SORT 32 (1) January-June 2008, 93-112

Canonical non-symmetrical correspondence analysis: an alternative in constrained ordination

Willems, Priscila M.[1] and Galindo Villardon, M. Purificación[2]

[1] *Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina*
[2] *Departamento de Estadística, Universidad de Salamanca, España.*

Abstract

Canonical non-symmetrical correspondence analysis is developed as an alternative method for constrained ordination, relating external information (e.g., environmental variables) with ecological data, considering species abundance as dependant on sites. Ordination axes are restricted to be linear combinations of the environmental variables, based on the information of the most abundant species. This extension and its associated unconstrained ordination method are terms of a global model that permits an empirical evaluation of the impact that the environmental variables have on the community composition. Scores, contributions, qualities of representation, interpretation of dispersion graphs and an application to real vegetation data are presented.

MSC: 62H25, 62P12

Keywords: Biplot; canonical correspondence analysis; non-symmetrical correspondence analysis; species-environment relationship

1 Introduction

The study of ecological communities is generally based on the analysis of two data tables, one that contains information on species compositions at given sites (e.g., abundance, cover of species; table $\mathbf{Y}$), and another containing habitat measurements at those sites, information on environmental variables that affect the distribution of

[1] Instituto Nacional de Tecnología Agropecuaria (INTA), Estación Experimental Agropecuaria Bariloche, C.C. 277, (8400) S.C. de Bariloche, Río Negro, Argentina. Email Address: pwillems@bariloche.inta.gov.ar

[2] Universidad de Salamanca, Departamento de Estadística, C/Espejo 2, 37007 Salamanca, España. Email Address: pgalindo@usal.es

Received: March 2005
Accepted: October 2007

those species (table **Z**). The objective is the detection of species distribution patterns or the ordination of sites compatible with a given gradient, and the study of the relationship between these results and the measured environmental variables. To achieve this, unconstrained or constrained ordination methods are used.

These data tables contain multidimensional information, part of which is redundant. Multivariate techniques that arrange sites along axes on the basis of species composition data are called (unconstrained) ordination methods. They can be thought of as methods for matrix approximation to summarize that information or as methods to detect the latent structure of such tables (ter Braak 1987[b]). Correspondence analysis (CA, Benzecri 1973) or a modification of CA called detrended correspondence analysis (DCA, Hill and Gauch 1980), are examples of such techniques. Ordination axes can be interpreted in relation to environmental variables through correlation coefficients between them and the external variables. This two-stage process, in which environmental gradients are inferred from the ecological data, is known as indirect gradient analysis (Whittaker 1967).

External information can also play an active role in the analysis by imposing the restriction that ordination axes should be linear combinations of the environmental variables, in a direct gradient analysis context, in which case the process of identifying the latent structure is called constrained ordination. Redundancy analysis (RDA - Rao 1964, van der Wollenberg 1977) and canonical correspondence analysis (CCA - ter Braak 1986) are two common techniques of this type, being the latter more appropriate for this type of data due to the relationship between species abundance and external variables. A general approach for constrained analysis is introduced by Anderson and Willis (2003), which take into account the correlation structure among the variables in the species table, in contrast with the traditional methods, like RDA or CCA, and with the one presented here. Notice that the role played by the tables **Z** and **Y** is asymmetric. Chessel and Mercier (1993) and Dolédec and Chessel (1994) consider the study of covariation of both tables in a symmetric approach, known as co-inertia analysis.

Under the assumption that the relationship between species data and environmental variables follows a Gaussian response curve, Gauch *et al.* (1974) proposed an ordination technique called Gaussian ordination. Ter Braak (1985) demonstrated that CA (with a single gradient) and DCA (with two gradients), approximate the results of a Gaussian ordination when the "packing model conditions", described by Hill (1979), are satisfied.

However, one of the drawbacks of CA is the excessive weight given to sites that contain rare species, due to the use of the chi-square metric (see also Cuadras *et al*. 2006, for CA advantages and drawbacks). As a consequence, those sites are placed as atypical points on the extremes of the first ordination axis. By contrast, CCA minimizes this problem provided those sites are not atypical in terms of the environmental variables (ter Braak 1987[a]).

As an alternative to CA, Gimaret-Carpentier *et al*. (1998) proposed, for the analysis of species occurrences data, the use of non-symmetrical correspondence analysis (NSCA) developed by Lauro and D'Ambra (1984). This technique gives uniform weight

to species (considering them as depending on sites), and its results are based on the most abundant ones. Furthermore, in NSCA the possibility of appearance of the arch effect, typical in CA representations, is reduced, unless it results from inherent data features (Gimaret-Carpentier *et al*. 1998).

Starting from the species occurrence table, Pélissier *et al*. (2003) worked on NSCA and CA relating them to diversity indices, and mentioned that the binary table describing sites could be replaced by another table of arbitrary variables.

The objective of this work is to develop this approach with respect to NSCA, but working directly on the sites-by-species and sites-by-environmental variables tables. That is, we extend NSCA as a constrained ordination method in a direct gradient analysis context, calling it *canonical non-symmetrical correspondence analysis* (CNCA). Scores for sites, species and environmental variables, with their respective indicators of contribution and qualities of representation will be given. Site-profiles (used to identify floristic affinities between sites – Gimaret-Carpentier *et al*. 1999), which take species as dependent on sites, will be considered.

Even when unconstrained and constrained ordination methods can be seen as pointing out to different purposes (Økland 1996), they can also be seen as terms of a general model that partition the total variance of the species data into components of explained and unexplained variance through the external variables. This approach was developed by Takane and co-workers (e.g., Takane and Hunter 2001), principally oriented to other research areas, as a way of investigating the empirical validity of the hypotheses incorporated as external constraints. The proposed method (CNCA), together with its unconstrained counterpart (NSCA), are terms of the general model, so in this work both analyses are taken as complementary, aiming to investigate the species-environment relationship.

2 Canonical non-symmetrical correspondence analysis (CNCA)

Let $\mathbf{Y} = (y_{ik})$, of order $n \times q$, be a data table containing information on q species (e.g. abundance, biomass or cover) measured at n sites, and $\mathbf{Z} = (z_{ij})$, of order $n \times p$, a second data table with p quantitative environmental variables, with non-singular (weighted) covariance matrix, measured at the same sites, such that y_{ik} is the descriptor of the k-th species at site i, and z_{ij} the value of the j-th environmental variable at site i. Let $\mathbf{F} = y_{..}^{-1}\mathbf{Y}$ be the associated correspondence matrix, where $y_{..}$ is the grand total of $\mathbf{Y}$.

Without loss of generality, let $\mathbf{Z}$ be standardized, to make the estimates of their coefficients comparable. This is done with $\mathbf{D}_n$-weighted means and variances:

$$\sum_{i=1}^{n} f_{i.}\, z_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} f_{i.}\, z_{ij}^2 = 1, \quad j = 1, \ldots, p,$$

where $\mathbf{D}_n = \text{diag}\,(f_{1.}, f_{2.}, \ldots, f_{n.})$, the diagonal matrix of the vector $\mathbf{f}_n = \mathbf{F}\,\mathbf{1}_q$, the row (site) marginals of $\mathbf{F}$, where $\mathbf{1}_q$ is a vector of ones.

We develop CNCA by two different approaches. The first one, as an ordination technique with instrumental variables, and the second one, from the analysis of an inter-table $\mathbf{L}$ obtained from the two basic tables $\mathbf{Y}$ and $\mathbf{Z}$.

First CNCA approach. Within this framework, this technique consists, as other constrained ordination methods like RDA and CCA, in analysing the fitted values of species information by the respective ordination technique, with metrics and weights determined by the original data. These fitted values come from the orthogonal (with respect to metric $\mathbf{D}_n$) projection of the original values onto the subspace generated by the columns of $\mathbf{Z}$. With respect to RDA this approach is given in Rao (1964), and with respect to CCA, in Lebreton *et al*. (1988).

When NSCA is used to analyse the information on species composition (with sites in rows), the starting point is the centred row-profile matrix $\tilde{\mathbf{P}}$:

$$\tilde{\mathbf{P}} = \mathbf{P} - \mathbf{1}_n\,\mathbf{f}_q^{\mathsf{T}} = \mathbf{D}_n^{-1}\left(\mathbf{F} - \mathbf{f}_n\,\mathbf{f}_q^{\mathsf{T}}\right), \tag{1}$$

where $\mathbf{P} = \mathbf{D}_n^{-1}\,\mathbf{F}$, the row-profile matrix, and $\mathbf{f}_q = \mathbf{F}^{\mathsf{T}}\,\mathbf{1}_n$ the vector of q column (species) marginals $f_{.k}$, $k = 1, \ldots, q$. For each species, matrix $\tilde{\mathbf{P}}$ provides the magnitude of the difference between its participation in that site-profile and the average profile, indicating a greater or smaller relative abundance given that site.

This approach defines CNCA as a NSCA on the projected centred row-profile matrix $\tilde{\mathbf{P}}^*$, with Euclidean metric and site weights $f_{i.}$, $i = 1, \ldots, n$, such that:

$$\tilde{\mathbf{P}}^* = \mathbf{\Pi}\,\tilde{\mathbf{P}} \tag{2}$$

being $\mathbf{\Pi} = \mathbf{Z}\left(\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n$ the orthogonal projector with respect to metric $\mathbf{D}_n$. The asterisk indicates that a matrix contains projected values.

Then, CNCA is based on the analysis of the triplet $(\tilde{\mathbf{P}}^*, \mathbf{I}_q, \mathbf{D}_n)$. It follows from the generalized singular value decomposition (GSVD, Greenacre 1984, p. 40):

$$\tilde{\mathbf{P}}^* = \mathbf{R}\,\Lambda\,\mathbf{T}^{\mathsf{T}} \tag{3}$$

such that $\mathbf{R}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{R} = \mathbf{I}_v$ and $\mathbf{T}^{\mathsf{T}}\mathbf{T} = \mathbf{I}_v$, where $\mathbf{R}$ and $\mathbf{T}$ are the left and right singular vectors matrices of $\tilde{\mathbf{P}}^*$, respectively, with diagonal matrix Λ containing the respective singular values, such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_v$, where v (rank of $\mathbf{Z}$) is the maximum number of constrained axes.

In this first approach, notice that the difference between CNCA and CCA, as it is for NSCA and CA (Pélissier *et al.* 2003), is the metric considered on the site space (when analysing site-profiles). Nevertheless, as is discussed in Section 4, this difference has implications that go further than a simple algebraic change.

Second CNCA approach. From this perspective, CNCA is defined as the analysis of the inter-table $\mathbf{L}$, of order $q \times p$, such that:

$$\mathbf{L} = \tilde{\mathbf{F}}^{\top}\, \mathbf{Z}, \tag{4}$$

where $\tilde{\mathbf{F}} = \mathbf{F} - \mathbf{f}_n\, \mathbf{f}_q^{\top}$ is the matrix that contains in its *ik-th* element the magnitude of the departure of the observed value f_{ik} from the independence hypothesis. In our context, $\tilde{\mathbf{F}}$ measures the magnitude of the difference between the observed (relative) abundance and that which would result in the presence of a random distribution of species through sites.

The kj-th element of $\mathbf{L}$ is the weighted total difference on the j-th environmental variable between the sites that possess relative abundance of species k greater than the one expected in the presence of a random distribution of species, and those whose relative abundance is less than that value, the weights being the corresponding elements of $\tilde{\mathbf{F}}$. Then, matrix $\mathbf{L}$ gives greater weight to species that contribute more largely to the differentiation between sites, giving less participation to those of low presence.

To obtain score estimates invariant to non-singular linear transformations of the external variables, these variables are weighted by the inverse of their covariance matrix. Thus, the solution follows from the singular value decomposition of $\mathbf{L}_p = \left(\mathbf{Z}^{\top}\, \mathbf{D}_n\, \mathbf{Z}\right)^{-1/2} \mathbf{L}^{\top}$:

$$\mathbf{L}_p = \mathbf{A}\, \Lambda\, \mathbf{T}^{\top} \tag{5}$$

such that $\mathbf{A}^{\top}\, \mathbf{A} = \mathbf{I}_v$ and $\mathbf{T}^{\top}\, \mathbf{T} = \mathbf{I}_v$, where $\mathbf{A}$ and $\mathbf{T}$ contain in their columns the left and right singular vectors of $\mathbf{L}_p$, respectively, with Λ a diagonal matrix containing the respective singular values ($\mathbf{T}$ and Λ are equivalent to the ones in (3)).

- *Sites and species scores in CNCA*

Scores for rows (sites) and columns (species) are obtained by solving equation (3), rewritten as $\tilde{\mathbf{P}}^{*} = \mathbf{X}\, \mathbf{T}^{\top}$. Hence

$$\mathbf{X} = \mathbf{R}\, \Lambda = \tilde{\mathbf{P}}^{*}\, \mathbf{T} \tag{6}$$

contains in its columns the principal coordinates of sites, with covariance matrix Λ^2, where $\mathbf{R}$ contains the site scores in standard coordinates; while the columns of $\mathbf{T}$ contain species standard coordinates, of unit variance, and $\mathbf{U} = \mathbf{T}\, \Lambda = \tilde{\mathbf{P}}^{*\top}\, \mathbf{D}_n\, \mathbf{R}$ the species principal coordinates.

- *Environmental variable scores in CNCA*

Since site scores are linear combinations of environmental variables, canonical weights ($\mathbf{C}$) of these variables can be obtained through the following expression:

$$\mathbf{C} = \left(\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{X} \tag{7}$$

Depending on the chosen scaling, equation (7) can be written in terms of the principal coordinates **X** or the standard coordinates **R**. These canonical weights are partial regression coefficients of the multiple regression of site scores on the environmental variables (measuring conditional effects). Thus when external variables are correlated, giving rise to the known multicollinearity problem, the interpretation of the canonical weights must be done carefully.

Other environmental variable score estimates are obtained from equation (5). These estimates represent marginal effects, and therefore, independent of possible correlations between them. Rewriting that equation as $\mathbf{L}^{\mathsf{T}} = (\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z})^{1/2}\,\mathbf{A}\,\Lambda\,\mathbf{T}^{\mathsf{T}}$, we obtain those scores in principal coordinates as:

$$\mathbf{B} = \left(\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z}\right)^{1/2}\,\mathbf{A}\,\Lambda, \tag{8}$$

with a corresponding expression in standard coordinates in $\mathbf{B}^{\circ} = \left(\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z}\right)^{1/2}\,\mathbf{A}$.

Matrix **B** contains simple regression coefficients of site scores on each one of the environmental variables, so the length of each vector quantifies the rate of change of that variable in the observed distribution of sites. When **Z** is standardized, these coefficients are comparable, and so are their respective lengths. Thus, in this situation, environmental variables with vectors of greater lengths are more related to ordination axes. With standardized environmental variables, $\mathbf{B}^{\circ}$ contains intraset correlations, correlations between those variables and site scores (it also evaluates marginal effects). The interpretation of **B** and $\mathbf{B}^{\circ}$ is analogous to that of the respective estimators in CCA (ter Braak 1986, ter Braak and Verdonschot 1995).

From (5), an expression for canonical weights (equivalent to (7)), can be obtained as $\mathbf{C}_L = \left(\mathbf{Z}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{Z}\right)^{-1/2}\,\mathbf{A}$. And, from these canonical weights, **X** can be estimated (equivalent to (6)).

- *Transition relationships*

  Transition formulas that relate site scores with species scores are:

$$\mathbf{U} = \tilde{\mathbf{P}}^{*\mathsf{T}}\,\mathbf{D}_n\,\mathbf{R} = \tilde{\mathbf{P}}^{\mathsf{T}}\,\mathbf{D}_n\,\mathbf{R} \tag{9}$$

$$\mathbf{X} = \tilde{\mathbf{P}}^{*}\,\mathbf{T} \tag{10}$$

Equations (9) (first two members) and (10) show the usual transition relationships of NSCA, taking as coefficients of the linear combination those coming from the columns and rows of $\tilde{\mathbf{P}}^{*}$, respectively. To make those relations interpretable in terms of the original data, the coefficients must come from $\tilde{\mathbf{P}}$, that is, from the data before projection

as in the last term in (9). This latter concept is not applicable to **X** as it is for **U**. By definition, site scores are constrained to be linear combinations of the external variables.

However, an alternative set of site scores can be calculated through coefficients resulting from $\tilde{\mathbf{P}}$ once **U** is obtained: $\mathbf{X}_{sp} = \tilde{\mathbf{P}}\,\mathbf{U}^{\circ}$. The correlations between analogous columns of this last score matrix and those in **X** (coming from environmental variables), are called species-environment correlations (same concept as in RDA and CCA, with their respective definitions of site scores). The square of each correlation is equal to the coefficient of determination of the multiple regression of each column of $\mathbf{X}_{sp}$ on the external variables in **Z**, where the respective column of **X** contains the predicted values of that regression.

Joint representation-biplot projections

Joint interpretations of species/sites and species/environmental variables, both representations superimposed in the same graph, are made by biplot rules, since each one of them is a biplot representation (Gabriel 1971). As CNCA starts from site-profiles, studying site distribution according to their species composition restricted by environmental conditions, we propose the used of site-conditional scaling (site and environmental scores in principal coordinates and species scores in standard coordinates).

The inner products of species and site scores are the fitted values of the centred site-profiles (see (2)), or their approximations if only the first axes are considered. As in NSCA species projected toward the positive side of the vector joining the origin with each one of the site scores, indicate those species with greater increase in probability of having important values of fitted abundance in that site (that is, with estimated values of its relative abundance superior to the marginal average). On the other hand, if that projection is found on the negative side, that shows a decrease in probability of having important values. Given the relationship through scalar products, a species can be represented far away from a site in terms of Euclidean distance, and due to its projection, be of relative importance at that site.

A small distance between two sites (well represented) indicates they have similar fitted species distribution. In these interpretations, the effect of two approximations must be taken into account, one due to the restriction done by regression, and the other due to reduction of the restricted space.

The inner product between species and environmental variable scores are the values of matrix **L** (see (4)), or their approximations if only the first axes are considered. Then, for each species, the approximation is the difference that exists in that environmental variable between sites with relative abundance superior to the one expected in the presence of a random distribution of species, and sites with values inferior to this reference quantity.

In site-conditional scaling, site scores (6) and environmental variable scores (8) are not a biplot representation. Anyway, the direction of each vector representing an environmental variable identifies sites where the values of the variable become greater.

Standard coordinates of environmental variables are a biplot representation of their correlation matrix. The scalar product between any pair of vectors defined by those coordinates is the correlation between the corresponding two variables. From the dispersion graph, the sign of these correlations can be deduced from the angle determined by each pair of vectors: acute, positive correlation; obtuse, negative.

Contributions and qualities of representation

Contributions and qualities of representation are measures that give information about the elements (sites or species), showing those with greater contribution to the orientation of factorial axes and evaluating the quality of their representation. Graffelman (2001) presented quality measures in CCA, especially for axes, some of them from a different perspective.

a) Contributions in CNCA:

These measures, $C_{\alpha,m}$ (which represent the contribution of the m-th element in the α-th factor determination), express the proportion of variability of that factor (determined by its eigenvalue) accounted for by that element. It takes values between zero (without contribution) and one (which would indicate that axis α is determined only by that element).

Their expressions for CNCA are (subscript m becomes i for sites and k for species):

– For sites: $C_{\alpha,i} = \dfrac{f_{i.}\, x_{i\alpha}^2}{\lambda_\alpha^2} \quad i = 1, \ldots, n; \quad \alpha = 1, \ldots, v,$

where $x_{i\alpha}$ is the principal coordinate of site i on axis α.

– For species: $C_{\alpha,k} = \dfrac{u_{k\alpha}^2}{\lambda_\alpha^2} \quad k = 1, \ldots, q; \quad \alpha = 1, \ldots, v,$

where $u_{k\alpha}$ is the principal coordinate of species k on axis α.

b) Qualities of representation in CNCA:

Table 1 presents measures of these qualities for factorial axes and for elements. They are given related to two different spaces, with respect to: a) the projected space, that is, the total inertia of the fitted values (indicated with superscript *PS*), and b) the original space, total inertia of the observed values (indicated with superscript *OS*).

Table 1: *Qualities of representation for factorial axes and elements (species and sites) scores, with respect to the projected space (PS) and to the original space (OS). ($\alpha = 1, \ldots, v$)*

		With respect to the projected space (superscript PS)	*With respect to the original space (superscript OS)*
Relative to factorial axes		$I_\alpha^{(PS)} = \dfrac{\lambda_\alpha^2}{\sum_{\alpha=1}^{v} \lambda_\alpha^2}$ Proportion of inertia explained by axis α, with respect to the total inertia of the projected data.	$I_\alpha^{(OS)} = \dfrac{\lambda_\alpha^2}{TI(NSCA)}$ Proportion of inertia explained by axis α, with respect to the total inertia (TI) of the original data.
Relative to elements	*Sites*	$Q_{\alpha,i}^{(PS)} = \dfrac{x_{i\alpha}^2}{d_{\mathrm{Pr},i}^2}$	$D_{\alpha,i} = \dfrac{x_{i\alpha}^2}{d_{O,i}^2}$ (1)
	Species	$Q_{\alpha,k}^{(PS)} = \dfrac{u_{k\alpha}^2}{d_{\mathrm{Pr},k}^2}$	$Q_{\alpha,k}^{(OS)} = \dfrac{u_{k\alpha}^2}{d_{O,k}^2}$

(1) $D_{\alpha,i}$ could be greater than one. See text for more details.

$d_{\mathrm{Pr},k}^2, d_{\mathrm{Pr},i}^2, d_{O,k}^2, d_{O,i}^2$: Square distances in the projected (Pr) and original (O) spaces, of species (k) and sites (i), respectively.

Those given for elements are denoted $Q_{\alpha,m}$ and take values between zero and one, being a squared cosine (one indicates an exact representation). These indicators represent the proportion of the *m-th* element variance that is explained by axis α, where this variance is evaluated as the squared distance of that element to the centroid.

For species, the variance of the fitted values is smaller, or equal, to the variance of the original values (because of regression properties). But this is not the case for sites, since the square of the distance to the respective centroid of the projected values could be greater than its own distance in the original space. The statistic for sites relative to the original space is denoted $D_{\alpha,i}$(to differentiate it from those which measure quality). It is not a squared cosine, and it will indicate for site i the ratio between the square of the distance accounted for by axis α with respect to the square of the total distance of that site in the original space. The statistic $Q_{\alpha,k}^{(OS)}$ indicates how well the environmental variables explain each one of the species.

3 Application to real data

Floristic data on 45 samples (sites) were analyzed, obtained in meadows of Río Negro and Neuquén provinces (Argentina), aiming to determine the influence of the environmental

conditions on the vegetation distribution. Each sample information consisted of the list of observed species with their visual cover estimate (Braun-Blanquet 1950). Species with very low frequency were removed. Thirty-two species were considered; some of which appeared in few sites with cover values less than 3% (nine species), or with only one important cover value and the others with negligible importance (four species). At those sites, five environmental variables were measured: annual mean precipitation (Z_1), pH at the superficial cap soil, 0-20 cms. (Z_2), watertable depth (Z_3), electrical conductivity of the superficial cap soil (Z_4), percentage of bare soil (Z_5).

The data were first analyzed by the constrained ordination method developed here (CNCA), and then by its indirect analysis counterpart (NSCA), in order to evaluate the impact that the chosen environmental variables have on the community composition. At last, the information was analyzed by CCA, to give a brief comparison between CNCA results and those of CCA. Species data were transformed by taking logarithm (log[cover+1]), because of their skewed distribution and to down-weight high values (also, as CNCA and NSCA are based on the most abundant species, this transformation tends to give more participation to the species with lower values than the one they have with their original values). The analysis was performed using procedure IML of the SAS System (Version 8).

Table 2 contains the cumulated percentages of inertia explained by the factorial axes of CNCA (it includes also those for NSCA and CCA). For CNCA, the value 40.7%, cumulated by the five constrained axes, indicates that an acceptable proportion of the total inertia of the original data is explained by the axes obtained from the fitted values, being 84.6% of this percentage explained by the first two axes (which were chosen as the reduced solution space).

Table 2: *Cumulated percentages of inertia for meadow data explained by the first five factorial axes, for unconstrained (NSCA) and constrained (CNCA and CCA) ordination.*

Axes	*Unconstrained Ordination*	*Constrained Ordination*			
	NSCA	*CNCA*		*CCA*	
		Cum. $I_\alpha^{(PS)(*)}$	*Cum.* $I_\alpha^{(OS)(**)}$	*Cum.* $I_\alpha^{(PS)(*)}$	*Cum.* $I_\alpha^{(OS)(**)}$
1	30.8	65.9	26.7	55.2	15.1
2	46.6	84.6	34.4	73.5	20.1
3	59.5	95.6	38.9	88.6	24.3
4	66.0	98.7	40.2	97.0	26.6
5	71.0	100.0	40.7	100.0	27.4

$^{(*)}$ With respect to the total inertia of the fitted values in the projected space.

$^{(**)}$ With respect to the total inertia of the observed values in the original space.

Table 3: *Species that contribute to the orientation of the first two factorial axes ($C_{\alpha,k}$), in at least one of the three analyses, CNCA, CCA or NSCA (in boldface those that contribute in that particular case).*

Species	*Abbrev.*	*Constrained Ordination*				*Unconstrained Ordination*	
		CNCA		*CCA*		*NSCA*	
		Axis 1	***Axis 2***	***Axis 1***	***Axis 2***	***Axis 1***	***Axis 2***
Azorella trifurcata	AzoTrif	0.00	**10.90**	0.00	**12.55**	0.04	**10.45**
Berberis heterophylla	BerbHet	0.21	0.03	**2.74**	0.53	0.10	0.00
Boopis sp.	Boopis	0.49	0.00	**2.74**	0.00	0.39	0.00
Carex gayana	CareGay	0.75	**5.67**	1.75	**8.47**	0.73	1.09
Chuquiraga erinacea	ChuqEri	0.52	0.06	**5.08**	0.62	0.26	0.02
Cortaderia araucana	CorArau	0.22	0.00	0.44	0.50	0.00	**12.80**
Distichlis sp	DistSp	**54.84**	0.74	**27.97**	1.38	**59.80**	0.44
Eleocharis albibracteata	EleAlb	**6.43**	**4.95**	**6.02**	**3.94**	**5.10**	**2.23**
Festuca pallescens	FestPal	1.53	**44.90**	1.13	**31.12**	1.39	**54.48**
Holcus lanatus	HolcLan	0.90	1.46	**2.09**	**2.39**	0.89	0.68
Juncus balticus	JunBal	**6.71**	**10.35**	**2.34**	**2.64**	**7.48**	**5.04**
Lycium repens	LycRepe	1.29	0.68	**5.95**	**3.11**	1.21	0.08
Nitrophylla australis	NitrAus	**6.45**	**2.11**	**11.59**	**4.43**	**4.90**	0.19
Plantago marítima	PlanMar	0.29	0.30	**2.56**	**3.57**	0.29	0.01
Poa lanuginosa	PoaLanu	0.66	0.00	**3.73**	0.11	0.62	0.12
Poa pratensis	PoaPra	**3.46**	**7.66**	**3.19**	**4.70**	**2.68**	**3.07**
Samolus spatulatus	SamSpat	0.18	0.70	0.96	**2.44**	0.25	0.45
Stipa speciosa var *major*	StipMajo	0.21	1.01	1.49	**4.42**	0.10	0.72
Stipa speciosa var *spec.*	StipSpec	0.54	0.10	**2.91**	0.13	0.44	0.02
Taraxacum officinalis	TarOff	**9.62**	**4.70**	**4.77**	**2.66**	**9.00**	0.56
Trifolium repens	TriRepe	**4.15**	1.76	**4.56**	1.33	**4.02**	**3.22**

In Table 3, species with greater contributions to the orientation of the first two factorial axes are described (for CNCA, and also for CCA and NSCA). Contributions related to sites are not presented. They were distributed among most of the sites.

Figure 1 represents the first CNCA factorial plane. Species quality of representation with respect to the projected space (those indicated with superscript *(PS)* in Table 1, values not presented here), showed that eight of them were not well represented. Considering their qualities with respect to the original variances (indicated with superscript *(OS)* in Table 1), species represented in Fig. 1.a are those with that quality greater than 0.30 in that plane. *Azorella trifurcata* was also included (even though its *(OS)* indicator

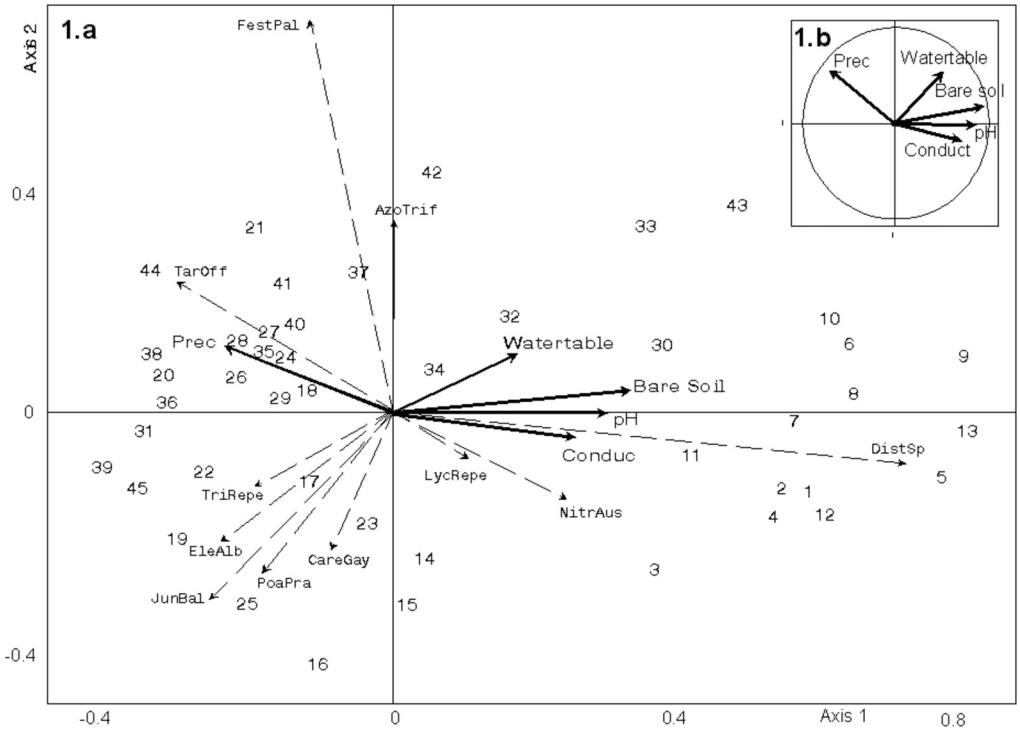

Figure 1: ***1.a.*** *First factorial plane of the CNCA (site-conditional) ordination diagram. Axes for sites (which are indicated by numbers) and environmental variables (indicated by solid vectors) should be scaled by 0.5. Dashed vectors represent species (abbreviations in Table 3) that contribute most to the orientation of the axes and/or have* $Q_{\alpha,k}^{(OS)}$ *in that plane greater than 0.3. Percentages of explained inertia are shown in Table 2.* ***1.b.*** *Intraset correlations between environmental variables and the first two CNCA factorial axes.*

is lower than that value), because of its contribution to the orientation of the second axis, furthermore its location being coherent with the data. These species, which are better described by the external variables, turn out to be the most representative of this type of ecosystem and are those that characterized the communities mentioned below. About the sites, their *(PS)* qualities were in general good, while the statistic $D_{\alpha,i}$ gave relatively low values in 30% of them.

In Fig. 1.a, from right to left, there is a gradient from the lowest recorded values of rainfall related to the highest pH, watertable depth, conductivity and percentage of bare soil values, towards the highest rainfall values and the lowest ones of the other environmental variables. This can also be seen in Fig. 1.b, which represents intraset correlations (this figure is also a biplot representation of the correlations between the environmental variables). The mentioned gradient separates mainly a community with moderate to abundant relative presence of *Distichlis sp*. (Community 1), combined in some sites with *Nitrophylla australis* and *Lycium repens*, which characterizes areas dominated by plant species that indicate subhumid, slightly alkaline and saline sites.

The second axis shows a smaller gradient, characterized by comparatively higher values of precipitation and watertable depth towards the positive side of the axis. This gradient is associated with the spatial distribution of plant species in the meadows due to moisture changes from their periphery towards their centre, and also because of altitude differences determined by topography. It differentiates two communities on the left sector of the graphic, one characterized mainly by *Festuca pallescens* (Community 2) and other characterized by the combination, in different proportions, of *Eleocharis albibracteata*, *Juncus balticus*, *Poa pratensis*, *Taraxacum officinalis* and *Trifolium repens* (Community 3).

Fig. 1.a also shows the joint interpretation of species and environmental variables (biplot representation of **L**). For example, for *Distichlis sp.*, precipitation values are lower where it is present, compared with sites without it (its projection is located at the negative extreme of the precipitation axis – see meaning of l_{kj} element given in Section 2).

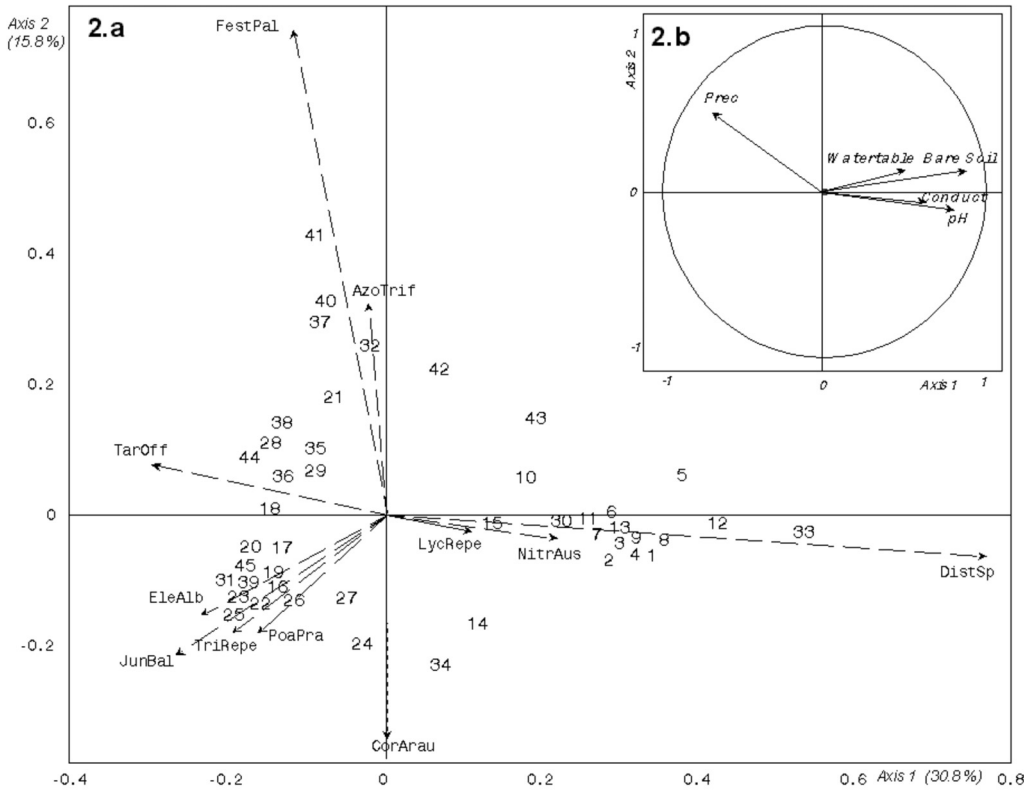

Figure 2: ***2.a.*** *First factorial plane of the NSCA (site-conditional) ordination diagram. Sites are indicated by numbers. Dashed vectors represent species (abbreviations in Table 3) that contribute most to the orientation of the axes and/or have representation quality in that plane greater than 0.3. Percentages of explained inertia are shown in Table 2.* ***2.b.*** *Projection of environmental variables on the first NSCA factorial plane as supplementary variables (correlations with those axes).*

As it was mentioned, data were also analyzed by NSCA. Figure 2.a shows its first factorial plane and Figure 2.b represents the environmental variables as supplementary ones (the coordinates are correlation coefficients between them and the site scores). The first axis explains 30.8% of the total inertia, with a 46.6% explained by the first plane (see Table 2).

The relative contributions of sites to the determination of the first factorial plane were well distributed between them, with a general good quality of representation. Fig. 2.a contains the species that most contributed to the orientation of the first two NSCA axes (see Table 3), whose qualities were greater than 0.30 in that plane. Even though *Cortaderia araucana* had a lower quality value, it was also included because of its contribution to axis 2 and it helped to differentiate one community where it was the characteristic plant species (Community 4-sites 14, 24, 27 and 34).

The results by NSCA showed the same three communities found by CNCA, with the addition of the one mentioned in the previous paragraph. By comparing the results

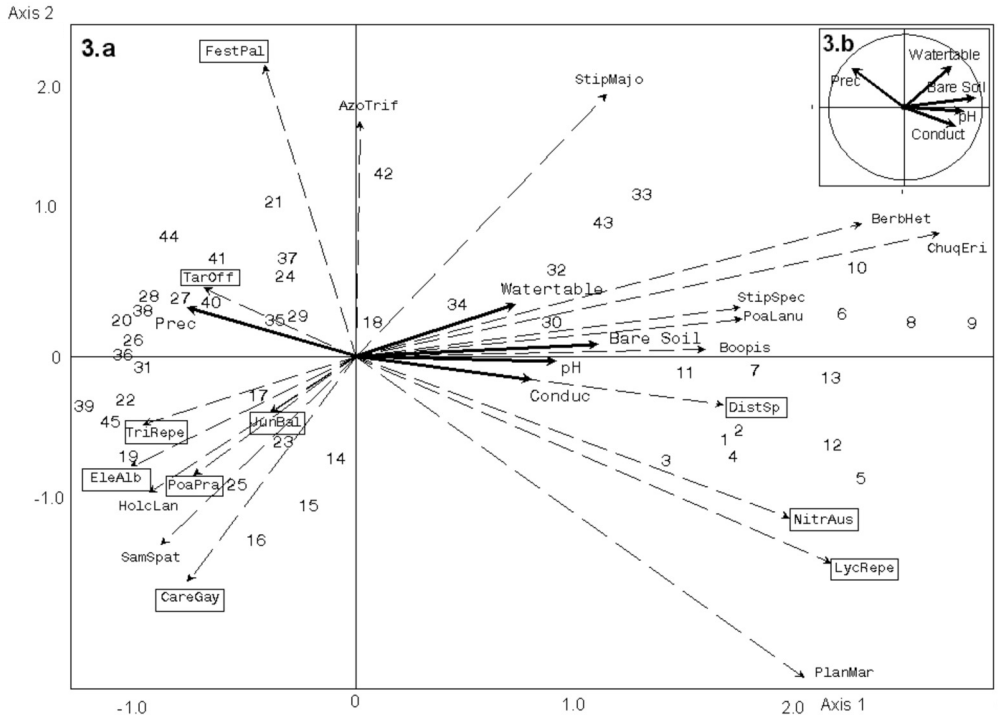

Figure 3: 3.a. *First factorial plane of the CCA (site-conditional) ordination diagram. Axes for sites (indicated by numbers) and environmental variables (indicated by solid vectors) should be scaled by 0.67. Dashed vectors represent species (abbreviations Table 3) that contribute most to the orientation of the axes, being squared those with* $Q_{\alpha,k}^{(OS)}$ *in that plane greater than 0.3. Percentages of explained inertia are shown in Table 2.* ***3.b.*** *Intraset correlations between environmental variables and the first two CCA factorial axes.*

of both analyses, it can be observed that: a) the measured external variables adequately allowed to characterize three of the four identified communities, but were not able to characterize the last one, with *Cortaderia araucana* (even when considering the third axis); b) with respect to the gradients, a similar environmental interpretation of the first axis was achieved by both analyses; but the interpretation as a gradient of spatial location for the second axis was not evident in the unconstrained analysis.

Finally, these data were also analyzed by CCA as a comparison with CNCA results. It is remarked that the implications that give rise from this comparison, should be checked later by simulation studies. Figure 3.a shows the CCA first factorial plane in site-conditional scaling. The first axis explains 55.2% of the total inertia of the fitted values in the projected space, with a 73.5% explained by the first plane (see Table 2). These percentages were 15.1% and 20%, respectively, with respect to the inertia of the original values.

Table 3, which contains the already mentioned contributions, shows that species like *Chuquiraga erinacea*, *Poa lanuginosa*, *Stipa speciosa* var. *major* (these three present in few sites, with low frequencies except for one) and a few other species (with low frequencies in few sites), had more contribution in CCA than in CNCA. But even when they had good $Q_{\alpha,k}^{(PS)}$ values (not shown here), they were not well explained by the external variables (they had low $Q_{\alpha,k}^{(OS)}$ values). By contrast, species squared in Fig. 3.a., which happened to be the same as the ones represented in Fig. 1.a, were the species which not only contributed to the orientation of the first CCA plane, but have acceptable $Q_{\alpha,k}^{(OS)}$ values.

With respect to sites, their distribution in Fig. 1.a. and Fig. 3.a. showed differences, but the global position of most of them was similar. This gave rise to the similarity shown by the intraset correlations represented in Figs. 1.b. and 3.b. Thus, the gradient interpretations in terms of the considered environmental variables, for this example, are quite similar in both constrained analysis. But it should be noticed that the role of low frequency species, taken from their contribution values were different, even though in this particular case, they did not influence the conclusions of the analysis.

4 Discussion and conclusions

In this work CNCA is introduced as an extension of NSCA into a constrained ordination context. Couteron *et al*. (2003) showed an illustration of this extension, without theoretical development, according to the first CNCA approach introduced in Section 2. In their work, the graph interpretation was made in a different way, without analyzing the relation between sites and species through biplot representations.

Even when a comparison between CCA and CNCA results was carried out in the above section, it was suggested that such a comparison between the performances of

each one of them should be carried out through a simulation study. However, there are some aspects that could be pointed out towards a theoretical comparison.

First of all, as it was said in Section 2, the difference between CCA and CNCA is the metric for species, as it is between their respective unconstrained ordination methods, CA and NSCA (see Pélissier *et al*. 2003), respectively. In CCA, it is the chi-square metric, while in CNCA it is the Euclidean one, which contains uniform species weights.

In general, those different metrics have strong effects on the results of NSCA and CA, making them quite different. However, when comparing CNCA and CCA, even though their respective metrics do not change (from NSCA to CNCA, and from CA to CCA), because of projection, the role of rare species in the original space, tends to change in the projection space. If these species are not found in atypical sites (atypical in terms of the values of the environmental variables), they have small importance in the projected space. This minimizes the effect of the chi-square metric concerning its weighting scheme when compared with the Euclidean metric. Then, their results, *depending on data*, might not differ in the same magnitude as those expected between NSCA and CA.

Another point to be mentioned is the relationship between the two species distances to their centroids, each one in its particular space according to the definition under a NSCA or a CA model. Those distances are proportional, even though the total inertias are different. In constrained ordination, the same relationship is kept in the projection space. As a consequence, given a set of environmental variables, the *maximum* proportion of inertia that can be explained for *each species* is the same for both constrained techniques. However, since the total inertias and their partitions in principal axes are not equal, the proportion of explained variance by the first axes of each one of the two constrained method, might differ for each species.

Looking now at the global structure of the data as it is proposed by, for example, Takane and Hunter 2001, CNCA and NSCA are terms of a global model that partitions the total variance of the data, into orthogonal components of explained and unexplained variance. From this, a comparison between their results allows for the evaluation of the influence of the external variables on the species behaviour, as it was done in the meadow example. Although this global model and its partition applies also to CCA and CA, the difference in the effect of the chi-square metric when applied to the observed or to the projected data, as it was said above, might affect such a comparison. Because of this, it tends to be accomplished with DCA (see Palmer 1993 or ter Braak 1986) – DCA, with its detrending and rescaling processes, generates an important controversy (e.g., Wartenberg *et al*. 1987; Jackson and Somers 1991).

In contrast, the only difference between an analysis done by NSCA and by CNCA is the value in the site-by-species table, since metric and site weights are identical in both procedures. The first analyzes the information about observed species composition (see (1)), while the second, the projected values of such information (see (2)). Then, the detected differences in the results should mostly be the consequence of considering

such environmental variables as those which explain the species distribution (through the proposed model). The higher the similarity between both results, the greater the association (causal or not) between such variables and the vegetation behaviour.

One last aspect to be mentioned in this comparison is about the inter-tables that CNCA and CCA analyze. When considering CNCA as the analysis of the inter-table **L**, it is deduced that species also offer information from their absences and importance values. In contrast, the inter-table considered in a CCA is a matrix of weighted averages of the environmental variables (say, matrix **W**), weighted by the species profiles. So, zeros in the species table **Y** do not contribute to the elements of **W** (Dray *et al.* 2003, Pélissier *et al*. 2003).

When zeros are taken into account, they should be real records and not a consequence of possible omitted species. Thus, this (different) meaning of zero values is related, in part, to the sampling scheme. Dray *et al*. (2003) and Pélissier *et al*. (2003) mentioned this concept in relation to CA/CCA and to NSCA. Related to CNCA, as the absences give information, the values of the environmental variables at those sites are considered as not favourable for that plant species. Then, in long gradients, this interpretation might be inappropriate. This is an important point that shows that the consequence of what at first appeared to be a simple change in CCA metric (looking at CNCA from its first approach), goes further than would be expected. So, at present, CNCA, even NSCA, are not recommended in long gradient situations.

In spite of the different meaning of the absences, both extensions as constrained ordination techniques could present difficulties in expressing the presence of limiting environmental factors, since both of them weight by site richness.

In this paper, only quantitative environmental variables were considered. In a simple way, adjustments can be made to handle mixed external variables. Furthermore, the first approach, the extension of a technique to a constrained ordination context based on projections onto instrumental variables, easily permits the generalization of these concepts to other situations.

Acknowledgement

The authors wish to thank the referees and editors for their constructive criticisms and helpful comments which have lead to a substantial improvement of the paper.

References

Anderson, M.J. and Willis, T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, 84, 511-525.

Benzécri, J.P. (1973). *L'Analyse des Données: Tome I: La Taxinomie. Tome 2: L'Analyse des Correspondance*. París: Ed. Dunod.

Braun-Blanquet, J.J. (1950). *Sociología vegetal. Estudio de las comunidades vegetales.* Ed. Acme Agency. Bs. As.

Chessel, D. and Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environment. In: J.D. Lebreton, B. Asselain, (Eds.), *Biométrie et Environment*, Mason: Paris, 15-44.

Couteron, P., Pélissier, R., Mapaga, D., Molino, J.-F. and Tellier, L. (2003). Drawing ecological insights from a management-oriented forest inventory in French Guiana. *Forest Ecology and Management*, 172, 89-108.

Cuadras, C.M., Cuadras, D. and Greenacre, M. (2006). A comparison of different methods for representing categorical data. *Communications in Statistics-Simulation and Computation*, 35, 447-450.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, 31, 277-294.

Dray, S., Chessel, D. and Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological tables. *Ecology*, 84, 3078-3089.

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

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

Gimaret-Carpentier, C., Chessel, D. and Pascal, J.-P. (1998). Non-symmetric correspondence analysis: an alternative for species occurrences data. *Plant Ecology*, 138, 97-112.

Gimaret-Carpentier, C., Chessel, D., Pascal, J.-P. and Ramesh, B.R. (1999). Advantages of non-symmetric correspondence analysis in identifying multispecific spatial patterns in the rain forest of the Western Ghats. In: Y. Laumonier, B. King, C. Leggs, K. Rennols (Eds.) *Data Management and Modelling using Remote Sensing and GIS for Tropical Forest Land Inventory*. Jakarta: Rodeo International Publishers, 397-411.

Graffelman, J. (2001). Quality statistics in canonical correspondence analysis. *Environmetrics*, 12, 485-497.

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

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

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

Jackson, D.A. and Somers, K.M. (1991). Putting things in order: the ups and downs of detrended correspondence analysis. *The American Naturalist*, 137, 704-712.

Lauro, N. and D'Ambra, L. (1984). L'Analyse non symetrique des correspondances. *Data analysis and informatics III*, Diday *et al.* (Eds.), North-Holland, 433-446.

Lebreton, J.D., Chessel, D., Prodon, R. and Yoccoz, N. (1988). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta oecologica-Oecologica Generalis*, 9, 53-67.

Økland, R.H. (1996). Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science*, 7, 289-292.

Palmer, M.W. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology*, 74, 2215-2230.

Pélissier, R., Couteron, P., Dray, S. and Sabatier, D. (2003). Consistency between ordination techniques and diversity measurements: two strategies for species occurrence data. *Ecology*, 84, 242-251.

Rao, R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya*, Sec. A, 26, 329-358.

SAS (1999). *Version 8*. Cary NC, USA: SAS Institute Inc.

Takane, Y. and Hunter, M.A. (2001). Constrained principal component analysis: a comprehensive theory. *Applicable Algebra in Engineering, Communication and Computing*, 12, 391-419.

ter Braak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics*, 41, 859-873.

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. (1987[a]). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69, 69-77.

ter Braak, C.J.F. (1987[b]). Ordination. In: Jongman, R.H.G.; Ter Braak, C.J.F.; Van Tongeren, O.F.R. (Eds.). *Data analysis in community and landscape ecology*. Netherlands: Pudoc Wageningen, 91-173.

ter Braak, C.J.F. and Verdonschot, P.F.M. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57, 255-289.

van den Wollenberg, A.L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42, 207-219.

Wartenberg, D., Ferson, S. and Rohlf, F.J. (1987). Putting things in order: a critique of detrended correspondence analysis. *The American Naturalist*, 129, 434-448.

Whittaker, R.H. (1967). Gradient analysis of vegetation. *Biological Reviews*, 49, 207-264.