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).

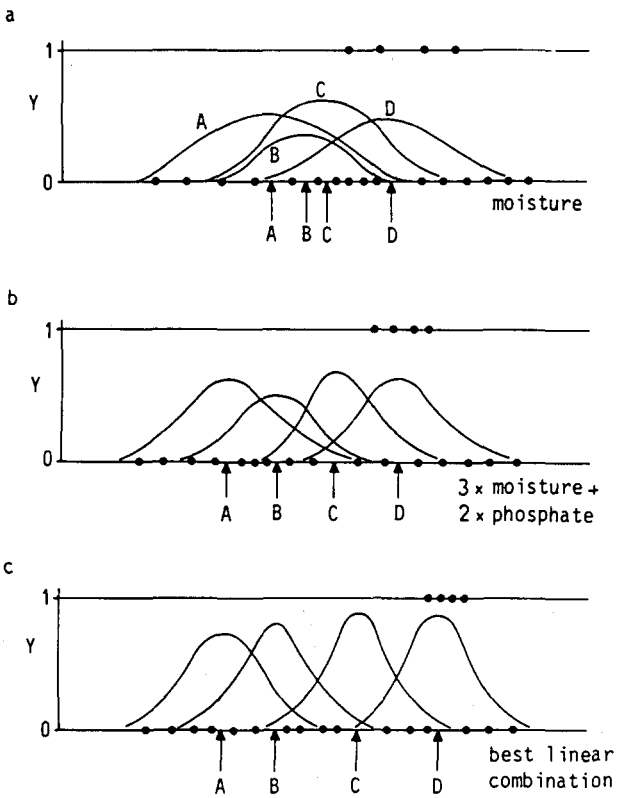


*Fig. 1.* Artificial example of single peaked response curves of four species (A–D) with respect to standardized environmental variables showing different degrees of separation of the species curves: (a) moisture; (b) a linear combination of moisture and phosphate, chosen apriori; (c) the best linear 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 $\delta$ 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.

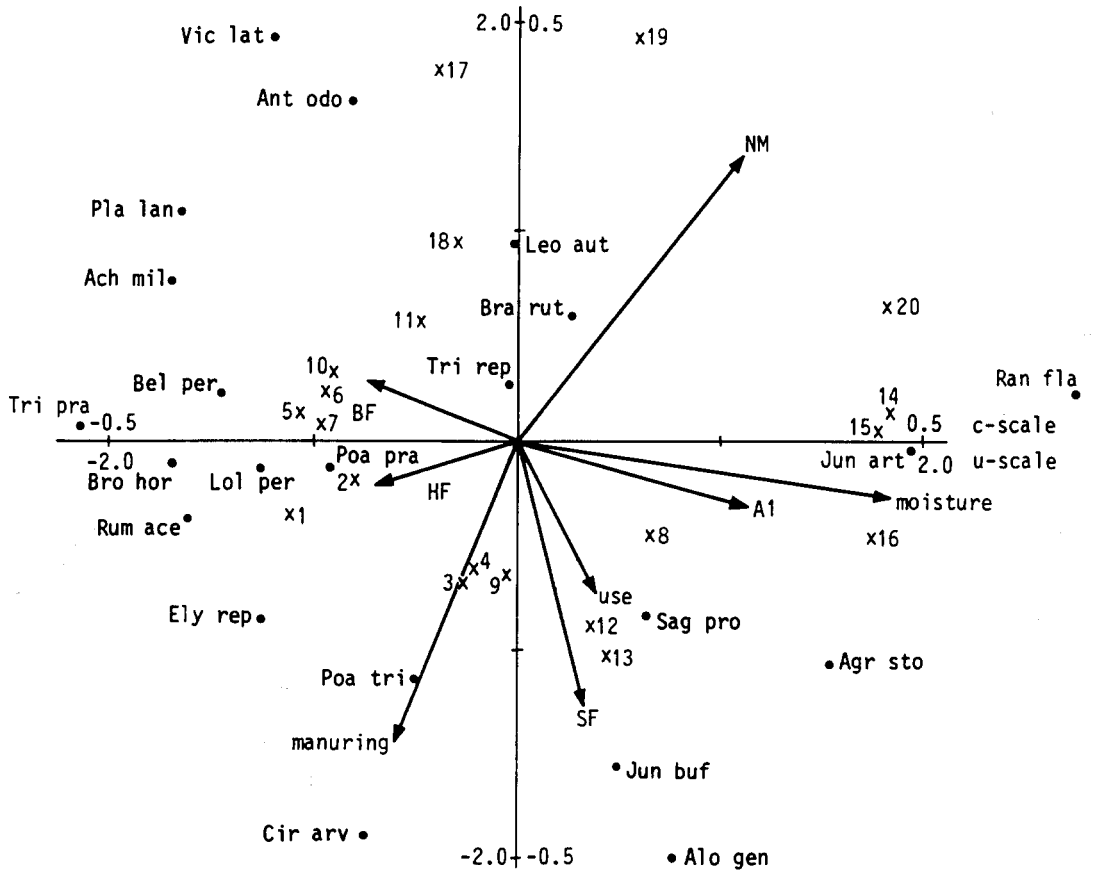


*Fig. 2.* Dune meadow data: CCA ordination diagram with relevés (x), plant species ( • ) and environmental variables (arrow); first axis horizontally, second axis vertically. For relevé numbers see Table 1. Abbreviations are given as *underlining* in full names in Table 1. The c-scale applies to the 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

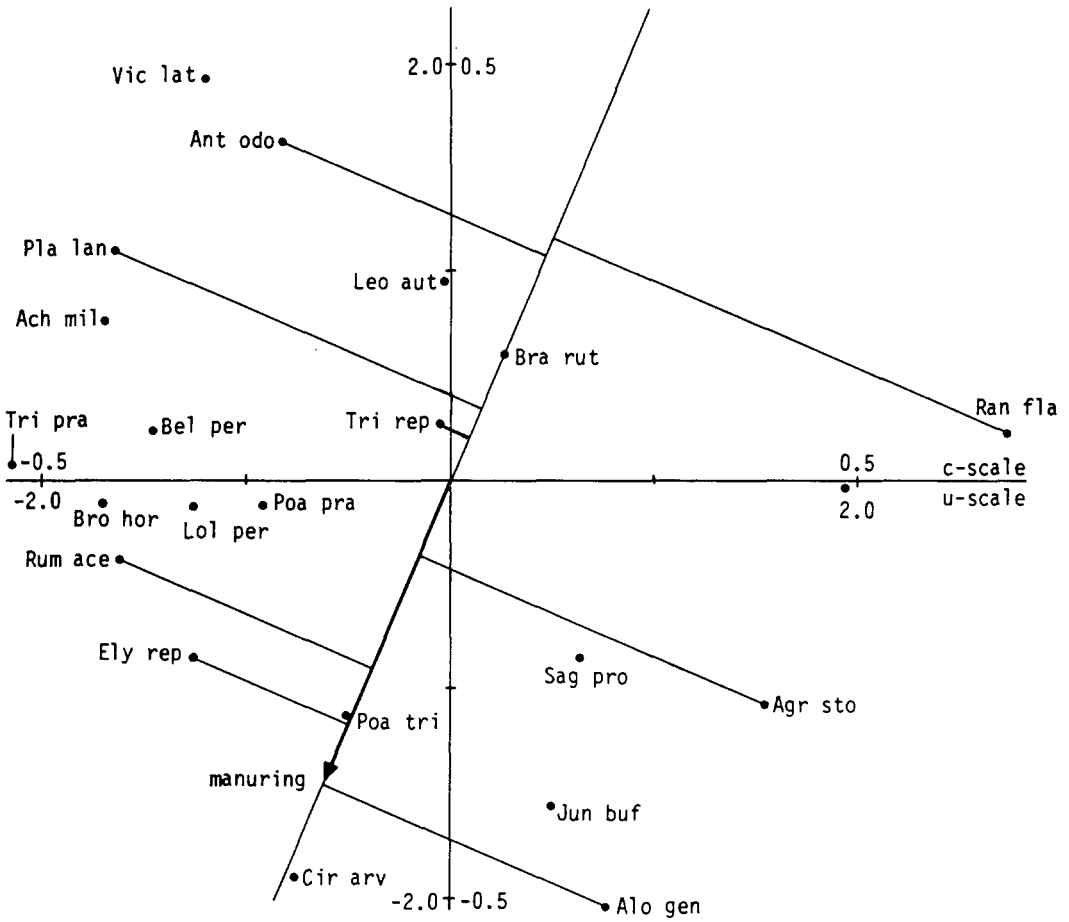


*Fig. 3.* Inferred ranking of the species along the variable quantity of manuring based on the biplot interpretation of Fig. 2. For explanation see the Ordination diagram section.

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<br>1 111 11 11112<br>51670217834923894560 |
|---|---|
| *Tri*folium *pra*tense | 2-52---------- |
| *Ach*illea *mil*lefolium | 212243-2---------- |
| *Bro*mus *hor*deaceus | 2--244----3--------- |
| *Pla*ntago *lan*ceolata | 5-553-323---------- |
| *Rum*ex *ace*tosa | 5-63-------22------- |
| *Bel*lis *per*ennis | 2---23--222--------- |
| *Ely*mus *rep*ens | 44---4---446 ------- |
| *Lol*ium *per*enne | 2766657-2652--4----- |
| *Vic*ia *lat*hyroides | ----1-2-1---------- |
| *Poa pra*tensis | 243444413544-24----- |
| *Ant*hoxanthum *odo*ratum | 4-324--4-------4---- |
| *Cir*sium *arv*ense | ----------2--------- |
| *Poa tri*vialis | 624547---655494---2- |
| *Tri*folium *rep*ens | 2-52653-2213322261-- |
| *Leo*ntodon *aut*umnalis | 3-3335525222223622-2 |
| *Bra*chythecium *rut*abulum | 2-622-4-62224-23-444 |
| *Jun*cus *buf*onius | ---2-------443------ |
| *Sag*ina *pro*cumbens | ------2---524223---- |
| *Alo*pecurus *gen*iculatus | -----2---723855---4- |
| *Hyp*ochaeris *rad*icata | ------22-------5---- |
| *Aira praec*ox | -------2-------3---- |
| *Sal*ix *rep*ens | --------2------3---5 |
| *Agr*ostis *sto*lonifera | ---------483454-4475 |
| *Jun*cus *art*iculatus | ----------4--4--334 |
| *Che*nopodium *alb*um | ----------1------ |
| *Emp*etrum *nig*rum | ----------2---- |
| *Ran*unculus *fla*mmula | ----------22-2224 |
| *Ele*ocharis *pal*ustris | ----------4-4584 |
| *Cal*liergonella *cus*pidata | ----------4-33 |
| *Pot*entilla *pal*ustris | ----------22-- |
| thickness *A1* | 40100001211133117930 |
| *moisture* | 11112112122445555555 |
| quantity of *manuring* | 24231210044123311131 |
| agricultural *use* | 12231231122122313231 |
| Standard *F*arming | 01000000011011000010 |
| *B*io-dynamic *F*arming | 00001110000000000000 |
| *H*obby *F*arming | 10110000000100100000 |
| *N*ature *M*anagement | 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 ($\lambda_1 = 0.46$ and $\lambda_2 = 0.29$) 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 $\delta$ 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

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

Hence,

$$\delta = y_{++}^{-1}\mathbf{u}'\mathbf{S}_{11}\mathbf{u} = y_{++}^{-1}\mathbf{b}'\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{b} \tag{A.2}$$

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

$$y_{++}^{-1}\mathbf{b}'\mathbf{S}_{22}\mathbf{b} = I \tag{A.3}$$

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

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

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 $\delta$ leads to the first axis of CCA as defined in Ter Braak (1986). Further, maximizing $\delta$ 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.

$$\mathbf{W} = \mathbf{S}_{11}^{-1}\mathbf{Y}'\mathbf{Z} \tag{A.5}$$

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

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

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

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

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

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

The pre- and post-multiplication factors involving $y_{++}$ 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.