4) in Eq. (1) we obtain

$$\mathbf{u} = \mathbf{S}_{11}^{-1}\mathbf{Y}'\mathbf{Z}\mathbf{b} = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} \qquad \text{(A.1)}$$

Hence,

$$\delta = y_{+\cdot}^{-1}\mathbf{u}'\mathbf{S}_{11}\mathbf{u} = y_{+\cdot}^{-1}\mathbf{b}'\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} \qquad \text{(A.2)}$$

which must be maximized with respect to $\mathbf{b}$, subject to Eq. (2). By inserting Eq. (4) in Eq. (2), we obtain $\mathbf{b}'\mathbf{Z}'\mathbf{R}\mathbf{1}_n = 0$, which is satisfied trivially because of the centering of Z, and

$$y_{+\cdot}^{-1}\mathbf{b}'\mathbf{S}_{22}\mathbf{b} = 1 \qquad \text{(A.3)}$$

The solution of this maximization problem is known to be the first eigenvector of the eigenvalue equation

$$(\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12} - \lambda\mathbf{S}_{22})\,\mathbf{b} = \mathbf{0} \qquad \text{(A.4)}$$

with $\delta = \lambda$ (see, for instance, Mardia *et al.*, 1979, theorem A.9.2). Eq. (A.4) is the centered version of Eq. (A5) in Ter Braak (1986). The latter equation has a trivial solution ($\lambda = 1$, $\mathbf{x} = \mathbf{1}_n$) and its nontrivial solutions satisfy Eq. (A.4) and Eq. (2). Therefore, maximizing δ leads to the first axis of CCA as defined in Ter Braak (1986). Further, maximizing δ subject to the constraint that the second axis is uncorrelated with the first axis (using weights y_{i+}, as in Eq. (2)) leads to the second eigenvector of (A.4), which is therefore identical to the second axis of CCA as defined in Ter Braak (1986), and so on for subsequent axes.

Let $\mathbf{W}$ be a $m \times p$ matrix containing the weighted averages of the species with respect to the environmental variables, i.e.

$$\mathbf{W} = \mathbf{S}_{11}^{-1}\mathbf{Y}'\mathbf{Z} \qquad \text{(A.5)}$$

The weighted principal components analysis of $\mathbf{W}$ described in the main text follows from the singular value decomposition

$$\mathbf{S}_{11}^{1/2}\,\mathbf{W}\,\mathbf{S}_{22}^{-1/2} = \mathbf{S}_{11}^{-1/2}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1/2} = \mathbf{P}\,\mathbf{\Lambda}^{1/2}\,\mathbf{Q}' \qquad \text{(A.6)}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are orthonormal $m \times p$ and $p \times p$ matrices and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. For convenience of notation it is assumed here that $p \leq m$. This singular value decomposition is just another way to solve (A.4) (see Mardia *et al.*, 1979, chapter 10). The coordinates of species k in the ordination diagram are given by the k-th row of the matrix

$$\mathbf{U} = y_{+\cdot}^{1/2}\,\mathbf{S}_{11}^{-1/2}\,\mathbf{P}(\mathbf{I} - \mathbf{\Lambda})^{-1/2}, \qquad \text{(A.7)}$$

and the coordinates of environmental variable j by the j-th row of the matrix

$$\mathbf{B}_e = y_{+\cdot}^{-1/2}\,\mathbf{S}_{22}^{1/2}\,\mathbf{Q}\,\mathbf{\Lambda}^{1/2}\,(\mathbf{I} - \mathbf{\Lambda})^{1/2} \qquad \text{(A.8)}$$

The pre- and post-multiplication factors involving $y_{+\cdot}$ and $(\mathbf{I}-\mathbf{\Lambda})$ in Eqs. (A.7) and (A.8) are not essential for the biplot; they are included to obtain the scaling used in DECORANA (Hill, 1979, section 4.5). In Hill's scaling the coordinates of the sites are weighted averages of the species coordinates and the (weighted) variance of the coordinates of species present at a site is equal to 1 on average. Hill's scaling is used in Fig. 2.

References

Austin, M. P., 1971. Role of regression analysis in plant ecology. Proc. Ecol. Soc. Austr. 6: 63–75.

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.

Campbell, N. A. & Atchley, W. R., 1981. The geometry of canonical variate analysis. Syst. Zool. 30: 268–280.

Carleton, T. J., 1984. Residual ordination analysis: a method for exploring vegetation-environment relationships. Ecology 65: 469–477.
Cramer, W. & Hytteborn, H., 1987. The separation of fluctuation and long-term change in the vegetation dynamics of a rising sea-shore. Vegetatio 69: 157–167.
Dargie, T. C. D., 1984. On the integrated interpretation of indirect site ordinations: a case study using semi-arid vegetation in south-eastern Spain. Vegetatio 55: 37–55.
Finney, D. J., 1964. Statistical methods in biological assay. Griffin, London, 668 pp.
Forsythe, W. L. & Loucks, O. L., 1972. A transformation for species response to habitat factors. Ecology 53: 1112–1119.
Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58: 453–467.
Gauch, H. G., 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge.
Gauch, H. G. & Stone, E. L., 1979. Vegetation and soil pattern in a mesophytic forest at Ithaca, New York. Am. Midl. Nat. 102: 332–345.
Gauch, H. G. & Wentworth, T. R., 1976. Canonical correlation analysis as an ordination technique. Vegetatio 33: 17–22.
Gittins, R., 1968. Trend-surface analysis of ecological data. J. Ecol. 56: 845–869.
Gittins, R., 1985. Canonical analysis. A review with applications in ecology. Springer Verlag, Berlin.
Hill, M. O., 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237–249.
Hill, M. O., 1979. DECORANA – A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics, Cornell University, Ithaca, New York.
Hill, M. O. & Gauch, H. G., 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42: 47–58.
Israëls, A. Z., 1984. Redundancy analysis for qualitative variables. Psychometrika 49: 331–346.
Jolliffe, I. T., 1982. A note on the use of principal components in regression. Appl. Statist. 31: 300–303.
Loucks, O. L., 1962. Ordinating forest communities by means of environmental scalars and phytosociological indices. Ecol. Monogr. 32: 137–166.
Mardia, K. V., Kent, J. T. & Bibby, J. M., 1979. Multivariate analysis. Academic Press, London.
Nishisato, S., 1980. Analysis of categorical data: dual scaling and its applications. University of Toronto Press, Toronto.
Ter Braak, C. J. F., 1983. Principal components biplots and alpha and beta diversity. Ecology 64: 454–462.
Ter Braak, C. J. F., 1985a. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41: 859–873.
Ter Braak, C. J. F., 1985b. CANOCO – A FORTRAN program for canonical correspondence analysis and detrended correspondence analysis. IWIS-TNO, Wageningen, The Netherlands.
Ter Braak, C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167–1179.
Ter Braak, C. J. F. & Looman, C. W. N., 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65: 3–11.
Whittaker, R. H., 1967. Gradient analysis of vegetation. Biol. Rev. 42: 207–264.
Yarranton, G. A., 1970. Towards a mathematical model of limestone pavement vegetation. III. Estimation of the determinants of species frequencies. Can. J. Bot. 48: 1387–1404.

Accepted 21.8.1986.

PARTIAL CANONICAL CORRESPONDENCE ANALYSIS

Cajo J.F. TER BRAAK

TNO Institute of Applied Computer Science, Box 100, 6700 AC Wageningen, The Netherlands

Canonical correspondence analysis is (multiple) correspondence analysis in which the ordination axes are constrained to be linear combinations of external, explanatory variables. We consider the case where the set of explanatory variables is subdivided in two sets, a set of covariables and a set of variables-of-interest. This leads to partial canonical correspondence analysis. Its ordination diagram displays the unimodal relationships between a set of response variables and the variables-of-interest after the effects of the covariables have been partialed out. The derivation shows that the response data can be incidence data, count data, compositional data or nominal data.

1. INTRODUCTION

Canonical correspondence analysis is a multivariate analysis technique to display unimodal relationships between a set of response variables and a set of explanatory variables in a low-dimensional space, called an ordination diagram [20,21]. Canonical correspondence analysis has been used in ecology as a simple form of constrained multidimensional unfolding [4,10,12] to relate the occurrences or abundances of a number of species to environmental variables [22]. Applied to nominal variables, canonical correspondence analysis is identical to redundancy analysis of qualitative variables [14] used, for example, to relate nominal welfare variables to social background variables. Here we consider the case where the set of explanatory variables is subdivided in two sets, a set of p covariables and a set of q variables in the effects of which one is particularly interested. Stated informally, we want an ordination diagram of the unimodal relationships between the response variables and the q variables of interest after eliminating the effects of the p covariables. The object is thus to partial out the effects of the covariables, hence the name partial canonical correspondence analysis. Ter Braak [20] derived canonical correspondence analysis as an approximation to canonical Gaussian ordination. Here we define partial canonical Gaussian ordination, derive partial canonical correspondence analysis as an approximation and give an example. Our derivation starts from a constrained generalized linear model and shows that the technique can be applied to nominal data (multi-way contingency data), compositional data, count data and incidence data, with quantitative or qualitative explanatory variables. Related work on partial analysis is given in [3,15,25].

2. THEORY

Let Y and Z be real matrices of order n×m and n×(p+q), containing n observations of m nonnegative response variables and p+q explanatory variables, respectively. The p+q explanatory variables are subdivided in p covariables (including the vector 1_n) and q variables of interest and $Z = (Z_1, Z_2)$ is partitioned accordingly. The response variables can be

incidences (1/0) or counts of animals or plants in regions, or fractions of constituents in a composition. For nominal variables, Y is a multivariate indicator matrix [7,9] with as many columns as categories. The elements of a matrix B are denoted by b_{ij}, the j-th column of B by $\underline{b}_j$ and the i-th row of B by $\underline{b}_{(i)}$, a column vector, and a generalized inverse of B by B^-. The symbol E denotes expectation.

We now define partial canonical Gaussian ordination as a constrained, generalized linear model.

Definition: For any integer $r<q$, the model of partial canonical Gaussian ordination is

$$\text{link}\,(E\underline{y}_{ik}) = \phi_i + a_k - \tfrac{1}{2}\,(\underline{z}_{(i)}-\underline{u}^*_{(k)})'M(\underline{z}_{(i)}-\underline{u}^*_{(k)}) \qquad (1)$$

where link is a natural link function (Table 1) [16:p.24] and M is constrained to

$$M = DD' \text{ with } D = \begin{pmatrix} F & G \\ 0 & C \end{pmatrix} \qquad (2)$$

with F, G and C parameter matrices of order p×p, p×r and q×r, respectively, and 0 is a matrix of order q×p with zeroes; $\underline{u}^*_{(k)}$ is a (p+q)-vector representing the optimum of response variable k, a_k is a scalar related to the maximum expected response, and ϕ_i is an incidental parameter for sampling unit i, which takes care of the constant-sum constraint, if present [16:p. 106, p. 142].

Table 1 shows for various types of data the appropriate link function, error distribution and ϕ_i. A statistical interpretation of partial canonical Gaussian ordination is that the m response variables (in Y) are explained by two sets of explanatory variables [in $Z = (Z_1, Z_2)$] by a generalized linear model (GLM) [16] with as predictor a quadratic form in the explanatory variables. It is a unimodal regression model (Fig. 1) with constraints. The difference with standard GLM, which is applied to each response variable separately, is that the parameter matrix M is identical for all response variables and that M is constrained to be positive semi-definite of rank at most p+r, so as to allow an r-dimensional representation of the partial effects of the q variables of interest on the response variables. This becomes clear by writing the model as a constrained ordination model. By setting $\underline{x}_{(i)} = D'\underline{z}_{(i)}$ and $\underline{u}^*_{(k)} = D'\underline{u}^*_{(k)}$, the model is transformed to the canonical form (Fig. 1)

$$\text{link}\,(E\underline{y}_{ik}) = \phi_i + a_k - \tfrac{1}{2}\,(\underline{x}_{(i)}-\underline{u}_{(k)})'(\underline{x}_{(i)}-\underline{u}_{(k)})\,. \qquad (3)$$

By this transformation, the n×(p+q) matrix Z is transformed to a n×(p+r) matrix X, whose i-th row is $\underline{x}_{(i)}$. In terms of variables (the columns of Z and X), the p+q explanatory variables are transformed to p+r axes of a new coordinate system, called ordination axes, by

$$\underline{x}_s = Z_1\underline{f}_s \qquad (1 \leq s \leq p) \qquad (4a)$$

$$\underline{x}_s = Z_1\underline{g}_s + Z_2\underline{c}_s \qquad (p < s \leq p+r) \qquad (4b)$$

Table 1. Types of response which can be analysed by model (1) which is the basis of partial canonical correspondence analysis (ϕ = incidental parameter, l = index of the L nominal variables with, in total m, categories, ref = references for related models).

type of response	example	link	error*	ϕ	ref
incidence	artifacts in graves pick-any-out-of-m data	logit	Bernoulli	0	13,19 2
abundance	species in regions	log	Poisson	0	13,19
compositions	pollen data electrophoresis data	log	multinomial	ϕ_i	13,16 23
nominal	multiple-choice data	log	multinomial	ϕ_{il}	1,8

*) including extensions to quasi-likelihood models [16].

i.e. the first p ordination axes are a linear combination of the p covariables and the last r axes are a linear combination of all p+q explanatory variables. Model (3) without the constraints in (4) is the Gaussian ordination model [6,11,19] and contains Ihm & Van Groenewoud's [13] generalized logit model. If $a_k = a$ and $\phi_i = 0$, the model shows shifted single-peakedness [11].

In the sequel we focus on the estimation of a basis for the column space of the matrix D and on the estimation of the optima after transformation $\underline{u}_{(k)} = (u_{k1}, u_{k2}, \ldots, u_{k(p+r)})'$. Under the assumption that the $\{y_{ik}\}$ are either independent Bernoulli variables when link(.) = logit(.), or independent Poisson or multinomial variables when link(.) = log(.) (Table 1), with expectations defined by (1), the maximum likelihood equations for u_{ks} and the elements of D become, after some rearrangement [20]

$$u_{ks} = \Sigma_i\, y_{ik}x_{is}/y_{+k} - [\Sigma_i\, (x_{is}-u_{ks})(Ey_{ik})/y_{+k}] \qquad (5)$$

$$\Sigma_i\, z_{ij}[\Sigma_k\, y_{ik}(x_{is}-u_{ks})] = \Sigma_i\, z_{ij}[\Sigma_k\, (x_{is}-u_{ks})Ey_{ik}] \qquad (6)$$

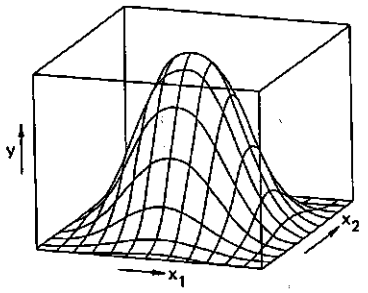
for k=1, ..., m; j=1, ..., p+q and s=1, ..., p+r, where $y_{+k} = \Sigma_i\, y_{ik}$.

We now derive partial canonical correspondence analysis as an approximation to Eqs (5)-(6) under the following simplifying conditions:

C1. the maxima are equal (a_k=a, k=1, ..., m), or random and independent of the optima $\underline{u}_{(k)}$,
C2. the optima $\underline{u}_{(k)}$ are uniformly distributed over a hypercube A with sides parallel to the ordination axes and of length much larger than 1,
C3. the sampling points $\underline{x}_{(i)}$ are uniformly distributed over a 'large' hypercube B that is contained in A and that has the origin as centroid,
C4. m and n are large so that the optima and sampling points are densely spaced. For nominal variables, the number of classes per variable should be large.

Under these conditions, Ey_{ik} is approximately symmetric about x_{is} and about u_{ks} for each s [19], so that we may use the approximations

Fig. 1. A unimodal relationship between a response variable y and two regressors (Eq. (3) with p+r = 2 and link = log).



$$\Sigma_k\ (x_{is}-u_{ks})Ey_{ik} = 0 \qquad (7)$$

$$\Sigma_i\ (x_{is}-u_{ks})Ey_{ik} = -\lambda_s^* u_{ks} y_{+k}\ . \qquad (8)$$

The proportionality constant λ_s^* comes in because the unimodal response surfaces are the more truncated the more their optima lie towards or beyond the edge of the sampling region [19,20]. Using Approximation (8) and the equation $\lambda_s = 1-\lambda_s^*$, we obtain from (5)

$$\lambda_s u_{ks} = \Sigma_{i=1}^{n}\ y_{ik}x_{is}/y_{+k}\ . \qquad (9)$$

By inserting (4) in (6) and using Approximation (7) we obtain

$$(Z_1'RZ_1)\underline{f}_s = Z_1'R\underline{x}_s^* \qquad (1 \leq s \leq p) \qquad (10)$$

$$(Z'RZ)\underline{d}_s = Z'R\underline{x}_s^* \qquad (p < s \leq p+r) \qquad (11)$$

where $R = \mathrm{diag}\ (y_{i+})$ with $y_{i+} = \sum_i y_{ik}$ and $\underline{x}_s^* = (x_{1s}^*, \ldots, x_{is}^*, \ldots, x_{ns}^*)'$ with

$$x_{is}^* = \Sigma_{k=1}^{m}\ y_{ik}u_{ks}/y_{i+} \qquad (12)$$

Equations (4) and (9)-(12) can be solved in a similar way as the transition formulae of canonical correspondence analysis [20].

Because Z_2 contains the variables of interest, it would be convenient to solve for the last r ordination axes without having to extract the first p ordination axes. Fortunately, this can be achieved by making the partitioning of Z in Z_1 and Z_2 explicit. By solving (11) for the component $\underline{c}$ in $\underline{d}_s' = (\underline{g}_s', \underline{c}_s')$ and using the standard formula for the inverse of a partitioned matrix [18,p.33], we obtain for $s > p$

$$\underline{c}_s = (\tilde{Z}_2'R\tilde{Z}_2)^{-}\ \tilde{Z}_2'R\underline{x}_s^*, \text{ where} \qquad (13)$$

$$\tilde{Z}_2 = (I-Z_1^{\circ})Z_2 \qquad (14)$$

and where the notation B° is used to denote $B(B'RB)^{-}B'R$, the projection operator on V(B), the column space of B, in the metric defined by R. Further $\underline{x}_s$ is the projection of $\underline{x}_s^*$ on $Z = (Z_1, Z_2)$ as follows from (4b) and (11), so that

$$\underline{x}_s = Z_1\underline{x}_s^* + \tilde{Z}_2\underline{x}_s^* \quad . \tag{15}$$

But, in canonical correspondence analysis the last r ordination axes are required to be orthogonal to the first p ordination axes [20], so that

$$Z_1'\underline{x}_s = Z_1'\underline{x}_s^* = 0 \qquad (p < s \leq p+r) \tag{16}$$

because $V(Z_1) = V(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p)$. Therefore,

$$\underline{x}_s = Z_2\underline{x}_s^* = Z_2(Z_2'RZ_2)^{-}Z_2'R\underline{x}_s^* = Z_2\underline{a}_s \qquad (p < s \leq p+r) \quad . \tag{17}$$

The last r ordination axes can thus be obtained from (9), (12), (13) and (17). These equations form the transition formulae of partial canonical correspondence analysis and define an eigenvalue problem akin to that of canonical correspondence analysis [20]. This can be verified by inserting consecutively in (13) the equations (12), (9) and (17), giving

$$(S_{21}K^{-1}S_{12} - \lambda S_{22})\underline{a}_s = 0 \tag{18}$$

where $S_{21} = Z_2'Y$, $S_{12} = Y'Z_2$, $S_{22} = Z_2'RZ_2$ and $K = \mathrm{diag}(y_{+k})$.

In summary, partial canonical correspondence is a canonical correspondence analysis technique whereby p+r orthogonal axes are constructed. The first p axes are linear combinations of the p covariables only and the subsequent r axes are linear combinations of the p covariables and the q variables of interest. As the covariables are of less interest in the analysis, the first p axes are usually ignored. The subsequent r axes are considered as the first r ordination axes of partial canonical correspondence analysis. They give a low-dimensional representation of the unimodal relationships according to model (1) with constraint (2) between the variables of interest and the response variables after partialing out the effects of the covariables. Technically, the only difference with canonical correspondence analysis is that the matrix of explanatory variables is replaced by the matrix $\tilde{Z}_2$ of residuals of a multivariate multiple regression of Z_2 on Z_1 (14).

Special cases of partial canonical correspondence analysis are:

1. Canonical correspondence analysis [14,20] if Z_1 is a n×1 matrix of 1's (a single trivial covariable only).
2. Partial correspondence analysis if Z_2 is a n×n identity matrix (no variables of interest) or any arbitrary n×(n-1) matrix of rank n-1 (too many variables of interest [21]).
3. Multiple correspondence analysis [7,9] if Z_1 and Z_2 are as specified in 1 and 2 above (no explanatory variables or too many of them).
4. Weighted averaging ordination [6,21] if p = 0 and q = 1 (a single variable of interest).

Our definition of partial correspondence analysis differs from that by Yanai [15].

3. ORDINATION DIAGRAM

As in correspondence analysis, the results can be presented in an ordination diagram in which the rows and columns of Y are represented by points at locations $\underline{x}_{(i)}$ and $\underline{u}_{(k)}$. To the extent that the analysis approximates the

fitting of Gaussian surfaces (1), the points for response variables are approximately the optima of these surfaces; hence, the value of Ey_{ik} decreases with the distance between the points of sampling unit i and response variable k (Fig. 1). The estimated values are, of course, conditional on the values of the covariables.

In partial canonical correspondence analysis the ordination diagram can be supplemented with arrows for the variables of interest (Fig. 2). This is done in such a way that, in conjunction with the points for response variables, the arrows give a weighted least squares approximation of the elements of the m×q matrix $W = K^{-1}Y'\tilde{Z}_2$. The (k, j)-th element of W is the weighted average of response variable k with respect to variable of interest j, after this variable is adjusted for the covariables. In a unimodal model, the weighted average indicates the centre of a response curve. So the matrix W summarizes unimodal relationships, like a matrix of partial correlation coefficients summarizes linear relationships. In the approximation of W, response variables are given weights proportional to their total y_{+k}. The coordinates of the supplementary arrows can be obtained by a multivariate regession of W on $U = \{u_{ks}\}$, i.e. by

$$C_r = W'KU(U'KU)^{-1}. \qquad (19)$$

The approximation to W is then given by the bilinear model UC_r'. The plot of points for response variables and arrows for variables of interest is thus a biplot [5], termed the species-environment biplot in [20]. This plot is not just supplementary, as it can be made central to (partial) canonical correspondence analysis [22].

4. EXAMPLE

The example is taken from H. Smit (in prep.). Smit studied the abundances of diatom species in dykes in the province of Zuid Holland (The Netherlands), with special reference to the effects of water pollution. A sample of 402 dykes was taken, which contained in total 330 species. Variables that indicate pollution were compounds with phosphorus (P) and nitrogen (N), and biological oxygen demand (BOD). Apart from variation in pollution, the sample showed strong natural variation due to the season of sampling and due to a gradient from fresh to brackish water. This natural variation was partialed out by specifying a season indicator variable and the chloride concentration (Cl) as covariables. Partial canonical correspondence analysis on diatom species with 24 variables-of-interest showed a first axis (λ_1 = 0.10) that was a clear pollution gradient as indicated by the arrows for P, BOD and N in the ordination diagram (Fig. 2). The second axis (λ_2 = 0.05) revealed the importance of other natural variation, notably soil type and dyke width. Species of polluted waters are represented on the right hand side of the diagram (Fig. 2), e.g. Navicula accomoda and N. subminuscula, whereas species of unpolluted waters lie on the left hand side, e.g. Eunotia pectinalis. Species in the middle have their optimum at intermediate pollution levels or are indifferent [20]. Which possibility is most likely can be decided upon by plotting the abundance values on the ordination diagram. Despite their occurrence at high values of P and BOD, two species of brackish waters, Melosira jürgensii and Navicula diserta, are displayed on the left hand side of the diagram, because brackish waters naturally have

high P- and BOD- values. This illustrates that Fig. 2 displays partial effects.

5. DISCUSSION

In this paper partial canonical correspondence analysis is derived as an approximation to maximum likelihood estimation of a particular unimodal model. But it does not maximize a likelihood. What is being maximized is the least-squares criterion of multiple correspondence analysis [7,11,12,21,24] with the additional constraints (a) that the axes are linear combinations of all explanatory variables and (b) that the axes are orthogonal to the covariables. We note that the orthogonality constraints do not follow necessarily from the maximum likelihood approach (see below Eq. (15)). They are not sufficient either; we conjecture that when the Guttman effect [9,19] crops up, the transition formulae have solutions close to solutions of the maximum likelihood equations that correspond to local maxima. Such solutions can be excluded by "detrending" [6,19,23] or by deleting explanatory variables [21]. Other loss-functions are considered in [10,12,17].

In the dual scaling approach to correspondence analysis [9], category scores form the optimal quantification [7] of the corresponding nominal variables. This paper gives reason to interpret category scores as optima of underlying response curves (termed trace lines in [24]). The properties of correspondence analysis in terms of a unimodal model were explored earlier by Torgerson [24: point items], Heiser [10,11] and Ihm & Van Groenewoud [13].

For data with a constant-sum constraint ($\phi_i \neq 0$ in Table 1), model (1) can be rewritten as

$$\log (Ey_{ik}) = \phi_i^* + a_k^* + z_{(i)}' M u_{(k)}^* \qquad (20)$$

where ϕ_i^* and a_k^* have absorbed the quadratic forms in $z_{(i)}$ and $u_{(k)}^*$ in Eq. (1), respectively. Model (20) with p=1 and r=1 is the qualitative logistic regression model, from which Anderson [1] developed his regression method for ordinal response variables (cf. [8]). The results of this paper can be used to show that his method can be approximated by canonical correspondence analysis with ordinal constraints, as in Gifi [7], on the category scores. It is surprising that for nominal and compositional data the unimodal unfolding model (1) can be reexpressed as a generalized bilinear model (20)!

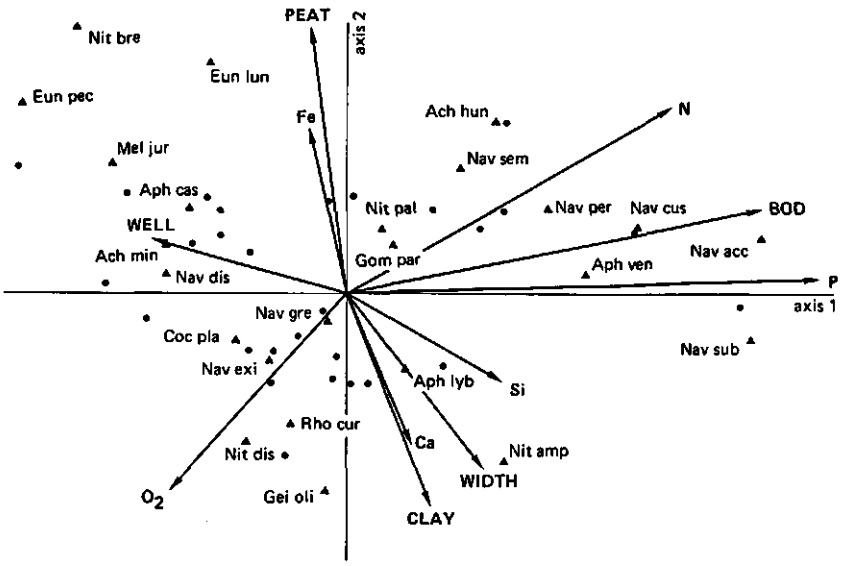


Fig. 2. Ordination diagram of a partial canonical correpondence analysis of diatom species (▲) in dykes with as explanatory variables 24 variables-of-interest (arrows) and 2 covariables (chloride concentration and season). The diagram is symmetrically scaled [23] and shows selected species and standardized variables and, instead of individual dykes, centroids (●) of dyke-clusters. The variables-of-interest shown are: BOD = biological oxygen demand, Ca = calcium, Fe = ferrous compounds, N = Kjeldahl-nitrogen, O_2 = oxygen, P = ortho-phosphate, Si = silicium-compounds, WIDTH = dyke width, and soil types (CLAY, PEAT). All variables except BOD, WIDTH, CLAY and PEAT were transformed to logarithms because of their skew distribution. The diatoms shown are: Ach hun = Achnanthes hungarica, Ach min = A. minutissima, Aph cas = Amphora castellata Giffen, Aph lyb = A. lybica, Aph ven = A. veneta, Coc pla = Cocconeis placentula, Eun lun = Eunotia lunaris, Eun pec = E. pectinalis, Gei oli = Gomphoneis olivaceum, Gom par = Gomphonema parvulum, Mel jur = Melosira jürgensii, Nav acc = Navicula accomoda, Nav cus = N. cuspidata, Nav dis = N. diserta, Nav exi = N. exilis, Nav gre = N. gregaria, Nav per = N. permitis, Nav sem = N. seminulum, Nav sub = N. subminuscula, Nit amp = Nitzschia amphibia, Nit bre = N. bremensis v. brunsvigensis, Nit dis = N. dissipata, Nit pal = N. palea, Rho cur = Rhoicosphenia curvata.
(Adapted from H. Smit, province of Zuid Holland, in prep.)

6. REFERENCES

[1] Anderson, J.A., Regression and ordered categorical variables. J. R. Statist. Soc. B 46 (1984) 1-30.

[2] Coombs, C.H., A theory of data. Wiley, New York, 1964.

[3] Daudin, J.J., Partial association measures and an application to qualitative regression. Biometrika 67 (1980) 581-590.

[4] DeSarbo, W.S., and Rao, V.R., GENFOLD2: a set of models and algorithms for the general unfolding analysis of preference/dominance data. J. of Classification 1 (1984) 147-186.

[5] Gabriel, K.R., The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (1971) 453-467.
[6] Gauch, H.G., Multivariate analysis in community ecology. Cambridge Univ. Press, Cambridge, 1982.
[7] Gifi, A. Nonlinear multivariate analysis. Dept. of Data Theory, Univ. of Leiden, Leiden, 1981.
[8] Goodman, L.A., Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. Int. Statist. Rev. 54 (1986) 243-309.
[9] Greenacre, M.J., Theory and applications of correspondence analysis. Academic Press, London, 1984.
[10] Heiser, W.J., Unfolding analysis of proximity data. Thesis. University of Leiden, Leiden, 1981.
[11] Heiser, W.J., Undesired nonlinearities in nonlinear multivariate analysis. In: Data Analysis and Informatics 4 (E. Diday et al, eds.), North-Holland, Amsterdam, 1986, pp. 455-469.
[12] Heiser, W.J., Joint ordination of species and sites: the unfolding technique. In: New developments in numerical ecology, (P. Legendre and L. Legendre, eds.), Springer-Verlag, Berlin, 1987, in press.
[13] Ihm, P. and Van Groenewoud, H., Correspondence analysis and Gaussian ordination, COMPSTAT lectures 3 (1984) 5-60.
[14] Israëls, A.Z., Redundancy analysis for qualitative variables. Psychometrika 49 (1984) 331-346.
[15] Lauro, N. and D'Ambra, L., Multiple and partial non symmetrical correspondence analysis, this volume.
[16] McCullagh, P. and Nelder, J.A., Generalized linear models. Chapman and Hall, London, 1983.
[17] Meulman, J., and Heiser, W.J., Constrained multidimensional scaling: more directions than dimensions. COMPSTAT 1984, Physica-Verlag, Vienna, 1984, pp. 137-142.
[18] Rao, C.R., Linear statistical inference and its applications. 2nd ed. Wiley, New York, 1973.
[19] Ter Braak, C.J.F., Correspondence analysis of incidence and abundance data: properties in terms of a unimal response model. Biometrics 41 (1985) 859-873.
[20] Ter Braak, C.J.F., Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67 (1986) 1167-1179.
[21] Ter Braak, C.J.F., The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetation 69 (1987), 69-77.
[22] Ter Braak, C.J.F., Ordination. In: Data analysis in community and landscape ecology (R.H.G. Jongman, C.J.F. Ter Braak and O.F.R. Van Tongeren, eds.). Pudoc, Wageningen, 1987.
[23] Ter Braak, C.J.F., CANOCO - a FORTRAN program for canonical community ordination by [partial][detrended][canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). TNO Institute of Applied Computer Science, Wageningen, 1987.
[24] Torgerson, W.S., Theory and methods of scaling. Wiley, New York, 1958.
[25] Yanai, H., Some generalizations of correspondence analysis in terms of projection operators. In: Data Analysis and Informatics 4. (E. Diday et al, eds.), North-Holland, Amsterdam, 1986, pp. 193-207.

Vegetatio 69: 79–87, 1987

Ecological amplitudes of plant species and the internal consistency of Ellenberg's indicator values for moisture*

Cajo J. F. Ter Braak[1] & Niek J. M. Gremmen[2]**
[1]*TNO Institute of Applied Computer Science, 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, Gaussian logit curve, Indicator value, Logit regression, Maximum likelihood, Optimum, Tolerance, Unimodal response curve, Weighted averaging

Abstract

Two methods for estimating ecological amplitudes of species with respect to Ellenberg's moisture scale are discussed, one based on weighted averaging and the other on maximum likelihood. Both methods are applied to phytosociological data from the province of Noord-Brabant (The Netherlands), and estimate the range of occurrence of species to be about 4–6 units on the moisture scale. Due to the implicit nature of Ellenberg's definition of moisture, it is impossible to improve the indicator values in a statistically sound way on the basis of floristic data only. The internal consistency of the Ellenberg indicator values is checked by using Gaussian logit regression. For 45 out of the 240 species studied the indicator value is inconsistent with those of the other species. The same method is used to estimate the optima and amplitudes of species considered moisture-indifferent and of some species not mentioned by Ellenberg. Some of these 'indifferent' species show a remarkably narrow amplitude.

It is concluded that the Ellenberg indicator values for moisture form a reasonably consistent system.

Introduction

Ellenberg (1979) summarized the ecology of the Central-European vascular plants, by assigning to each species indicator values for light, temperature, moisture, nitrogen and acidity.

Ellenberg's indicator values are used to estimate the value of any of these environmental factors at a particular site by averaging the indicator values for this factor of all species present (e.g. Ellenberg, 1979, 1983; Persson, 1981; Smeets, Werger & Tevonderen, 1980; Böcker, Kowarik & Bornkamm, 1983). Plants often reflect temporally integrated environmental conditions and are therefore particularly useful indicators when values averaged over time are needed. When the value of an environmental factor in the past is required, the only possible approach may be to base it on historical vegetation data.

During the development of a model simulating the effects of withdrawal of groundwater on the disappearance of plant species (Gremmen *et al.*, 1985; Reijnen & Wiertz, 1984), we wished to know:

(1) do Ellenberg's indicator values for moisture and nitrogen correctly represent the optima of species for these factors in our study area,

(2) what is the ecological amplitude of each species for these factors, including species not mentioned by Ellenberg (1979)?

We will only discuss moisture values here. Clear-

* Nomenclature follows Heukels-Van der Meijden (1983), Flora van Nederland, 20th ed.

** We would like to thank M. J. S. M. Reijnen and J. Wiertz for the discussions that gave us the idea for this research. We are grateful to J. de Bree, C. Hengeveld and the referees for comments on the manuscript. Part of this research was supported by the Commissie Grondwaterwet Waterleidingbedrijven, the Keuringsinstituut van Waterleidingartikelen, the Landinrichtingsdienst, Staatsbosbeheer, and the Ministerie van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer.

ly, the same reasoning can be applied for other factors.

Ellenberg (1979) placed each species on a 12-point ordinal scale according to its distribution with respect to moisture (Table 1). It is not clear which characteristic(s) of the moisture regime (e.g. groundwater level, soil moisture content, and soil moisture deficit) were used in the definition of these classes. In practice the indicator values of Ellenberg's 'intuitive' scale seem to work well, however.

The implicit nature of Ellenberg's definition of moisture makes it impossible to check the correctness of the indicator values against actual measurements. Nevertheless, it is possible to check the internal consistency by comparing the indicator values of species that occur together: when a species mainly occurs together with species with higher (lower) indicator values, its indicator value is in comparison with those of the other species too low (too high). (When species have extreme indicator values this intuitive idea needs modification.) Alternatively, the consistency of the Ellenberg moisture values could be checked by studying the distribution of each species with respect to moisture. In this approach the moisture value of a site is calculated by averaging the indicator values of the species present. The indicator value of a particular species is clearly inconsistent with those of the other species when it deviates considerably from the center of the distribution of this species. This distribution also contains information on the ecological amplitude of the species for moisture.

In this paper this simple method is developed further and compared with a more sophisticated maximum likelihood method, in which the species' distributions are modelled by Gaussian logit curves (Ter Braak & Looman, 1986). Both methods are applied to phytosociological (presence/absence) data from a diluvial part of The Netherlands to answer the questions stated above, the first of which being reformulated as: 'are Ellenberg's indicator values internally consistent in our study area?'.

Table 1. Definition of Ellenberg's moisture values (Ellenberg, 1979).

1	on extremely dry soils, e.g. bare rocks
2	in-between 1 and 3
3	on dry soils
4	in-between 3 and 5
5	on fresh soils, i.e. under intermediate conditions
6	in-between 5 and 7
7	on moist soils which do not dry out
8	in-between 7 and 9
9	on wet, often not well aerated soils
10	on frequently inundated soils
11	water plant with leaves mostly in contact with the open air
12	underwater plant, mostly totally immersed in water
×	indifferent

Methods

Type of response curve

The relationship between the occurrence of a species and moisture may be shown in a *presence-absence response curve*, in which the probability $p(x)$ of occurrence of the species is plotted against moisture (x). Response curves may differ in shape and vary in complexity, but the response curves of species with respect to environmental variables are usually unimodal (Ellenberg, 1983; Whittaker, 1956). In this study we assume a unimodal response curve for each species with respect to moisture. In such curves, the width of the curve is proportional to the ecological amplitude and the position of its maximum is the indicator value. These two concepts lose their meaning in other response curves, such as bimodal or sigmoid curves.

Weighted averaging method

In the method of weighted averaging the *indicator value and ecological amplitude* of a species are defined as the *mean* (M) and *standard deviation* (SD) of the species' response curve. Thus, these characteristics are defined as if a response curve $p(x)$ were a statistical probability distribution (see Ter Braak & Barendregt, 1986). The moisture value of a relevé is estimated here as the average of Ellenberg's indicator values for moisture of all the species present in the relevé. Simplistic estimates of a species' indicator value and ecological amplitude would then be the sample mean and the sample standard deviation, respectively, of the moisture values of all relevés containing the species (Ter Braak & Looman, 1986). The newly calculated indicator values might then be compared with the indicator values given by Ellenberg (1979) to provide an informal test on the internal consistency of the latter. However, these estimates are too simple, because they neglect the distribution of the moisture values and their results may be misleading (Ter Braak & Looman, 1986). In an attempt to correct for the distribution of the moisture values, the moisture scale is divided into twelve classes, and the number of relevés, n_j, in each class j is counted. For any species a rough estimate of its response curve can then be obtained by calculating the fraction of relevés in each class that contain the species. These fractions can be displayed in a *response histogram* (Fig. 1). Improved estimates for the indicator value and ecological amplitude are then the mean and standard deviation of the response histogram. In this study the ecological amplitude is estimated in a slightly more subtle way, namely by

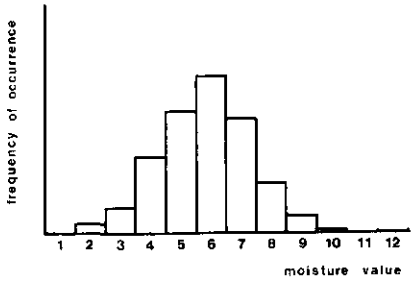


Fig. 1. Response histogram of a species with respect to moisture. The bars show the fraction of relevés in each moisture class which contain the species.

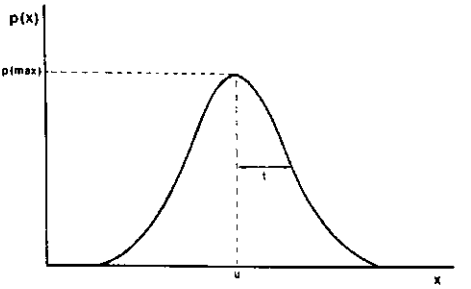


Fig. 2. Gaussian-logit response curve (p(x) = probability of occurrence of the species at value x, p(max) = maximum probability of occurrence, x = environmental variable, t = tolerance, u = optimum).

using Ellenberg's indicator value of the species instead of the sample mean in the formula for the standard deviation:

$$\widetilde{SD}^2 = \sum_{i=1}^{n} \frac{y_i}{n_j}(\bar{x}_i - M_o)^2 / \sum_{i=1}^{n} \frac{y_i}{n_j} \qquad (1)$$

where n is the number of relevés, y_i=1 or 0 depending on whether the species is present or absent in relevé i, $\bar{x}_i$ is the estimated moisture value and j the class of relevé i, n_j is the number of relevés in class j and M_o is Ellenberg's (1979) indicator value of the species. The latter is used in equation (1), instead of any newly computed indicator value, to avoid underestimation of the ecological amplitude. We also used some variants of equation (1), but the differences in the results did not seem to be of practical importance.

Maximum likelihood method

Ter Braak & Looman (1986) proposed to model the presence-absence response curve of a species by the Gaussian logit curve, in which the logit-transform of probability is a quadratic function. According to this model the probability p_{ik} that species k occurs in relevé i is (Fig. 2)

$$p_{ik} = 1/[1 + c_k \exp\ [½\ (x_i - u_k)^2/t_k^2]] \qquad (2)$$

where u_k is the optimum (the value of x with highest probability of occurrence of species k) and t_k is the tolerance (a measure of ecological amplitude) of species k and x_i is the moisture value of relevé i. The maximum probability of occurrence of species k is $1/(1 + c_k)$. The Gaussian logit curve is symmetric. Its optimum is therefore identical to its mean. Also, its tolerance is almost identical to its standard deviation when the maximum of the curve is small (Ter Braak & Looman, 1986). The range of occurrence of a species is largely restricted to an interval of length $4t$ (Fig. 2).

The idea behind the maximum likelihood method is to fit Gaussian logit curves to the relevé data. This is done by varying the parameter values of the model in order to maximize the likelihood. The likelihood of a set of parameter values is defined as the probability of collecting the same data when this set of values were the true set of parameter values. In the present case the likelihood is taken to be the product of $p^y(1-p)^{1-y}$ over all relevés and species, with $p=p_{ik}$ and y=1 or 0 depending on whether species k is present or absent in relevé i. Logistic regression as utilized by Ter Braak & Looman (1986) is a special case of the maximum likelihood method, in which the species parameters (u_k, t_k and c_k) are estimated from data on species occurrence and known values of x_i. We could apply logistic regression here, using the moisture values from the weighted averaging method. However, in estimating the tolerances of the species it is more natural to assume, as in equation (1), that the optima are known, namely, that they are equal to Ellenberg's indicator values. From this assumption maximum likelihood estimates are derived for the moisture values of the relevés as well as for the tolerances and maxima of the species. The maximum likelihood estimates are obtained with an iterative algorithm:

(1) Start with the moisture values obtained by weighted averaging.

(2) Estimate the tolerance and maximum of each species from that species' data and the current moisture values.

(3) Estimate a new moisture value for each relevé from the floristic data, the species' optima and the current values for the tolerances and maxima of the species.

(4) Check whether the moisture values have changed, and if so, go back to step (2), otherwise stop.

In step (2) and step (3) the likelihood is maximized for each species and each relevé separately and, as a result, the total likelihood increases with each step. Step (2) resembles a Gaussian logit regression, but differs in that the optimum is given instead of being estimated. Step (3) of the maximum likelihood procedure has the attractive property that species with a small tolerance will have a greater effect on the estimation of the moisture value of a relevé than species with a large tolerance (cf. Ter Braak & Barendregt, 1986).

With the maximum likelihood method one can test statistically whether a species' optimum as specified by Ellenberg's indicator value is consistent with the indicator values of the other species. In this test the likelihood calculated above is compared with

a likelihood that is maximized also with respect to the value of the species' optimum (cf. Ter Braak & Looman, 1986). When the difference in residual deviance (= −2 log-likelihood) is larger than the critical value of a chi-square distribution with 1 degree of freedom, the species' optimum is shown to differ significantly from the value specified by Ellenberg (1979) and is therefore inconsistent with the indicator values of the other species. In principle this test can be carried out for each species in turn. However, in the present case, the test is very laborious because of the large number of parameters in the model. Because it is unlikely that the moisture values of the relevés will change much, when the second likelihood is maximized, they may just as well be kept fixed. Then, the statistical test amounts to comparing a species' indicator value with its optimum as estimated by a Gaussian logit regression of the data of this particular species on fixed moisture values. Instead of testing by deviance, we checked whether Ellenberg's indicator value lay within the 95%-confidence interval for the optimum. The construction of this interval is described by Ter Braak & Looman (1986). Such intervals were only constructed for species occurring in more than five relevés.

Data

In this study, 1041 relevés (all from 1980–1982) were used representing the vegetation of the diluvial area in the western part of the province of Noord-Brabant, The Netherlands (Gremmen *et al.*, 1985) as follows: 323 relevés of woodland, 312 grassland, 250 marsh and ditch vegetation, 94 heathland and bog, and 62 other types. Quadrat size ranged from 4 m^2 in bog and grassland to 200 m^2 in woodlands.

Trees, large shrubs, and species that occurred less than 3 times were excluded. A total of 311 species remained, on average 13 per relevé; 280 of them had been assigned indicator values for moisture (Ellenberg, 1979). Most species have indicator values that are in the middle range (5–9). Of the species with more extreme moisture values 12% have an indicator value of 4 or less, and 16% have one above 9.

Results

The moisture values of the relevés estimated by the weighted averaging method showed a markedly uneven distribution, with many more 'wet' than 'dry' relevés (Table 2). These moisture values were strongly correlated (r=0.94) with those estimated by the maximum likelihood method, but as shown in Table 2, the estimated values for any single relevé may differ considerably (30% of the relevés differed by more than 0.5 unit, and 9% of the relevés by more than 1 unit).

Table 2. Comparison of the estimates of the moisture values of the relevés resulting from the weighted averaging method (x_{WA}) and the maximum likelihood method (x_{ML}). Entries refer to number of relevés.

x_{WA} / x_{ML}	1	2	3	4	5	6	7	8	9	10	11	12	Total
1				1									1
2					3								3
3			1	11	5	2							19
4				1	13	1							15
5				1	67	36	1						105
6					28	220	61						309
7						15	122	13					150
8							43	99	6				148
9							16	108	76	4			204
10								3	24	10			37
11									7	19	11	1	38
12											10	2	12
Total	0	0	1	14	116	274	243	223	133	33	21	3	1041

Table 3. Comparison of the estimates of the species amplitudes from the weighted averaging method (SD, Equation (1)) and the maximum likelihood method (t, Equation (2)). Entries refer to number of species.

SD / t	0.0 – 0.5	0.5 – 1.0	1.0 – 1.5	1.5 – 2.0	2.0 – 2.5	2.5 – 3.0	3.0 – 3.5	≥3.5	Total
0.0 – 0.5	6	30	1						37
0.5 – 1.0		30	20	1					51
1.0 – 1.5		18	51	7	4				80
1.5 – 2.0		1	32	27	2				62
2.0 – 2.5			7	13	4	1			25
2.5 – 3.0			2	3		2			7
3.0 – 3.5			1	2	2				5
≥3.5		1	1	5	1	1	1	3	13
Total	6	80	115	58	13	4	1	3	280

The simplistic estimate of a species' amplitude, that is the sample standard deviation (*SD*) of the moisture values of the relevés in which the species occurs, showed low correlation (0.2) with the more subtle estimate of *SD* by equation (1), which was on average 1.3 moisture scale unit. The maximum likelihood method tended to result in somewhat larger estimates of the amplitude than *SD* (Table 3). Species with indicator values of 11 and 12 had on average a markedly smaller tolerance than other species. This may be so because they are water plants.

In general the maximum probability of occurrence of a species estimated by the maximum likelihood method, was quite small; for only 23 (8%) of the species the maximum exceeded 0.50 and for 154 (55%) it was less than 0.10. Thus, the occurrence of most species cannot be predicted with confidence from the moisture value of the site alone.

Figure 3 shows some typical examples of the response histograms and Gaussian logit curves fitted by the maximum likelihood method and by Gaussian logit regression.

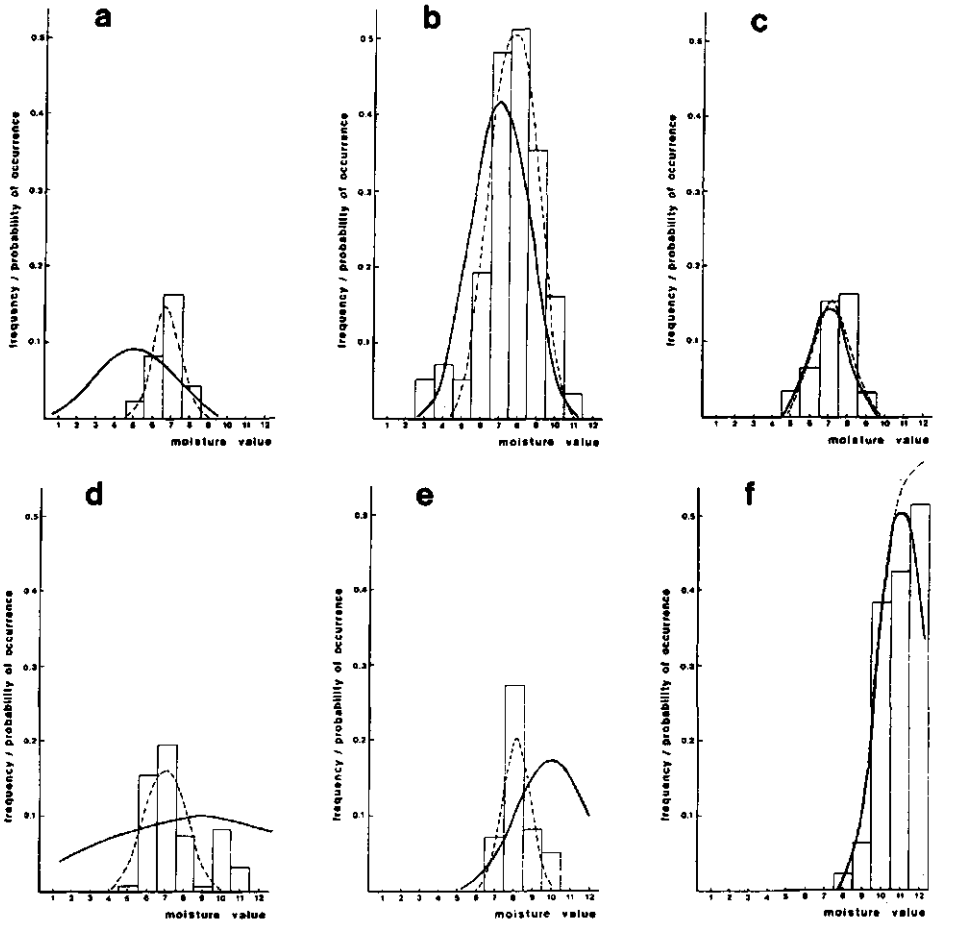


Fig. 3. Some examples of response histograms (bars) and estimated response curves. (— = response curve estimated by the maximum likelihood method; – – – = response curve estimated by Gaussian logit regression). F = Ellenberg moisture value, t = tolerance / estimated by the maximum likelihood method, SD = ecological amplitude estimated by the weighted averaging method. (a) *Heracleum sphondylium* $F = 5$; $t = 2.1$; $SD = 1.7$ (b) *Juncus effusus* $F = 7$; $t = 1.4$; $SD = 1.4$ (c) *Juncus subuliflorus* $F = 7$; $t = 1.0$; $SD = 0.9$ (d) *Alopecurus geniculatus* $F = 9$; $t = 5.5$; $SD = 2.2$ (e) *Iris pseudacorus* $F = 10$; $t = 1.8$; $SD = 1.8$ (f) *Lemna minor* $F = 11$; $t = 1.0$; $SD = 1.0$.

In some cases the mean of the response histogram deviates strongly from the indicator value of the species (Fig. 3a, e). In those cases the curve fitted by maximum likelihood with the species' indicator value taken as a fixed optimum, also deviates strongly from both the response histogram and the curve fitted by Gaussian logit regression. By using Gaussian logit regression 95%-confidence intervals for the optimum could be constructed for 175 (=73%) of the 240 species occurring in more than five relevés. For 45 (=26%) of these, Ellenberg's (1979) indicator value for moisture lay more than 0.5 unit outside this confidence interval. The extra 0.5 unit was used to allow for the fact that Ellenberg (1979) reports whole numbers. Thus for instance, an indicator value of 6.45 would be reported as 6. The indicator values of these species therefore are inconsistent with those of the other species. Table 4A gives information on species with an extreme deviation (≥ 1.7) between the Ellenberg moisture value and the estimated optimum. When no 95%-confidence interval could be calculated, the relationship between moisture and probability of occurrence was either non-significant (21 species) or sigmoid rather than unimodal (44 species), as judged by the deviance test at the 5%-level (cf. Ter Braak & Looman, 1986). No great inconsistencies in indicator value could be shown for species with a sigmoid relationship, because those with an Ellenberg indicator value of less than 7, showed a decreasing fitted response curve and those with an indicator value of 7 or more showed an increasing response curve (cf. Fig. 3f). It should be noted that a nonsignificant relationship or optimum may be due to a low frequency of a species in our data set and does not necessarily point to inconsistencies in Ellenberg's indicator values.

Gaussian logit regression was also used to check whether species Ellenberg (1979) considered indifferent, were also indifferent in our data set. For 28 of the 38 such species that occurred in 6 or more relevés, a 95%-confidence interval for the optimum could be calculated, and for 14 species the estimated tolerance was even less than 1.0 unit. Table 4B lists the species with the narrowest ecological amplitude ($t < 0.9$).

Our data set contained only three herbaceous species not mentioned by Ellenberg (1979) that occurred in more than 5 relevés; their indicator values were estimated by Gaussian logit regression (Table 4C).

Table 4. Ellenberg moisture value (*F*), estimated optimum, 95%-confidence interval for the optimum and estimated amplitude (tolerance) of a number of species. A. Species with a large discrepancy between Ellenberg moisture value and estimated optimum. B. Species with a narrow ecological amplitude, although regarded as indifferent by Ellenberg. C. Species not mentioned by Ellenberg.

Species name	*F*	optimum	interval	tolerance
A.				
Ornithopus perpusillus	2	4.1	3.5 – 4.3	0.5
Stellaria graminea	4	6.0	5.0 – 6.4	1.4
Alopecurus geniculatus	9	6.9	6.6 – 7.1	1.1
Iris pseudacorus	10	8.2	8.0 – 8.4	0.7
B.				
Anemone nemorosa	x	6.1	6.0 – 6.3	0.2
Melampyrum pratense	x	6.3	4.6 – 7.3	0.5
Bellis perennis	x	6.6	6.3 – 6.8	0.6
Prunella vulgaris	x	7.1	6.8 – 7.5	0.6
Ranunculus acris	x	6.9	6.7 – 7.0	0.7
Capsella bursa-pastoris	x	4.6	3.5 – 5.0	0.8
C.				
Eleocharis multicaulis	?	9.1	9.0 – 9.3	0.3
Epilobium obscurum	?	7.0	6.9 – 7.3	0.6
Myosotis laxa	?	7.8	7.5 – 8.1	0.7

Discussion

The ordinal scale of Ellenberg's indicator values

Ellenberg's indicator values are ordinal (strictly speaking values 11 and 12 are nominal); from the values in Table 1 we may infer which of two species prefers wetter conditions, but not the magnitude of the difference. But, in the methods applied here, the indicator values are treated as if they were quantitative, that is, as if they were measured on an interval scale. Durwen (1982) raised objections against such a quantitative treatment. In our opinion the ordinal nature of Ellenberg's moisture scale is far less important than the shape of the response curves, which should be symmetric (cf. Ter Braak & Barendregt, 1986). In the maximum likelihood method, a particular symmetric response curve was assumed – although response curves that are monotone by truncation, could also be dealt with. This condition of symmetry is equally important in the weighted averaging method, as mean and standard deviation are only useful characteristics for response curves that are more or less symmetric. After inspecting the response histograms of all species (cf. Fig. 3) we

concluded that the assumption of symmetry was not unreasonable, except, of course, for species with extreme optima. Therefore, we used the moisture indicator values of Ellenberg without transformation.

Comparison of the two methods

The weighted averaging method has three major problems. Firstly, as the number of relevés in each moisture class is not equal (Table 2), the estimates of the probability of occurrence in a class are not equally precise for all classes. The estimate of *SD* in equation (1) is closely related to the *SD* of the response histogram (Fig. 1), and it would seem reasonable to give less weight to classes with relatively few relevés. However, any such weighting policy, would make the estimator for *SD* again dependent on the distribution of the relevés, and thus cause bias.

A second problem is caused by relevés of extremely wet or extremely dry sites. The moisture values of these relevés will always be too low and too high, respectively, because only a few species indicate extreme conditions and many more species indicate conditions that are less extreme. Just by their numbers the probability of species of the latter group occurring at extreme sites is higher than of species indicating extreme conditions. This results in a general trend towards more moderate moisture values for extreme relevés, and this also results in a bias in the estimates for *SD*. Thirdly, the response histograms of species with an extreme indicator value will be truncated (cf. Fig. 3f) and it is not clear how the *SD* value of such species should be interpreted. The problem is partly one of definition, that is, when the response curve is truncated because more extreme conditions do not exist, it is not clear how *SD* should be defined, and partly one of estimation, namely when the response curve is truncated because more extreme conditions were not sampled, it is not clear how *SD* should be estimated. We do not know how to solve this problem in the weighted averaging method.

In the maximum likelihood method a specific model has to be adopted, in our case the Gaussian logit model. This is a disadvantage, since we do not really know the correct model. When the model is correct, the resulting estimates are better than in the weighted averaging method, but when it is incorrect, the meaning and quality of the estimates are unknown. We investigated the goodness-of-fit of the Gaussian logit curves obtained from the regressions with the usual chi-square test on the basis of observed and expected numbers of presence and absence in the 12 moisture classes. At the 5% level 72 species (=27%) showed significant lack-of-fit. An example is *Alopecurus geniculatus* (Fig. 3d). The response histogram suggests gross deviations from the Gaussian logit curve in moisture classes 10 and 11, but these are due to only four occurrences. The important deviation is the low frequency of occurrence in moisture class 9. Despite the deviations, we believe that for our purpose and data the Gaussian logit model is a good compromise between model complexity and goodness-of-fit.

The problems in the weighted averaging method are largely solved automatically in the maximum likelihood method, where a truncated response curve is assumed to be part of a full Gaussian logit curve. However, an unexpected new problem arose, namely that the distribution of the moisture values of the relevés showed local minima near integer values.

This artifact (which is not apparent in Table 2) is because the Ellenberg (1979) indicator values are all integer values and in our method form the optima of the species' response curves. The maximum likelihood estimate of the moisture value of a relevé is based both on the species present and the species absent. When a species is present, it forces the estimate in the direction of the species' indicator value, whereas, when a species is absent, it forces the estimate away from the species' indicator value. Absence of a species usually has far less influence than presence, that is, when the maximum probability of occurrence of the species is low (Ter Braak & Barendregt, 1986). But the number of species absent in a relevé is large compared to the number of species present. If, for instance, the true moisture value of a relevé is 6.0, all species with an indicator value of 6 that are absent will force the estimate away from the value 6.0 and this force cannot be counteracted by the presence of a small number of species with this same indicator value. The maximum likelihood estimate thus tends to avoid the integer values. We believe that in the present study this artifact is not a very serious problem. Because the average width of the response curves is large as compared to the scale of these irregularities, the fitting of curves will still give a reasonable estimate of the species tolerance.

The maximum likelihood method has the additional advantage over the weighted averaging method by giving approximate standard errors of estimates, which makes it possible to test the internal consistency of the Ellenberg indicator values.

Improving the indicator values by ordination?

Clausman (1980) attempted to improve indicator values by an iterative procedure; he calculated moisture values for the relevés from the indicator values and then new indicator values from the moisture values, and then new moisture values from the new indicator values, and so on. This procedure is essentially an ordination method. For example, when weighted averaging is used in each calculation, the method amounts to reciprocal averaging. By consequence, the original meaning of the indicator values may get lost.

We applied detrended correspondence analysis (Hill & Gauch, 1980), to our data and found practically no correlation between the (initial) moisture values of the relevés and the (final) scores on the first axis ($r=0.01$). The first axis turned out to be highly correlated ($r=0.99$) with the nitrogen values of the relevés, estimated by averaging the Ellenberg indicator values for N, whereas the second axis was highly correlated ($r=0.99$) with the moisture values. Applied to our data, Clausman's (1980) method would have changed the Ellenberg's indicator values for moisture into indicator values for nitrogen, which is clearly unwanted! Consequently, ordination cannot be used to improve indicator values, except in the hypothetical case that it is certain that the main variation in the species data corresponds exactly to the factor one wants to improve the indicator values of. Therefore, we kept the indicator values fixed in both our methods and tested each species separately to see if its value was consistent with the indicator values of the other species.

Due to the implicit nature of Ellenberg's definition of moisture, it is impossible to improve the moisture values in a statistically sound way on the basis of floristic data only.

On generalizing the results

Our results show the ecological amplitude (*SD* or tolerance) of a species to be about 1.0 to 1.5 units on Ellenberg's moisture scale. Consequently, the range of a species' occurrence is estimated to be on average 4–6 units. It is difficult to say how these results are affected by conditions specific to our study area. The detrended correspondence analysis showed nitrogen to be the environmental variable that is most important for explaining the floristic variation in our data. Consequently, the assumption in the maximum likelihood method of independence of the species is incorrect. Fortunately, nitrogen was practically uncorrelated with moisture, and therefore unlikely to have distorted the results to a large extent. The fact that moisture is shown to be the second most important environmental variable in our data set also gives some confidence in the results. In different geographical regions, the environmental variables that are most important for explaining the species distribution may differ. Especially when these factors are correlated with moisture, the estimates of the amplitude of a species with respect to moisture may differ because of distortion by these factors. In principle, the problem of other influential variables can be overcome in the maximum likelihood method by analysing more than one variable simultaneously. We may attempt this in the future.

Conclusion

The use of Ellenberg's moisture values on floristic data in estimating site moisture is an example of environmental calibration. Ellenberg's method of environmental calibration assumes a simple model of the responses of plant species to moisture: symmetric, unimodal response curves and equal amplitudes. This model does not include interaction effects of other environmental variables with moisture. A more precise calibration system necessarily has to include such interactions. Such a system could be derived from actual measurements of environmental variables and associated floristic data (Ter Braak & Barendregt, 1986), but would lose the simplicity and supposed general applicability of the Ellenberg system. May our results serve to increase the confidence with which Ellenberg's indicator values for moisture are used.

References

Böcker, R., Kowarik, I. & Bornkamm, R., 1983. Untersuchungen zur Anwendung der Zeigerwerte nach Ellenberg. Verh. Ges. Ökol. 9: 35–56.

Clausman, P. H. M. A., 1980. Ecologische interpretatie van vegetatieopnamen m.b.v. een computer. WLO-Meded. 7: 92–98.

Durwen, K.-J., 1982. Zur Nutzung von Zeigerwerten und artspezifischen Merkmalen der Gefäszpflanzen Mitteleuropas für Zwecke der Landschaftsökologie und -planung mit Hilfe der EDV-Voraussetzungen, Instrumentarien, Methoden und Möglichkeiten. Arbeitsber. Lehrst. Landschaftsökol. Münster 5: 1–138.

Ellenberg, H., 1979. Zeigerwerte der Gefäszpflanzen Mitteleuropas. 2nd ed., Scripta Geobotanica 9, Göttingen.

Ellenberg, H., 1983. Vegetation Mitteleuropas mit den Alpen in ökologischer Sicht. 3rd ed., Ulmer Verlag, Stuttgart.

Gremmen, N. J. M., Reijnen, M. J. S. M., Wiertz, J. & Van Wirdum, G., 1985. Modelling for the effects of ground-water withdrawal on the species composition of the vegetation in the Pleistocene areas of The Netherlands. In: Ann. Rep. 1984. Research Institute for Nature Management, Arnhem, pp 89–111.

Hill, M. O. & Gauch, H. G., 1980. Detrended correspondence analysis: an improved ordination technique. Vegetatio 42: 47–58.

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

Reijnen, M. J. S. M. & Wiertz, J., 1984. Grondwater en vegetatie: een nieuw systeem voor kartering en effectvoorspelling. (Engl. Summary) Landschap 1: 261–281.

Smeets, P. J. A. M., Werger, M. J. A. & Tevonderen, H. A. J., 1980. Vegetation changes in a moist grassland under altered water conditions. Biol. Conserv. 18: 123–142.

Ter Braak, C. J. F. & Barendregt, L. G., 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Biosci. 78: 57–72.

Ter Braak, C. J. F. & Looman, C. W. M., 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65: 3–11.

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

Accepted 20.8.1986.

A THEORY OF GRADIENT ANALYSIS

Cajo J.F. ter Braak[1]) and I. Colin Prentice[2])

[1]) TNO Institute of Applied Computer Science, Statistics Department Wageningen. The Netherlands

[2]) Institute of Ecological Botany, Uppsala University, Uppsala, Sweden

CONTENTS PAGE

I. INTRODUCTION

All species occur in a characteristic, limited range of habitats; and within their range, tend to be most abundant around their particular environmental optimum. The composition of biotic communities thus changes along environmental gradients. Successive species replacements occur as a function of variation in the environment, or (analogously) with successional time (Pickett, 1980; Peet and Loucks, 1977). The concept of niche space partitioning also implies the separation of species along "resource gradients" (Tilman, 1982). Gradients do not necessarily have physical reality as continua in space or time, but are a useful abstraction for explaining the distributions of organisms in space and time (Austin, 1985). Austin's review explores the interrelationships between niche theory and the concepts of ecological continua and gradients.

Our review concerns data analysis techniques that assist the interpretation of community composition in terms of species' responses to environmental gradients in the broadest sense. Gradient analysis sensu lato includes direct gradient analysis, in which each species' abundance (or probability of occurrence) is described as a function of measured environmental variables; the converse of direct gradient analysis, whereby environmental values are inferred from the species composition of the community; and indirect gradient analysis, sensu Whittaker (1967), in which community samples are displayed along axes of variation in composition that can subsequently be interpreted in terms of environmental gradients. There are close relationships among these three types of analysis. Direct gradient analysis is a regression problem - fitting curves or surfaces to the relation between each species'abundance or probability of occurrence (the response variable) and one or more environmental variables (the predictor variable(s)) (Austin, 1971). Inferring environmental values from species composition when these relationships are known is a calibration problem. Indirect gradient analysis is an ordination problem, in which axes of variation are derived from the total community data. Ordination axes can be considered as latent variables, or hypothetical environmental variables, constructed in such a way as to optimize the fit of the species data to a particular (linear or unimodal) statistical model of how species abundance varies along gradients (Ter Braak, 1985, 1987a). These latent variables are constructed without reference to environmental measurements, but they can subsequently be compared with actual environmental data if available. To these three well-known types of gradient analysis we add a fourth, constrained ordination, which has its roots in the psychometric literature on multidimensional scaling (Bloxom, 1978; De Leeuw and Heiser, 1980; Heiser, 1981). Constrained ordination also constructs axes of variation in overall community composition, but does so in such a way as to explicitly optimize the fit to supplied environmental data (Ter Braak, 1986a, 1987c). Constrained ordination is thus a multivariate generalization of direct gradient analysis, combining aspects of regression, calibration and ordination. Table 1 gives an arbitrary selection of literature references, chosen simply to illustrate the wide range of ecological problems to which each of the four types of gradient analysis has been applied; the reader is also referred to Gauch (1982), who includes an extensive bibliography, and to Gittins (1985).

Standard statistical methods that assume linear relationships among variables exist for all four types of problems (regression, calibration, ordination and constrained ordination) but have found only limited application in gradient analysis because of the generally non-linear,

Table 1. Selected applications of gradient analysis

	taxa	environmental variables	purpose of study
Regression			
Alderdice (1972)	marine fish	salinity, temperature	defining ranges
Peet (1978)	trees	elevation, moisture, latitude	biogeography
Wiens & Rotenberry (1981)	birds	vegetation structure	niche characterization
Austin et al. (1984)	*Eucalyptus* spp.	climatic indices	habitat characterization
Bartlein et al. (1986)	plant pollen types	temperature, precipitation	Quaternary palaeoecology
Calibration			
Chandler (1970)	benthic macro-invertebrates	water pollution	water quality management
Imbrie & Kipp (1971)	foraminifera	sea surface temperature	palaeoclimatic reconstruction
Sládecek (1973)	freshwater algae	organic pollution	ecological monitoring
Balloch et al. (1976)	benthic macro-invertebrates	water pollution	ecological monitoring
Ellenberg (1979)	terrestrial plants	soil moisture, N, pH	bioassay from vegetation
Van Dam et al. (1981)	diatoms	pH	acid rain effects
Böcker et al. (1983)	terrestrial plants	soil moisture, N, pH	bioassay from vegetation
Bartlein et al. (1984)	plant pollen types	temperature, precipitation	palaeoclimatic reconstruction
Battarbee (1984)	diatoms	pH	acid rain effects
Charles (1985)	diatoms	pH	acid rain effects
Atkinson et al. (1986)	beetles	summer temperature, annual range	palaeoclimatic reconstruction
Ordination[a]			
Van der Aart & Smeenk-Enserink (1975)	spiders	micro-environmental features	habitat characterization
Kooijman & Hengeveld (1979)	beetles	lutum content, elevation	habitat characterization
Wiens & Rotenberry (1981)	birds	vegetation structure	niche characterization
Prodon & Lebreton (1981)	birds	vegetation structure	niche characterization
Kalkhoven & Opdam (1984)	birds	habitat and landscape features	habitat characterization
Macdonald & Ritchie (1986)	plant pollen types	vegetation regions	Quaternary palaeoecology
Constrained ordination			
Webb & Bryson (1972)	plant pollen types	climate variables, airmass frequencies	palaeoclimatic reconstruction
Gasse & Tekaia (1983)	diatoms	pH classes	palaeolimnology
As (1985)	beetles	vegetation types	niche theory
Cramer & Hytteborn (1987)	terrestrial plants	time, elevation	land uplift effects
Purata (1986)	tropical trees	successional boundary conditions	study of secondary succession
Willén & Fängström (1986)	phytoplankton	physical/chemical variables	environmental monitoring

[a] [FOOTNOTE] excluding vegetation studies, where ordination is used routinely: see Gauch (1982) for a review.

non-monotone response of species to environmental variables. Ecologists have independently developed a variety of alternative techniques. Many of these techniques are essentially heuristic, and have a less secure theoretical basis. These heuristic techniques can nevertheless give useful results, and can be understood as approximate solutions to statistical problems similar to those solved by standard methods, but formulated in terms of a unimodal (Gaussian or similar) response model instead of a linear one. We present here a theory of gradient analysis, in which the heuristic techniques are integrated with regression, calibration, ordination and constrained ordination as distinct, well-defined statistical problems.

The various techniques used for each type of problem are classified into families according to their implicit response model and the method used to estimate parameters of the model. We consider three such families (Table 2). First we treat the family of standard statistical techniques based on the linear response model, because these are conceptually the simplest and provide a basis for what follows, even though their ecological application is restricted. Second, we outline a family of somewhat more complex statistical techniques which are formal extensions of the standard linear techniques and incorporate unimodal (Gaussian-like) response models explicitly. Finally we consider the family of heuristic techniques based on weighted averaging. These are not more complex than the standard linear techniques, but implicitly fit a simple unimodal response model rather than a linear one. Our treatment thus unites such apparently disparate data analysis techniques as linear regression, principal components analysis, redundancy analysis, Gaussian ordination, weighted averaging, reciprocal averaging, detrended correspondence analysis and canonical correspondence analysis in a single theoretical framework.

II. LINEAR MODELS

Species abundances may seem to change linearly through short sections of environmental gradients, so a linear response model may be a reasonable basis for analysing quantitative abundance data spanning a narrow range of environmental variation.

A. Regression

If a plot of the abundance (y) of a species against an environmental variable (x) looks linear, or can easily be transformed to linearity, then it is appropriate to fit a straight line by linear regression. The formula $y = a + bx$ describes the linear relation, with a the intercept of the line on the y-axis and b the slope of the line, or regression coefficient (Fig. 1). Separate regressions can be carried out for each of m species.

We are usually most interested in how the abundance of each species changes with a change in the environmental variable, i.e. in the slopes b_k (the index k refers to species k). If we first centre the data - by subtracting the mean of each species' abundances from the species data and the mean of the environmental values from the environmental data - the intercept disappears. Then if y_{ki} denotes the centred abundance of species k in the i-th out of n sites, and x_i the centred environmental value for that site, the response model for fitting the straight lines becomes

$$y_{ki} = b_k x_i + e_{ki} \qquad (1)$$

Table 2. Classification of gradient analysis techniques by type of problem, response model and method of estimation.

	RESPONSE MODEL: linear		unimodal
METHOD OF ESTIMATION:	least-squares	maximum likelihood	weighted averaging
TYPE OF PROBLEM:			
regression	multiple regression	Gaussian regression	weighted averaging of site scores (WA)
calibration	linear calibration; "inverse regression"	Gaussian calibration	weighted averaging of species scores (WA)
ordination	principal components analysis (PCA)	Gaussian ordination	correspondence analysis (CA); detrended correspondence analysis (DCA)
constrained[1] ordination	redundancy analysis (RDA)[4]	Gaussian canonical ordination	canonical correspondence analysis (CCA); detrended CCA
partial ordination[2]	partial components analysis	partial Gaussian ordination	partial correspondence analysis; partial DCA
partial constrained ordination[3]	partial redundancy analysis	partial Gaussian canonical ordination	partial canonical correspondence analysis; partial detrended CCA

1) = constrained multivariate regression
2) = ordination after regression on covariables
3) = constrained ordination after regression on covariables = constrained partial multivariate regression
4) = "reduced-rank regression" = "PCA of y with respect to x"

where e_{ki} is an error component with zero mean and variance v_{ki}. The standard estimator for the slope in equation (1) is

$$\tilde{b}_k = \sum_{i=1}^{n} y_{ki} x_i / s_x^2 \qquad (2)$$

where $s_x^2 = \sum_{i=1}^{n} x_i^2$. This is the least-squares estimator, which is the best linear unbiased estimator when errors are uncorrelated and homogeneous across sites ($v_{ki} = v_k$). It is also the maximum likelihood (ML) estimator when the errors are also normally distributed (the maximum likelihood method is a statistical method with well-established optimal properties (see e.g. Cox and Hinkley, 1974); a ML-estimator is the value for which, if it were the true value, the probability of the observed data is highest). The fitted lines can be used to predict the abundances of species in a site with a known value of the environmental variable simply by reading off the graph.

B. Calibration

We now turn to the inverse problem, calibration. When the relationship between the abundances of species and the environmental variable we are interested in is known, we can infer values of that environmental variable for new sites from the observed species abundances. If we took into account the abundance of only a single species, we could simply read off the graph, starting from a value on the vertical axis (Fig. 1). However, another species may well give a different estimate. We therefore need a good and unambigous estimator that combines the information from all m species. In terms of Eq. (1), the b_k are now assumed to be known and x_i is unknown. The role of the b_k and x_i have been interchanged. By interchanging their roles in Eq. (2) as well, we obtain

$$\tilde{x}_i = \sum_{k=1}^{m} y_{ki} b_k / s_b^2 \qquad (3)$$

where $s_b^2 = \sum_{k=1}^{m} b_k^2$. This is the least-squares estimator (and also the ML-estimator) when the errors follow a normal distribution and are independent and homogeneous across species ($v_{ki} = v_i$).

A problem with equation (3) is that these conditions are likely to be unrealistic, because effects of other environmental variables can cause correlation between the abundances of different species even after the effects of the environmental variable of interest have been removed. Further, the residual variance v_{ki} may be different for different species. If these conditions do not apply, we also need to take the residual correlations and variances into account. (In practice, the residual correlations and variances are estimated from the residuals of the regressions used for estimating the b_k's.) Searching for the maximum of the likelihood with respect to x_i then leads to a general weighted least-squares problem (Brown 1979, Brown 1982) that can be solved by using standard algorithms.

C. Ordination

After having fitted a particular environmental variable to the species data by regression, we might ask whether another environmental variable

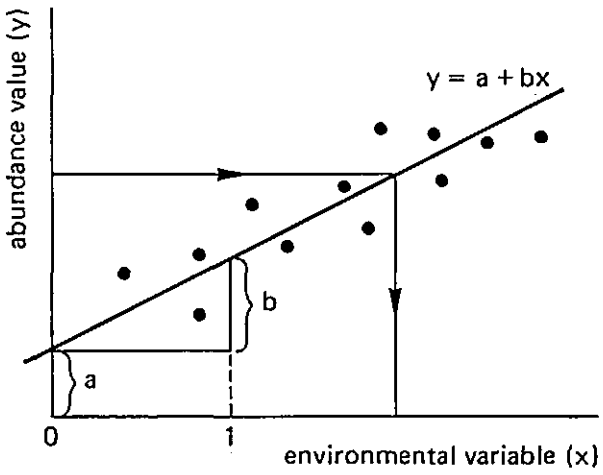


Fig. 1. A straight line displays the linear relation between the abundance value (y) of a species and an environmental variable (x), fitted to artifical data (•). (a = intercept; b = slope or regression coefficient).

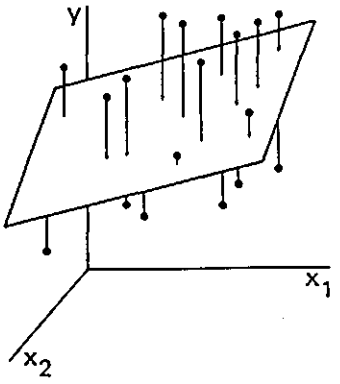


Fig. 2. A plane displays the linear relation between the abundance value (y) of a species and two environmental variables (x_1 and x_2), fitted to artifical data (•).

would provide a better fit. For some species one variable may fit better, and for other species another variable. To get an overall impression we might judge the goodness-of-fit (explanatory power) of an environmental variable by the "total regression sum of squares", the sum over all species of the regression sum of squares for each species (the sum of squares of the fitted values - see e.g. Montgomery and Peck, 1982). The question then arises: what is the best possible fit that is theoretically obtainable with the straight line model of Eq. (1)?

This question defines an ordination problem, i.e. to construct the single "hypothetical environmental variable" that gives the best fit to the species data according to Eq. (1). This hypothetical environmental variable is termed the latent variable, or simply the (first) ordination axis. Principal components analysis (PCA) provides the solution to this ordination problem. In Eq. (1), x_i is then the score of site i on the latent variable, b_k is the slope for species k with respect to the latent variable (also called the species loading or species score) and the eigenvalue of the first PCA axis is equal to the goodness-of-fit, i.e. the total sum of squares of the regressions of the species abundances on the latent variable. PCA provides the least-squares estimates of the site and species scores: these estimates are also ML estimates if the errors are independently and normally distributed with constant variance ($v_{ki} = v$).

PCA is usually performed using a standard computer package, but several different algorithms can be used to do the same job. The following algorithm, known as the power method (Gourlay and Watson, 1973), shows that PCA can be obtained by an alternating sequence of linear regressions and calibrations:

Step 1. Start with some (arbitrary) initial site scores $\{x_i\}$ with zero mean.
Step 2. Calculate new species scores $\{b_k\}$ by linear regression (Eq. (2)).
Step 3. Calculate new site scores $\{x_i\}$ by linear calibration (Eq. (3)).
Step 4. Remove the arbitrariness in scale by standardizing the site scores as follows: new x_i = old $x_i\sqrt{n}/s_x$, with s_x as defined beneath Eq. (2).
Step 5. Stop on convergence, i.e. when the newly obtained site scores are close to the site scores of the previous cycle of iteration, else go to step 2.

The final scores obtained in this way do not depend on the initial scores.

D. Extension to more than one environmental variable

Species experience the effect of more than one environmental variable simultaneously, so more than one variable may be required to account for variation in species abundances.

The joint effect of two environmental variables on a species can be analysed by multiple regression (see e.g. Montgomery and Peck, 1982). For two environmental variables the linear response model is

$$y_{ki} = a_k + b_{k1}x_{i1} + b_{k2}x_{i2} + e_{ki} \qquad (4)$$

with a_k the intercept for species k, x_{i1} the value of variable 1 at site i and b_{k1} the (partial) regression coefficient for the effect on species k. For variable 2, x_{i2} and b_{k2} are defined analogously, and y_{ki} and e_{ki} are defined as before. This model specifies a plane in three dimensions (Fig.

2). Standard computer packages are available to obtain least-squares (ML) estimates for the regression coefficients. Multiple regression provides for each variable a regression coefficient that takes into account the effect of the other variables: hence the term "partial" regression coefficient. Only when the two environmental variables are uncorrelated will the partial regression coefficients be identical to the coefficients estimated by separate regressions using Eq. (1).

The inverse problem, multiple calibration - inferring values of more than one environmental variable simultaneously - has been given surprisingly little attention in the literature. However, Williams (1959) derived the necessary formulae from the ML-principle; see also Brown (1982).

The ordination problem for the two-dimensional linear model turns out to be relatively simple, compared with the regression and calibration problems. The solution does not need an alternating sequence of multiple regressions and calibrations, because the latent variables can always be chosen in such a way that they are uncorrelated; and if the latent variables are uncorrelated, then the multiple regressions and calibrations reduce to a series of separate linear regressions and calibrations. PCA provides the solution to the linear ordination problem in any number of dimensions; one latent variable is derived first, as in the one-dimensional case of Eq. (1), and the second latent variable can be obtained next by applying the same algorithm again but with one extra step - after step 3, the trial scores are made uncorrelated with the first latent variable. On denoting x_{i2} simply by x_i, this orthogonalization is computed by

Step 3b: Calculate $f = \Sigma_i \, x_i x_{i1}/n$,
Calculate new x_i = old $x_i - f \, x_{i1}$.

(Further latent variables (ordination axes) may be derived analogously.) As in the one-dimensional case, PCA provides the ML-solution to the multi-dimensional linear ordination problem if the errors are independently and normally distributed with constant variance across species and sites. The power algorithm for PCA as described above makes its relationship to regression and calibration clear in a way that the usual textbook treatment, in terms of singular value decomposition of inner product matrices, does not; it also facilitates comparison with correspondence analysis, which we discuss later. Jolliffe (1986) reviews the theory and applications of PCA.

E. The environmental interpretation of ordination axes (indirect gradient analysis)

In indirect gradient analysis the species data are first subjected to ordination, e.g. using PCA, to find a few major axes of variation (latent variables) with a good fit to the species data. These axes are then interpreted in terms of known variation in the environment, often by using graphical methods (Gauch, 1982). A more formal method for the second step in indirect gradient analysis would be to calculate correlation coefficients between environmental variables and each of the ordination axes. This analysis is similar to performing a multiple regression of each separate environmental variable on the axes (Dargie, 1984), because the axes are uncorrelated. But the result is still not an analysis of the combined effects of all environmental variables. Such a joint analysis can be carried out by multiple regression of each ordination axis on the environmental variables, i.e. estimating the coefficients c in the model

$$x_i = c_0 + \sum_{j=1}^{q} c_j z_{ij} \qquad (5)$$

in which x_i is the score of site i on that one ordination axis, z_{ij} denotes the value at site i of the j-th out of q actual environmental variables, and c_j is the corresponding regression coefficient. (For later reference, the error term in Eq. (5) is not shown.) The multiple correlation coefficient R measures how well the environmental variables explain the ordination axis.

F. Constrained ordination (multivariate direct gradient analysis)

Indirect gradient analysis, as outlined above, is a two-step approach to relate species data to environmental variables. A few ordination axes that summarize the overall community variation are extracted in the first step; then in the second step one may calculate weighted sums (linear combinations) of the environmental variables that most closely fit each of these ordination axes. However, the environmental variables that have been studied may turn out to be poorly related to the first few ordination axes, yet may be strongly related to other, "residual" directions of variation in species composition. Unless the first few ordination axes explain a very high proportion of the variation, this residual variation can be substantial, and strong relationships between species and environment can potentially be missed.

In constrained ordination this approach is made more powerful by combining the two steps into one. The idea of constrained ordination is to search for a few weighted sums of environmental variables that fit the data of all species best, i.e. that give the maximum total regression sum of squares. The resulting technique, redundancy analysis (Rao, 1964; Van den Wollenberg, 1977), is an ordination analysis in which the axes are constrained to be linear combinations of the environmental variables. These axes can be found by extending the algorithm of PCA described above with one extra step, to be performed directly after step 3 (Ter Braak, 1987a):

Step 3a: Calculate a multiple regression of the site scores $\{x_i\}$ on the environmental variables (Eq. (5)), and take as new site scores the fitted values of this regression.

The regression is thus carried out within the iteration algorithm, instead of afterwards. On convergence, the coefficients $\{c_j\}$ are termed canonical coefficients and the multiple correlation coefficient in step 3a can be called the species-environment correlation.

Redundancy analysis is also known as reduced-rank regression (Davies and Tso, 1982), PCA of y with respect to x (Robert and Escoufier, 1976) and two-block mode C partial least squares (Wold, 1982). It is intermediate between PCA and separate multiple regressions for each of the species: it is a constrained ordination, but it is also a constrained form of (multivariate) multiple regression (Davies and Tso, 1982; Israëls, 1984). By inserting Eq. (5) into Eq. (1), it can be shown that the 'regression' coefficient of species k with respect to environmental variable j takes the simple form $b_k c_j$. With two ordination axes this form would be, in obvious notation, $b_{k1}c_{j1} + b_{k2}c_{j2}$. With two ordination axes, redundancy analysis thus uses $2(q+1)$ parameters to describe the species data, whereas the multiple regressions use $m(q+1)$ parameters (cf. Eq. (4)). One of the attractive features of redundancy analysis is that it leads to an ordination diagram that simultaneously displays (i) the main pattern of community variation as far as this variation can be explained by the environmental variables, and (ii) the main pattern in the correlation coefficients between the species and each of the environmental variables. We give an example of such a diagram later on.

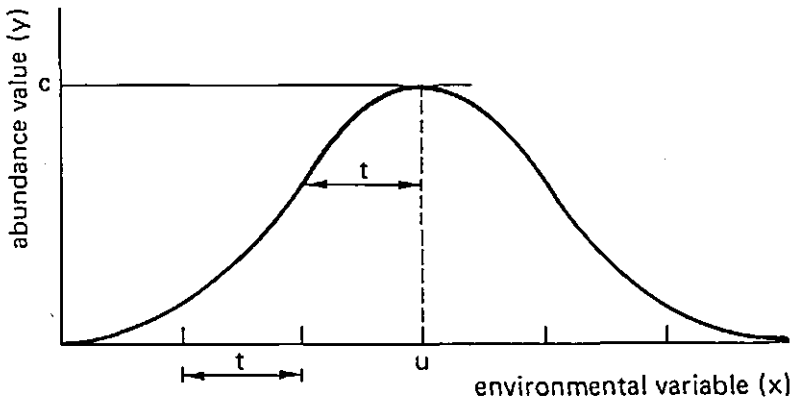


Fig. 3. A Gaussian curve displays a unimodal relation between the abundance value (y) of a species and an environmental variable (x). (u = optimum or mode; t = tolerance; c = maximum = exp(a)).

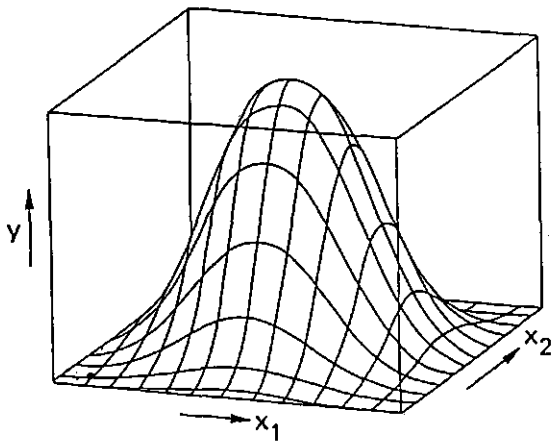


Fig. 4. A Gaussian surface displays a unimodal relation between the abundance value (y) of a species and two environmental variables (x_1 and x_2).

Redundancy analysis is much less well known than canonical correlation analysis (Gittins, 1985), which is the standard linear multivariate technique for relating two sets of variables (in our case, the set of species and the set of environmental variables). The most important practical difference is that redundancy analysis can analyse any number of species whereas in canonical correlation analysis the number of species (m) must be less than n-q; the latter restriction is often a nuisance.

Canonical variates analysis, or multiple discriminant analysis, is simply the special case of canonical correlation analysis in which the "environmental" variables are a series of dummy variables reflecting a single-factor classification of the samples. A similar restriction on the number of species thus also applies to canonical variates analysis. Redundancy analysis with dummy variables provides an alternative to canonical variates analysis, evading this restriction.

III. NONLINEAR (GAUSSIAN) METHODS

A. Unimodal response models

Linear methods are appropriate to community analysis only when the species data are quantitative abundances (with few zeroes) and the range of environmental variation in the sample set is narrow. Alternative analytical methods can be derived from unimodal models.

A unimodal response model for one environmental variable can be obtained by adding a quadratic term (x_i^2) to the linear model, changing the response curve from a straight line into a parabola. But this quadratic model can predict large negative values, whereas species abundances are always zero or positive. A simple remedy for the problem of negative values is provided by the Gaussian response curve (Gauch and Whittaker, 1972) in which the logarithm of species abundance is a quadratic in the environmental variable:

$$\begin{aligned} \log y &= b_0 + b_1 x + b_2 x^2 \\ &= a - {}^1/_2\ (x-u)^2/t^2 \end{aligned} \qquad (6A)$$

where $b_2 < 0$ (otherwise the curve would have a minimum instead of a mode). The coefficients b_0, b_1 and b_2 are most easily interpreted by transformation to u, t, and a (Fig. 3) -- u being the species' optimum (the value of x at the peak), t being its tolerance (a measure of response breadth or ecological amplitude), and a being a coefficient related to the height of the peak (Ter Braak and Looman, 1986).

A closely related model can describe species data in presence-absence form. In analysing presence-absence data, we want to relate probability of occurrence (p) to environment. Probabilities are never greater than 1, so rather than using Eq. (6A) we use the Gaussian logit model,

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x + b_2 x^2 \qquad (6B)$$

which is very similar to the Gaussian model unless the peak probability is high (> 0.5); then Eq. (6B) gives a curve that is somewhat flatter on top. The coefficients b_0, b_1 and b_2 can be transformed as before into coefficients representing the species' optimum, tolerance and maximum probability value.

Although real ecological response curves are still more complex than implied by the Gaussian and Gaussian logit models, these models are nevertheless useful in developing statistical descriptive techniques for data showing mostly unimodal responses, just as linear models are useful in statistical analysis of data that are only approximately linear.

With two environmental variables, Eqs. (6A) and (6B) become full quadratics with both square and product terms (Alderdice, 1972). For example, the Gaussian model becomes

$$\log y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_2 + b_4x_2^2 + b_5x_1x_2 \ . \qquad (7)$$

If $b_2 + b_4 < 0$, and $4b_2b_4 = b_5^2 > 0$ then Eq. (7) describes a unimodal surface with ellipsoidal contours (Fig. 4). If one of these conditions is not satisfied then Eq. (7) describes a surface with a minimum, or with a saddle point (e.g. Davison, 1983). Provided the surface is unimodal, its optimum (u_1, u_2) can be calculated from the coefficients in Eq. (7) by

$$u_1 = (b_5b_3 - 2b_1b_4)/d$$
$$u_2 = (b_5b_1 - 2b_3b_2)/d \qquad (8)$$

where $d = 4b_2b_4 - b_5^2$. When $b_5 \neq 0$, the optimum with respect to x_1, depends on the value of x_2; the environmental variables are then said to show interaction in their effect on the species. In contrast, when $b_5 = 0$ the optimum with respect to x_1 does not depend on the value of x_2 (no interaction) and Eq. (8) simplifies considerably (Ter Braak and Looman, 1986).

The unknown parameters of nonlinear response models in the context of regression, calibration or ordination can (at least in theory) be estimated by the maximum likelihood principle, however difficult this may be in a particular situation. Usually iterative methods are required, and initial parameter values must be specified. The likelihood function may have local maxima, so that different sets of initial parameter values may result in different final estimates. It cannot be guaranteed that the global maximum has been found. Further, all kinds of numerical problems may occur. However, the special cases of Gaussian and Gaussian logit response models do allow reasonably practical solutions, which we consider now.

B. Regression

The regression problems of fitting Gaussian or Gaussian logit curves are relatively straightforward, since these models are special cases of the Generalized Linear Model (for details see Austin and Cunningham, 1981; Dobson, 1983). If the data are abundances (which may include zeroes), the Gaussian model is fitted by specifying a Poisson error distribution and a logarithmic link function. If the data are presence-absence, the Gaussian logit model is fitted by specifying a Bernoulli error distribution and a logit link function. Alternatively, any statistical package that will do logit (= logistic) regression can be used to fit the Gaussian logit model (Ter Braak and Looman, 1986). No initial estimates are needed and local maxima do not arise, so these techniques are quite practical for direct gradient analysis.

The most common complication arises when the optimum for a species is estimated well outside the sampled range of environments, or if the fitted curve shows a minimum rather than a peak. These conditions suggest that the

regression is ill-determined and that it might be better to fit a monotone curve by setting $b_2 = 0$ in Eqs. (6); a statistical test can be used to determine whether this simplification is acceptable (Ter Braak and Looman, 1986). Such cases are bound to arise in practice because any given set of samples will include some species that are near the edge of their range.

C. Calibration

The calibration problem of inferring environmental values at sites from species data and known Gaussian (logit) curves by ML (Ter Braak and Barendregt, 1986) is feasible by numerical optimization, but no easy-to-use computer programs are available at present. Local maxima may occur in the likelihood, when the tolerances of the species are unequal, and one needs to specify an initial estimate. The assumption of independence of species responses is required, but might not be tenable in practice; it remains to be studied how important this assumption is. Dependency among species could most obviously be caused by the effects of additional, unconsidered environmental variables, in which case the best remedy would be to identify these variables and include them in the analysis.

D. Ordination

Ordination based on Gaussian (logit) curves aims to construct a latent variable such that these curves optimally fit the species data. This problem involves the ML estimation of site scores $\{x_k\}$ and the species' optima $\{u_k\}$, tolerances $\{t_k\}$ and maxima $\{a_k\}$, usually by an alternating sequence of Gaussian (logit) regressions and calibrations. This kind of ordination has been investigated by Gauch, Chase and Whittaker (1974), Kooijman (1977), Kooijman and Hengeveld (1979), Goodall and Johnson (1982) and Ihm and Van Groenewoud (1975, 1984). The numerical methods required are computationally demanding; and in the general case, when the tolerances of the species are allowed to differ, the likelihood function typically contains many local maxima.

E. Extension to more than one environmental variable

The effects of two environmental variables can be modelled by Gaussian or Gaussian logit surfaces (see Eq. (7)), which can be fitted by Generalized Linear Modelling or by logit regression (Austin and Cunningham,1981; Austin et al., 1984; Bartlein et al., 1986). Inferring the values of more than one environmental variable simultaneously on the basis of several such response surfaces is also possible in principle, but has not been applied as far as we know.

Kooijman (1977) and Goodall and Johnson (1982) reported numerical problems in their attempts to perform ML ordination using two-dimensional Gaussian-like models. A simple model with circular contours ($b_2 = b_4$ and $b_5 = 0$) may be amenable in practice, especially if b_2 is not allowed to vary among species (Kooijman, 1977). This model is equivalent to the "unfolding model" used by psychologists to analyse preference data (Coombs, 1964; Heiser, 1981; Davison, 1983; DeSarbo and Rao, 1984). But with more than two latent variables the Gaussian (logit) model with a second- degree polynomial as linear predictor contains so many parameters that it is likely to be difficult to get reliable estimates of them, even if all the interaction terms are dropped.

F. Constrained ordination

The constrained ordination problem for Gaussian-like response models is to construct ordination axes that are also linear combinations of the environmental variables, such that Gaussian (logit) surfaces with respect to these axes optimally fit the data. As in redundancy analysis (section II F), the joint effects of the environmental variables on the species are "channelled" through a few ordination axes which can be considered as composite environmental gradients influencing species composition. Ter Braak (1986a) refers to this approach as Gaussian canonical ordination, the word canonical being chosen in analogy with canonical correlation analysis. The estimation problem is actually simpler than in unconstrained Gaussian ordination, and is more easily soluble in practice because the number of parameters to be estimated is smaller: instead of n site scores one has to estimate q canonical coefficients. (Meulman and Heiser (1984) have applied similar ideas in the context of nonmetric multidimensional scaling.) Gaussian canonical ordination can also be viewed as multivariate Gaussian regression with constraints on the coefficients of the polynomial. In multivariate Gaussian regression each species has its own optimum in the q-dimensional space formed by the environmental variables; the constraints imposed in Gaussian canonical ordination amount to a requirement that these optima lie in a low-dimensional subspace. If the optima lie close to a plane then the most important species-environment relationships can be depicted graphically in an ordination diagram.

IV. WEIGHTED AVERAGING METHODS

Ecologists have developed alternative, heuristic methods that are simpler but have essentially the same aims as the methods of the previous section based on Gaussian-type models. Each method in the Gaussian family has a counterpart in the family of heuristic methods based on weighted averaging (WA). These methods have been used extensively, and even re-invented in different branches of ecology.

A. Regression

As a regression technique, WA is a method of estimating species' optima with respect to known environmental variables. When a species shows a unimodal relationship with environmental variables, the species' presences will be concentrated around the peak of this function. One intuitively reasonable estimate of the optimum is the average of the values of the environmental variable over those sites in which the species is present. With abundance data, WA applies weights proportional to species abundance; absences still carry zero weight. The estimate of the optimum for species k is thus

$$\bar{u}_k = \sum_{i=1}^{n} y_{ki} x_i / y_{k+} \qquad (9)$$

where y_{ki} is from now onwards the abundance (not centred) or presence/absence (1/0) of species k at site i, y_{k+} is the species total ($y_{k+} = \Sigma_i y_{ki}$) and x_i is the value of the environmental variable at site i.

As a follow-up to an investigation of the theoretical properties of this estimator (Ter Braak and Barendregt, 1986), Ter Braak and Looman (1986) showed by simulation of presence-absence data that WA estimates the optimum of a Gaussian logit curve as efficiently as the ML technique of Gaussian logit regression provided:

Condition 1a: The site scores $\{x_i\}$ are equally spaced over the whole range of occurrence of the species along the environmental variable.

WA also proved to be only a little less efficient whenever the distribution of the environmental variable among the sites was reasonably homogeneous (rather than strictly equally spaced) over the whole range of species occurrences, or more generally for species with narrow ecological amplitudes. But the estimate of the optimum of a rare species may be imprecise, because the standard error of the estimate is inversely proportional to the square root of the number of occurrences. So for efficiency, we also need

Condition 1b: the site scores $\{x_i\}$ are closely spaced in comparison with the species' tolerance.

B. Calibration

WA is also used in calibration, to estimate environmental values at sites from species' optima - which in this context are often called indicator values ('Zeigerwerte', Ellenberg, 1979) or scores (Whittaker, 1956). When species replace one another along the environmental variable of interest, i.e. have unimodal response functions with optima spread out along that variable, then species with optima close to the environmental value of a site will naturally tend to be represented at that site. Intuitively, to estimate the environmental value at a site, one can average the optima of the species that are present. With abundance data, the corresponding intuitive estimate is the weighted average,

$$\tilde{x}_i = \sum_{k=1}^{m} y_{ki} u_k / y_{+i} \qquad (10)$$

where y_{+i} is the site total ($y_{+i} = \Sigma_k y_{ki}$).

Ter Braak and Barendregt (1986) showed that WA estimates the value x_i of a site as well as the correponding ML techniques if the species show Gaussian curves and Poisson-distributed abundance values (or, for presence-absence data, show Gaussian logit curves), and provided:

Condition 2a: The species' optima are equally spaced along the environmental variable over an interval that extends for a sufficient distance in both directions from the true value x_i;
Condition 3: The species have equal tolerances;
Condition 4: The species have equal maximum values.

These conditions amount to a "species packing model" wherein the species have equal response breadth and equal spacing (Whittaker et al., 1973). The conditions may be relaxed somewhat (Ter Braak and Barendregt, 1986) without

seriously affecting the efficiency of the WA-estimate. When the optima are uniformly distributed instead of being equally spaced, the efficiency is still high if the maximum probabilities of occurrence are small (< 0.5). The species' maximum values may differ, but they must not show a trend along the environmental variable (for instance, leading to species- rich samples at one end of the gradient and species-poor samples at the other end). The efficiency of WA is worse if the tolerances substantially differ among species; a tolerance weighted version of WA, as suggested by Zelinka and Marvan (1961) and Goff and Cottam (1967), would be more efficient since it would give greater weight to species of narrower tolerance, which are more informative about the environment.
Under conditions 2a-4 above, the standard error of the estimate of $\bar{x}_i$ is approximately $t/\sqrt{y_{+i}}$, where t is the (common) species-tolerance. For the weighted average to be practically useful, the number of species encountered in a site should therefore not be too small (not less than five). We therefore need the extra condition (cf. Section 5 in Ter Braak and Barendregt, 1986):

Condition 2b: The species' optima must be closely spaced in comparison with their tolerances.

An alternative heuristic method of calibration is by "inverse regression". This is simply multiple linear regression of the environmental variable on the species abundances (Brown, 1982): the environmental variable is treated as if it were the response variable and the species abundances, possibly transformed, as predictor variables. The regression coefficients can be estimated from the training set of species abundances and environmental data, the resulting equations being applied directly to infer environmental values from further species abundance data. When applied to data on percentage composition, e.g pollen spectra or diatom assemblages (Bartlein et al., 1984; Charles, 1985), the method differs from WA calibration only in the way in which the species optima are estimated, since the linear combination of percentage values used to estimate the environmental value is by definition a weighted average of the regression coefficients.

C. Ordination

Hill (1973) turned weighted averaging into an ordination technique by applying alternating WA regressions and calibrations to a species-by-site data table. The algorithm of this technique of "reciprocal averaging" is similar to that given earlier for PCA:

Step 1. Start with arbitrary, but unequal, initial site scores $\{x_i\}$.
Step 2. Calculate new species scores $\{u_k\}$ by WA (Eq. (9)).
Step 3. Calculate new site scores $\{x_i\}$ by WA (Eq. (10)).
Step 4. Remove the arbitrariness in scale by standardizing the site scores by new $x_i = \{\text{old } x_i - z\}/s$ where $z = \Sigma_i y_{+i} x_i / \Sigma_i y_{+i}$ and

$$s^2 = \Sigma_i y_{+i} (x_i - z)^2 / \Sigma_i y_{+i} \qquad (11)$$

Step 5. Stop on convergence, else go to step 2.

As in PCA, the resulting site and species scores do not depend on the initial scores. The final scores produced by this reciprocal averaging algorithm form the first eigenvector or ordination axis of correspondence

analysis (CA), an eigenvector technique that is widely used especially in the French-language literature (Laurec et al., 1979; Hill, 1974). As with the power algorithm for PCA, the reciprocal averaging algorithm makes clear the relationship between CA and regression and calibration - this time, with WA regression and calibration. The method of standardization chosen in step 4 is arbitrary, but chosen for later reference. On convergence, s in step 4 is equal to the eigenvalue of the first axis, and lies between 0 and 1.

Correspondence analysis has many applications outside ecology. Nishisato (1980), Greenacre (1984) and Gifi (1981) provide a variety of different rationales for correspondence analysis, each adapted to a particular type of application. Heiser (1986) and Ter Braak (1985, 1987c) develop rationales for correspondence analysis that are particularly relevant to ecological applications.

Ter Braak (1985) showed that CA approximates ML Gaussian (logit) ordination under Conditions 1 to 4 listed above, i.e. under just these conditions for which WA is as good as ML-regression and ML-calibration. In practice CA can never be exactly equivalent to ML ordination, because Condition 1a implies that the range of site scores is broad enough to include the ranges of all of the species whereas Condition 2a implies that there must be species with their optima situated beyond the edge of the range of site scores. These conditions cannot both be satisfied if the range of site scores is finite. As a result, CA shows an edge effect: the site scores near the ends of the axes become compressed relative to those in the middle (Gauch, 1982). This effect becomes less strong, however, as the range of site scores becomes wider and the spacing of the site scores and species scores becomes closer relative to the average species' tolerance.

Conditions 1-4 also disallow "deviant" sites and rare species. CA is sensitive to both (Hill, 1974; Feoli and Feoli Chiapella, 1979; Oksanen, 1983). This sensitivity may be useful in some applications, but is a nuisance if the aim is to detect major gradients. Deviant sites (and, possibly, the rarest species) should therefore ideally be removed from the data before analysis by CA.

As in PCA, further ordination axes can be extracted in CA by adding an extra step after Step 3, making the trial scores on the second axis uncorrelated with the (final) scores on the first axis. (In the calculation of f in Step 3b (see section II D) the sites are weighted proportional to the site total y_{+i}. This weighting is implicitly applied from now on.) However, there is a problem with the second and higher axes in CA. The problem is the well-known but hitherto not well-understood "arch effect" (Hill, 1974). If the species data come from an underlying one-dimensional Gaussian model the scores on the second ordination axis show a parabolic ("arch") relation with those of the first axis; if the species data come from a two-dimensional Gaussian model in which the true site and species scores are located homogeneously in a rectangular region in 2D-space (the extension to two dimensions of Conditions 1a and 2a), the scores of the second ordination axis lie not in a rectangle but in an arched band (Hill and Gauch, 1980). The arch effect arises because the axes are extracted sequentially in order of decreasing "variance". Suppose CA has succeeded in constructing a first axis, such that species appear one after the other along that axis as in a species packing model. Then a possible second axis is obtained by folding the first axis in the middle and bringing the ends together so that it is a superposition of two species packing models, each with half the gradient length of the first axis. This folded axis is a candidate for becoming the second axis, because it has no linear correlation with the first CA-axis yet has as much as half the gradient length of the

first axis (Ter Braak, 1987a). The folded axis by itself thus "explains" a part of the variation in the species data, even though when taken jointly with the first axis it contributes nothing. Even if there is a strong second gradient, CA will not associate it with the second axis if it separates the species less than a folded first axis. As a result of the arch effect, the two-dimensional CA-solution is generally not a good approximation to the ML-solution (two-dimensional Gaussian ordination).

Hill and Gauch (1980) developed detrended correspondence analysis (DCA) as a heuristic modification of CA designed to remedy both the edge effect and the arch effect. The edge effect is removed in DCA by nonlinear rescaling of the axis. Assuming a species packing model with randomly distributed species' optima, Hill and Gauch (1980) noted that the variance of the optima of the species present at a site (the 'within-site variance') is an estimate of the average response curve breadth of those species (they used the standard deviation as a measure of breadth, which is about equal to tolerance as we define it). Because of the edge effect, the species curves before rescaling are narrower near the ends of the axis than in the middle, and the within-site variance is correspondingly smaller in sites near the ends of the axis than in sites in the middle. The rescaling therefore attempts to equalize the within-site variance at all points along the ordination axis by dividing the axis into small segments, expanding the segments with sites with small within-site variance, and contracting the segments with sites with large within-site variance. The site scores are then calculated as weighted averages of the species scores and the scores are standardized such that the within-site variance is equal to 1.

Hill and Gauch (1980) defined the length of the ordination axis to be the range of the site scores. This length is expressed in 'standard-deviation units' (SD). The tolerance of the species' curves along the rescaled axis are therefore close to 1, and each curve rises and falls over about 4 SD. Sites that differ by 4 SD can thus be expected to have no species in common. This interpretation of the length of the ordination axis is extremely useful. Even if nonlinear rescaling is not used, one can still set the average within-site variance of the species scores along a CA-axis equal to 1 by linear rescaling (Hill, 1979), so as to ensure that the length of the ordination axis still has approximately this interpretation.

The arch effect, a more serious problem in CA, is removed in DCA by the heuristic method of "detrending-by-segments". This method ensures that at any point along the first ordination axis, the mean value of the site scores on subsequent axes is approximately zero. In order to achieve this, the first axis is divided into a number of segments and the trial site scores are adjusted within each segment by subtracting their mean after some smoothing across segments. Detrending-by-segments is built into the reciprocal averaging algorithm, and replaces Step 3b. Subsequent axes are derived similarly by detrending with respect to each of the existing axes.

DCA often works remarkably well in practice (Hill and Gauch, 1980; Gauch et al., 1981). It has been critically evaluated in several recent simulation studies. Ter Braak (1985) showed that DCA gave a much closer approximation to ML Gaussian ordination than CA did, when applied to simulated data based on a two-dimensional species packing model in which species have identically shaped Gaussian surfaces and the optima and site scores are uniformly distributed in a rectangle. This improvement was shown to be mainly due to the detrending, *not* to the *nonlinear rescaling of axes*. Kenkel and Orlóci (1986) found that DCA performed substantially better than CA when the two major gradients differed in length, but also noted that DCA sometimes "collapsed and distorted" CA results when there were (a) few species per

site and (b) the gradients were long (we believe (a) to be the real cause of the collapse). Minchin (1987) further found that DCA can flatten out some of the variation associated with one of the underlying gradients. He ascribed this loss of information to an instability in the detrending-by-segments method. Pielou (1984, p. 197) warned that DCA is "overzealous" in correcting the "defects" in CA, and "may sometimes lead to the unwitting destruction of ecologically meaningful information". Minchin's (1987) results indicate some of the conditions under which such loss of information can occur.

DCA is popular among practical field ecologists, presumably because it provides an effective approximate solution to the ordination problem for a unimodal response model in two or more dimensions - given that the data are reasonably representative of sections of the major underlying environmental gradients. Two modifications might increase its robustness with respect to the problems identified by Minchin (1987). First, nonlinear rescaling aggravates these problems; since the edge effect is not too serious, we advise against the routine use of nonlinear rescaling. Second, the arch effect needs to be removed (as Heiser, 1986, also noted), but this can be done by a more stable, less "zealous" method of detrending which was also briefly mentioned by Hill and Gauch (1980): namely detrending-by-polynomials. Under the one-dimensional Gaussian model, it can be shown that the second CA-axis is a quadratic function of the first axis, the third axis is a cubic function of the first axis, and so on (Hill, 1974; Iwatsubo, 1984). Detrending-by-polynomials can be incorporated into the reciprocal averaging algorithm by extending step 3b such that the trial scores are not only made uncorrelated with the previous axes, but are also made uncorrelated with polynomials of the previous axes. The limited experience sofar suggests that detrending up to fourth-order polynomials should be adequate. In contrast with detrending-by-segments, the method of detrending-by-polynomials removes only specific defects of CA that are now theoretically understood.

D. Constrained ordination

Just as CA/DCA is an approximation to ML Gaussian ordination, so is canonical correspondence analysis (CCA) an approximation to ML Gaussian canonical ordination (Ter Braak, 1986a). CCA is a modification of CA in which the ordination axes are restricted to be weighted sums of the environmental variables, as in Eq. (5). CCA can be obtained from CA as redundancy analysis was obtained from PCA. An algorithm can be obtained by adding to the CA algorithm an extra multiple regression step. The only difference from Step 3a of redundancy analysis (section II F) is that the sites must be weighted in the regression proportional to their site total y_{+i} (Ter Braak, 1986a). CCA can also be obtained as the solution of an eigenvalue problem (Ter Braak, 1986a). It is closely related to "redundancy analysis for qualitative variables" (Israëls, 1984) but has a different rationale and is applied to another type of data.

In constrained ordination the constraints always become less strict as more environmental variables are included. If $q \geq n-1$, then there are no real constraints, and CA and CCA become equivalent. As in CA the edge effect in CCA is a minor problem that is best left untreated. Detrending may sometimes be required to remove the arch effect - i.e. to prevent CCA from selecting weighted sums of environmental variables that are approximately polynomials of previous axes. Detrending-by-segments does not work very well here for technical reasons; detrending-by-polynomials is better-founded and more appropriate (see Appendix and Ter Braak, 1987b). However, the arch

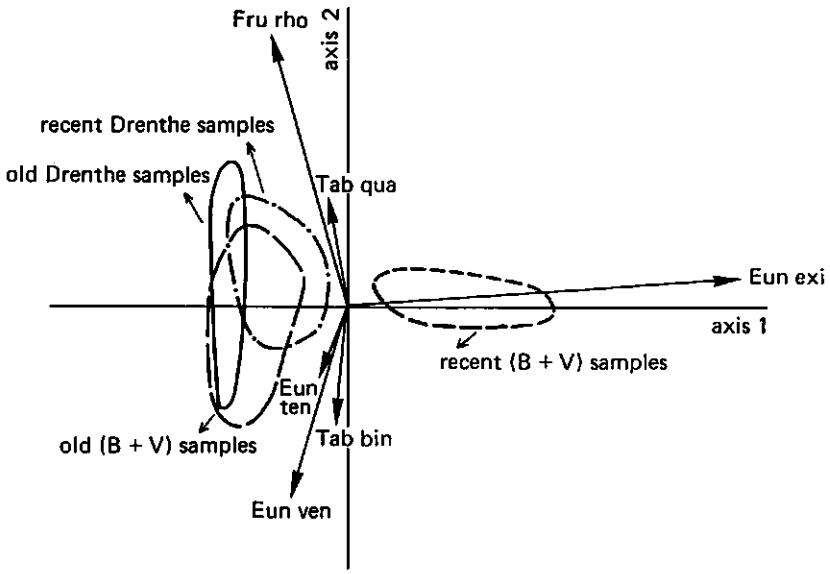


Fig. 5.

Biplot based on principal components analysis of diatom assemblages from Dutch moorland pools (schematic after from Van Dam et al. 1981). The arrows for the six most frequent species and the regions where different categories of samples lie jointly display the approximate community compostion in each of the regions (old = ca. 1920, recent = 1978; B+V = from the province of Brabant and the Veluwe). Abbreviations: Eun exi = Eunotia exigua, Eun ten = Eunotia tenella, Eun ven = Eunotia veneris, Fru rho = Frustulia rhomboides var. saxonica, Tab bin = Tabellaria binalis, Tab qua = Tabellaria quadriseptata.

effect in CCA can be eliminated much more elegantly, simply by dropping superfluous environmental variables (Ter Braak, 1987c). Variables that are highly correlated with the "arched" axis (often the second axis) are the most likely to be superfluous. If the number of environmental variables is small enough for the relationship of individual variables to the ordination axes to be significant, the arch effect is not likely to occur at all.

CCA can be sensitive to deviant sites, but only when they are outliers with regard to both species composition and environment. When realistically few environmental variables are included, CCA is thus more robust than CA in this respect too.

CCA leads to an ordination diagram that simultaneously displays (a) the main patterns of community variations, as far as these reflect environmental variation, and (b) the main pattern in the weighted averages (not correlations as in redundancy analysis) of each of the species with respect to the environmental variables (Ter Braak, 1986a, 1987c). CCA is thus intermediate between CA and separate WA calculations for each species. Geometrically, the separate WA calculations give each species a point in the q-dimensional space of the environmental variables, which indicates the centre of the species' distribution. CCA attempts to provide a low-dimensional representation of these centres; CCA is thus also a constrained form of WA, in which the weighted averages are restricted to lie in a low-dimensional subspace.

Like redundancy analysis, CCA can be used with dummy "environmental" variables to provide an ordination constrained to show maximum separation among pre-defined groups of samples. This special case of CCA is described, for example, by Feoli and Orlóci (1979) under the name of "analysis of concentration", by Greenacre (1984, section 7.1) and by Gasse and Tekaia (1983).

V. ORDINATION DIAGRAMS AND THEIR INTERPRETATION

The linear ordination techniques (PCA and redundancy analysis) and the ordination techniques based on WA (CA/DCA and CCA) represent community data in substantially different ways. We focus on two-dimensional ordination diagrams, as these are the easiest to construct and to inspect, and illustrate the interpretation of each type of diagram with an example.

A. Principal components: biplots

PCA fits planes to each species' abundances in the space defined by the ordination axes. The species' point (b_{k1}, b_{k2}) may be connected with the origin (0,0) to give an arrow. Such a diagram, in which sites are marked by points and species by arrows is called a "biplot" (Gabriel, 1971). There is a useful symbolism in this use of arrows: the arrow points in the direction of maximum variation in the species' abundance, and its length is proportional to this maximum rate of change. Consequently, species on the edge of the diagram (far from the origin) are the most important; species near the centre are of minor importance. (Ter Braak (1983) provides more detailed, quantitative rules for interpreting PCA ordination diagrams.)

Van Dam et al. (1981) applied PCA to data consisting of diatom assemblages from 16 Dutch moorland pools, sampled in the 1920's and again in 1978, to investigate the impact of acidification on these shallow water bodies. Ten clearwater (non-humic) pools were situated in the province of Brabant and on the Veluwe and six brownwater (humic) pools in the province of Drenthe. The arrow of Eunotia exigua in the biplot (Fig. 5) indicates

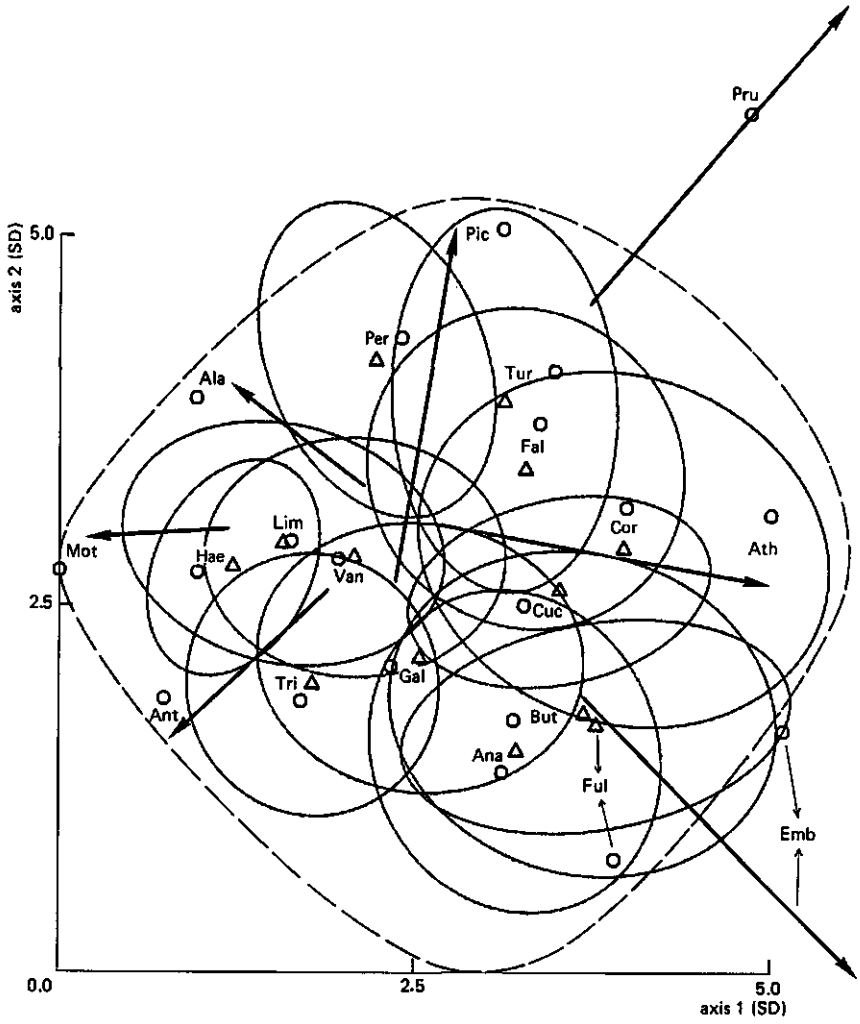


Fig. 6.

Joint plot based on detrended correspondence analysis (DCA) of bird species communities in the Rhine vallay near Amerongen, the Netherlands (data from Opdam et al., 1984), displaying the major variation in bird species composition across the landscape. This plot shows the DCA-scores (o) of the 20 most frequent species and the region in which the samples fall (---). Also shown are optima (Δ) and lines of equal probability for the 13 species whose probability surfaces had clear maxima (as fitted by Gaussian logit regression), and arrows representing directions of increase for the seven species whose probability surfaces were monotonic.

that this species increases strongly along the first principal component: E. exigua is abundant in the recent Brabant and Veluwe samples, which lie on the right-hand side of the diagram, and rare in the remaining samples, which lie more to the left. The second axis accounts for some of the difference among the old and recent samples from Drenthe. These groups differ in the abundances of Frustulia rhomboides var. saxonica, Tabellaria quadriseptata, Eunotia tenella, Tabellaria binalis, and Eunotia veneris, as shown by the directions of the arrows for these species in Fig. 5. As E. exigua is acidobiontic and the first principal component is strongly correlated with the sulphate concentration of the 1978 samples, this component clearly depicts the impact of acidification of the moorland pools in Brabant and the Veluwe (and to a smaller extent also in Drenthe). Thus Van Dam et al. (1981) used PCA to summarize the changes in diatom composition between the 1920's and 1978, and also to show how the nature of the change differed among provinces.

B. Correspondence analysis: joint plots

In CA and DCA both sites and species are represented by points, and each site is located at the centre of gravity of the species that occur there. One may therefore get an idea of the species composition at a particular site by looking at "nearby" species points. Also, as far as DCA approximates the fitting of Gaussian (logit) surfaces, the species points are approximately the optima of these surfaces; hence the abundance or probability of occurrence of a species decreases with distance from its location in the diagram (Fig. 4).

Fig. 6 illustrates this interpretation of the species points as optima in ordination space. DCA was applied to presence-absence data on 51 bird species in 526 contiguous, 100 m × 100 m grid-cells in an area with pastures and scattered woodlots in the Rhine valley near Amerongen, the Netherlands (Opdam et al., 1984). Fig. 6 shows the scores of the 20 most frequent species, and the outline of the region in which the grid-cells fall (the individual grid-cells are not shown, to avoid crowding). Opdam et al. (1984) interpreted the first axis, of length 5.7 SD, as a gradient from open to closed landscape and the second axis, of length 5.5 SD, as a gradient from wet to drier habitats.

To show that the species' scores were indeed close to their optima, we also fitted a response surface for each species by logit regression using Eq. (7) with the first and the second DCA-axes as the predictor variables x_1 and x_2. For 13 of the 20 bird species, the fitted surface had a maximum. For each of these species the optimum was calculated by Eq. (8) and plotted in Fig. 6, together with the contour within which that bird species occurs with more than half of its maximum probability. The fitted optima of the species lie close to their DCA-scores.

For the remaining seven species, the fitted surface had a minimum or saddle point suggesting that their optima are located well outside the sampled range. For these species we fitted a "linear" logit surface by setting b_2, b_4 and b_4 in Eq. (7) to zero. The direction of steepest increase of each of the fitted surfaces is indicated in Fig. 6 by an arrow through the centroid of the site points; the beginning and end points of each arrow correspond to fitted probabilities of 0.1 and 0.9 respectively. These arrows point more or less in the same direction as the DCA-scores of the corresponding species.

In contrast to the PCA-diagram, the species points on the edge of the DCA-diagram are often rare species, lying there either because they prefer

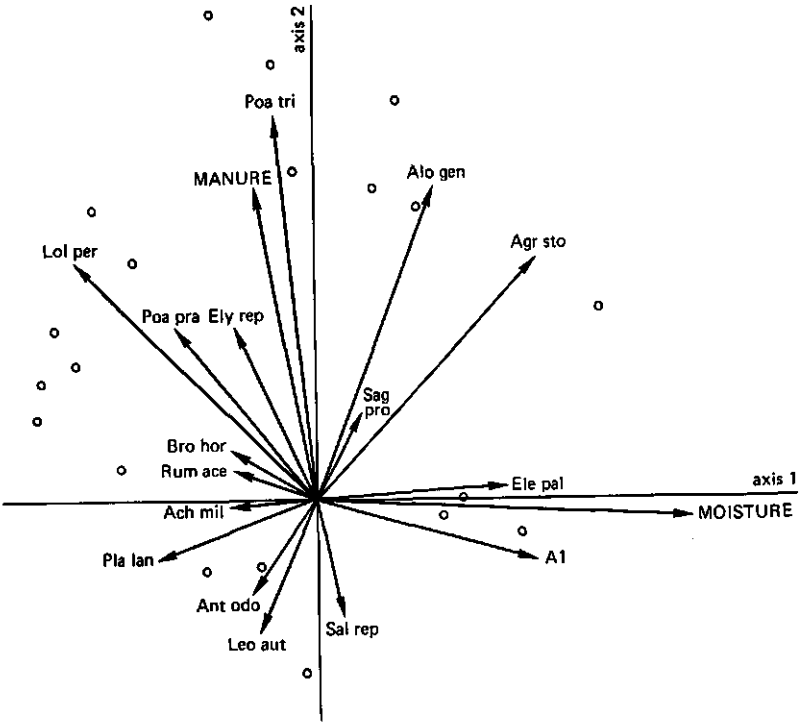


Fig. 7.

Biplot based on redundancy analysis of vegetation with respect to three environmental variables (quantity of manure, soil moisture and thickness of the A1 horizon) in dune meadows (o) on the island of Terschelling, The Netherlands. The arrows for plant species and environmental variables display the approximate (linear) correlation coefficients between plant species and the environmental variables. Abbreviations: Ach mil = Achillea millefolium, Agr sto = Agrostis stolonifera, Alo gen = Alopecurus geniculatus, Ant odo = Anthoxanthum odoratum, Bro hor = Bromus hordaceus, Ele pal = Eleocharis palustris, Ely rep = Elymus repens, Leo aut = Leontodon autumnalis, Lol per = Lolium perenne, Pla lan = Plantago lanceolata, Poa pra = Poa pratensis, Poa tri = Poa trivialis, Rum ace = Rumex acetosa, Sag pro = Sagina procumbens, Sal rep = Salix repens.

extreme (environmental) conditions or (very often) because their few occurrences by chance happen to fall in sites with extreme conditions; one cannot decide between these possibilities without additional data. Such peripheral species have little influence on the analysis and it is often convenient not to display them at all. Further, species near the centre of the diagram may be ubiquitous, unrelated to the ordination axes, bimodal, or in some other way not fitting a unimodal response model - or they may be genuinely specific with a habitat-optimum near the centre of the sampled range of habitats. The correct interpretation may be found most easily just by plotting the species' abundance in the ordination space.

C. Redundancy analysis

In redundancy analysis sites are indicated by points, and both species and environmental variables are indicated by arrows whose interpretation is similar to that of the arrows in the PCA biplot. The pattern of abundance of each species among the sites can be inferred in exactly the same way as in a PCA biplot, and so may the direction of variation of each environmental variable. One may also get an idea of the correlations between species' abundances and environmental variables. Arrows pointing in roughly the same direction indicate a high positive correlation, arrows crossing at right angles indicate near-zero correlation, and arrows pointing in opposite directions indicate high negative correlation. Species and environmental variables with long arrows are the most important in the analysis; the longer the arrows, the more confident one can be about the inferred correlation. (It is assumed here that for the purpose of the ordination diagram the environmental variables have been standardized to zero mean and unit variance, so as to make the lengths of arrows comparable.) Ter Braak (1987a) provides more quantitative rules for interpreting the ordination diagrams derived in this way from redundancy analysis.

The data we use to illustrate redundancy analysis were collected to study the relation between the vegetation and management of dune meadows on the island of Terschelling, The Netherlands (M. Batterink and G. Wijffels, unpublished). Fig. 7 displays the main variation in the vegetation in relation to three environmental variables (thickness of the A1 horizon, moisture content of the soil and quantity of manuring). The arrows for Poa trivialis and Elymus repens make small angles with the arrow for manuring; these species are inferred to be positively correlated with manuring. Salix repens and Leontodon autumnalis have arrows pointing in directions roughly opposite to that of manuring, and are inferred to be negatively correlated with manuring. Correlations of species with moisture and thickness of the A1 horizon can be inferred in a similar way.

D. Canonical correspondence analysis

In CCA, since species are assumed to have unimodal response surfaces with respect to linear combinations of the environmental variables, the species are logically represented by points (corresponding to their approximate optima in the two-dimensional environmental subspace) and the environmental variables by arrows indicating their direction and rate of change through the subspace.

Purata (1986, and unpublished results) applied CCA to plant species abundance data from 40 abandoned cultivation sites within Mexican tropical rain forest. Data were available for 24 of these sites on the regrowth age (A), the length of the cropping period in the past (C), and the proportion

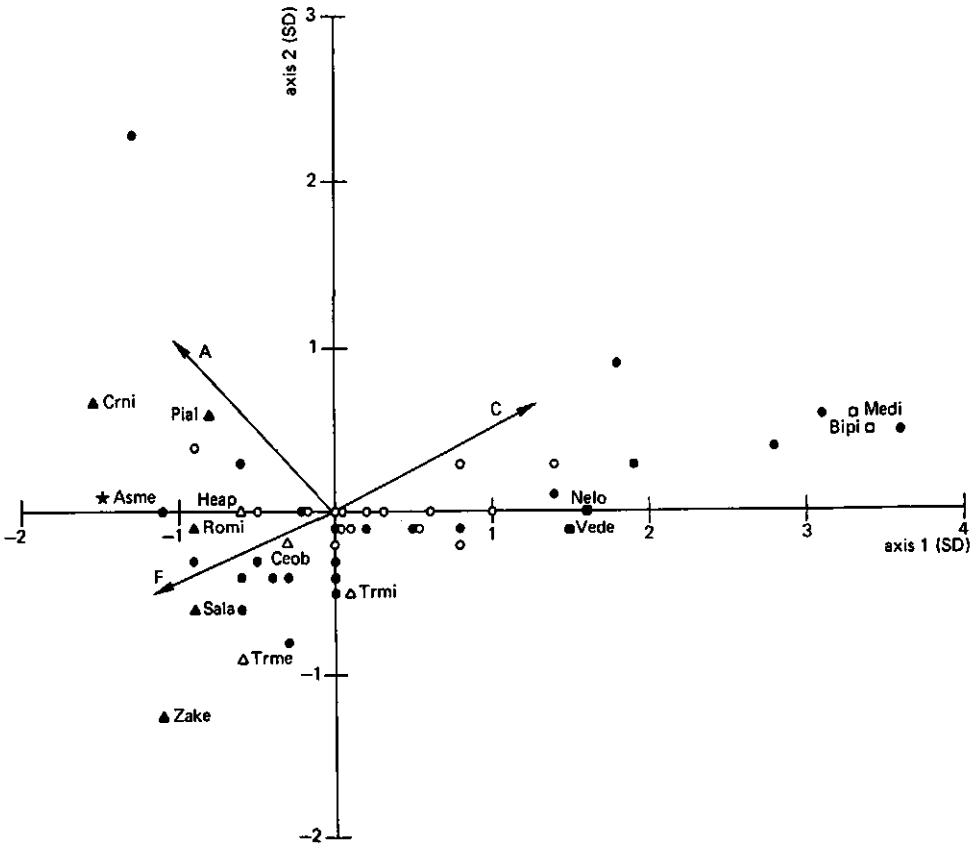


Fig. 8.

Ordination diagram based on canonical correspondence analysis of successional plant communities with respect to three environmental variables (regrowth age A, length of cropping period C, and extent of forested perimeter F) on abandoned cultivation sites within Mexican tropical rain forest (Purata, 1986 and unpublished). ●, sites with environmental data; o, sites added "passively" on the basis of floristic composition. The species shown are a selection among the 285 included in the analysis. □ denotes ruderals, ■ pioneer shrubs, Δ pioneer trees, ▲ late-secondary canopy trees and * an understory palm. Abbreviations: Bipi = Bidens pilosa, Medi = Melampodium divaricatum, Nelo = Neurolaena lobata, Vede = Vernonia deppeana, Trmi = Trema micrantha, Ceob = Cecropia obtusifolia, Heap = Heliocarpus appendiculatus, Trme = Trichospermum mexicanum, Pial = Piper amalago, Romi = Robinsonella mirandae. Sala = Sapium lateriflorum, Zake = Zanthoxylum kellermanii, Crni = Croton nitens, Asme = Astrocaryum mexicanum.

of the perimeter that had remained forested (F). These three variables were used as environmental variables in CCA. The remaining 16 sites were entered as "passive" sites, to be positioned with respect to the CCA axes according to their floristic composition in relation to the "active" sites.

Fig. 8 illustrates the results. The first axis, with length 3.3 SD, was interpreted as an indicator of the general trend of secondary succession. The direction of the arrow for regrowth age shows that this trend runs broadly from right to left. The species' locations are consistent with their life history characteristics: the trend of succession runs from ruderals (to the right), through pioneer shrubs and trees, to late-secondary canopy dominants and shade-tolerant understory species (to the left). The directions of the other two arrows in relation to axis 1 show that a long cropping period delays succession, while an extensive forested perimeter accelerates succession. Axis 2 (3.0 SD) may (more speculatively) differentiate species whose establishment is favoured by the presence of mature forest around the site from those that simply require a long time to grow.

Purata (1986) first tried indirect gradient analysis - DCA followed by multiple regression of the first DCA axis on the three environmental variables - but did not succeed in showing a significant effect of the environmental variables. However, their effect expressed on the first CCA axis was shown to be significant by using a Monte Carlo permutation test (Ter Braak, 1987b).

CCA also allows the computation of unconstrained, "residual" axes summarizing floristic variation that remains after the effect of the environmental variables has been taken out. In Purata's study, the successive eigenvalues of the first three (constrained) CCA axes were 0.49, 0.34 and 0.18. (There can be no more constrained axes than environmental variables.) The first residual axis gave an eigenvalue of 0.74, showing that at least as much floristic variation was not explained by the environmental variables. In our experience, terrestrial community data commonly give a residual eigenvalue as large as the first constrained eigenvalue, however carefully the environmental variables were chosen. Thus DCA and CCA tend to give different ordinations, and CCA - as in this example - is more powerful in detecting relationships between species composition and environment.

VI. CHOOSING THE METHOD

A. Which response model?

Regression methods can fit response models with a wide variety of shapes. The linear and Gaussian-like models are convenient starting points; more complex shapes can be fitted by adding further parameters, if the data are sufficiently detailed to support it. Other species may be used as additional explanatory variables if the specific aim is to detect species interactions (Fresco, 1982). The shapes of the response functions may be made even more general by applying Box-Cox transformations to the explanatory variables (Bartlein et al., 1986) or still more general by fitting splines (Smith, 1979). Even with all these modifications, regression can still be done with standard packages for Generalized Linear Modelling.

After species response curves or surfaces have been fitted by regression, calibration based on the maximum likelihood principle can be used to make inferences about the environment from community data. If the surfaces fitted by regression have complex shapes, then calibration by numerical maximization of the likelihood may be problematic. But even then,

if there are only a few environmental variables involved, the "most likely" combination of environmental values can be searched for on a grid across the environmental space (Atkinson et al., 1986; Bartlein et al., 1986). So the type of response model used in both regression and calibration should generally be guided by the characteristics and resolution of the data, and inspection of the data should show whether the model being used is adequate for the purpose.

In contrast to regression and calibration, the ordination problem requires the simultaneous estimation of large numbers of parameters and cannot be solved practically without some constraints on the structure one wants to fit. That these constraints may seem unduly restrictive simply shows that there are limits to what ordination can achieve. The number of ordination axes to be extracted must be small, and the type of response model must be restricted, in order to permit a solution. For example, it seems necessary to disregard the possibility of bimodal species distributions (Hill, 1977). (Certainly bimodal distributions sometimes occur, but ordination has to assume that species "on average" have simple distributions - otherwise, the problem would be insoluble; the utility of ordination techniques depends on them being robust with respect to departures from the simple models they are based on.) The Gaussian model seems to be of the right order of complexity for ordination of ecological data, but the full second-degree model of Eq. (7) is already difficult to fit (Kooijman, 1977; Goodall and Johnson, 1982). The Gaussian model with circular contour lines and equal species tolerances, i.e. the unfolding model, might provide a good compromise between practical solubility and realism in ordination. Promising algorithms for unfolding are developed by Heiser (1986) and DeSarbo and Rao (1984). DCA provides a reasonably robust approximation to ML Gaussian ordination and requires far less computing time. Similarly, ML Gaussian canonical ordination is technically feasible but CCA provides a practical and robust approximation to it.

Nonlinear methods are appropriate if a reasonable number of species have their optima located within the data set. If the gradient length is reduced to less than about 3 SD, the approximations involved in WA become worse and ultimately (if the gradient length is less than about 1.5 SD) the methods yield poor results because most species are behaving monotonically over the observed range. Thus if the community variation is within a narrow range, the linear ordination methods - PCA and redundancy analysis - are appropriate. If the community variation is over a wider range, nonlinear ordination methods - including DCA and CCA - are appropriate.

B. Direct or indirect?

Direct gradient analysis allows one to study the part (large or small) of the variation in community composition that can be explained by a particular set of environmental variables. In indirect gradient analysis attention is first focused on the major pattern of variation in community composition; the environmental basis of this pattern is to be established later. If the relevant environmental data are to hand, the direct approach - either fitting separate response surfaces by regression for each major species, or analysing the overall patterns of the species-environment relationship by constrained ordination - is likely to be more effective than the traditional indirect approach. However, indirect gradient analysis does have the advantage that no prior hypothesis is needed about what environmental variables are relevant. One does not need to measure the environmental variables in advance, and one can use informal field knowledge

to help interpret the patterns that emerge - hence the emphasis in the literature on ordination as a technique for "hypothesis generation", the implication being that experimental or more explicit statistical approaches can be used for subsequent hypothesis testing. This distinction is not hard and fast, but it does draw attention to the strengths and limitations of indirect gradient analysis.

In Section VD, we showed in passing how an indirect gradient analysis can be carried out after a direct gradient analysis in order to summarize the community variation that remains after known effects have been removed. When the known environmental variables are not the prime object of study, they are called concomitant variables (Davies and Tso, 1982) or covariables. It would be convenient to solve for the residual (unconstrained) axes without having to extract all the constrained axes first. Fortunately, this is straightforward. In the iterative algorithm for PCA and CA, one simply extends step 3b such that the trial scores are not only made uncorrelated with any previous axis (if present) but are also made uncorrelated with all specified covariables (see Appendix for details.) In this way the effects of the covariables are partialled out from the ordination; hence the name "partial ordination". The theory of "partial components analysis" and "partial correspondence analysis", as we call these extensions of PCA and CA, is given by Gabriel (1978, theorem 3) and Ter Braak (1988), respectively. Swaine and Greig-Smith (1980) used partial components analysis to obtain an ordination of within-plot vegetation change in permanent plots. Partial correspondence analysis, or its detrended form, would be more appropriate if the gradients were long.

C. Direct gradient analysis: regression or constrained ordination?

Whether to use constrained ordination (multivariate direct gradient analysis) instead of a series of separate regressions (the traditional type of direct gradient analysis) depends on whether or not there is any advantage in analysing all the species simultaneously. Both constrained and unconstrained ordination assume that the species react to the same composite gradients of enviromental variables, while in regression a separate composite gradient is constructed for each species. Regression can therefore allow more detailed descriptions and more accurate prediction and calibration, if properly carried out (with due regard to its statistical assumptions) and if sufficient data are available. However, ecological data that are collected over a large range of habitat variation require non-linear models, and building good non-linear models by regression is demanding in time and computation. In CCA the composite gradients are linear combinations of environmental variables and the non-linearity enters through a unimodal response model with respect to a few composite gradients, taken care of in CCA by the procedure of weighted averaging. Constrained ordination is thus easier to apply, and requires less data, than regression; it provides a summary of the species-environment relationship, and we find it most useful for the exploratory analysis of large data sets.

Constrained ordination can also be carried out after regression, in order to relate the residual variation to other environmental variables. This type of analysis, called "partial constrained ordination", is useful when the explanatory (environmental) variables can be subdivided in two sets, a set of covariables - the effects of which are not the prime object of study - and a further set of environmental variables whose effects are of particular interest.

For example, in the illustration of section VC, the study was initiated

to investigate differences in vegetation among dune meadows that were exploited under different management regimes (standard farming, bio-dynamical farming, nature management, among others). Standard CCA showed systematic differences in vegetation among management regimes. A further question is then whether these differences can be fully accounted for by the environmental variables moisture, quantity of manure and thickness of the A1 horizon, whose effects are displayed in Fig. 7, or whether the variation that remains after fitting the three environmental variables (three constrained ordination axes) is systematically related to management regimes. This question can be tackled by using partial constrained ordination, with the three environmental variables as covariables, and a series of dummy variables (for each of the management regimes) as the variables-of-interest.

Technically, partial constrained ordination can be carried out by any computer program for constrained ordination. The usual environmental variables are replaced by the residuals obtained by regressing each of the variables of interest on the covariables (see Appendix). Davies and Tso (1982) gave the theory behind partial redundancy analysis; Ter Braak (1987b) derived partial canonical correspondence analysis as an approximation to "partial Gaussian canonical ordination".

Partial constrained ordination has the same essential aim as Carleton's (1984) residual ordination, i.e. to determine the variation in the species data that is uniquely attributable to a particular set of environmental variables, taking into account the effects of other (co-) variables; however Carleton's method is somewhat less powerful, being based on a pre-existing DCA which may already have removed some of the variation of interest. Partial constrained ordination is, by contrast, a true direct gradient analysis technique which seems promising e.g. for the analysis of permanent plot data (effects of time, with location and/or environmental data as covariables), and a variety of other applications in which effects of particular environmental variables are to be sorted out from the "background" variation imposed by other variables.

VII. CONCLUSIONS

Regression, calibration, ordination and constrained ordination are well-defined statistical problems with close interrelationships. Regression is the tool for investigating the nature of individual species' response to environment, and calibration is the tool for (later) inferring the environment from species composition at an individual site. Both tools come in various degrees of complexity. The simplest are linear and WA regression and calibration. The linear methods are applicable over short ranges of environment, where species' abundance appears to vary monotonically with variation in the environment. The WA methods are applicable over wider ranges of environment; WA regression is a crude method to estimate each species' optimum, and WA calibration just averages the optima of the species that are present. WA works with presence-absence data. If abundances are available, they provide the weights. These WA techniques can be shown to give approximate estimates of the species optima and environmental values when the species response surfaces (the relationships between the species' abundance, or probability of occurrence, and the environmental variables) are Gaussian (or for probabilities, Gaussian-logit) in form. Gaussian regression and calibration are also possible, but the WA techniques are simpler and are approximations to the Gaussian methods.

These simple tools are suitable when there are many species of interest

and the exact form of the response surface is not critical, and they are very easy to use. If the form of the response surfaces is critical, more complex models can be fitted by using Generalized Linear Modelling (for regression) and maximum likelihood techniques (for calibration). These more complex tools are becoming important in the theoretical study of species-environment relationships (Austin, 1985) and environmental dynamics (Bartlein et al., 1986). Naturally, they require skilled users who are aware of their statistical assumptions, limitations and pitfalls.

Ordination and constrained ordination can be related to the simpler methods of regression and calibration. Ordination is the tool for exploratory analysis of community data with no prior information about the environment. Constrained ordination is the equivalent tool for the analysis of community variation in relation to environment. Both implicitly assume a common set of environmental variables and a common response model for all of the species. (Without these simplifying assumptions, they could not work; such major simplifications of data can only be achieved at the expense of some realism.) The basic ordination techniques are PCA and CA. PCA constructs axes that are as close as possible to a linear relationship with the species. These axes can be found by a converging sequence of alternating linear regressions and calibrations. Each axis after the first is obtained by partialling out linear relationships with the previous axis. CA is mathematically related to PCA, but has a very different effect. CA axes can be found by a converging sequence of WA regressions and calibrations. In CA, axes after the first are obtained analogously with PCA; in DCA they are obtained by removing all trends, linear or nonlinear, with respect to previous axes. CA suffers from the arch effect, which DCA eliminates. DCA is a reasonably robust approximation to Gaussian ordination, in which the axes are constructed so that the species response curves with respect to the axes are Gaussian in form. Gaussian ordination is feasible but not convenient. DCA is much more practical. But there are problems with the detrending, and the method can break down when the connections between sites are too tenuous. Some modifications - including an improved method of detrending - may improve DCA's robustness; alternatively, some forms of nonmetric multidimensional scaling may be more robust (Kenkel and Orlóci, 1986; Minchin, 1987).

Constrained ordination methods have the added constraint that the ordination axes must be linear combinations of environmental variables. This constraint can be implemented as an extra multiple regression step in the general iterative ordination algorithm. PCA then becomes redundancy analysis (a more practical alternative to canonical correlation), Gaussian ordination becomes Gaussian canonical ordination, and CA becomes CCA (Table 2). The constraint makes Gaussian canonical ordination somewhat more stable than its unconstrained equivalent, but still CCA provides a much more practical alternative. All these constrained methods are most powerful if the number of environmental variables is small compared to the number of sites. Then the constraints are much stronger than in normal ordination, and the common problems of ordination (such as the arch effect, the need for detrending and the sensitivity to deviant sites) disappear.

Often, community-environment relationships have been explored by "indirect gradient analysis" - ordination, followed by interpretation of the axes in terms of environmental variables. But if the environmental data are to hand, constrained ordination ("multivariate direct gradient analysis") provides a more powerful means to the same end. Hybrid (direct/indirect) analyses are also possible. In partial ordination and partial constrained ordination, the analysis works on the variation that remains after the

effects of particular environmental, spatial or temporal "covariables" have been removed.

The choice between linear and nonlinear ordination methods is not a matter of personal preference. Where gradients are short, there are sound statistical reasons to use linear methods. Gaussian methods break down, and edge effects in CA and related techniques become serious; the representation of species as arrows becomes appropriate. As gradient lengths increase, linear methods become ineffective (principally through the "horseshoe effect", which scrambles the order of samples along the first axis as well as creating a meaningless second axis); Gaussian methods become feasible, and CA and related techniques become effective. The representation of species as points, representing their optima, becomes informative. The range 1.5 - 3 SD for the first axis represents a "window" over which both PCA and CA/DCA, or both redundancy analysis and CCA, can be used to good effect.

VIII. APPENDIX

A general iterative algorithm can be used to carry out the linear and weighted-averaging methods described in this review. The algorithm is essentially the one used in the computer program CANOCO (Ter Braak, 1987b). It operates on response variables, each recording the abundance or presence/absence of a species at various sites, and on two types of explanatory variables: environmental variables and covariables. By environmental variables we mean here explanatory variables of prime interest, in contrast with covariables which are "concomitant" variables whose effect is to be removed. When all three types of variables are present, the algorithm describes how to obtain a partial constrained ordination. The other linear and WA techniques are all special cases, obtained by omitting various irrelevant steps.

Let $Y = [y_{ki}]$ ($k = 1, \ldots, m$; $i = 1, \ldots, n$) be a species-by-site matrix containing the observations of m species at n sites ($y_{ki} \geq 0$) and let $Z_1 = [z_{1li}]$ ($l = 0, \ldots, p$; $i = 1, \ldots, n$) and $Z_2 = [z_{2ji}]$ ($j = 1, \ldots, q$; $i = 1, \ldots, n$) be covariable-by-site and environmental variable-by-site matrices containing the observations of p covariables and q environmental covariables at the same n sites, respectively. The first row of Z_1, with index $l = 0$, is a row of 1's, which is included to account for the intercept in Eq. (5). Further, denote the species and site scores on the s-th ordination by $\underline{u} = [u_k]$ ($k = 1, \ldots, m$) and $\underline{x} = [x_i]$ ($i = 1, \ldots, n$), the canonical coefficients of the environmental variables by $\underline{c} = [c_j]$ ($j = 1, \ldots, q$) and collect the site scores on the $(s - 1)$ previous ordination axes as rows of the matrix A. If detrending-by-polynomials is in force (Step A10), then the number of rows of A, s_A say, is greater than s-1. In the algorithm we use the assign statement " := ", for example a := b means "a is assigned the value b". If the left hand side of the assignment is indexed by a subscript, it is assumed that the assignment is made for all permitted subscript values: the subscript k will refer to species ($k = 1, \ldots, m$), the subscript i to sites ($i = 1, \ldots, n$) and the subscript j to environmental variables ($j = 1, \ldots, q$).

Preliminary calculations

P1. Calculate species totals $\{y_{k+}\}$, site totals $\{y_{+i}\}$ and the grand total y_{++}. If a linear method is required, set

$$r_k := 1, \; w_i := 1, \; w_i^* := \frac{1}{n} \tag{A.1}$$

and if a weighted averaging method is required, set

$$r_k := y_{k+}, \; w_i := y_{+i}, \; w_i^* := y_{+i}/y_{++} \tag{A.2}$$

P2. Standardize the environmental variables to zero mean and unit variance. For environmental variable j calculate its mean $\bar{z}$ and variance v

$$\bar{z} := \sum_i w_i^* z_{2ji}, \; v := \sum_i w_i^* (z_{2ji} - \bar{z})^2 \tag{A.3}$$

and set $z_{2ji} := (z_{2ji} - \bar{z})/\sqrt{v}$

P3. Calculate for each environmental variable j the residuals of the multiple regression of the environmental variable on the covariables, i.e.

$$\underline{c}_j^* := (Z_1 W Z_1')^{-1} Z_1 W z_{2j} \tag{A.4}$$

$$\bar{z}_{2j} := z_{2j} - Z_1' \underline{c}_j^* \tag{A.5}$$

where $z_{2j} = (z_{2j1}, \ldots, z_{2jn})'$, $W = \text{diag}\ (w_1, \ldots, w_n)$ and $\underline{c}_j^*$ is the (p+1)-vector of the coefficients of the regression of z_{2j} on Z_1. Now define $\bar{Z}_2 = [\bar{z}_{2ji}]$ $(j = 1, \ldots, q,\ i = 1, \ldots, n)$.

Iteration algorithm

Step A0 Start with arbitrary, but unequal site scores $\underline{x} = [x_i]$. Set $x_i^0 = x_i$.

Step A1 Derive new species scores from the site scores by

$$u_k := \sum_i y_{ki} x_i / r_k. \tag{A.6}$$

Step A2 Derive new site scores $\underline{x}^* = [x_i^*]$ from the species scores

$$x_i^* := \sum_k y_{ki} u_k / w_i. \tag{A.7}$$

Step A3 Make $\underline{x}^* = [x_i^*]$ uncorrelated with the covariables by calculating the residuals of the multiple regression of $\underline{x}^*$ on Z_1:

$$\underline{x}^* := \underline{x}^* - Z_1'(Z_1 W Z_1')^{-1} Z_1 W \underline{x}^*. \tag{A.8}$$

Step A4 If $q \leq s_A$, set $x_i := x_i^*$ and skip Step A5.

Step A5 If $q > s_A$, calculate a multiple regression of $\underline{x}^*$ on $\bar{Z}_2$

$$\underline{c} := (\bar{Z}_2 W \bar{Z}_2')^{-1} \bar{Z}_2 W \underline{x}^*, \tag{A.9}$$

and take as new site scores the fitted values:

$$\underline{x} := \bar{Z}_2' \underline{c}. \tag{A.10}$$

Step A6 If $s > 0$, make $\underline{x} = [x_i]$ uncorrelated with previous axes by calculating the residuals of the multiple regression of $\underline{x}$ on A:

$$\underline{x} := \underline{x} - A'(AWA')^{-1} AW\underline{x} \tag{A.11}$$

Step A7 Standardize $\underline{x} = [x_i]$ to zero mean and unit variance by

$$\bar{x} := \sum_i w_i^* x_i, \quad s^2 := \sum_i w_i^* (x_i - \bar{x})^2, \tag{A.12}$$

$$x_i := (x_i - \bar{x})/s.$$

Step A8 Check convergence, i.e. if

$$\sum_i w_i^* (x_i^0 - x_i)^2 < 10^{-10} \tag{A.13}$$

goto Step A9, else set $x_i^0 := x_i$ and goto Step A1.

Step A9 Set the eigenvalue λ equal to s in (A.12) and add $\underline{x} = [x_i]$ as a new row to the matrix A.

Step A10 If detrending-by-polynomials is required, calculate polynomials of $\underline{x}$ up to order 4 and first order polynomials of $\underline{x}$ with the previous ordination axes,

$$x_{2i} := x_i^2, \; x_{3i} := x_i^3, \; x_{4i} := x_i^4, \; x_{(b)i} := x_i a_{bi} \tag{A.14}$$

where a_{bi} are the site scores of a previous ordination axis ($b = 1, \ldots, s-1$). Now perform for each of the (s+2)-variables in (A.14) the Steps A3-A6 and add the resulting variables as new variables to the matrix A.

Step A11 Set $s := s+1$ and goto Step A0 if required and if further ordination axes can be extracted, else stop.

At convergence, the algorithm gives the solution with the greatest real value of λ to the following transition formulae [where $R = \mathrm{diag}(r_1, \ldots, r_m)$ and $W = \mathrm{diag}(w_1, \ldots, w_n)$ and where the notation B^0 is used to denote $B'(BWB')^{-1}BW$, the projection operator on the row space of a matrix B in the metric defined by the matrix W]

$$\underline{y} = R^{-1} Y \underline{x} \tag{A.15}$$

$$\underline{x}^* = (I - Z_1^0) W^{-1} Y' \underline{y} \tag{A.16}$$

$$\underline{c} = (\tilde{Z}_2 W \tilde{Z}_2')^{-1} \tilde{Z}_2 W \underline{x}^* \tag{A.17}$$

$$\lambda \underline{x} = (I - A^0) \tilde{Z}_2' \underline{c}. \tag{A.18}$$

The wiggle above Z_2 is there as a reminder that the original environmental variables were replaced by residuals of a regression on Z_1 in (A.5) i.e. in terms of the original variables

$$\tilde{Z}_2' = (I - Z_1^0) Z_2' \tag{A.19}$$

Remarks

1. Note that u_k in the algorithm takes the place of b_k in section II.

2. Special cases of the algorithm are: constrained ordination: p = 0; partial ordination: q = 0; (unconstrained) ordination: p = 0, q = 0; linear calibration and weighted averaging: p = 0, q = 1; (partial) multiple regression: m = 1. The corresponding transition formulae follow from (A.15) - (A.18) with the proviso that, if q = 0, Z_2 in (A.19) must be replaced by the n×n identity matrix and generalized matrix inverses are used. Note that, if p = 0, Z_1 is a 1×n matrix containing 1's; Z_1 renders the centring of the species data in the linear methods in section II redundant.

3. The standardization in P2 removes the arbitrariness in the units of measurement of the environmental variables, and makes the canonical coefficients comparable among each other, but does not influence the values of λ, $\underline{u}$ and $\underline{x}$ to be obtained in the algorithm.

4. Step A6 simplifies to step 3b of the main text if the rows of A are W-orthonormal. The steps A3-A6 form a single projection of $\underline{x}^*$ on the column space of $(I - A^0)\tilde{Z}_2$ if and only if A defines a subspace of the rowspace of $\tilde{Z}_2$. As each ordination axis defines such a subspace, this is trivially so without detrending. The method of detrending-by-polynomials as defined in step A10, ensures that A defines also the relevant subspace if detrending is in force. The transition formulae (A.15) - (A.18) define an eigenvalue equation of which all eigenvalues are real nonnegative (Ter Braak, 1987b).

5. If a particular scaling of the biplot or the joint plot is wanted, the ordination axes may require linear rescaling. With linear methods one can choose between a Euclidean distance biplot and a covariance biplot, which focus on the approximate Euclidean distances between sites and correlations among species, respectively (Ter Braak, 1983). With weighted averaging methods it is customary to use the site scores $\underline{x}^*$ (A.16) and the species scores $\underline{u}$ (A.15) to prepare an ordination diagram after a linear rescaling so that the average within-site variance of the species scores is equal to 1 (cf. section IV C), as is done in DECORANA (Hill, 1979) and CANOCO (Ter Braak, 1987b).

ACKNOWLEDGEMENTS

We thank Dr. M.P. Austin, Dr. P.J. Bartlein, Professor L.C.A. Corsten, J.A. Hoekstra, Dr. P. Opdam and Dr. H. van Dam for comments on the manuscript. Our collaboration was supported by a Netherlands Science Research Council (ZWO) grant to I.C.P. and a Swedish Natural Science Research Council (NFR) grant to the project "Simulation of Natural Forest Dynamics". We also thank Dr. S.E. Purata V. for supplying unpublished results.

REFERENCES

Alderdice, D.F. (1972). Factor combinations: responses of marine poikilotherms to environmental factors acting in concert. In: "Marine Ecology," (O. Kinne, ed.), vol 1, part 3, p. 1659-1722, Wiley, New York.

As, (1985). Predictability and density compensation of carabid assemblages on islands. MS.

Atkinson, T.C., Briffa, K.R., Coope, G.R., Joachim, M.J., and Perry, D.W. (1986). Climatic calibration of coleopteran data. In "Handbook of Holocene Palaeoecology and Palaeohydrology," (B.E. Berglund, ed.), pp. 851-858, Wiley, Chichester.

Austin, M.P. (1971). Role of regression analysis in plant ecology. Proc. ecol. Soc. Austr. 6, 63-75.

Austin, M.P. (1985). Continuum concept, ordination methods, and niche theory. Ann. Rev. Ecol. Syst. 16, 39-61.

Austin, M.P., and Cunningham, R.B. (1981). Observational analysis of environmental gradients. Proc. ecological Soc. Austr. 11, 109-119.

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

Balloch, D., Davies, C.E., and Jones, F.H. (1976). Biological assesment of water quality in three British rivers: The North Esk (Scotland), The Ivel (England) and the Taf (Wales). Wat. Pollut. Control 75, 92-114.

Bartlein, P.J., Webb, T. III, and Fleri, E. (1984). Holocene climatic changes in the Northern Midwest: pollen-derived estimates. Quat. Res. 22, 361-374.

Bartlein, P.J., Prentice, I.C., and Webb, T. III (1986). Climatic response surfaces from pollen data for some eastern North American taxa. J. Biogeogr. 13, 35-57.

Battarbee, R.W. (1984). Diatom analysis and the acidification of lakes. Philos. Trans. Roy. Soc. London Ser. B 305, 451-477.

Bloxom, B. (1978). Constrained multidimensional scaling in N spaces. Psychometrika 43, 397-408.

Böcker, R., Kowarik, I., and Bornkamm, R. (1983). Untersuchungen zur Anwendung der Zeigerwerte nach Ellenberg. Verh. Ges. Oekol. 11, 35-56.

Brown, G.H. (1979). An optimization criterion for linear inverse estimation. Technometrics 21, 575-579.

Brown, P.J. (1982). Multivariate calibration. J. R. Statist. Soc. B 44, 287-321.

Chandler, J.R. (1970). A biological approach to water quality management. Water Pollut. Control 69, 415-421.

Charles, D.F. (1985). Relationships between surface sediment diatom assemblages and lakewater characteristics in Adirondack lakes. Ecology 66, 994-1011.

Coombs, C.H. (1964). "A theory of data". Wiley, New York.

Cox, D.R., and Hinkley, D.V. (1974). "Theoretical Statistics." Chapman and Hall, London.

Cramer, W., and Hytteborn, H. (1987). The seperation of fluctuation and long-term change in vegetation dynamics of a rising sea-shore. Vegetatio 69, 157-167.

Dargie, T.C.D. (1984). On the integrated interpretation of indirect site ordinations: a case study using semi-arid vegetation in southeastern Spain. Vegetatio 55, 37-55.

Davies, P.T., and Tso, M.K-S. (1982). Procedures for reduced-rank regression. Appl. Statist. 31, 244-255.
Davison, M.L. (1983). "Multidimenional scaling." Wiley, New York.
De Leeuw, J., and Heiser, W. (1980). Multidimensional scaling with restrictions on the configuration. In: "Multivariate Analysis-V," (P.R. Krishnaiah, ed.), pp. 501-522, North-Holland Publ., Amsterdam.
DeSarbo, W.S., and Rao, V.R. (1984). GENFOLD2: a set of models and algorithms for the general unfolding analysis of preference/dominance data. J. Class. 1, 147-186.
Dobson, A.J. (1983). "Introduction to statistical modelling." Chapman and Hall, London.
Ellenberg, H. (1979). Zeigerwerte der Gefässpflanzen Mitteleuropas. Scripta Geobotanica 9, 1-121.
Feoli, E., and Feoli Chiapella, L. (1980). Evaluation of ordination methods through simulated coenoclines: some comments. Vegetatio, 42, 35-41.
Feoli, E., and Orlóci, L. (1979). Analysis of concentration and detection of underlying factors in structured tables. Vegetatio, 40, 49-54.
Fresco, L.F.M. (1982). An analysis of species response curves and of competition from field data: some results from heath vegetation. Vegetatio 48, 175-185.
Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453-467.
Gabriel, K.R. (1978). Least squares approximation of matrices by additive and multiplicative models. J.R.Statist.Soc. B 40, 186-196.
Gasse, F., and Tekaia, F. (1983). Transfer functions for estimating paleoecological conditions (pH) from East African diatoms. Hydrobiologia 103, 85-90.
Gauch, H.G. (1982). "Multivariate analysis in community ecology." Cambridge Univ. Press, Cambridge.
Gauch, H.G., and Whittaker, R.H. (1972). Coenocline simulation. Ecology 53, 446-451.
Gauch, H.G., Chase, G.B., and Whittaker, R.H. (1974). Ordination of vegetation samples by Gaussian species distributions. Ecology 55, 1382-1390.
Gauch, H.G., Whittaker, R.H., and Singer, S.B. (1981). A comparative study of nonmetric ordinations. J. Ecol. 69, 135-152.
Gifi, A. (1981). "Nonlinear multivariate analysis." Department of Data Theory, University of Leiden, Leiden.
Gittins, R. (1985). "Canonical analysis. A review with applications in ecology". Springer-Verlag, Berlin.
Goff, F.G., and Cottam, G. (1967). Gradient analysis: The use of species and synthetic indices. Ecology 48, 793-806.
Goodall, D.W., and Johnson, R.W. (1982). Non-linear ordination in several dimensions: a maximum likelihood approach. Vegetatio 48, 197-208.
Gourlay, A.R., and Watson, G.A. (1973). "Computational methods for matrix eigen problems." Wiley, New York.
Greenacre, M.J. (1984). "Theory and applications of correspondence analysis." Acad. Press, London.
Heiser, W.J. (1981). "Unfolding analysis of proximity data". Thesis. University of Leiden, Leiden.
Heiser, W.J. (1986). Joint ordination of species and sites: the unfolding technique. In: "New developments in numerical ecology," (P. Legendre and L. Legendre, eds.), Springer-Verlag, Berlin, in press.
Hill, M.O. (1973). Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61, 237-249.

Hill, M.O. (1974). Correspondence analysis: a neglected multivariate method. Appl. Statist. 23, 340-354.
Hill, M.O. (1977). Use of simple discriminant functions to classify quantitative phytosociological data. In: "First International Symposium on Data Analysis and Informatics," (E. Diday, L. Lebart, J.P. Pages & R. Tomassone, eds.), Vol. 1, pp. 181-199, INRIA, Chesnay.
Hill, M.O. (1979). "DECORANA - a FORTRAN program for detrended correspondence analysis and reciprocal averaging." Section of Ecology and Systematics, Cornell University, Ithaca, New York.
Hill, M.O., and Gauch, H.G. (1980). Detrended correspondence analysis: an improved ordination technique. Vegetatio 42, 47-58.
Ihm, P., and Van Groenewoud, H. (1975). A multivariate ordering of vegetation data based on Gaussian type gradient response curves. J. Ecol. 63, 767-777.
Ihm, P., and Van Groenewoud, H. (1984). Correspondence analysis and Gaussian ordination. COMPSTAT Lectures 3, 5-60.
Imbrie, J., and Kipp, N.G. (1971). A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. In: "The late Cenozoic glacial ages," (K.K. Turekian, ed.), pp. 71-181, Yale University Press, New Haven.
Israëls, A.Z. (1984). Redundancy analysis for qualitative variables. Psychometrika 49, 331-346.
Iwatsubo, S. (1984). The analytical solutions of eigenvalue problem in the case of applying optimal scoring method to some types of data. In: "Data Analysis and Informatics 3", (E. Diday et al., eds.) pp. 31-40, North- Holland Publ., Amsterdam.
Jolliffe, I.T. (1986). "Principal component analysis." Springer-Verlag, Berlin.
Kalkhoven, J., and Opdam, P. (1984). Classification and ordination of breeding bird data and landscape attributes. In: "Methodology in Landscape Ecological Research and Planning," (J. Brandt and P. Agger, eds.), Vol. 3, Theme 3, pp. 15-26, Roskilde Universitetsforlag GeoRue, Roskilde.
Kenkel, N.C., and Orlóci, L. (1986). Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67, 919-928.
Kooijman, S.A.L.M. (1977). Species abundance with optimum relations to environmental factors. Ann. Syst. Res. 6, 123-138.
Kooijman, S.A.L.M., and Hengeveld, R. (1979). The description of a non-linear relationship between some carabid beetles and environmental factors. In: "Contemporary Quantitative Ecology and Related Econometrics," (G.P. Patil, and M.L. Rossenzweig, eds.), pp. 635-647, Intern. Co-operative Publ. House, Fairland, Maryland.
Laurec, A., Chardy, P., de la Salle, P., and Rickaert, M. (1979). Use of dual structures in inertia analysis: ecological implications. In: "Multivariate Methods in Ecological Work," (L. Orlóci, C.R. Rao, and W.M. Stiteler, eds.), pp. 127-174. Intern. Co-operative Publ. House, Fairland, Maryland.
Macdonald, G.M., and Ritchie, J.C. (1986). Modern pollen spectra from the western interior of Canada and the interpretation of Late Quaternary vegetation development. New Phytol. 103, 245-268.
Meulman, J., and Heiser, W.J. (1984). Constrained multidimensional scaling: more directions than dimensions. COMPSTAT 1984, Physica-Verlag, Vienna, pp. 137-142.

Minchin, P. (1987). An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69, 89-107.
Montgomery, D.C. and Peck, E.A. (1982). "Introduction to linear regression analysis." Wiley, New York.
Nishisato, S. (1980). "Analysis of categorical data: dual scaling and its applications." University of Toronto Press, Toronto.
Oksanen, J. (1983). Ordination of boreal heath-like vegetation with principal component analysis, correspondence analysis and multidimensional scaling. Vegetation 52, 181-189.
Opdam, P.F.M., Kalkhoven, J.T.R., and Phillippona, J. (1984). "Verband tussen broedvogelgemeenschappen en begroeiing in een landschap bij Amerongen". Pudoc, Wageningen.
Peet, R.K. (1978). Latitudinal variation in southern Rocky Mountain forests. J. Biogeogr. 5, 275-289.
Peet, R.K., and Loucks, O.L. (1977). A gradient analysis of southern Wisconsin forests. Ecology 58, 485-499.
Pickett, S.T.A. (1980). Non-equilibrium coexistence of plants. Bulletin of the Torrey Botanical Club 107, 238-248.
Pielou, E.C. (1984). "The interpretation of ecological data." Wiley, New York.
Prodon, R., and Lebreton, J.-D. (1981). Breeding avifauna of a Mediterranean succession: the holm oak and cork oak series in the eastern Pyrenees, 1. Analysis and modeling of the structure gradient. Oikos 37, 21-38.
Purata, S.E. (1986). Studies on secondary succession in Mexican tropical rain forest. Acta Univ. Ups. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science 19, Almqvist and Wiksell Internationsl, Stockholm.
Rao, C.R. (1964). The use and interpretation of principal components analysis in applied research. Sankhya A 26, 329-358.
Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Appl.Statist. 25, 257-265.
Sládecek, V. (1973). System of water quality from the biological point of view. Arch. Hydrobiol. Beiheft 7, 1-218.
Smith, P.L. (1979). Splines as a useful and convenient statistical tool. Amer. Stat. 33, 57-62.
Swaine, M.D., and Greig-Smith, P. (1980). An application of principal components analysis to vegetation change in permanent plots. J. Ecol. 68, 33-41.
Ter Braak, C.J.F. (1983). Principal components biplots and alpha and beta diversity. Ecology 64, 454-462.
Ter Braak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of a unimodal reponse model. Biometrics 41, 859-873.
Ter Braak, C.J.F. (1986a). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167-1179.
Ter Braak, C.J.F. (1987a). Ordination. In: "Data Analysis in Community and Landscape Ecology.", (R.H.G. Jongman, C.J.F. ter Braak, and O.F.R. Van Tongeren, eds.). Pudoc, Wageningen.
Ter Braak, C.J.F. ter (1987b). "CANOCO - a FORTRAN program for canonical community ordination by [partial][detrended][canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1)." TNO Institute of Applied Computer Science, Wageningen.

Ter Braak, C.J.F. (1987c). The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69, 69-77.
Ter Braak, C.J.F. (1988). Partial canonical correspondence analysis. In: "Classification and related methods of data analysis", (H.H. Bock, ed.), North-Holland Publ., Amsterdam.
Ter Braak, C.J.F., and Barendregt, L.G. (1986). Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Biosci. 78, 57-72.
Ter Braak, C.J.F., and Looman, C.W.N. (1986). Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65, 3-11.
Tilman, D. (1982). "Resource competition and community structure." Princeton University Press., Princeton, New York.
van der Aart, P.J.M., and Smeenk-Enserink, N. (1975). Correlations between distribution of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Neth. J. Zool. 25, 1-45.
van den Wollenberg, A.L. (1977). Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42, 207-219.
van Dam, H., Suurmond, G., and Ter Braak, C.J.F. (1981). Impact of acidification on diatoms and chemistry of Dutch moorland pools. Hydrobiologia 83, 425-459.
Webb, T., III, and Bryson, R.A. (1972). Late- and postglacial climatic change in the northern Midwest, USA: quantitative estimates derived from fossil spectra by multivariate statistical analysis. Quat. Res. 2, 70-115.
Whittaker, R.H. (1956). Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26, 1-80.
Whittaker, R.H. (1967). Gradient analysis of vegetation. Biol. Rev. 49, 207-264.
Whittaker, R.H., Levin, S.A., and Root, R.B. (1973). Niche, habitat and ecotope. Amer. Natur. 107, 321-338.
Wiens, J.A., and Rotenberry, J.T. (1981). Habitat associations and community structure of birds in shrubsteppe environments. Ecol. Monogr. 51, 21-41.
Willén, E., and Fängström, I. (1986). Analysis of phytoplankton and environmental data from 23 Swedish lakes by canonical correspondence analysis (in prep.).
Williams, E.J. (1959). "Regression analysis." Wiley, New York.
Wold, H. (1982). Soft modeling. The basic design and some extensions. In: "Systems under indirect observation. Causality-structure-prediction," (K.G. Jöreskog and H. Wold, eds.), Vol. 2, pp. 1-54, North-Holland Publ., Amsterdam.
Zelinka, M., and Marvan, P. (1961). Zür Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. Arch. Hydrobiol. 57, 389-407.

APPENDIX

Short description of CANOCO (version 2.1)

Aim

A common problem in community ecology and ecotoxicology is to discover how a multitude of species respond to external factors such as environmental variables, pollutants and management regime. Data are collected on species composition and the external variables at a number of points in space and time. Statistical methods sofar available to analyse such data either assumed linear relationships or were restricted to regression analysis of the response of each species separately. To analyse the generally non-linear, non monotonic response of a community of species, one had to resort to the data-analytic methods of ordination and cluster analysis - "indirect" methods that are generally less powerful than the "direct" statistical method of regression analysis. Recently, regression and ordination have been integrated into techniques of multivariate direct gradient analysis, called canonical (or constrained) ordination. The use of canonical ordination greatly improves the power to detect the specific effects one is interested in. One of these techniques, canonical correspondence analysis, avoids the assumption of linearity and is able to detect unimodal relationships between species and external variables. The computer program CANOCO is designed to make these techniques available to ecologists studying community responses. CANOCO can carry out most of the multivariate techniques described in chapter 9 by using a general iterative ordination algorithm.

Researchers in other fields may find CANOCO useful as well, for example, to analyse percentage data/compositional data, nominal data or (dis)- similarity data in relation to external explanatory variables. Such use is explained in separate sections in the manual (ter Braak, 1987). CANOCO is particularly suited if the number of response variables is large compared to the number of objects.

Techniques covered

1. CANOCO is an extension of DECORANA (Hill, 1979). CANOCO formerly stood for canonical correspondence analysis (chapter 5) and included weighted averaging, [multiple] correspondence analysis, detrended correspondence analysis and canonical correspondence analysis. The program has been extended to cover also principal components analysis (PCA) and the canonical form of PCA, called redundancy analysis (RDA). Redundancy analysis (van den Wollenberg, 1977; Israëls, 1984) is also known under the names of reduced-rank regression (Davies and Tso, 1982), PCA of y with respect to x (Robert and Escoufier, 1976) and mode C partial least squares (Wold, 1982). For these linear methods there are options for centring/standardization by species and by sites and for the method of scaling the species and site scores for use in the biplot. The eigenvalues reported in PCA/RDA are fractions of the total variance in the species data (percentage variance accounted for). Principal coordinates analysis and canonical variate analysis (= linear discriminant analysis) are also available.

2. CANOCO can also carry out "partial" analyses in which the effects of particular environmental, spatial or temporal "covariables" are eliminated from the ordination. A partial analysis allows one to display the residual variation in the species data and to relate the residual variation to the variables one is specifically interested in. Partial

canonical correspondence analysis is the appropriate technique for the analysis of permanent plot data or for the joint analysis of data from several locations.

3. CANOCO allows one to test statistically whether the species are related to supplied environmental variables. The test provided is a Monte Carlo permutation test (Hope, 1968). The effect of a particular environmental variable can be tested after elimination of possible effects of other (environmental) variables by specifying the latter as covariables. For the analysis of randomized-block experiments or data from several locations, there is an option to restrict the permutation to permutations among samples-within-blocks or samples-within-locations.

4. CANOCO provides an alternative method of detrending which is intended to solve the problems reported to occur with the method used in DECORANA. CANOCO allows one to remove polynomial relations between ordination axes (up to order 4). Use of the old method of detrending by segments (Hill and Gauch, 1980) in partial and canonical analyses is not recommended.

5. CANOCO has an option for nonstandard analyses. In one possibility, the reciprocal averaging algorithm is modified so that at each iteration the species and/or site scores are replaced by ranks. This procedure circumvents what is known as the "deviant sample/rare species problem" in correspondence analysis.

Data input

CANOCO can read species data, environmental variables and covariables that are either in Cornell condensed format or in full format. The machine readable copy of the analysis can be used again as input for subsequent analyses. This possibility allows one, for example:

- to use principal components extracted from environmental data as input for a later canonical analysis of species data,
- to extract more than four ordination axes - simply by supplying the extracted ordination axes as covariables in a subsequent analysis.

Output options

CANOCO can supply:

- means, variances and correlations of environmental variables,
- eigenvalues, the percentages of variance accounted for by the biplot of species-environment relations,
- scores of species and sites on the ordination axes,
- canonical coefficients or regression coefficients of environmental variables with associated t-values,
- correlations of environmental variables with the ordination axes,
- scores of environmental variables for constructing the arrows in the species-environment biplot,
- centroids (weighted averages) of environmental variables in the ordination diagram (for variables with positive values). In particular, classes of nominal environmental variables are more naturally displayed by their centroid in the ordination diagram than by arrows. This option is also useful for displaying the results of a cluster analysis in an ordination diagram.

CANOCO allows interactive data analysis: results of an analysis can be displayed at the terminal and after inspection the analysis can be pursued, for example,

- by changing from an indirect gradient analysis to a direct gradient analysis,
- by dropping environmental variables,
- by reading other environmental variables to be related to the current ordination axes or to be used in further canonical analyses,
- by changing detrending options,
- by changing scaling options of the ordination scores.

Practical information

CANOCO is written in standard FORTRAN 77 and can be supplied on 5.25 inch diskette for IBM-compatible PC's, on 3.5 inch diskette for ATARI-ST PC's or Apple Macintosh, on magnetic tape (800/1600 bpi, ASCII-code) or via BITNET/EARN. On an IBM-compatible PC with 640 Kb, CANOCO can analyse ca. 750 samples, 600 species, 60 environmental variables and 100 covariables. The one-time costs are ca. Dfl 300 for educational institutes and ca. Dfl 600 for other institutes (prices may change without notice). Researchers from countries with valuta problems may send in a request for a free copy. The comprehensive manual will be sent with the program.

References

Davies, P.T., and Tso, M.K.S. (1982). Procedures for reduced-rank regression. Appl. Statist. 31, 244-255.

Hill, M.O. (1979). DECORANA: a FORTRAN program for detrended correspondence analysis and reciprocal averaging. Section of Ecology and Systematics, Cornell University, Ithaca, New York.

Hill, M.O. and Gauch, H.G. (1980). Detrended correspondence analysis, an improved ordination technique. Vegetatio 42, 47-58.

Hope, A.C.A. (1968). A simplified Monte Carlo significance test procedure. J. Roy. Statist. Soc., Ser. B, 30, 582-598.

Israëls, A.Z. (1984). Redundancy analysis for qualitative variables. Psychometrika, 49, 331-346.

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Appl. Statist. 25, 257-265.

ter Braak, C.J.F. (1987). CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). TNO-Report 87 ITI A 11, TNO Institute of Applied Computer Science, Wageningen, 95 pp.

van den Wollenberg, A.L. (1977). Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42, 207-219.

Wold, H. (1982). Soft modeling. The basic design and some extensions. Pages 1-54 in: Systems under indirect observation. Causality-structure-prediction, Vol. II (Jöreskog, K.G. and Wold, H. eds.) North-Holland Publishers, Amsterdam, 343 pp.

SUMMARY

To assess the impact of environmental change on biological communities knowledge about species-environment relationships is indispensable. Ecologists attempt to uncover the relationships between species and environment from data obtained from field surveys. In the survey, species are scored on their presence or their abundance at each of several sampling sites and environmental variables that ecologists believe to be important are measured.

The research that led to this thesis aimed to unravel the assumptions required for the application of statistical methods that are popular among ecologists to analyse such data. From a statistical point of view, species data are difficult to analyse:

- there are many species involved (10 - 500),
- many species occur at a few sites only. So the data contain numerous zeroes,
- relations between species and environmental variables are not linear, but unimodal: a plant, for example, preferably grows under for that species optimal moisture conditions and is encountered less frequently at drier or wetter sites. A mathematical model for a unimodal relationship is the Gaussian response model.

Standard statistical methods such as linear regression, principal components analysis and canonical correlation analysis are often inappropriate for analysing species data because they are based on linear relationships. One of the methods that ecologists use instead is correspondence analysis. This thesis contributes to the understanding of the underlying response model.

With correspondence analysis, species and sites are arranged to discover the structure in the data (ordination) and the arrangement is subsequently related to environmental variables. It is an indirect method to detect relations between species and environment, hence R.H. Whittaker's term "indirect gradient analysis".

Correspondence analysis has been invented around 1935 but did not receive interest from ecologists before 1973 when M.O. Hill derived the technique once more as the repeated application of "weighted averaging" - a method that was familiar to ecologists ever since 1930. Weighted averaging has the advantage of being simple to apply. The method can be used for two different aims: (1) to estimate the optimum of a species for an environmental variable and (2) to estimate the value of an environmental variable at a site from known optima of the species present (calibration).

In chapter 2, estimating optima by weighted averaging is compared with the results of non-linear regression on the basis of the Gaussian response model. Under particular conditions, both methods agree precisely. In other cases, weighted averaging gives a biased estimate of the optimum and non-linear regression is the method to be preferred. An additional advantage of non-linear regression is that it can also be used to fit response models with more than one environmental variable. In chapter 3, weighted averaging to estimate the value of an environmental variable is compared with calibration on the basis of the Gaussian response model. Also in this context the techniques are sometimes equivalent. Chapter 4 deals with correspondence analysis. It is shown that, under particular conditions, correspondence analysis approximates ordination on the basis of the Gaussian response model, which is computationally much more complicated.

To detect relations, indirect methods have an important disadvantage. The impact of some environmental variables on the species composition can be so large that the impact of other interesting environmental variables may fail to be detected. This problem can be overcome by using non-linear regression, but with many species and environmental variables this is laborious. In chapter 5, a simpler "direct" method is proposed, canonical correspondence analysis. In chapter 6, canonical correspondence analysis turns out to be a multivariate extension of weighted averaging. The results can be displayed graphically. In chapter 7, an extension with "covariables" is discussed, which leads to partial canonical correspondence analysis. Chapter 7 also shows that Gaussian models and, hence, canonical correspondence analysis are relevant to the analysis of contingency tables.

Chapter 8 describes a study to estimate ecological amplitudes of plant species with respect to Ellenberg's moisture scale from species data alone. The question that is addressed as well, is how consequent Ellenberg's moisture indicator values are.

Finally, chapter 9 cross-tabulates various gradient-analysis techniques by the type of problem (regression, calibration, ordination, etc.) and the response model (linear or unimodal). Furthermore, improvements are proposed for detrended correspondence analysis. A computer program, named CANOCO, is written which can perform most of the methods discussed.

Samenvatting

Bij de theoretische onderbouwing van natuurbeheer en milieu-effect-rapportage moeten de gevolgen worden getaxeerd van milieu-ingrepen op levensgemeenschappen. Kennis over de relatie tussen milieuvariabelen en het voorkomen van soorten is daarbij onontbeerlijk. Ecologen proberen die relaties te achterhalen door op verschillende monsterplekken soorten te inventariseren (op aan/afwezigheid of abundantie) en tevens huns inziens relevante milieuvariabelen te meten.

Het onderzoek, dat tot dit proefschrift heeft geleid, richtte zich op het ontrafelen van de vereiste veronderstellingen van statistische methoden, die vaak door ecologen worden toegepast en op het ontwikkelen van een nieuwe techniek.

Vanuit klassiek statistisch oogpunt zijn soortgegevens moeilijk te verwerken:

- er zijn veel soorten bij betrokken (10 - 500),
- heel wat soorten komen maar op weinig plekken voor, dus de gegevens zitten vol nullen,
- verbanden tussen soorten en milieuvariabelen zijn vaak niet rechtlijnig, maar ééntoppig: een plant bijvoorbeeld groeit bij voorkeur onder een voor die soort optimale vochtconditie en wordt zowel op drogere als op nattere monsterplekken minder aangetroffen. Een wiskundig model voor een eentoppig verband is het Gaussische responsiemodel.

Klassieke methoden als lineaire regressie, hoofdcomponenten-analyse en canonische correlatie-analyse kunnen niet zinnig worden gebruikt, omdat ze van rechtlijnige verbanden uitgaan. Eén van de methoden, waar ecologen wel mee werken, is correspondentie-analyse. Het inzicht in het achterliggende responsiemodel hiervan liet tot voor kort te wensen over. Via correspondentie-analyse wordt een ordening in soorten en monsterplekken aangebracht (ordinatie) om de structuur in de gegevens te laten zien. De ordening wordt vervolgens aan de milieuvariabelen gekoppeld. Het is een indirecte methode om relaties op te sporen, ofwel een methode voor indirecte gradienten-analyse.

Correspondentie-analyse werd omstreeks 1935 ontwikkeld, maar staat bij ecologen pas in de belangstelling sinds 1973. Toen leidde M.O. Hill de techniek opnieuw af als het herhaald toepassen van gewogen middelen - een methode waar ecologen al sinds 1930 mee vertrouwd zijn. Gewogen middelen heeft het voordeel van de eenvoud bij toepassing op ecologische gegevens. Deze techniek kan voor twee verschillende doelstellingen worden gebruikt. Ten eerste kan het optimum van een soort voor een milieuvariabele ermee geschat worden. Ten tweede kan bij bekende optima de waarde van een milieuvariabele op een monsterplek worden geschat (calibratie) aan de hand van de soortensamenstelling (dit is ook de methode die Ellenberg aanbeveelt voor gebruik van zijn milieu-indicatiegetallen).

In hoofdstuk 2 wordt het schatten van optima met gewogen middelen vergeleken met de resultaten van niet-lineaire regressie op basis van het Gaussische responsiemodel. Onder bepaalde voorwaarden blijken deze twee methoden precies overeen te komen. In andere gevallen schat men door gewogen middelen het optimum onzuiver en verdient niet-lineaire regessie de voorkeur. Bovendien kunnen met niet-lineaire regressie responsiemodellen met meer dan één milievariabele worden aangepast. In hoofdstuk 3 wordt het schatten van de waarde van een milieuvariabele via gewogen middelen afgezet tegen calibratie op basis van het Gaussische responsiemodel. Ook hier blijken de technieken soms equivalent te zijn. Hoofdstuk 4 gaat in op correspondentie-analyse. Er wordt aangetoond, dat correspondentie-analyse onder bepaalde voorwaarden een

benadering geeft van ordinatie op basis van het Gaussische responsiemodel, wat qua rekentechniek veel ingewikkelder is.

Indirecte methoden voor het opsporen van relaties hebben een belangrijk nadeel. Een aantal milieuvariabelen kan de soortensamenstelling zo sterk beïnvloeden, dat het effect van andere interessante milieuvariabelen niet meer te achterhalen is. Alleen directe methoden als niet-lineaire regressie ondervangen dit probleem, maar niet-lineaire regressie met veel soorten en milieuvariabelen is zeer bewerkelijk. In hoofdstuk 5 wordt een veel eenvoudiger directe methode voorgesteld, canonische correspondentie-analyse. In hoofdstuk 6 blijkt canonische correspondentie-analyse een multivariate uitbreiding van gewogen middelen te zijn. De resultaten kunnen grafisch weergegeven worden. In hoofdstuk 7 wordt een uitbreiding met covariabelen besproken, wat leidt tot partiële canonische correspondentie-analyse. Er wordt tevens op gewezen dat Gaussische modellen en canonische correspondentie-analyse kunnen worden toegepast op afhankelijkheidstabellen.

Hoofdstuk 8 beschrijft onderzoek om ecologische amplitudes van planten ten opzichte van de vochtschaal van Ellenberg te bepalen op basis van alleen soortgegevens. Hoe consequent de vochtindicatie-getallen zijn is ook onderzocht.

Hoofdstuk 9 tenslotte geeft een kruisklassificatie van mogelijkheden voor gradiënten-analyse. Het type probleem (regressie, calibratie, ordinatie, enz.) en het responsiemodel (lineair of unimodaal) zijn hierbij de ingangen. Verder worden verbeteringen voorgesteld voor "detrended correspondence analysis". Er is een computerprogramma ontwikkeld, CANOCO, waarmee het merendeel van de behandelde technieken kan worden uitgevoerd.

CURRICULUM VITAE

Carl Johan Frederik ter Braak werd geboren op 21 juli 1954 te Dodewaard. In 1972 behaalde hij het diploma Gymnasium B aan de Rijksscholengemeenschap "Hamaland" te Winterswijk. Hij studeerde Biologie aan de Rijksuniversiteit Utrecht. In 1975 legde hij cum laude het kandidaatsexamen Biologie af met als tweede hoofdvak Wiskunde. De studie Biologie werd in april 1977 afgerond met als hoofdvakken Ethologie en Mathematische Statistiek en als bijvak Didaktiek van de Biologie. Hij verkreeg tevens onderwijsbevoegdheid in de Biologie en de Wiskunde.

In augustus 1977 trad hij als consulterend biometricus in dienst van het Instituut TNO voor Wiskunde, Informatieverwerking en Statistiek (afdeling Wageningen). In die hoedanigheid werkt hij sinds 1978 (met onderbreking van één studiejaar) voor het Rijksinstituut voor Natuurbeheer te Arnhem en Leersum. Het studiejaar (1979/1980) werd hem verleend om de intern begonnen opleiding tot consulterend biometricus te voltooien. In dat jaar studeerde hij aan de Universiteit van Newcastle upon Tyne (Engeland). Onder leiding van dr. P.J. Diggle verrichtte hij er onderzoek op het gebied van de Ruimtelijke Statistiek. Dit onderzoek mondde uit in de thesis "Binary mosaics and point quadrat sampling in ecology" op basis waarvan hij de graad van Master of Science verkreeg (december 1980). Na een reorganisatie in 1986 maakte hij deel uit van de afdeling Statistiek Landbouw van het Instituut voor Toegepaste Informatica TNO. Deze afdeling wordt eind 1987 door TNO overgedragen aan de Directie Landbouwkundig Onderzoek van het Ministerie van Landbouw en Visserij. Dit proefschrift komt voort uit zijn werkzaamheden voor het Rijksinstituut voor Natuurbeheer.