Community Ecology 18(3): 260-279, 2017
1585-8553 © Akadémiai Kiadó, Budapest
DOI: 10.1556/168.2017.18.3.5

Recoding and multidimensional analyses of vegetation data: a comparison

S. Camiz[1,3], P. Torres[2] and V. D. Pillar[3]

[1] *Dipartimento di Matematica, Sapienza Università di Roma, Roma, Italy. Corresponding author. E-mail: sergio@camiz.net*
[2] *Facultad de Ciencias Agrarias, Universidad de Rosario, Rosario, Argentina*
[3] *Departamento de Ecologia, Universidade Nacional do Rio Grande do Sul, Porto Alegre, Brasil*

Keywords: Braun-Blanquet coding, Comparisons, Correspondence Analysis, Non-metric Multidimensional Scaling, Principal Component Analysis, Recoding.

Abstract: Two simulated coenoclines and a real data set were differently recoded with respect to the Braun-Blanquet coding (including presence/absence) and analysed through the most common multidimensional scaling methods. This way, we aim at contributing to the debate concerning the nature of the Braun-Blanquet coding and the consequent multidimensional scaling methods to be used. Procrustes, Pearson, and Spearman correlation matrices were computed to compare the resulting sets of coordinates and synthesized through their Principal Component Analyses (PCA). In general, both Procrustes and Pearson correlations showed high coherence of the obtained results, whereas Spearman correlation values were much lower. This proves that the main sources of variation are similarly identified by most of used methods/transformations, whereas less agreement results on the continuous variations along the detected gradients. The conclusion is that Correspondence Analysis on presence/absence data seems the most appropriate method to use. Indeed, presence/absence data are not affected by species cover estimation error and Simple Correspondence Analysis performs really well with this coding. As alternative, Multiple Correlation Analysis provides interesting information on the species distribution while showing a pattern of relevés very similar to that issued by PCA.

Abbreviations: BBc – Braun-Blanquet coding; CA – Correspondence Analysis; DCA – Detrended Correspondence Analysis; GPA – Generalized Procrustes Analysis; MCA – Multiple Correlation Analysis; MDS – nonmetric MultiDimensional Scaling; PCA – Principal Component Analysis.

Introduction

Braun-Blanquet is acknowledged as the father of vegetation science (Podani 2006). His most important work (Braun-Blanquet 1932) has had lasting influence in the field work and the scientific thinking of the majority of the vegetation scientists all over the world. An enormous amount of vegetation data (see, e.g. Chytrý et al. 2016) has been collected according to his protocols. Yet, the method has been criticized (Camiz 1993, Podani 2006), in particular for what concerns the phytosociological sampling, the coding, and the way data are being analysed in practice. Indeed, Braun-Blanquet protocols were conceived when the principles of statistics were neither sufficiently developed nor used in vegetation yet. At these times, multidimensional analysis methods were in theoretical development, with very limited applications due to technical difficulties resulting by non-automatic computations. Only starting the 1960', the development of vegetation data analysis improved paralleling that of the available multidimensional analysis programs.

A vegetation data table is basically a list of plant species observed in a series of plots, the relevés, with an estimation of both species abundance/cover and of sociability in the relevé. The recording of abundance/cover is based on 7 ordered symbols (the Braun-Blanquet coding, in the following BBc), ranging from rare to 75-100% cover (see the first two columns of Table 1), and the sociability is recorded according to a 5-level nominal scale. We focus on the first index, as the second received little attention so far. Born when no idea of numerical automatic treatment existed, the problem of its transformation into measures was raised and several recoding schemes were suggested to fit the requirements of quantitative treatment (see van der Maarel 1979, for a list): among the various, class midpoint, the van der Maarel (1966)'s abundance/cover, and presence/absence (Table 1) will be considered in this work. In addition, data transformation was applied in a style analogous to that so diffused in biostatistics that the Koordinaten-Schweinerei appeared in Bräuer (1982) to denunciate this exaggeration. It deserves being observed that all recoding is based on researcher's estimate of BBc, thus all suffer from its intrinsic limitation, its approximate estimation.

Braun-Blanquet's protocols generated a large database of vegetation plots, an invaluable knowledge base that may not be neglected. Thus, the identification of the most suitable methodology to analyse these data is welcome. We remind here that the phytosociological sampling, as it is essentially descriptive of the different syntaxa, is not random: thus, it prevents any statistical inference. On the other side, his cover-abundance scale has been used not only for phytosocio-

Table 1. The Braun-Blanquet (1932) abundance/cover coding, its meaning, the mean values, the van der Maarel (1966) recoding, the 8- and 3-level ordinal scales, and presence/absence coding, considered in this paper.

Braun-Blanquet coding	cover meaning	mean value	van der Maarel recoding	8-level scale	3-level scale	presence absence
void (0)	absent	0	0	1	1	0
R (+)	rare	0.01	1	2	2	1
+	<1%	0.1	2	3	2	1
1	1-5%	2.5	3	4	2	1
2	5-25%	17.5	5	5	2	1
3	25-50%	37.5	7	6	3	1
4	50-75%	62.5	8	7	3	1
5	75-100%	87.5	9	8	3	1

logical purposes but also for the description of both randomly and systematically selected plots. Here we concentrate on it and the suitable data analysis methods, including its recoding. As it is conceived, the BBc is a scale character, because it is based on a rough estimate of a species cover percentage, with no constant difference between levels. Podani (2006) states very clearly that both this and its van der Maarel (1966) recoding are not measures but merely orders; thus, he argues that they deserve being treated by specific methods and he (Podani 2005) suggests an association coefficient (Podani 1997) to be used with *Non-Metric Multidimensional Scaling* (commonly abbreviated as NMMDS: note that for brevity sake in the following, we shall write it as MDS). MDS was first introduced by Kruskal (1964a,b) as a tool to reduce dimensionality while minimizing the deviation from the original order.

This argument contrasts with the currently used multidimensional analysis methods, such as *Principal Component* and *Correspondence Analyses* (in the following, PCA and CA, respectively, Benzécri 1982, Orlóci 1978, Legendre and Legendre 2012), which may use either quantitative data or frequencies. Indeed, MDS has been suggested long time ago in the literature (Orlóci 1978, Kenkel and Orlóci 1986) as better performing with respect to the other methods, but these keep several advantages, such as *i)* they are easier to be found and used; *ii)* they do not need a pre-fixed dimension solution; and *iii)* they deal simultaneously with both species and relevés. In fact, in these methods based on eigenanalysis, the so-called *transition formulas* (Lebart et al. 1984, 2006) allow to deal with relevés and species at the same time, since the same meaning may be attributed to the factors influencing relevés and species, a property that does not exist in MDS, whose independent results for the two sets of objects are non-comparable.

To understand the matter, we briefly remind here that species and relevés may be represented in two separate geometrical spaces. Both PCA and CA may deal simultaneously with both of them, but MDS may treat a set at a time. On the opposite, MDS may deal with any kind of data, provided a suitable association index is chosen, whereas PCA and CA may only deal with measures and frequencies, respectively, with presence/absence as a possible alternative for both and the Spearman's rank correlation coefficient as an alternative for PCA in place of Pearson's correlation. Based on this scheme, Podani (2006) severely criticized the application of both PCA or CA to BBc-based data, too often done without considering the caveats imposed by their nature. Such a criticism strongly contrasts with the fifty years and more of practice, in which attention of vegetation scientists was driven more to a good estimate of the position of both relevés and species on a supposed environmental gradient revealed by the methods, than to observe the methods' *mode d'emploi* and their limits.

It must be emphasized that this estimation is impossible through the use of these methods, since they belong to the exploratory analysis methods *sensu* Tukey (1977). In this framework estimations are impossible because no model is built and consequently neither errors nor their distributions may be obtained. Nevertheless, *Detrended Correspondence Analysis* (DCA, Hill and Gauch 1980) was introduced aiming at better estimating the main ecological gradient underlying a vegetation table. The method is still largely used, notwithstanding its poor consistence: indeed, its artificial de-trending by blocks does not affect the first principal component, that is expected to be identical to that issued by CA, up to a rescaling of the extremes, but heavily biases the second one that this way loses any meaning (Camiz 2005). Strong inertia against the quest for alternative more suitable methods is still present: neither criticism (Camiz 1991, 1993, 1994, 2005) nor alternative methods based on more consistent rationale received due attention in literature. Yet, they do exist, both based on Gaussian distributions: Ihm and van Groenewoud (1975, 1984) propose an eigenvalues method to study seriations, whereas Johnson and Goodall (1980) and Goodall and Johnson (1982) propose a maximum likelihood one for dealing with vegetation data, including the distribution of absences too.

Due to our aim, to concentrate the attention to the use of BBc and its transformations, in this paper we limit attention to the exploratory methods. They are the most used and best known in the vegetation environment and thus it is in this framework that our examination will take place.

As expected, the Podani (2005, 2006) statements raised strong reactions, in particular from Ricotta and Avena (2006) who quoted some "topological distance": a mathematical nonsense. Yet, they defend the current practice, because its results seem to correspond to the vegetation scientist's thought. A better argument might be that the coding is a rough transformation of a measure, say a way of conventionally measuring species cover. This argument may be contrasted by the too high uncontrolled error introduced by the transformation, including the one due to the researcher's error in the estimation. As an alternative, a recent study by Wilson (2012) discusses the advantages to limit attention to presence/absence with respect to quantitative coding, an issue already proposed by Camiz (1994, 2002). Wilson claims that presence/absence is "quicker to collect, sufficient and may better represent the communities of the area", and his practical and theoretical conclusions are in favour of its use.

Our present aim is to discuss how the BBc data tables may be analysed at their best, according to the limits set by the combination of coding, transformations, and exploratory ordination methods in search of a suitable way to unravel the intricacies between their uncontrolled use. In particular, we tried to understand how these different combinations may affect the results, in order to check to what extent the discussion that took place among Podani (2006) and Ricotta and Avena (2006) may be solved. Eventually, we would aim at finding a univocal combination of recoding and scaling methods to suggest for a general use, at least as the first step of a study.

To understand the influence of data recoding and data analysis methods, we set in an exploratory framework; we dealt with two simulated and one real data tables - with one and two independent gradients and a real transect, respectively - that we transformed to get 5 different recoding, and we ran five different methods: we considered metric methods, such as Principal Components, Simple and Multiple Correspondence Analyses (PCA, CA, and MCA, respectively, Benzécri 1982, Orlóci 1978, Lebart et al. 1984, Legendre and Legendre 2012), Detrended Correspondence Analysis (DCA, Hill and Gauch 1980), and Non-Metric Multidimensional Scaling (MDS, Kruskal 1964a,b, Orlóci 1978, Legendre and Legendre 2012), the latter through different association indexes. PCA and CA are the methods most used by vegetation scientists - the latter through its variant DCA - while MDS was proved to give better results by Kenkel and Orlóci (1986) and was suggested by Podani (2005) to be used with a coefficient suitable for scale data. We added MCA as a possible alternative to explore: in vegetation science it has been used only once by Romane (1972) and it might be suitable to deal with scale data. Indeed, in MCA the order is not taken into account, but its relation with the factors may be graphically inspected, and, in addition, the concurrent analysis of the sociability coding is possible.

As our aim is to identify the main gradients underlying the vegetation table, we remind that they may correspond to either the factors themselves or the curvilinear pattern found on the factor spaces (the Guttman effect, *arch, horseshoe*, see Guttman 1953, Camiz 2005). In the following these methods are described and the comparison of the results obtained by the various combinations of coding and method are reported.

Materials and methods

2.1 The data

As said, in this work we dealt with two sets of simulated and one of real data. The simulated data were generated according to Minchin (1987). The abundance y of a species in a given community unit according to an ecological gradient was modelled through the beta function

$$y = x^{\alpha}(1-x)^{\gamma} \qquad 0 \leq x \leq 1$$

in which y depends on the position x of the unit on the gradient, normalized to range in the interval (0,1), and the two parameters α and γ control the shape of the function. Minchin (1987) generalizes the beta function to adapt it to the simulation of unimodal species response over any range of x along an environmental gradient in terms of abundance y. This is based on α and γ, on the position x of each community, the position m where the species has its mode or maximum M, and the range r (niche breath) on the gradient within which the abundance is larger than zero. In this generalized formulation, the abundance becomes:

$$y = \begin{cases} \frac{M}{d}\left(\frac{x-m}{r}+b\right)^{\alpha}\left(1-\left(\frac{x-m}{r}+b\right)\right)^{\gamma} & m-br \leq x \leq r(1-b) \\ 0 & \text{otherwise} \end{cases}$$

where the two parameters b and d depend only on α and γ as

$$b = \frac{\alpha}{\alpha+\gamma}, \qquad d = b^{\alpha}(1-b)^{\gamma}$$

To simulate noise, random numbers were added to the so computed abundances y. For this task negative binomial distribution was adopted. It is widely used to model species abundances (White and Bennetts 1996, O'Hara and Kotze 2010, Pillar 2013), because it generates data more similar to real community data, with a large number of zeroes (absences) when the expected abundances y are low. Considering a number $\tilde{y}$ of successes in independent trials with fixed probability of success p, a random value was extracted from the negative binomial distribution with a number of successes $n = \tilde{y}*p/(1-\mathbf{p})$. Both generalized beta function and negative binomial random extraction were coded in C++. The compiled versions of the programs are available at http://ecoqua.ecologia.ufrgs.br/arquivos/Software/DataSimulation.

The analysed data sets are the following:

1) A simulated coenocline. It is a table with the abundance of 60 taxa in 30 simulated communities, based on one environmental gradient. The position on the gradient of each community x_i , $i = 1, \ldots, 30$, was a random number between 1 and 100. For each species, the modal abundance M was a random number between zero and 80, the modal position m was a

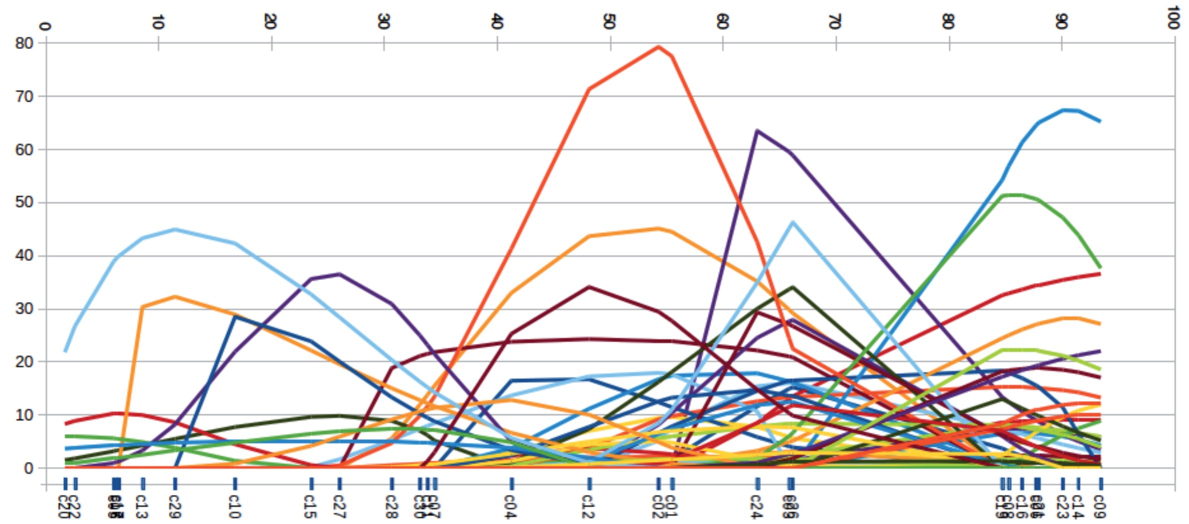

Figure 1. The abundance functions of the 60 taxa simulated along one gradient prior the introduction of noise and the position along the gradient of the 30 randomly simulated communities.

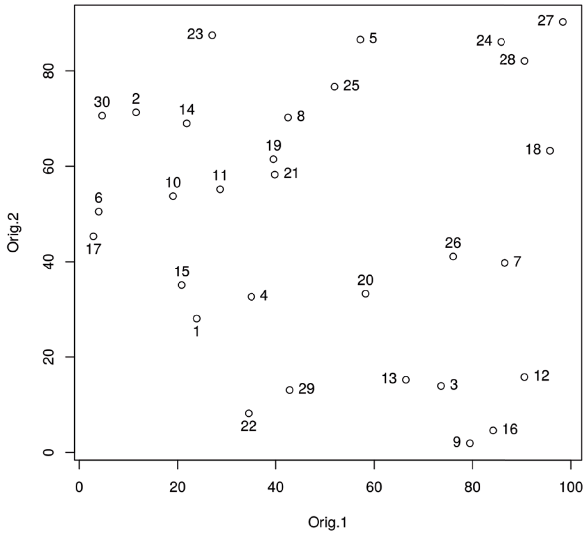

Figure 2. The position of the 30 relevés on the coenoplane issued by the two simulated gradients.

random number between zero and 100, the range was a random number between 20 and 60, and the α and γ parameters were random numbers between 0.1 and 4. All these random numbers were extracted from the uniform distribution. To add noise, the probability p = 0.8 of success was considered in the procedure. The abundance functions y_j, $i = 1, \ldots, 60$ of the simulated taxa and the positions x_i, $i = 1, \ldots, 30$ of the simulated communities along the coenocline without noise are represented in Figure 1.

2) *A simulated coenoplane*. It is a table with the abundance of 60 taxa in 30 simulated communities, based on two environmental gradients. All parameters were generated as before, separately for each gradient, and 30 simulated communities were built, by randomly pairing the values of both coenoclines. The position of the relevés on the coenoplane before the introduction of noise is reported in Figure 2.

3) *A real data table* with 27 relevés and 65 species of the vegetation observed in a study area situated in the valley of Saladillo river, around 10 km South of Sandford and Chabás, district of Caseros, in the Argentinian province of Santa Fe. Close to the river a *Stipa hyalina* Nees grassland is found (the so-called *flechillar*), then high *Spartina densiflora* Brongon pasturelands (*espartillar*), and further from the river, different types of *halophylous grasslands* (Carnevale et al. 1987, Carnevale and Torres 1990). Along a transect perpendicular to the river, a quadrat of 4 sq.m. was delimited every 25 m. and all species present into the quadrat were identified together with an estimate of cover/abundance of each according to BBc. Thus, we suppose that the sequential numbering of the relevés along the transect may be an reasonable approximation of both the hypothesized gradient and the rank. All original data tables are reported in Electronic Appendix A1.

2.2 The recoding

The simulated data (FR), that represent abundances, were recoded according to the rules described in Table 1. To transform them to cover percentages, we divided each one by their sum in each relevé. The obtained relative abundance of each species was coded according to the corresponding interval of the Braun-Blanquet scale. Thus, the different adopted coding schemes are the following:

- FR: the abundances, resulting as above;
- BB: the midpoint of each Braun-Blanquet class;
- VDM: the van der Maarel (1966)'s recoding;
- N8: an 8-levels scale, identical to van der Maarel (1966)'s recoding, but in which, for technical reasons, the scale levels are relabeled from 1 = void(0) to 8 = 75 − 100%;
- N3: a 3-levels scale, obtained by aggregating the levels of N8, corresponding to 1 = void(0), 2 = rare to 25% (from 1 to 5), 3 > 25% (from 6 to 8);
- PA: the presence/absence.

Note that, as the real transect was originally sampled according to Braun-Blanquet coding, no abundance data could be used in this case.

2.3 The ordination methods

Here we sketch briefly the main features of the considered exploratory methods.

Principal Component Analysis is the most known metric multidimensional scaling technique, based on both Singular Value Decomposition (SVD, Abdi 2007, Greenacre 2007) and Eckart and Young (1936) theorems; the principal components are orthogonal directions, linear combinations of the original characters, along which the inertia (that is, the sum of squared distances of units to their centroid, in our case the scattering of relevés) is maximum. As mean and variance are basic concepts to compute PCA, it may be applied only to quantitative (measure or frequency) data, but is currently accepted for presence/absence data, by giving sense to a weighed average between 0 and 1, which is actually a proportion of presences. In our case, it may be applied safely only to the mean values of the code levels and on presence/absence. To apply it correctly to van der Maarel (1966)'s coding, one might consider it a conventional measure, thus accepting implicitly intermediate, continuous values. Note that, when PCA is applied on ranks, its results are identical to those obtained by performing PCA on the Spearman correlation matrix (Lebart et al. 1984, 2006). Thus, as an alternative, the rank variant (in the following indicated SPE-PCA) may be safely considered, with the drawback that the simultaneous treatment of the species is not possible.

Simple *Correspondence Analysis* may be defined as a generalized SVD of the ratios of the squared differences between observed and predicted values of a contingency data table and the predicted ones, under independence hypothesis (Benzécri 1982, Lebart et al. 1984, 2006, Greenacre 2007). It is suitable for frequencies so that, if one deals with the number of plants of a species present in a relevé, CA is the method to use. In practice, it is admitted for transformations of counts and for this reason it proved effective for both van der Maarel (1966) recoded (a rough transformation of frequencies) and presence/absence data.

Detrended Correspondence Analysis is an artificial adjustment of CA that aims at fixing the Guttman effect by removing piecewise the arch effect from the second axis and by rescaling the extreme values (Oksanen et al. 2017). This way the interpretation is limited to the first dimension: for this reason the method is really inconsistent (Camiz 2005) and no more informative than CA. We included it in this study due to its large use in literature.

Multiple Correspondence Analysis (Benzécri 1982, Lebart et al. 1984, 2006, Greenacre 2007) is a generalization of CA to a set of nominal data. It consists in applying CA to an indicator table. It may be applied also to ordinal data, but the algorithm ignores the scale nature, that may be recovered the same in the graphics by joining the increasing levels positions with lines. In our case, it may be used for both 8- and 3-levels scales and it may constitute an interesting alternative application, in particular because the sociability index might be taken into account in the same study, though it will not be considered in this one.

Table 2. The pairing of the methods (by rows) and the coding (by column) used for the experimentation: * means that the pairing was applied to all data tables, the × refers only to the simulated data. The corresponding acronym results by joining the label of the coding with that of the method.

	FR	BB	VDM	PA
PCA	×	*	*	*
SPE-PCA	*		*	
CA	×	*	*	*
DCA	×	*	*	*
MCA			*	
MDS	×		*	

Non-Metric MultiDimensional Scaling (Kruskal 1964a,b) is a method aiming at reproducing at the best dissimilarities between items as Euclidean distances between points in a small predefined-dimensional space. Unlike the other methods that provide a global solution through SVD, MDS is based on the minimization of the stress (Kruskal and Carmone 1971) or any quadratic measure of the deviation from the expected distances of those measured on the space of representation. It is an iterative converging process that only drives to local optima, depending on both the starting configuration and the chosen dimension of the space of representation. In addition, the different dimensional solutions are not encapsulated nor the found coordinates have a specific meaning, so that they may be rotated at wish. MDS may be applied to any kind of data, but depends heavily on the dissimilarity chosen. It is sometimes preferred to eigenanalysis based methods for its robustness with respect to the Guttman effect (Kenkel and Orlóci 1986): unlike them, the lack of transition formulas prevents the corresponding construction of the species pattern, that must be performed independently, with non-comparable results. For this reason, no relation exists between the analyses carried out on species and relevés, not even they may be simultaneously used for the interpretation of the factors.

2.4 Data analyses

The association between coding and analysis applied is shown on Table 2. Note that MDS was used on frequencies according to:

1) *Euclidean distance* (EU-MDS):

$$d(x_i, x_j) = \sqrt{\sum_k (x_{ik} - y_{jk})^2}$$

2) *Chord distance* (CH-MDS) (Orlóci 1978, Legendre and Legendre 2012),

$$c(x_i, x_j) = \sqrt{2\left(1 - \frac{\sum_k x_{ik} y_{ik}}{\sqrt{\sum_k x_{ik}^2 \sum_k y_{jk}^2}}\right)}$$

and on van der Maarel recoding, using the three coefficients discussed by Podani (1997):

3) *Kendall (1938)'s tau index* adjusted by Diday and Simon (1976):

$$KE = \frac{2(a-b)}{\sqrt{(p(p-1)-2t_j)(p(p-1)-2t_k)}};$$

4) *Goodman and Kruskal (1954)'s gamma index*:

$$GK = \frac{a-b}{a+b};$$

5) *Podani (1997)'s discordance index*:

$$PO = 1 - \frac{2(a-b+t_{jk}-t_j-t_k)}{p(p-1)}.$$

in which

p = number of variables,

a = number of pairs of variables equally ordered by the objects j and k,

b = number of pairs of variables inversely ordered by j and k,

in *KE*, t_j = number of pairs of variables that are tied for j, and t_k = number of pairs of variables that are tied for k,

in *PO*, $t_j + t_k$ = number of pairs of variables that are tied for j, k, or both such that one, two or three scores are 0, these indicate contradiction of j and k; and t_{jk} = number of ties in both j and k corresponding to mutual presence or mutual absence, these indicate agreement of j and k

Remember that, due to the BBc coding of the real data, to them the analyses on frequencies could not be applied. Thus, 20 different analyses resulted for the simulated data, by combining coding and methods, and only 15 for the real data.

2.5 Comparison of results

In this study, we may recognize two main targets: *i)* to measure the ability of methods to identify the gradients underlying the data, and *ii)* to check the coherence of the different methods, concerning both the ordinations and the corresponding ranks. The comparison has been limited to the first two coordinates of the relevés issued by each method, because: *i)* as one or two gradients were expected, no further dimension was of interest; *ii)* in the case of one gradient only, the Guttman effect ensures on one side the existence of such gradient, but on the other one it influences the pattern of the following dimensions, whose interest may be reduced; and *iii)* we did not think that the exam of the results concerning the species could carry other relevant elements to the discussion, unless to distinguish between methods in which they are directly available too: however, some results concerning the species will be reported in the discussion, just to put in evidence the advantages of a joint treatment. In order to study both the large and the small variations, we considered both *i)* the vectors of coordinates issued by each method, that may inform on the large variations, and *ii)* the corresponding vectors of ranks that may better inform on the small ones. Then, to compare the methods we ran the following analyses:

1. *Generalized Procrustes Analysis* (GPA, Gower 1975, Gower and Dijksterhuis 2004, Camiz and Denimal 2011, Legendre and Legendre 2012, Lisboa et al. 2014). GPA was used to check the homogeneity of the methods' results. It is essentially the generalization to more configurations of the *Procrustes Analysis*. Given two clouds of pairwise corresponding points, the aim of Procrustes Analysis is to find the best Euclidean transformation (translation, scaling, mirror reflection, and rotation) of either cloud to match at the best the other, thus minimizing the sum of their pairwise distances. Generalizing to many clouds, an iterative process is set that converges to a consensus, that is a kind of average cloud that best approximates all others. This allows to graphically compare the various clouds to the consensus. From GPA the Procrustes pairwise correlation matrix results: the Procrustes correlation is an index of multidimensional matching of two clouds. It ranges within the interval [0, 1] with the same meaning of the common correlation and allows a numerical comparison of the patterns of the relevés issued by the different analyses. Note that, as GPA could reflect and/or rotate some cloud to optimize the fit, we applied reflections and/or rotations where GPA did, to get more consistent the visual comparison of the results.

2. *Pearson's correlation* matrix of the corresponding first and (in the coenoplane case) second coordinates of the different analyses. This was used to find the best methods, as those giving coordinates most correlated with the original data, thus informing on the agreement with respect to the large scale variation. Indeed, GPA may inform specifically on the coherence between methods but not appropriately with the original gradients, that in two cases are uni-dimensional.

3. *Spearman's correlation* matrix. It is a rank correlation that we used to fine tune the information provided by the Pearson's one. Indeed, the Pearson's correlation is not very sensitive to exchanges in the order of units nearby, while Spearman's informs on the small scale variation, thus contributing to better understand what happens when units are somehow originally clustered.

4. *Principal Component Analysis*. For sake of a graphical synthesis only, we ran PCA on all built correlation matrices: Procrustes', Pearson's, and Spearman's. This allows to easily visualize and understand the most relevant aspects, in particular the coherence of the methods between them and with the original gradient. Indeed, the coherence among methods may be measured by the first eigenvalue of these PCAs, whose proportion to the total inertia indicates the methods' coherence: the largest is such proportion the most coherent the methods are.

The computations of the association matrices to be used in MDS were done with the SYN-TAX 2000 program (Podani 2001); all other methods were run with R (R Core Team 2015), most through the Vegan package (Dixon 2003, Oksanen et al. 2017), but CA, MCA, and GPA, which were run through the *FactoMineR* function of R (Husson et al. 2017).

3. Results

The essential results, that is the correlations (according to Procrustes, Pearson, and Spearman) between all methods with the original data, are reported in Table 3. All other correlations are reported in the Electronic Appendix.

3.1 First simulated coenocline

In Figure 3, the patterns of the 30 simulated communities are shown on the plane spanned by the two main dimensions of the considered analyses. Note that, according to GPA, in order to get the patterns as much alike as possible, the sign

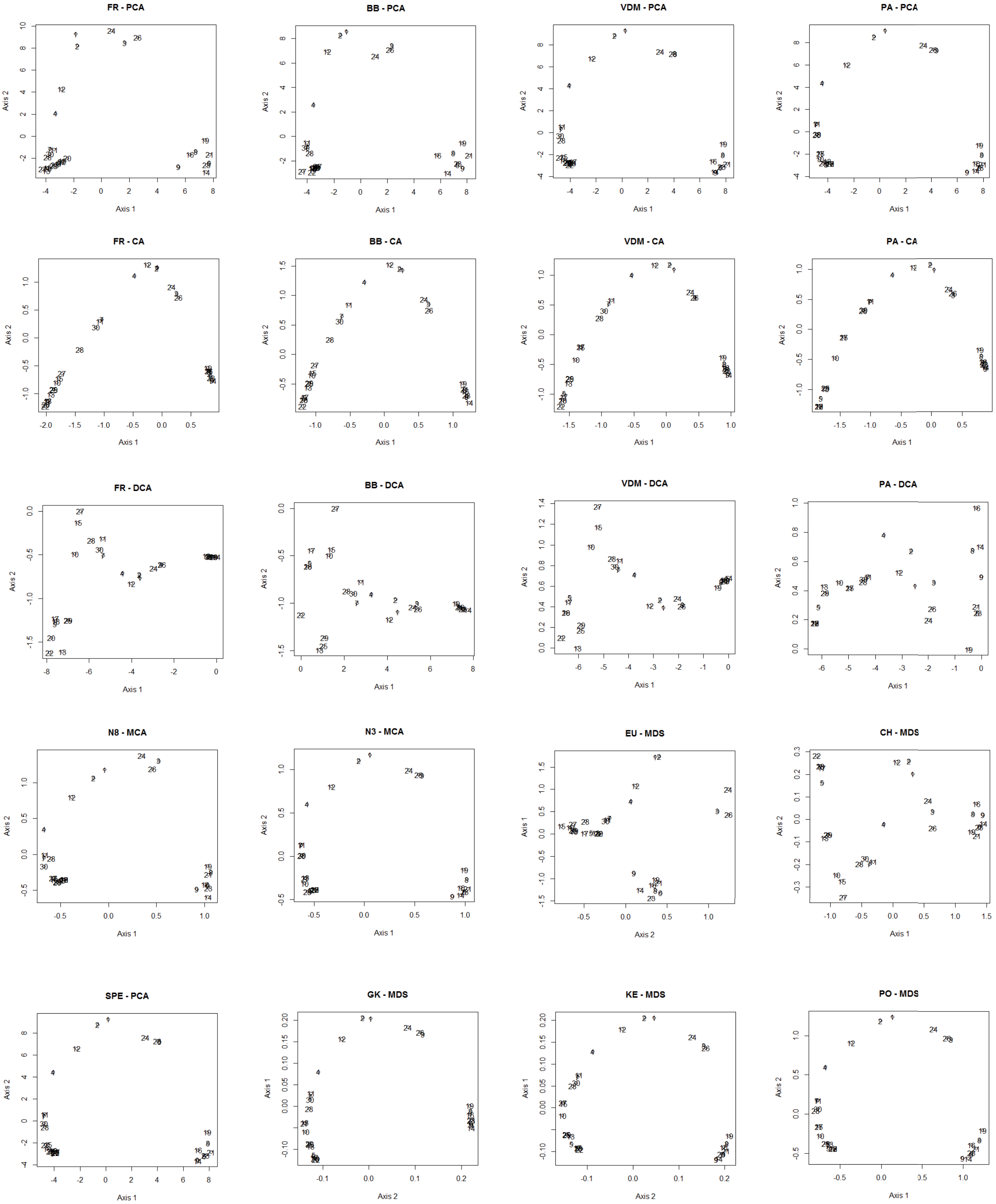

Figure 3. First simulated coenocline: representation of the 30 simulated communities on the plane spanned by the first two dimensions of the considered scaling methods. In the titles, the first part represents the coding and the second the used method.

of coordinates of some axes have been toggled (without any loss, as it is well known that in all scaling methods the sign of the coordinates is arbitrary), whereas for all MDS but CH-MDS the two axes had to be exchanged: once again, this is not relevant, since the MDS solutions are not encapsulated and are invariant under orthogonal transformation.

The first evidence is that most methods show a horseshoe effect, with some variations. This was expected, observed first by Guttman (1953) and continuously commented since then, in particular with the aspect of a horseshoe in PCA and of an arch in CA. Keep in mind that the position of the relevés in respect with the arch's width in CA informs about their richness in species with differently centred and/or larger/more narrow niches along the gradient (Camiz 2005).

Observing carefully Figure 3, one may note that the methods' pattern is not really influenced by the data recoding: the horseshoes of PCAs (top row of the figure) are rather confused, in particular towards both gradient's ends: the four different coding behave alike, with the presence/absence more regular than the others; the arches of CA (second row) have really minor variations within them; the two MCA coding (first two on the fourth row) are similar and repeat somehow the PCA patterns; the two MDS methods applied on abundance data (second two of the fourth row) do not show any effect, but a complicated pattern along the second axis: indeed, it must be emphasized that in CH-MDS the variation along this axis is very limited, in agreement with the unidimensional gradient. Thus, in this case, the second axis does not result of interest. An analogous pattern results from DCA methods, due to the removal of the arch from the CA solution, so that the second axis loses any meaning, due to the piecewise detrending. On the opposite, both PCA on Spearman's index and the MDSs on ordinal indexes (last row) show again horseshoes, much more regular than the ones issued from PCAs. Summarizing, three different patterns resulted from the methods: a horse-shoe, an arch, and a linear pattern. Note that, in the DCA case, variations along the second axis have no interpretable meaning.

To compare the obtained results, in Figure 4a the methods are represented on the circle of correlation issued by the PCA performed on the matrix of the Procrustes correlations between the two-dimensional solutions of the different methods, including the specified coordinates of the simulated communities on the gradient (labelled *Original* in the graphics). It is evident the agreement between methods, with the exception of both the original data, the four DCAs and *CH-MDS* due to the fact that the original data are limited to one coordinate only (thus, the second was set to zero in this comparison), the DCAs second axes are random in practice, and CH-MDS's second coordinate is always very small. Note also the worst performance of EU-MDS. In Figure 4b the pattern of each relevé as seen from each method is shown around the compromise position found by Procrustes method. Note that nearly all large differences with the consensus are with points close to the first axis, corresponding to the original data, the DCAs, and CH-MDS coordinates, as expected. To have a reason of these results one may observe the Procrustes correlation matrix in Electronic Appendix A2. It appears that correlations are really very strong within each group of method: within PCAs the minimum is 0.984, within CAs it is 0.9774, within DCAs it is 0.9885, within MCAs it is 0.9904, and within the MDS on scales it is 0.9899. An independent behaviour have the other two MDSs, but CH-MDS is most correlated, 0.9825, with the specified coenocline coordinates (without noise), a really relevant result, as the four DCAs (all over 0.9895) whereas the best other method is PA-CA with a correlation of 0.8265. Thus, we may appreciate the DCAs and CH-MDS for their outstanding performance and PA-CA, much simpler, that produces the highest result among the others. Indeed, all methods are well correlated within each other, with very few correlations lower than 0.70, all referring to either FR-PCA or EU-MDS, this latter the worst performing.

To better appreciate the methods, both Pearson's and Spearman's correlations between the first coordinates or ranks, respectively, have been computed. Both are reported

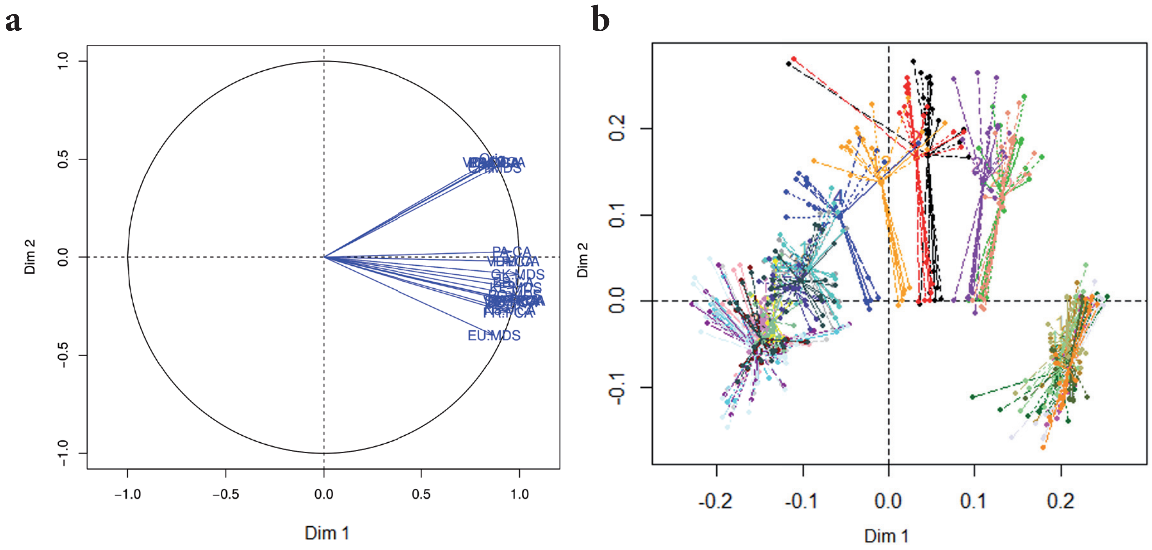

Figure 4. First simulated coenocline: circle of correlation issued by the PCA on the Procrustes correlation matrix (**a**) and the pattern of the samples seen by each method around each compromise (**b**). PCA inertia explained (**a**): Dim 1 = 87.8%, Dim 2 = 9.7%.

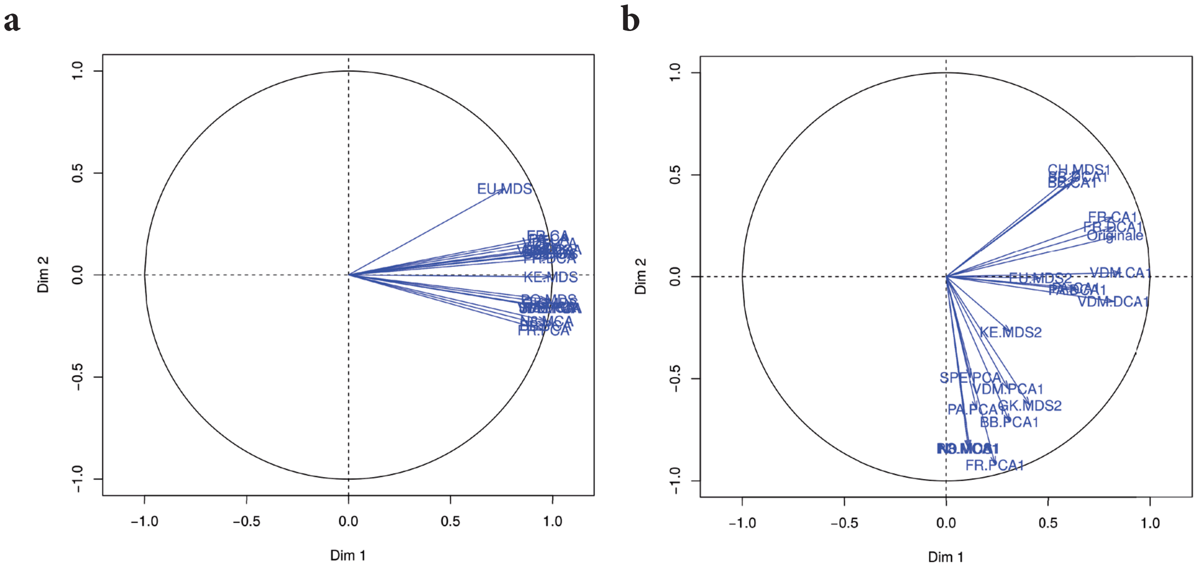

Figure 5. First simulated coenocline: circle of correlation issued by the PCA on the first coordinate of each method (**a**) and on the first corresponding rank (**b**). PCA inertia explained; (**a**): Dim 1 = 94.5%, Dim 2 = 3.3%, (**b**): Dim 1 = 39.2%, Dim 2 = 23.4%.

in Electronic Appendix A2. To visualize the correlations pattern, the methods are represented on the circle of correlations issued by the PCA performed on the first coordinates of each method (Fig. 5a). Note that the first principal component is accounted for 94.5% of total inertia, without any relevance of the second. Here, the original data appear strongly correlated with all methods but EU-MDS, with preference for the methods already quoted above. It must be noted that in this case, a distinction may be observed between the methods with the horseshoe, that are oriented below the first axis, and those with either arch or no effect, including the original data, oriented above it. Once again it may be noted that EU-MDS performs worse than all others, which keep rather well correlated between each other, in particular between the different coding dealt with the same method. Considering the Pearson's correlations between the first coordinates (or the second for four out of five MDSs, according to Procrustes results), we find that all methods but FR-PCA and EU-MDS have correlation with the specified coordinates not lower than 0.90, with the first coordinate of both CAs and DCAs always nearly totally correlated with the original data, ranging within 0.9885 and 0.9998, and also that of CH-MDS, whose correlation with the original data is 0.9974.

Very different values result for the ranks. The Spearman's correlations are much worst than the Pearson's, probably due to the folding of most methods at the extremes of the first axes or some exchanges among units. Note that the PCA of the Spearman correlation matrix gives up to three significant dimensions according to the Brokenstick method (Frontier 1976), that are accounted for 39.2, 23.4, and 11.3% of total inertia, respectively. In Figure 5b the ranks are plotted on the circle of correlations on the axes 1 and 2, summarizing 62.6% of total inertia. It is evident the apparent independence of all CAs, DCAs, and CH-MDS, the most correlated with the original order, from the others, more oriented towards the second axis. It is not a surprise that both CAs and DCAs perform much better than the others, since by projecting vertically the nearly-aligned plots onto the horizontal axis, the order may not change. In this case, FR-CA and BB-CA perform pretty well (0.84 and 0.74, respectively) and a little better do the corresponding DCAs (0.91 and 0.79, respectively, but not the others). Note also the very good performance of CH-MDS (0.77), the medium of PA-CA (0.56), probably due to the loss of fine information given by the different cover values, and the worst of VDM-CA.

3.2 *Second simulated coenoplane*

In Figure 6, the patterns of the 30 relevés issued by the 20 methods are represented on their first factor plane. In some cases, according to the GPA results, the two axes and/or their signs are exchanged to appreciate their agreement with the original pattern of the coenocline (Fig. 2). This may be due to the independence between the original gradients (correlation = –0.099) with equal inertia. Here, the first evidence is that most patterns are very different from the original one, in particular those issued by all PCAs, CAs, and DCAs with the exception of PA-CA and PA-DCA. Note that the patterns of CAs, except PA-CA, are very similar to an arch, thus in this case highly misleading, and the corresponding DCAs do not improve the representation. This ought to be expected, since the detrending destroys the ordination along the second axis. Thus, from simple inspection it is not really easy to appreciate the relations between methods and with the original pattern. Indeed, some analogy may be seen among all PCAs and MCAs, some among CAs and some among the last three MDSs on rank measures, but to check them out it is better to apply to the analyses results.

In Figure 7a, the pattern of the methods is represented on the circle of correlations issued from PCA run on the Procrustes correlation matrix between the methods (together with the other matrices in Electronic Appendix A3). This time the first axis is accounted for 82% of total inertia with 8.3% attributed to the second one. Once again, the same methods result close, though not as in the single gradient case, thus the coding does not have a strong influence on the results. This is not true for MDS, since the used indices are different. Indeed, GK-MDS, KE-MDS, VDM-CA, and PA-CA are the most correlated with the original data, whereas a certain scattering is visible along the second dimension. This is due to the

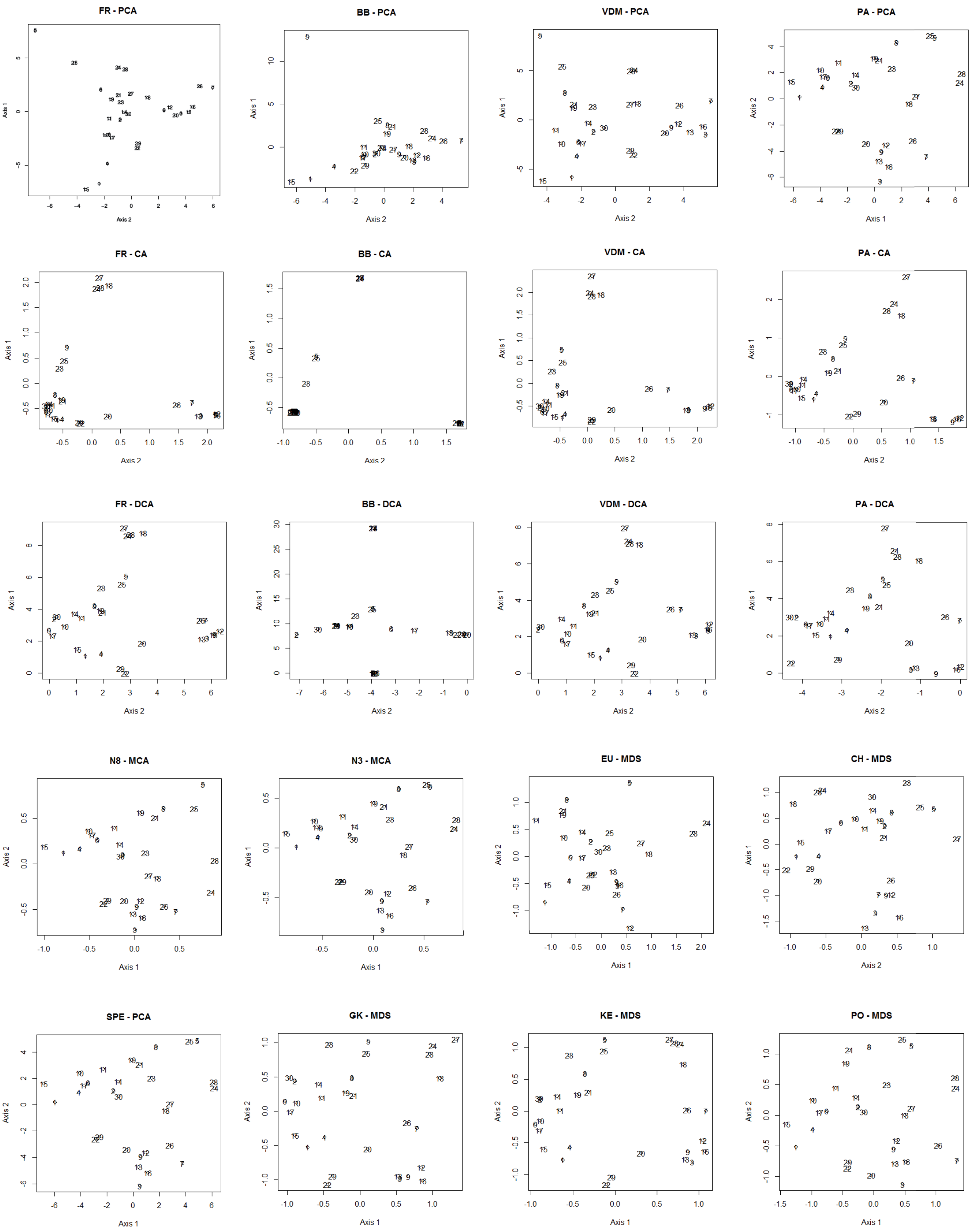

Figure 6. Second simulated coenoplane: representation of the 30 simulated communities on the plane spanned by the first two dimensions of the considered scaling methods. In the titles, the first part represents the coding and the second the used method; the bottom left graphics refers to the position of the simulated communities on the two coenoclines.

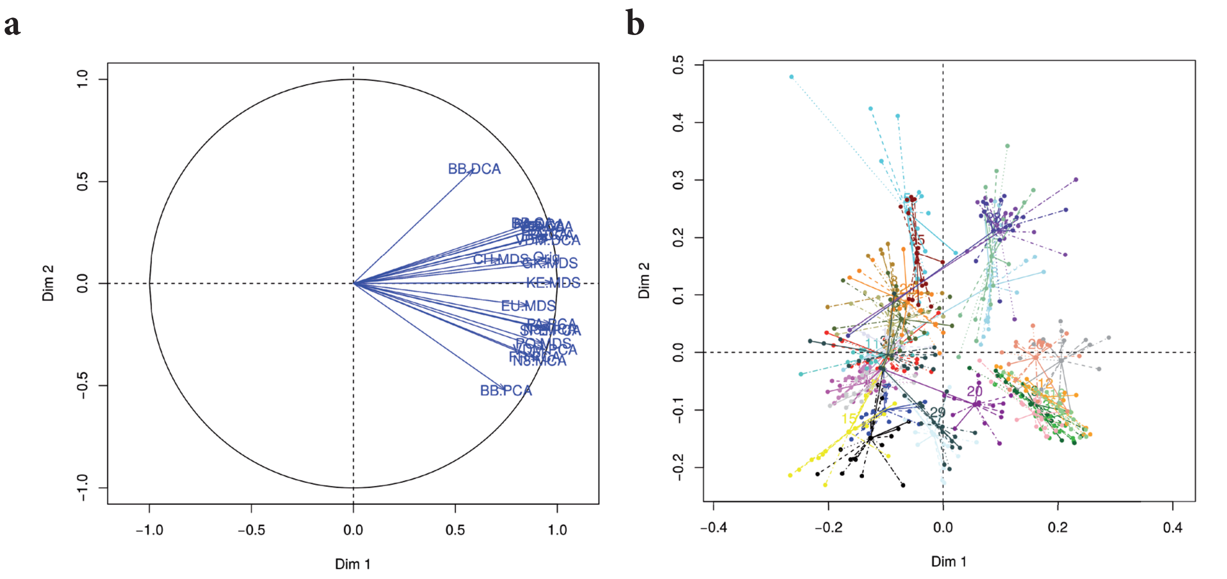

Figure 7. Second simulated coenoplane: circle of correlation issued by the PCA on the Procrustes correlation matrix (**a**) and the pattern of the samples seen by each method around each compromise (**b**). PCA inertia explained (**a**): Dim 1 = 82.0%, Dim 2 = 8.3%.

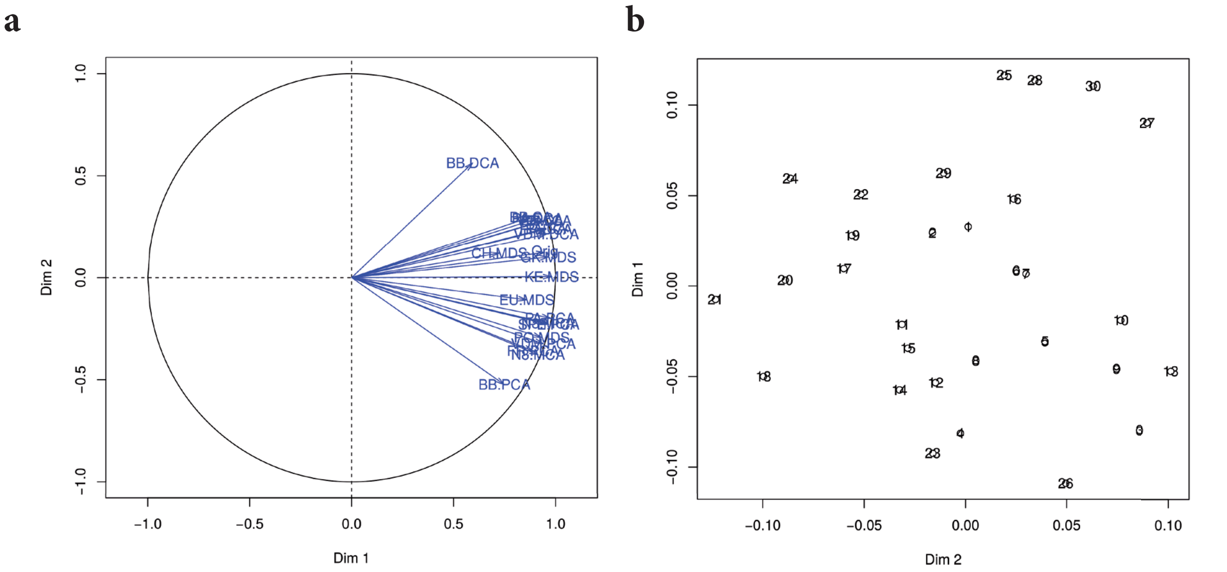

Figure 8. Second simulated coenoplane: circle of correlation issued by the PCA on the Procrustes correlation matrix computed on ranks (**a**) and the pattern of the consensus configuration on the common space of representation (**b**). PCA inertia explained (**a**): Dim 1 = 28.5%, Dim 2 = 9.2%.

bad performance of the other methods, that are opposed to the best ones (that include all CA-based methods). Note that in this case DCAs perform in a contradictory way, since VDM-DCA improves the CA performance, but PA-DCA performs worst than PA-CA and BB-DCA is really bad. In Figure 7b, the pattern of each relevé as seen from each method is shown around the compromise position found by GPA: unlike the first coenocline, here only few partial points are very distant from the consensus.

Looking at the Procrustes correlation matrix (in Electronic Appendix A3), within PCA correlations are higher than 0.79, within CA higher than 0.91, the DCAs, as said, are contradictory, the two MCAs have Procrustes correlation 0.97, and the MDSs highly various. The correlation with the original configuration ranges from 0.51 (BB-DCA) to pretty high: it is noteworthy the very good value of CA on both presence/absence and VDM (0.96 and 0.93, respectively) and of MDSs on both Kendall and Goodman-Kruskal (0.99 and 0.98, respectively). Indeed, the latter two scatter diagrams are pretty similar to the original configuration, whereas the PA-CA one does not seem, at least at first sight, really alike.

GPA was run also on the ranks corresponding to the first two coordinate sets of all methods and compared with the original ones. In Figure 8a, the methods are represented on the circle of correlation issued from the PCA run on the Procrustes correlations matrix computed on ranks and in Figure 8b the compromise position of the relevés is represented in the corresponding 2-dimensional space of representation. Note that in this case the representation of the partial positions of the samples was too confuse to be readable - and consequently interpretable - and thus we dropped it. This PCA resulted with only one significant dimension, albeit summarizing only 28.5% of the total inertia. This means that the ranks are not really coherent. Indeed, the Procrustes correlations with the original pair of ranks are not very high, the highest being 0.36 and 0.30 for BB-CA and FR-CA, re-

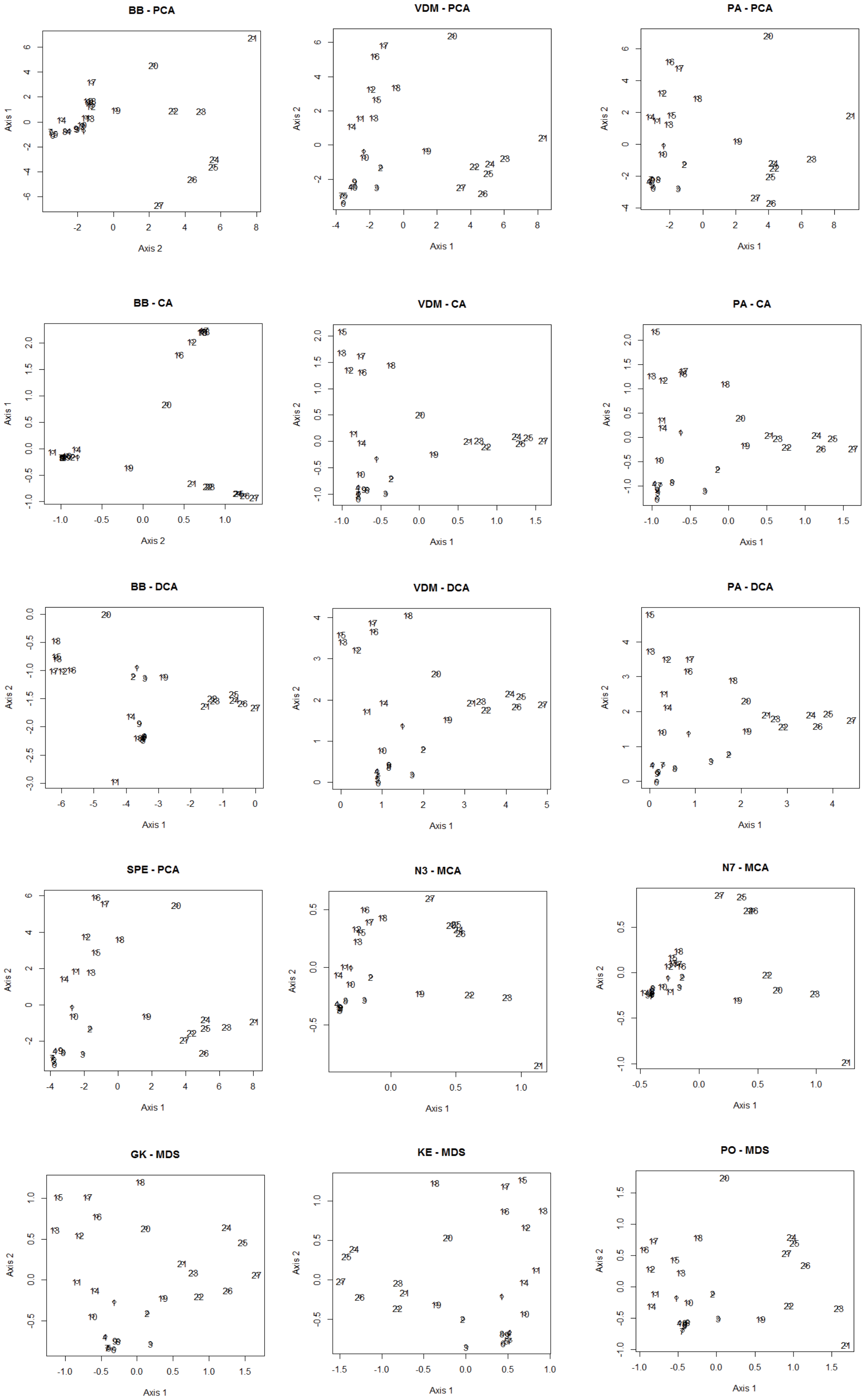

Figure 9. The real transect: representation of the relevés on the plane spanned by the first two axes issued from the 15 used methods.

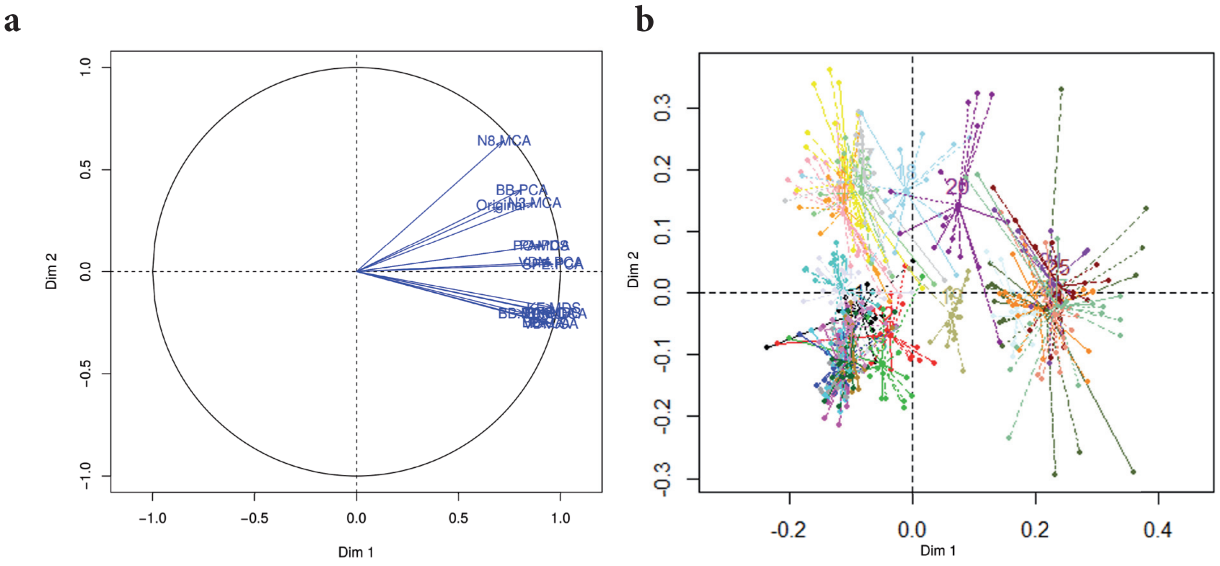

Figure 10. Analyses of the real transect: the original order and the 15 methods on the circle of correlations issued by PCA of the matrix of Procrustes correlations between methods (**a**) and the pattern of the samples seen by each method around each compromise (**b**). Dim 1 = 75.3%, Dim 2 = 7.0%.

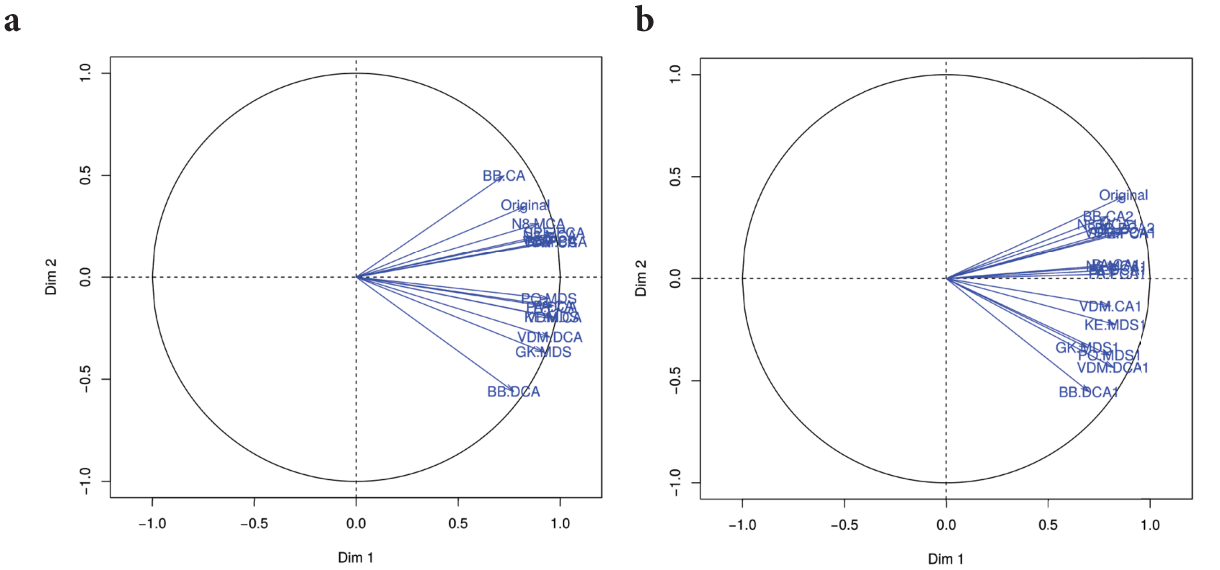

Figure 11. Analyses of the real transect: circles of correlation issued by the PCAs run on the first coordinates (**a**) and on the first ranks (**b**) of each method. PCA inertia explained: (**a**): Dim 1 = 84.8%, Dim 2 = 8.0%; (**b**): Dim1 = 66.7, Dim 2 = 8.3.

spectively, followed by PA-CA with 0.22 and both GK-MDS and KE-MDS with 0.21. As usual, some larger correlations result between different coding within the same methods, but not really relevant. Note also that DCAs this time perform systematically worst than the corresponding CAs.

The examination of the correlations between original and corresponding coordinates shows the highest values of both GK-MDS (both over 0.99) and KE-MDS (0.95 and 0.93), with PA-CA performing very well too, with correlation of 0.91 for the first and 0.86 for the second coordinate. As for the ranks, the Spearman correlations are very low and in general not significant: no method approaches both original orders, thus, we thought irrelevant to deepen this study.

3.3 The real transect

As the Argentinean transect consisted of a systematic survey of relevés taken at regular 25 m intervals, we could consider the relevés' sequence as a proxy for the gradient we were trying to reveal. As such, it could be used to compare the several gradients found in the different analyses. Note also that, as the species abundance was recorded with the Braun-Blanquet (1932) coding, the frequencies were not available and the corresponding analyses could not be done.

In Figure 9 the pattern of the relevés on the first factor plane issued by the 15 used methods is represented. Again, the usual toggle of sign and coordinates have been done according to GPA results. Unlike the other cases, this time the variations depending on the three coding (the abundance was not available in this case) within the same methods appear more different, nor a clear arch pattern appears, at least with the regularity seen in the first coenocline.

In Figure 10a the results of GPA are summarized by the representation of the methods on the first circle of correlation of the PCA of the Procrustes correlation matrix. Note that the first axis is accounted for 75.3% of total variance, with the second limited to nearly 7.0%. Unlike the previous studies, now the best performing methods appear to be N3-MCA and BB-PCA but it must be noted that the estimated original coordinates are not very well represented on the circle of correlations. This depends on the fact that only one set of original coordinates could be estimated. This may be confirmed looking at the Procrustes correlation matrix (in Electronic Appendix A4): it results that the average correlation between all methods is 0.80, with minimum 0.54. Once again it is confirmed the similar behaviour of all methods. On the opposite, the Procrustes correlation of the methods with the original gradient ranges within 0.69 of SPE-PCA and 0.67 of N3-MCA on one side and 0.43 of BB-DCA on the other, with an average of 0.64. Here, the DCAs perform a little better than the corresponding CAs, but in the case of BB-DCA. Indeed, PCA on Procrustes correlation matrix shows better the agreement within methods than the relation with the original data, as said due to the fact that only one gradient was considered as the original reference.

Some irregular pattern of vegetation along the transect may be the cause of this irregularity: this is put in evidence by the distribution of relevés on the common reference space issued by GPA, shown in Figure 10b, in which all the methods' partial representations are shown too. It is evident the gathering of the relevés in three well separated homogeneous groups, with the exception of 18, 19, and 20 that are far apart from the others, in some intermediate position. Within each group, no agreement with the original order appears, whereas the sequence of the three vegetation types in a kind of horseshoe form is visible as a whole. As in the case of the first coenocline, a better information may result by checking both Pearson's and Spearman's correlations limited to only one dimension: they too are reported in Electronic Appendix A4. Once again, exchanges of signs and coordinates result from GPA.

In Figure 11 the circles of correlations issued from the PCAs on these correlation matrices are represented. In both cases, the matrices are uni-dimensional, the following axes being accounted for little inertia, not significant in comparison with the brokenstick statistics. Looking at Figure 11a the pattern of methods in the Pearson's correlation is shown. Here the first factor is accounted for nearly 85% of the total inertia, whereas the second (non-significant) explains only nearly 8.0%. The axis is not really well correlated with the original sequence (only 0.83) and worse only with BB-CA and BB-DCA, whereas with all other methods the correlations range within 0.89 and 0.96. This result confirms the high homogeneity of the first factors of the different analyses. Along the second factor, five groups are visible: BB-CA, the most correlated (0.87) with the original transect that is aside, then a group composed by all PCAs and all MCAs, whose correlation with the original gradient ranges between 0.85 and 0.72 but whose inner correlation is always over 0.93, then most others, less correlated with the original, yet highly correlated between them, but GK-MDS, the less correlated with all others, and eventually BB-DCA alone. Note that in this case PA-CA belongs to the group of the poorly correlated, but its correlation with the transect is worth 0.79, the fourth best value. Once again, DCAs perform worst than CAs.

In Figure 11b, the pattern of methods in the Spearman's correlation is shown. Here the first factor is accounted for nearly 67% of the total inertia, whereas the second (non-significant) explains only 8.3%. Thus, the one-dimensionality is less marked than in the previous case: the methods are more scattered along the second axis, depending on their inner correlation: here, five groups are visible, progressively with decreasing correlation with the original order. Looking at the Spearman correlation matrix (in Electronic Appendix A4), it results that the most correlated with the original order are SPE-PCA, BB-PCA, and BB-CA (0.86, 0.85 and 0.84, respectively), whereas the worst are the three DCAs and the three MDS. This time the correlation of PA-CA is only 0.72, in a medium position.

4. Discussion

In Table 3, the correlations between the first coordinates issued by the methods and the original ones (two coordinates in the case of the coenoplane, one for the others) are reported: in the first three columns are found the Procrustes', Pearson's, and Spearman's correlations in the case of the first simulated coenocline; in the following six, the Procrustes' and two Pearson's referring to the coordinates of the second coenocline, then the Procrustes' and two Spearman's according to their ranks; in the last three the Procrustes', Pearson's, and Spearman's correlations referring to the third real transect. To summarize the results, we may observe the following:

First coenocline. Three main patterns are recognized with minor differences within them: the horseshoe, the arch, and the linear one. The methods showing the latter result better according to Procrustes correlation, due to their linear main structure. As for Pearson's, the best performing methods are PA-CA and CH-MDS based on the frequencies, followed by all other correspondence analyses, then all other methods, with EU-MDS by far the worst performing. Note that here the DCAs give their best, but not significantly better than the corresponding CAs. Considering ranks, FR-DCA and FR-CA are the best performing, followed by CH-MDS and the other CAs, with PA-CA in a medium position. Apart from these, only KE-MDS is correlated with the original order, while the others are in practice independent.

Second coenoplane. The original pattern is difficult to be recognized but from GK-MDS and KE-MDS, that are the best performing. The only other method that approaches them is PA-CA, but only numerically, since the configuration does not seem alike on visual inspection. The DCAs perform in a contradictory way: it was expected, due to the manipula-

Table 3. The Procrustes, Pearson, and Spearman correlations of the coordinates issued by the tested methods of ordination with the original ones. First three columns: rst coenocline; second three: coordinates of second coenoplane; third tree: ranks of the second coenoplane; last three: real transect. The asterisk indicates that the two following correlations are with the same axis issued by the analysis.

	First coenocline				Coordinates		Second coenocline	Ranks		Real Transect		
	Procrustes	Pearson	Spearman	Procrustes	Pearson 1	Pearson 2	Procrustes	Spearman 1	Spearman 2	Procrustes	Pearson	Spearman
FR.PCA	0.6774	0.8989	0.1217	0.7543	0.5183	0.5340	0.1738	* 0.2939	0.0945			
BB.PCA	0.7192	0.9148	0.1217	0.6181	0.6602	0.5555	0.1966	* 0.2610	0.1742	0.6291	0.8071	0.8535
VDM.PCA	0.7294	0.9315	0.2085	0.8290	0.6906	0.6398	0.0801	* 0.0532	0.0923	0.6567	0.8211	0.7973
PA.PCA	0.7313	0.9277	0.1279	0.8880	0.7469	0.7764	0.1172	0.1484	0.0398	0.6272	0.7862	0.7515
SPE.PCA	0.7294	0.9309	0.1795	0.8733	0.7254	0.7642	0.1954	0.3055	0.0443	0.6906	0.8562	0.8626
FR.CA	0.8169	0.9885	0.8394	0.9067	0.7481	0.6809	0.3031	0.2912	0.1368			
BB.CA	0.7631	0.9893	0.7433	0.8346	0.7560	0.6259	0.3552	0.4518	0.1742	0.5613	0.8723	0.8437
VDM.CA	0.8102	0.9910	0.5373	0.9330	0.7380	0.6644	0.1829	0.1871	0.1666	0.6088	0.7809	0.6325
PA.CA	0.8265	0.9922	0.5591	0.9558	0.9074	0.8581	0.2226	0.1786	0.1608	0.6055	0.7933	0.7216
FR.DCA	0.9895	0.9967	0.9088	0.9347	0.8278	0.7836	0.1930	* 0.2690	0.1235			
BB.DCA	0.9913	0.9972	0.7895	0.5169	0.0048	0.7077	0.2933	0.3544	0.2191	0.4377	0.4549	0.3736
VDM.DCA	0.9917	0.9987	0.4732	0.9653	0.8149	0.7204	0.2140	* 0.1551	0.2507	0.6432	0.7114	0.5592
PA.DCA	0.9953	0.9988	0.5813	0.8955	0.8876	0.9118	0.1952	* 0.3068	0.1079	0.6452	0.7778	0.7057
N8.MCA	0.6915	0.9107	0.0514	0.8122	0.6393	0.6294	0.1752	0.1204	0.2271	0.6354	0.7249	0.7277
N3.MCA	0.7302	0.9278	0.1279	0.8767	0.7534	0.7820	0.2120	0.1337	0.2734	0.6706	0.7825	0.7692
EU.MDS	0.5603	0.7313	0.1947	0.7787	0.7475	0.7437	0.1851	0.1172	0.2770			
CH.MDS	0.9825	0.9974	0.7726	0.6628	0.2191	0.8712	0.0688	0.0638	0.0723			
GK.MDS	0.8108	0.9421	0.2783	0.9907	0.9891	0.9925	0.2061	0.3317	0.0794	0.6245	0.6520	0.5244
KE.MDS	0.7657	0.9469	0.4256	0.9781	0.9541	0.9328	0.2095	0.2249	0.1849	0.6392	0.7503	0.6068
PO.MDS	0.7474	0.9207	0.0514	0.8361	0.8092	0.8072	0.1851	0.0380	0.3050	0.6144	0.6714	0.5409

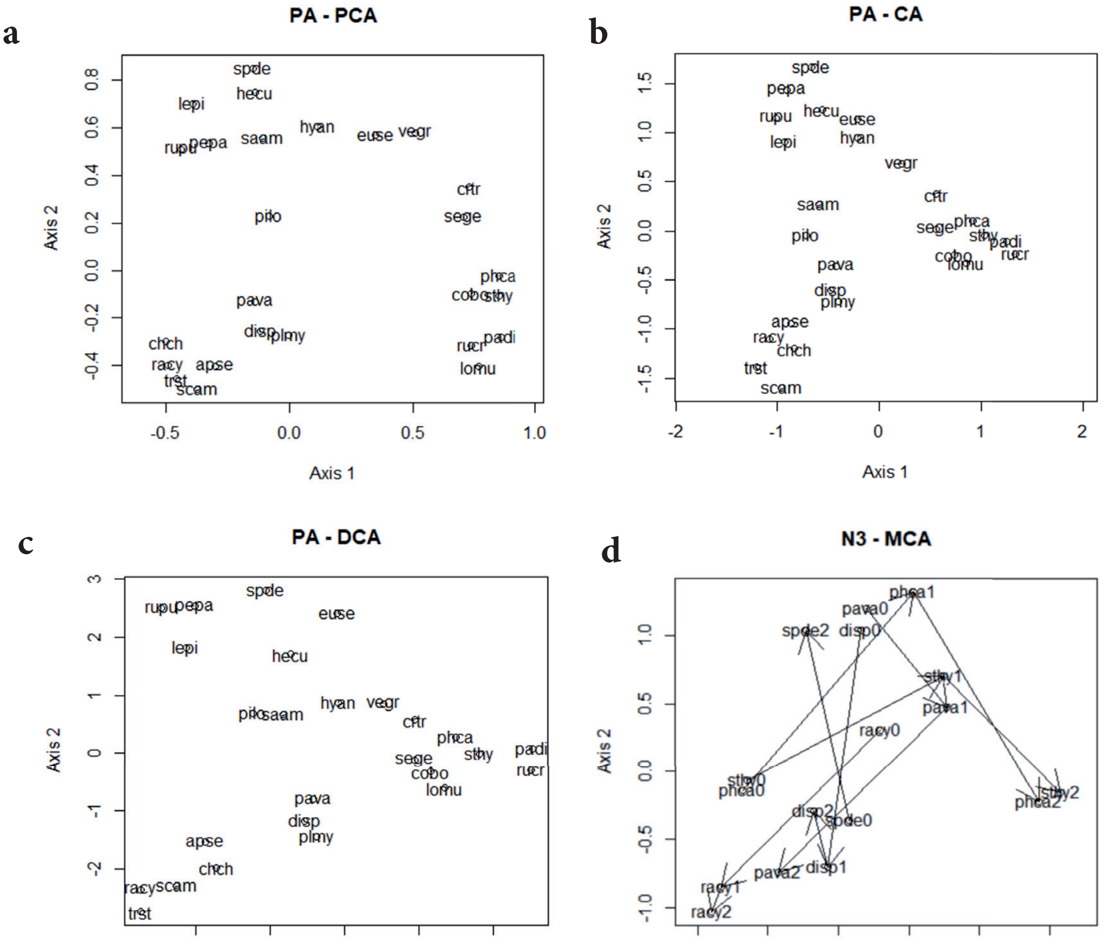

Figure 12. Analyses of the real transect: relevant species on the first factor plane issued by PA-PCA (**a**), PA-CA (**b**), PA-DCA (**c**), and N3-MCA (**d**).

tion of the second coordinate. All others perform worst. Considering ranks, the best performing were FR-CA and BB-CA, both with Procrustes correlation little above 0.30, while all others performed really poorly and without any apparent relation between them. Indeed, considering the two ordinations separately, no method resulted able to approximate both. Nevertheless, we may quote both FR-CA and BB-CA with a reasonable correlation at least for the first coordinate. The best pair of rank correlations is that of BB-DCA, which on the opposite returns a first coordinate with nearly zero correlation with the original first one.

Real transect. The patterns are really various and it is difficult to visually detect similarities among them. Nevertheless, there is a high Procrustes correlation between the same methods applied to the different coding. On the opposite the Pearson's correlations of methods with the original transect are within 0.69 of SPE-PCA and 0.60, with only BB-CA and BB-DCA lower (0.56 and 0.43, respectively). Higher Pearson's correlations result between the first coordinates, ranging within 0.87 of BB-CA and 0.65 of GK-MDS, with only BB-DCA worst (0.43); as for Spearman's, they range within 0.86 and 0.52 of the same methods. In both cases PA-CA performs reasonably well (0.79 and 0.72, respectively). Note that in this case PO-MDS performs better than GK-MDS.

All methods, applied to the different coding, provided similar results: in particular, all PCAs, all CAs, and all MCAs resulted highly coherent in all data sets with respect to the coordinates; larger differences result when dealing with the ranks. Note in this case that some PCAs' and DCAs' highest rank correlations with the two original coenoplane coordinates refer to the first found axis only. For this reason, the two independent gradients may not be detected by these methods. Among PCAs, the better performing resulted that on Spearman's correlation: a pity that it does not allow the simultaneous analysis of the species. The other coding perform alternatively, so that no true best method results. Among multiple correspondence analyses, N3-MCA performs systematically better than N8-MCA, showing a pattern very similar to PCA, with the advantage to represent the variation of species' densities along the gradient.

In what concerns the ordinations, CA, that shows an arch effect when dealing with one gradient, performs much better than all others which show a horseshoe: this was expected for PCA but not for MCA and MDS on ordinal data.

On the opposite, DCA results really poor: it equals and little overcomes CA on the simulated coenocline, but performs much worst on both the coenoplane and the transect. This was expected: considering the loss of meaning of the second component, it is obvious that a coenoplane may not be well represented through DCA, not even a real data set in which some minor gradients may appear, even if we did not consider them here. In our comparison, the better performance of the first coordinate with respect to CA, according to both data and ranks, should depend essentially on the rescaling at the extreme of the first axis, since theoretically their first set of coordinates ought to be identical. Note that the loss of meaning of the second dimension does not improve the interpretation of the first dimension in the corresponding CA, but rather prevents the identification of a second gradient, if any. In addition, the loss of information about species' both diversity and range within the relevés (provided in CA by the second component, see Camiz 2005) suggests to avoid the use of DCA.

As for MDS, CH-MDS performs pretty well on the simulated coenocline only, but the best performing on the coenoplane are GK-MDS and KE-MDS (showing a horseshoe on the coenocline) and none is outstanding in the case of real data. Considering also the lack of simultaneous treatment of the species in these analyses, one may really question the opportunity to adopt them in place of CA. Indeed, this is a major point that deserves a reflection: PCA, CA, and MCA provide graphics of species whose factors are interpreted together with those of the relevés, thanks to the transition formulas. This does not exist in MDS, whereas in DCA the manual destruction of the second axis may seriously compromise the interpretation.

In Figure 12 are reported the graphics concerning the species pattern issued by PA-PCA, PA-CA, PA-DCA, and N3-MCA when run on the real transect data. In each graphics, from left to right, the *salt grassland* is found first, represented by the most abundant species *Paspalum vaginatum* Sw. (in the graphics *pava*) and *Distichlis spicata* (L.) Greene (*disp*), together with *Ranunculus cymbalaria* Pursh (*racy*), *Chaetotropis chilensis* Kunth (*chch*), and *Apium sellowianum* H.Wolff (*apse*). The center of the transect is occupied by the *espartillar*, whose dominant species is *Spartina densiflora* Brogn. (*spde*) and where *Rumex pulcher* L. (*rupu*), *Heliotropium curassavicum* L. (*hecu*), and *Petunia parviflora* Juss. (*pepa*) may be found albeit less abundant. Eventually, on the right the *flecillar* species are found at the end of the transect, close to the stream, where its dominant species are found: *Phyla canescens* (Kunth) Greene (*phca*), *Stypa hyalina* Nees (*sthy*), and *Paspalum dilatatum* Poir. (*padi*). This pattern is clearly visible in the two graphics issued by both PA-PCA and PA-CA in which either the horseshoe or the arc lead the reader through the curvilinear pattern. Note also the different position of *Paspalum vaginatum* and *Heliotropium curassavicum*, probably more spread along the transect, in respect with *Ranunculus cymbalaria* and *Rumex pulcher*, probably more concentrated in the respective communities: this may be inferred, since the first are closer to the centre and the second at the extreme of the arch thickness.

Unlike those, in the graphics issued by PA-DCA this pattern is lost: the disappearing of the curve and the lack of meaning of the second axis cause a confusion in the intermediate positions. As a consequence, both *Paspalum vaginatum* and *Distichlis spicata* (belonging to the *salt grassland*) now follow *Spartina densiflora, Rumex pulcher, Heliotropium curassavicum,* and *Petunia parviflora*, all species belonging to the *espartillar*. Thus, along the first axis, the species belonging to either communities are mixed without the possibility to tell them apart.

As for the N3-MCA, the curvilinear pattern is analogous to PCA, but here the different species densities are represented: thus, *Ranunculus cymbalaria* appears exclusively into the *salt grassland, Paspalum vaginatum* is present there with high density, but with low density in the other communities too, *Phyla canescens* and *Stypa hyalina*, absent in the *salt grassland*, are little present in the *espartillar* and highly present in the *flechillar*. Indeed, this information in both PA-PCA and PA-CA is missing, whereas it may be of high interest for the study of vegetation distribution along gradients.

5. Conclusions

The original aim of this paper was to check to what extent different recoding of the Braun-Blanquet codes could affect the multidimensional analyses: the answer given by this experimentation is *very little* and the corresponding debate seems very little grounded too. In fact, this was confirmed by the high correlations found within the same methods between the coding. Thus, presence/absence might be preferred, this way avoiding all cover estimation errors: it may be wrong only if the researcher did not identify correctly a species, but it is not affected by the estimate error. Thus, our results and interpretations are in agreement with the conclusions of Wilson (2012) that state the better performance of presence/absence.

Should one wish to get small scale information concerning the relevés ordering, then the use of frequencies (even approximated by the Braun-Blanquet's classes mean values) may be worth. Note that this result may also prevent the idea of applying transformations to the abundances: their use aims at producing better results, but instead introduces arbitrary bias in the data.

Concerning the methods, once again differences are not really substantial for an experienced user, but their specific features lead the same to a choice. CA proved once again to be the best method to use: it is advantageous in respect with PCA, thanks to the arch pattern instead of the horseshoe: indeed, this allows to consider the component representing the arch length (usually the first) as a reasonable approximation of the underlying gradient, that in PCA is folded at the extremes. In addition, the transversal position of the relevés with respect to the arch mid-line may be interpreted as higher/lower diversity for the relevés and larger/thicker range for the species (Camiz 2005): a feature difficult to find in the other representations.

This argument leads to a reject of all DCA methods: they do not carry any new information in respect with CA, because the first component is only partially (and arbitrarily) rescaled, but instead they destroy the information carried by the transversal position, only visible in the factor plane. This may create problems in the interpretation of the communities, in particular when ubiquitous species exist (located close to the centroid), that will be mixed together with those in the intermediate communities (in CA located at the extreme of the second axis, but here flattened on the first one).

The behaviour of MDS does not really overcome that of CA, unless by using ordinal indices. The drawbacks are the loss of the correspondent ordination of the species and the finding of the horseshoe once again, unless a strong second gradient exists.

A particular remark deserves MCA: considering its ability to represent the different levels of species, it may be helpful to understand their pattern of distribution better than CA. Indeed, MCA performs in practice just as PCA, with the same horseshoe effect, but with the extra ability to represent the various levels of species density and to deal simultaneously with the sociability code. Thus, its experimentation may be an effective alternative to PCA, and it may be preferred if one wishes to study the structure of communities rather than gradients. Thus, this method certainly deserves being studied in deeper detail, maybe including the sociability index. The found results suggest to prefer the N3 coding instead of the VDM-like N8.

The argument concerning the use of order-based rather than measure-based methods remains partially open: theoretically order-based statistics ought to be used, unless dealing with abundances or presence/absence. Indeed, the best response of SPE-PCA and the good performances of both Goodman-Kruskal and Kendall MDSs on simulated data may encourage their use, but it is surprising that the theoretically very well grounded statement of Podani (2006)'s index, in our experimentation resulted in line with the other order based indices. A hypothesis may be that the large amount of noise in ecological data (Gauch 1982) may prevent the relevance of such an adjustment. However, both the presence of horseshoe and the lack of a corresponding representation of the species might discourage the use of the order based indices, at least as a first exploratory tool. We must admit that this result was far from our expectations.

Summarizing, we may state that Correspondence Analysis with presence/absence data is the best tool for a general purpose work, with some advantage of CA on frequencies or Braun-Blanquet's averages of classes, should one wish to have better information about low level ordering. Should one be interested in classification and species distribution, MCA limited to three classes of abundance (none, little, much), might be considered. Indeed, Noy-Meir (1971) already commented in favour of PCA, should one be interested in detecting communities structure: now, MCA provides results very close to PCA (including the horseshoe in place of the arch) for what concerns relevés, but, even with a very reduced number of levels, adds to the usual identification of the association species-relevés some elements of distribution that may contribute to a better understanding of the structure of the studied communities.

Note that the representation of species, even more than that of relevés, shows that dealing with the Guttman effect and following the pattern issued by the analyses is much more fruitful than arbitrarily flattening it, because this way relevant information provided by the first factor plane as a whole is lost. Camiz (2005) discusses the problem and suggests a simple redressing method that does not destroy the original pattern.

We remind here that we worked in an exploratory framework, without any aim to identify exactly the gradients' values in our data, a purpose that none of the experimented methods may fulfill. For this, some kind of modelling by taking into account the form of the species distribution might be much more appropriate: as a starting point, a serious experimentation on Ihm (Ihm and van Groenewoud 1975, 1984) and Goodall's models (Johnson and Goodall 1980, Goodall and Johnson 1982) would be a really interesting complement to our study.

Acknowledgements. For this work, Camiz and Pillar were supported by the Sapienza Università di Roma and Universidade Federal do Rio Grande do Sul grants for bilateral relations. Camiz was also granted by Sapienza's agreement with Argentina and by the Special Visiting Researcher Fellowship of Brazilian CNPq, under the Brazilian Scientific Mobility Program "Ciências sem Fronteiras", Process #: 314443/2014-2.

References

Abdi, H. 2007. Singular value decomposition (*SVD*) and generalized singular value decomposition (*GSVD*). In: N. Salkind (ed.), *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA. pp. 940–945.

Benzécri, J.-P. 1982. *L'analyse des données*. Dunod, Paris.

Bräuer, H. 1982. Koordinaten-Schweinerei. *Münchner Mediziner Wochenschrift* 124. quoted in *ROeS Nachristen*, Biometrische Gesellschaft Region Oesterreich-Schweiz, 15:15.

Braun-Blanquet, J. 1932. *Plant Sociology: The Study of Plant Communities*. McGraw-Hill, New York. English translation of *Pflanzensoziologie* by G.D. Fuller and H.S. Conard.

Camiz, S. 1991. Reflections on spaces and relationships in ecological data analysis: effects, problems, and possible solutions. *Coenoses* 6:3–13.

Camiz, S. 1993. Scopi e finalità dell'analisi della vegetazione e relativi schemi di rilevazione campionaria. In: S. Zani (ed.): *Metodi statistici per le analisi territoriali, Studi urbani e regionali.* Franco Angeli, Milano. pp. 301–322.

Camiz, S. 1994. A procedure for structuring vegetation tables. *Abstracta Botanica* 18:57–70.

Camiz, S. 2002. *Contributions, à partir d'exemples d'application, à la méthodologie en analyse des données. Chap. 4 - Les données de végétation*. Ph.D. thesis, Université Paris IX Dauphine.

Camiz, S. 2005. The Guttman effect: its interpretation and a new redressing method. *Data Analysis Bulletin* 5:7–34.

Camiz, S. and J.-J. Denimal. 2011. Procrustes analysis and stock markets. *Case Studies in Business, Industry and Government Statistics* 4:93–100.

Carnevale, N. and P. Torres. 1990. The relevance of physical factors on species distributions in inland salt marshes (Argentina). *Coenoses* 5:113–120.

Carnevale, N., P. Torres, S. Boccanelli and J.P. Lewis. 1987. Halophilous communities and species distributions along environmental gradients in southeastern Santa Fe Province, Argentina. *Coenoses* 2:49–60.

Chytrý, M. et al.. 2016. 'European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Appl. Veg. Sci.* 19:173–180.

Diday, E. and J.-C. Simon. 1976. Clustering analysis. In: K. Fu (ed.), *Digital Pattern Recognition.* Springer, Berlin. pp. 47–94.

Dixon, P. 2003. VEGAN, a Package of R functions for community ecology. *J. Veg. Sci.* 14:927–930.

Eckart, C. and G. Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1:211–218.

Frontier, S. 1976. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. *J. Exp. Mar. Biol. Ecol.* 25:67–75.

Gauch, H.G. 1982. Noise reduction by eigenvector ordinations. *Ecology* 63:1643–1649.

Goodall, D.W. and R. Johnson. 1982. Non-linear ordination in several dimensions. *Vegetatio* 48:197–208.

Goodman, L.A. and W.H. Kruskal. 1954. Measures of association for cross classifications. *J. Amer. Stat. Assoc.* 49:732–764.

Gower, J.C. 1975. Generalized procrustes analysis. *Psychometrika* 40:33–51.

Gower, J.C. and G.B. Dijksterhuis. 2004. *Procrustes Problems*. Oxford University Press, Oxford.

Greenacre, M. 2007. *Correspondence Analysis in Practice*. 2nd ed. Chapman and Hall/CRC, London.

Guttman, L. 1953. A note on Sir Cyril Burt's factorial analysis of qualitative data. *Brit. J. Stat. Psychol.* 6:21–24.

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

Husson, F., S. Lê, and J. Pagès. 2017. *Exploratory Multivariate Analysis by Example Using R*. CRC Press, Boca Raton, FL..

Ihm, P. and H. van Groenewoud. 1975. A multivariate ordering of vegetation data based on Gaussian type gradient response curves. *J. Ecol.* 63:767–777.

Ihm, P. and H. van Groenewoud. 1984. Correspondence analyses and Gaussian ordination. In: J.M. Chambers, J. Gordesch, A. Klas, L. Lebart, and P.P. Sint (eds.), *Compstat Lectures*. Physica-Verlag, Vienna. pp. 5–60.

Johnson, R.W. and D.W. Goodall. 1980. A maximum likelihood approach to non-linear ordination. *Vegetatio* 41:133–142.

Kendall, M.G. 1938. A new measure of rank correlation. *Biometrika* 30:81–89.

Kenkel, N.C. and L. Orlóci. 1986, Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology* 67:919–928.

Kruskal, J.B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1–27.

Kruskal, J.B. 1964b Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:115–129.

Kruskal, J.B. and F. Carmone. 1971. *How to Use MD-SCAL (Version 5M): And Other Useful Information*. Technical report, Bell Laboratories.

Lebart, L., A. Morineau, and K.M. Warwick. 1984. *Multivariate Descriptive Statistical Analysis; Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.

Lebart L., M. Piron, and A. Morineau. 2006. *Statistique exploratoire multidimensionnelle: visualisation et inférences en fouilles de données*. Dunod, Paris.

Legendre, P. and L.F.J. Legendre. 2012. *Numerical Ecology*. 3rd ed. Elsevier, Amsterdam.

Lisboa, F.J.G., P.R. Peres-Neto, G.M. Chaer, E. Da Conceio Jesus, R. J. Mitchell, S.J. Chapman, and R.L.L. Berbara. 2014. Much beyond Mantel: bringing Procrustes association metric to the plant and soil ecologist's toolbox. *PLoS ONE* 9(6):e101238

Minchin, P.R. 1987. Simulation of multidimensional community patterns: towards a comprehensive model. *Vegetatio* 71:145–156.

Noy-Meir, I.. 1971. Multivariate analysis of the semi-arid vegetation in Southeastern Australia: nodal ordination by component analysis. *Proc. Ecol. Soc. Aust*. 6:159–193.

O'Hara, R.B. and D.J. Kotze. 2010. Do not log-transform count data. *Meth. Ecol. Evol.* 1:118–122.

Oksanen, J. et al. 2017. *Vegan: Community Ecology Package. R-package version 2.4-5*. Technical report, URL http://CRAN. R-project.org/package= vegan.

Orlóci, L. 1978. *Multivariate Analysis in Vegetation Research*. 2nd ed. Junk, The Hague.

Pillar, V.D. 2013. How accurate and powerful are randomization tests in multivariate analysis of variance? *Community Ecol.* 14: 153–163.

Podani, J. 1997. A measure of discordance for partially ranked data when presence/absence is also meaningful. *Coenoses* 12:127–130.

Podani, J. 2001. SYN-TAX 2000. *Computer Programs for Data Analysis in Ecology and Systematics. User's Manual.* Scientia, Budapest.

Podani, J. 2005. Multivariate exploratory analysis of ordinal data in ecology: pitfalls, problems and solutions. *J. Veg. Sci.* 16:497–510.

Podani, J. 2006. Braun-Blanquet's legacy and data analysis in vegetation science. *J. Veg. Sci.* 17:113–117.

R Core Team. 2015. *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ricotta, C. and G. Avena. 2006. On the evaluation of ordinal data with conventional multivariate procedures. *J. Veg. Sci.* 17:839–842.

Romane, F. 1972. *Applications à la phytoécologie de quelques méthodes d'analyse multivariable.* Ph.D. thesis, Université des Sciences et Techniques du Languedoc, Montpellier.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

van der Maarel, E. 1966. *Over Vegetatiestructuren, Relaties en Systemen, in het Bijzonder in de Duingraslanden van Voorne*. Ph.D. thesis, Utrecht Universitaet, Utrecht.

van der Maarel, E. 1979. Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio* 39:97–114.

White, G.C. and R.E. Bennetts. 1996, Analysis of frequency count data using the negative binomial distribution. *Ecology* 77:2549–2557.

Wilson, J.B. 2012. Species presence/absence sometimes represents a plant community as well as species abundances do, or better. *J. Veg. Sci.* 23:1013–1023.

Received June 15, 2017
Revised November 12, 2017
Accepted December 10, 2017

Electronic Appendix

A1. Data files

Table A1.1. First coenocline data.

Table A1.2. Second coenoplane data.

Table A1.3. Real transect data.

A2. First coenocline matrices

Table A2.1. First coenocline: Procrustes correlation matrix resulting from the first two coordinates issued by the used methods.

Table A2.2. First coenocline: Pearson's correlations between the first coordinates issued by the used methods.

Table A2.3. First coenocline: Spearman's correlations between the ranks of the first coordinates issued by the used methods.

A3. Second coenoplane matrices

Table A3.1. Second coenoplane: Procrustes correlation based on the first two coordinates issued by the used methods.

Table A3.2. Second coenoplane: Procrustes correlation based on the ranks of the first two coordinates issued by the used methods.

A4. Real transect matrices.

Table A4.1. Real transect data: Procrustes correlation based on the first two coordinates issued by the used methods.

Table 4.2. Real transect data: Pearson's correlation based on the first coordinates issued by the used methods.

Table A4.3. Real transect data: Spearman's correlations based on the ranks of the first coordinates issued by the used methods.

The Appendix may be downloaded from www.akademiai.com.