Assessing the relative importance of methodological decisions in classifications of vegetation data

Attila Lengyel[1*] and János Podani[2,3]

[1]MTA Centre for Ecological Research, Institute of Ecology and Botany, Alkotmány u. 2-4., H-2163 Vácrátót, Hungary

[2]Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University, Pázmány P. s. 1/C, H-1117 Budapest, Hungary

[3]Ecology Research Group of the Hungarian Academy of Sciences, Pázmány P. s. 1/C, H-1117 Budapest, Hungary. E-mail: podani@ludens.elte.hu

*Corresponding author: lengyel.attila@okologia.mta.hu

Abstract

Questions: What is the relative importance of our methodological decisions concerning sampling (plot size) and data analysis (data transformation, resemblance coefficient, hierarchical clustering strategy and the number of clusters) in vegetation classification? Are there differences between the conclusions when the full range or only a more practical narrow range of methodological choices is tested? What is the difference between results for actual and random data?

Location: Rock grassland in Hungary.

Methods: The full procedure of vegetation classification was simulated using actual and random data. Variation in classification results was partitioned using distance-based redundancy analysis. The RDA models were subjected to variation partitioning to determine the relative importance of methodological decisions.

Results: RDA models explained more variation in classifications of random than in real data. Classification algorithm, cluster level, data transformation and mean plot size were always included among the most significant variables, however, the other variables also had considerable effect in certain situations.

Conclusions: As adjusted R-squared values suggest, the overall effect of methodological decisions on classifications is larger for randomly structured than actual data, due possibly to stronger clustering tendency in the latter. The clustering algorithm, cluster level, data transformation and plot size should be chosen most carefully before classification analyses, but any of the examined decisions can significantly affect the result. In addition to the mean, the range of plot sizes should also be carefully delimited during relevé selection for classification studies. The main decision about the classification algorithm is whether a chain-forming or group-forming method is used. The data transformation had more significant effect on real data than on simulations with random variation, thus supporting the ability of the application of different abundance scales in revealing different facets of biologically relevant patterns in community composition. The resemblance measure had relatively weak effect suggesting that it is not as influential as previously thought.

Keywords

Data transformation; Flexible clustering; Model selection; Multivariate analysis; Plot size; Resemblance measure

Abbreviations

db-RDA = distance-based redundancy analysis; GAM = generalized additive model; PCoA = principal coordinates analysis; RDA = redundancy analysis; SD = standard deviation

Running head

Relative importance of methodological decisions

Introduction

Classification of vegetation has long been the primary research objective in phytosociology and still represents an integral part of vegetation science in general (Whittaker 1973; Mucina 1997; Peet & Roberts 2013). It provides a firm reference basis for syntaxonomy, similarly to the classification of living organisms in biological systematics (or taxonomy). Scientific communication would be impossible without a common basis for recognizing, separating, describing, naming and mapping plant communities, that is, without classification of vegetation units. In addition to the fact that syntaxonomy is conditioned upon taxonomy, there is a fundamental difference between these two fields of biology. Whereas the basic observational units of classification in conventional systematics are natural entities (individuals), community classification requires the use of – more or less – arbitrarily delineated tracts from the vegetation continuum. Therefore, one is faced with a multitude of methodological choices that have to be made in the real *topographical* space (Podani 1984), that is, in the field. These include the appropriate selection of sampling criteria, that is, plot size, shape, number and arrangement (Kenkel et al. 1989). Syntaxonomy and taxonomy share only the problem of *conceptual* and *methodological* decisions which concern the variables to describe the study objects, measurement scale, resemblance coefficient and clustering algorithm to be used during data processing (Podani 1989). Tradition, fashion, practicability, comparability with others' results, availability of software and similar, more or less subjective considerations may guide the user in this methodological jungle. Nevertheless, since no absolute and universally valid criteria are available, all decisions remain unavoidably arbitrary in every step of the study. An important philosophical conclusion is that any attempt to find and define unique classifications in vegetation science will be illusory – which does not mean that the effect of sampling and analysis upon the results should be disregarded in community analysis.

The importance of such methodological choices in multivariate analysis has long been recognized by several authors (plot size: Kenkel & Podani 1991; Otýpková & Chytrý 2006; Dengler et al. 2009; measurement scale: Jensen 1978; van der Maarel 1979; Wilson 2012; resemblance measure: Green 1980; Hajdu 1981; Wolda 1981; Hubálek 1982; clustering method: Milligan & Cooper 1987; Belbin & McDonald 1993; Dale 1995; Lötter et al. 2013; cluster number: Milligan 1996; Aho et al. 2008; Tichý et al. 2010). It is fair to say, however, that vegetation classifications are not equally influenced by the above-mentioned factors, and that differences are always case-dependent. In this regard, the evaluation of the relative importance of decisions influencing the classifications may be extremely helpful. Ecological interpretation of results is greatly enhanced, for example, if we learn that switching from abundances to presence-absence data is more critical than either changing the plot size or selecting among various clustering algorithms. In order to draw such conclusions, we need comparative studies that allow quantifying the amount of variance in the results attributable to a particular factor changed. One such approach was suggested earlier by Podani (1989) in which classification results of the same objects, each obtained by a given combination of choices related to sampling and data analysis, were mapped into an ordination. Then, each ordination axis was identified by a given factor and the order of importance of these factors was determined based on the percentages of variance explained by the associated ordination dimensions. However, this method has limited applicability, because there is no guarantee that axes can be unambiguously identified with any of the factors modified. Furthermore, that approach required the use of all possible combinations of factors, which is a strong methodological limitation. A more general procedure is necessary which is able to partition total variation in the results into components which have one to one correspondence with the modified factors.

In this paper, we use an actual data set from dolomite grasslands and randomly simulated data to partition variation in the results attributable to plot size, data transformation, resemblance coefficient, hierarchical clustering strategy and, finally, to the cluster level (i.e. the number of clusters) obtained from the resulting dendrograms. The method involves random parametrization of these factors, followed by variation partitioning by distance-based redundancy analysis of classifications. Our expectation was that methodological decisions are more influential on classifications of random data than grassland data assuming that biological pattern involves some robustness thereby diminishing the effect of the changes in methods upon results. However, we had no *a priori* expectation about the order of importance of methodological decisions.

Materials and Methods

Data sets

Actual community data

This data set comes from an extensive study of rock grasslands on the dolomite bedrock of Sas Hill, lying within the city limits of Budapest, Hungary (Podani 1998). Eighty sample units were located in the grasslands, representing open rock grassland, closed grassland and slope steppe. Each sample unit consisted of a series of 8 nested quadrats with a common corner, the smallest being 0.5 m × 0.5 m, and the largest 4 m × 4 m, with 0.5 m side increments in between. Percentage cover of vascular plants was recorded within each plot for each size. The total number of species ranged from 79 (smallest quadrats) to 123 (largest quadrats). The eight data matrices can be ordered according to plot size, representing a logical order in the real topographical space, i.e., a spatial series.

Simulated spatial series data

Artificial data matrices were generated for 80 virtual quadrats containing up to 100 species. For each quadrat, a probability of occurrence for each species was generated based on the lognormal distribution (mean = 2, SD = 2 on the ln scale). A predefined number of plant individuals were distributed over the species based on these probabilities. The total number of individuals in the sample unit was used as a proxy for plot size, assuming that these two are proportional to each other. Applied virtual 'plot sizes' were 25, 100, 225, 400, 625, 900, 1225, 1600 individuals. Individuals were assigned to species such that those occurring in the smallest 'quadrat' were retained in all larger quadrats, thus providing a nested species composition similarly to the actual grassland data. In summary, simulated spatial series data were stored in a three-dimensional matrix with 80 locations, 100 species and 8 plot sizes.

Methodological decisions

The basic idea is that both actual and randomized data series serve as input for resampling, in order to generate 200 new matrices for the 80 quadrats. In each of these matrices, quadrats have various sizes determined as described below, and each matrix is subjected to classification based on a random combination of data transformation, resemblance coefficient, hierarchical clustering algorithm and number of clusters to be derived from the resulting dendrogram. It means that 200 classifications are obtained for the actual and for the random data as well. Then, in each case the 200 classifications are compared in every possible pair to yield a distance matrix which serves as the input for distance-based RDA (Legendre & Anderson 1999). In this, constraining variables were those reflecting our decisions on plot size, data transformation etc. The resulting RDA models were subjected to variation partitioning to determine the relative importance of plot size, data transformation, resemblance coefficient, hierarchical clustering algorithm and number of clusters upon the classifications.

Resampling and the matter of plot sizes

The size of each quadrat in each of the sample data matrices was chosen randomly according to the following design. An 8-point scale corresponding to the sampled plot sizes was used for random number generation. First, *M*, a mid-point of the interval from which the plot sizes would be selected was drawn. Then, it is supplied with a half-range value, *d,* in order to control the spread of the plot sizes within the sample. *d* could take values from 1 to 4 randomly. The actual range from which the plot sizes are selected for each location is the interval [min(*M*-*d*, 1); max(*M*+*d*, 8)], 1 referring to the first (smallest) and 8 to the eighth (largest) plot size. For the 'full-range' analysis, *M* could take values on the range [1; 8], while it was limited to [1; 4] for the 'narrow-range' scenarios. The narrow-range design simulates the situation when only a limited range of plot sizes is useful only for classification. In the modelling experiments, the mean and the standard deviation of quadrat sizes are used as explanatory variables.

Data transformation and resemblance measures

After obtaining a data matrix comprising 80 plots of different sizes, abundance values were transformed by Clymo's function (van der Maarel 1979, Podani 2000) given by

$$x'_{ij} = (1 - e^{-cx_{ij}})/(1 - e^{-c})$$

in which x_{ij} is the relative percentage cover value for species *i* in quadrat *j* ranging from 0 to 1, and *c* is a parameter falling in the range [-∞, ∞] such that *c*=0 is not allowed. This procedure allows for weighting abundances differently by adjusting the *c* parameter. In cases with high positive *c*, transformed data approximate the presence/absence situation, thus giving more weight to less abundant species. Large negative values of *c* lead to overweighting the dominant species. If *c* is very close to 0, the relative abundance differences of species remain practically unaffected. However, in real situations data transformation is rather used for downweighting dominant species, therefore, we made separate 'full-range' analyses and 'narrow-range' analyses by changing the value of *c* within [-16; +16] or (0; +16]. Note that *c* must not equal 0.

From the transformed data, dissimilarity matrices were calculated. The resemblance measure was randomly chosen from four indices commonly applied in community ecology: Euclidean, Manhattan, Bray-Curtis and Marczewski-Steinhaus indices (Podani 2000), all of them selected with equal frequency, i.e. 50 times out of 200 trials. The Bray-Curtis and Marczewski-Steinhaus indices are the abundance versions of the dissimilarity forms of the Sørensen and Jaccard coefficients for presence-absence data, respectively. All but one measures, the exception being the Bray-Curtis index, satisfy the metric axioms.

Classification algorithm

A hierarchical classification was obtained from the dissimilarity matrix by agglomerative clustering. The fusion algorithm was the beta-flexible method because it allows for reproducing classifications of different grouping mechanisms by adjusting its β parameter within the interval [-1; 1] (Lance & Williams 1967; see also Podani 2000). Values of β close to 1 tend to emphasise a chained group structure (similarly to the single link or nearest neighbour method), while negative β values lead to increased grouping tendency (as observed for complete link or farthest neighbour algorithms). In each trial, the value of β was chosen randomly from -1 to 1 ('full range'). However, in practice 'group-forming' methods are preferred, therefore β values were drawn from [-1; 0] for 'narrow-range' analyses. The cluster level (simulating the case of an ’optimal non-hierarchical classification’) was randomly chosen between 2 and 8. The hierarchical classification was ’cut’ at this level and hereafter only this non-hierarchical clustering was used.

Data analysis

The 200 trials of the randomization resulted in 200 classifications of the same spatial series. From each classification, an incidence matrix, ***C***, was calculated in which c_{ij} is 1 if objects *i* and *j* in the same cluster and 0 otherwise. Euclidean distances were calculated between all pairs of incidence matrices. This method is also called ‘PAIRBONDS’ (Arabie & Boorman 1973; Podani 2000). These distances were then summarized into another distance matrix based on which principal coordinates analysis was computed. In the resulting ordination all points correspond to a non-hierarchical classification. Then, the following explanatory variables were fitted to the ordination diagram: mean and standard deviation of plot sizes, resemblance measure, *c* of Clymo’s transformation, *β* of the flexible classification and the number of clusters. Trend surfaces of numerical variables were fitted onto the scatter plots by generalized additive models, while average scores were calculated for the resemblance measures. The relative importance of the explanatory variables was tested by constrained ordination: the Euclidean distances obtained earlier were subjected to a distance-based redundancy analysis (db-RDA, Legendre & Anderson 1999). When mean plot size, Clymo’s *c* and *β* were scaled on full-range, their squared terms were also included in the model as explanatory variables. Low (<2) values of generalized variance inflation factors (GVIF, Fox & Monette 1992) indicated negligible collinearity between model terms. The models were evaluated by comparing *F* ratios of the model terms vs. residual variation, by calculating adjusted *R*-squared measures and by visual observation of fitted explanatory variables on the PCoA diagrams. During the evaluation of db-RDA models, predictors with *F* ratios with a type I error rate of $P<0.01$ were considered significant.

Our variation partitioning approach relies on the basic assumption that db-RDA models can properly explain the variation among classifications attributed to the different methodological decisions. In order to validate our modelling technique, we applied a simulation test. The above described simulation analysis with narrow-range variables, starting from the sample selection and ending at calculation of explained variances was repeated many times. However, instead of the fully random parametrization of the six variables representing methodological decisions, some of them were ‘fixed’, i.e. they were given zero variance. For example, if plot size was fixed, only plots of the same size were selected from each location in all of the 200 classifications that were entered in each db-RDA. Of course, in such cases, the fixed variable was not included as an explanatory variable of the db-RDA, since it had no variation. The number of fixed variables was increased from zero to five in six steps and for each number of fixed variables, 100 trials were performed. Then, average explained variation, unexplained and total variation were plotted against the number of fixed variables. We expected that explained variation would decrease with increases in the number of fixed variables because reducing the possible outcomes of methodological decisions should also reduce the variation among classification they account for. If unexplained variation also decreased with the increased number of fixed variables, we could conclude that variation caused by methodological decisions was not properly explained by the db-RDA model. On the contrary, approximately constant unexplained variation obtained for different numbers of fixed variables would mean that independently from the methodological decisions and the explanatory variables, there is a certain amount of inherent variation in the compositional data.

All analyses were performed by the R software environment (version 2.14.1, R Development Core Team, www.r-project.org) using the packages vegan (Oksanen et al., http://CRAN.R-project.org/package=vegan, vegdist(), cmdscale(), capscale(), vif.cca(), ordistep(), anova.cca() and RsquareAdj() functions) and cluster (Maechler et al., http://cran.r-project.org/web/packages/cluster/, agnes() function).

Results

Distance-based RDA models of simulated and grassland data sets explained different proportions of the total variation among classifications. The adjusted R^2 values were higher for the simulated data sets (full-range: 0.466, narrow-range: 0.258) than the grassland data (full-range: 0.260, narrow-range: 0.157). In the model of the simulated data set with full-range variables flexible β (F=121.388), cluster level (F=26.437), mean plot size (F=6.592), Clymo's *c* (F=5.827) and SD of plot sizes (F=3.455) proved to have a significant effect at p<0.01 (Table 1). Mean plot size, Clymo's *c*, flexible β and cluster number showed a good fit on the first two dimensions of the PCoA ordination (P=4.1e-11, P=9.2e-7, P=3.4e-88 and P=1.19e-14, respectively; Fig. 1). Values of flexible β changed gradually along the first PCoA axis with increasing β values in the positive direction, while mean plot size and Clymo's *c* showed a gradient along the second axis. A non-linear pattern was found for cluster number. Centroids of classifications with different resemblance measures fell close to each other.

In the narrow-range analyses on the simulated data set, five predictors had significant effect (Table 2). The flexible β and the cluster level again explained the largest variation (F=36.524 and F=24.538, respectively), followed by mean and SD of plot sizes (F=2.984 and F=2.564) and, finally, Clymo's *c* (F=2.300). The four most important variables fitted relatively well to the first two PCoA axes (P=1.23e-31, P=1.71e-15, P=2.7e-6, P=7.2e-4; Fig. 2). Flexible β

increased along the first dimension, while mean plot size correlated positively with the second axis.

Five predictors had a significant effect on the variation between partitions in the model of the grassland data set with full-range variables (Table 3). Flexible *β* obtained by far the highest *F*-value (*F*=43.651), while the other model terms showed lower and gradually decreasing explanatory power, like cluster level (*F*=9.865), Clymo's *c* (*F*=7.793), Clymo's *c* squared (*F*=3.678) and mean plot size (*F*=2.206). The resemblance measure, the SD of plot sizes and the squared form of the flexible beta showed no significant effect at the pre-set level of α, but were significant at α=0.05. The *β* parameter, Clymo's *c* and cluster number were fitted well onto the ordination diagram (*P*=4e-75, *P*=3.8e-33, *P*=5.1e-17, respectively; Fig. 3). The values of the first correlated positively with Axis 1, while those of Clymo's *c* with Axis 2. The pattern of cluster number on these two dimensions was non-linear again. Different resemblance measures seemed more separated than in the simulations. The fits of the other model terms were weak.

After narrowing the range of explanatory variables, five terms had significant effect (Table 4). Cluster level proved by far the most influential variable (*F*=28.336). Clymo's *c* (*F*=3.847), flexible *β* (*F*=2.841), mean plot size (*F*=2.678) and resemblance measure (*F*=1.391) had lower but still significant effect. Only the two most important variables showed significant fit on the ordination diagram (*P*=1.57e-27, *P*=2.5e-17; Fig. 4). Cluster number decreased along the first axis, while Clymo's *c* showed a gradient along Axis 2.

In the simulation test to examine the validity of our modelling approach, variation explained by db-RDA models decreased monotonically and significantly as more variables were fixed, while unexplained variation showed small changes with no clear trend (Figure 5).

Discussion & Conclusions

At the outset, we put forward the hypothesis that adjusted *R*-squared values would be higher, for simulated data with random structure than for actual grassland data. In the first case, variation among classifications would only be attributed to the differences in the methodological decisions, as superimposed on random variation, while in the second robustness of biological pattern would resist changes in methodology. Our findings confirmed this expectation.

The order of importance of the predictors was not the same in all experiments, while some general trends did appear. Flexible *β*, cluster level, Clymo's *c* and mean plot size were always among the significant model terms, and in many cases they were given the highest rank. Obvious interpretation is that decisions about clustering process, including the chaining algorithm and the number of clusters, influence most strongly the outcome of numerical classification of compositional data. Nevertheless, the other variables were also critical at least in one of the four scenarios.

The decision of how large sample units should be is an often highlighted problem in the ecological literature (Kenkel & Podani 1991; Reed et al. 1993). Mean plot size was among the most influential variables in all trials and the SD of plot size also had a significant effect in the model in the simulations. Simulated data lacked biological pattern contrary to the grassland data, thus plot size can be accountable for a false discovery of non-existing pattern in multivariate data with random structure. During classification of phytosociological data comprising different plot sizes, it is advised to check the distribution of plot sizes among clusters *a posteriori*. Mean plot size had an effect regardless whether ‘full’ or ‘narrow’ range of parameters was used. In the narrow-range analysis of the grassland data, plot sizes varied within a range that is typical or even narrower than usual in phytosociological studies of dry grasslands (2 to 4 m^2; see recommendation e.g. by van der Maarel 2009 or basic statistics of databases by Dengler et al. 2011). Although in this trial mean plot size was just the fourth most important predictor of the model, it was still significant. It implies that the influence of plot size should not be overlooked even within its recommended standard range. This result supports the recommendations by Chytrý & Otýpková (2003) who argued that for a comprehensive investigation of a vegetation type, analyses should be done separately for each plot size. The final definition of vegetation types should be elaborated based on this series of classifications. The difficulties caused by the uneven distribution of relevés in the space or among vegetation types should be handled by acquiring new data or by appropriate resampling methods (Knollová et al. 2005; Lengyel et al. 2011).

Through the four scenarios, data transformation affected classifications of the grassland data set more strongly than the simulated scenarios. This finding is in line with earlier views that data transformation can reveal significantly different but biologically relevant patterns of the same data set (van der Maarel 1979; Podani 1989). Since the effect of data transformation was higher for the grassland data, we conclude that the choice of the optimal abundance scale is crucial for understanding the multiple facets of biological variation in real data sets. Thus, much care should be taken before transforming abundance data.

The resemblance measure showed weaker effect than plot size and data transformation, however, it was still significant in the narrow-range analysis of the grassland data set, and it was near the pre-set significance level in the full-range trial of the same data. The matter of choosing among resemblance measures is more deeply investigated compared to other methodological decisions, and many papers highlight the differences of the available indices (Campbell 1978; Legendre & De Cáceres 2013). Without questioning that different resemblance measures can be appropriate for specific purposes, and the choice between them had to be taken carefully, our results suggest that the importance of this decision may be over-emphasized in comparison with other decisions. Thus, we consider the importance of the resemblance measure as a good reference to assess the significance of the other explanatory variables. Nevertheless, it must be noted that we employed only four indices that are very popular among vegetation ecologists.

The β parameter of the flexible clustering was the most significant predictor in three cases. Its value with full range was more influential than with narrow range, which clearly indicates that decision on the classification method is most critical between chain-forming (β>0) and group-forming (β<0) methods, while differences within group-forming algorithms are not that substantial. This difference is the most striking with the grassland data, for which its effect is dropped from the 1st to the 3rd most important model term if compared to the full-range scenario. In recent works of numerical syntaxonomy (for example, Havlová 2006; Knollová & Chytrý 2004), of the distance-based methods chain-forming algorithms have received much fewer applications than group-forming ones which include the flexible method with negative β values applied here. Much more widespread is Ward’s agglomerative method (more precisely, incremental sum of squares) which also has a preference for spherical group shapes. The good performance of flexible method with β=-0.25 and the Ward’s method was also indicated by Lötter et al. (2013) but one is warned that groups show up apparently clearly in the resulting dendrograms even if in fact they do not exist in the data (Podani 2000). Another very popular hierarchical method is TWINSPAN (Hill 1979; Rolecek et al. 2009), however, its weaknesses are pointed out in several papers (Belbin & McDonald 1993; Dufrene & Legendre 1997; Lötter et al. 2013). The significant effect of clustering algorithm implies that during the comparison and revision of existing vegetation classifications the applied clustering methods should be taken into account carefully. Large differences between classifications of the same vegetation units of a certain area can be attributed to the different methods used, and therefore comparison of classification prepared by different algorithms may even be meaningless.

Cluster level was the second most significant model term in three of the four scenarios and the most important one for the grassland data set with narrow-range variables. In classification studies, the number of clusters is usually determined by an expert-based, i.e. a rather subjective method (but see Botta-Dukát et al. 2005 or Illyés et al. 2007). Cluster validation, including the choice of the optimal ‘cut level’, is the most data-specific decision among those we studied here, therefore the only general recommendation that we could stress is to investigate and to use quantitative measures for this purpose instead of subjective assessment (for example, Milligan 1996; Aho et al. 2009; Tichý et al. 2010; Tichý et al. 2011). The validation tools are so numerous that their comparative study focusing on specific requirement for numerical syntaxonomy would be timely.

In the modelling approach applied here, two crucial assumptions were made in order to quantify the effect of methodological decisions on the classifications. The first assumption was that the PAIRBONDS method expresses appropriately the dissimilarities between pairs of classifications. This index gives the square-root of the number of pairs of plots in the same group in one classification but separated in the other classification. This is a Euclidean measure of distance and its suitability to our variation partitioning approach is also supported by the R-squared values (ca. 18-48%). In ecological modeling studies, in general, lower explanatory power is often considered meaningful (Møller & Jennioins 2002). It is to be noted that PAIRBONDS is relatively sensitive to cluster structure, i.e. the number and the sizes of groups. With this measure, two classifications with different numbers of clusters can never be at zero distance from each other, therefore any differences in cluster number are immediately mirrored by the distance matrix. In contrast, certain other dissimilarity indices (e.g. Cramér’s *V*, Cramér 1946; Goodman-Kruskal’s *Λ*, Goodman & Kruskal 1954) control for the numbers of clusters, thus giving standardized measures of similarity between non-hierarchical classifications. However, we consider these types of indices misleading in our situation because in practice two classifications of the same data set are rarely interpreted identically if the numbers of clusters differ. Our preliminary analyses showed that the use of Cramér’s *V* or Goodman-Kruskal’s *Λ* would attribute lower effect to flexible *β* and cluster level, nevertheless, it would result in much weaker overall model performance as well.

The second assumption was that the db-RDA model captured relevant information on variation among classifications. The first part of db-RDA was PCoA known to preserve the original distance structure of the input matrix. Then, the PCoA axes, as transformed variables of between-classification distances, were related to the explanatory variables (i.e. the methodological decisions) by usual RDA method. At this step, even patterns that are non-linear functions of the explanatory variables are decomposed into separate components for which the explanatory variables can be linearly related. To account for eventual non-linear relationships that cannot be revealed by this procedure, we included squared terms into the models and the distribution of the explanatory variables over the first two PCoA axes were also mapped by a flexible fitting method (GAM). These trend surfaces revealed that cluster number can show a non-linear pattern along the first two axes. However, this pattern can likely to be accounted for by db-RDA because cluster number came out as a highly significant predictor in all cases. In our analysis to validate the appropriateness of our modelling approach, we found that the amount of unexplained variation of our models is not related to the number of fixed and randomized variables, that is, it is independent from the methodological decisions. This suggests that the variation caused by the random parametrization of the classifications is satisfactorily explained by the db-RDA models. Therefore, we do not suspect a significant amount of unexplained variation due to non-linear effects or interactions among methodological decisions. The unexplained variation may have several different origins. The most trivial reason is that the data set has a certain degree of robustness which explains low sensitivity to methodological changes. Robustness is obviously higher for the grassland data set that contains biologically interpretable patterns. Nevertheless, it is also present in the simulated data set since randomized data do not lack variation completely but this variation is comparable to what is expected by chance. Another possible source is the individual 'fate' of plots in the analysis. Two classifications can be identically parameterized in terms of the selected plot sizes but the sample to be analysed can still differ because it is not fixed which plot size should be selected from a certain location.

The few most important variables identified by the variation partitioning approach using db-RDA in most cases showed good fit to the first two axes of the PCoA ordination. However, their pattern was not always linear, therefore they could not be detected by simply checking the correlation between ordination axes and the tested variables.

Acknowledgements

We thank Miquel De Cáceres, Milan Chytrý, Zoltán Botta-Dukát, Otto Wildi, Philip Dixon and two anonymous referees for their helpful comments. The work of A.L. was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP-4.2.4.A/ 2-11/1-2012-0001 'National Excellence Program'. This work was partially funded by the Hungarian Scientific Research Grant (OTKA K106177).

References

Aho, K., Roberts, D. W. & Weaver, T. 2008. Using geometric and non-geometric internal evaluators to compare eight vegetation classification methods. *Journal of Vegetation Science* 19: 549–562.

Arabie, P. & Boorman, S. A. 1973. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology* 10: 148–203.

Belbin, L. & McDonald, C. 1993. Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* 4: 341–348.

Botta-Dukát, Z., Chytrý, M., Hájková, P. & Havlová, M. 2005. Vegetation of lowland wet meadows along a climatic continentality gradient in Central Europe. *Preslia* 77: 89–111.

Campbell, B.M. 1978. Similarity coefficients for classifying relevés. *Vegetatio* 37: 101–109.

Chytrý, M. & Otýpková, Z. 2003. Plot sizes used for phytosociological sampling of European vegetation. *Journal of Vegetation Science* 14: 563–570.

Cramér, H. 1946. *The elements of probability theory and some of its applications.* Wiley, New York, US.

Dale, M.B. 1995. Evaluating classification strategies. *Journal of Vegetation Science* 6: 437–440.

Dengler, J., Jansen, F., Glockler, F., Peet, R.K., De Cáceres, M., Chytrý, M., Ewald, J. Oldeland, J., Lopez-Gonzalez, G., Finckh, M., Mucina, L., Rodwell, J.S., Schaminée, J.H.J. & Spencer, N. 2011. The global index of vegetation-plot databases 1 (GIVD): a new resource for vegetation science. *Journal of Vegetation Science* 22: 582–597.

Dengler, J., Löbel, S. & Dolnik, C. 2009. Species constancy depends on plot size – a problem for vegetation classification and how it can be solved. *Journal of Vegetation Science* 20: 754–766.

Dufrene, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs* 67: 345–366.

Fox, J. & Monette, G. 1992. Generalized collinearity diagnostics. *Journal of the American Statistical Association* 87: 178–183.

Goodman, L. & Kruskal, W. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association* 49: 732–764.

Green, R.H. 1980. Multivariate approaches in ecology: The assessment of ecological similarity. *Annual Review of Ecology and Systematics* 11: 1–14.

Hajdu, L.J. 1981. Graphical comparison of resemblance measures in phytosociology. *Vegetatio* 48: 47–59.

Havlová, M. 2006. Syntaxonomical revision of the Molinion meadows in the Czech Republic. *Preslia* 78: 87–101.

Hill, M.O. 1979. *TWINSPAN. A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes.* Cornell University, Ithaca, NY, US.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews* 57: 669–689.

Illyés, E., Chytrý, M., Botta-Dukát, Z., Jandt, U., Škodová, I., Janišová, M., Willner, W. & Hájek, O. 2007. Semi-dry grasslands along a climatic gradient across Central Europe: Vegetation classification with validation. *Journal of Vegetation Science* 18: 835–846.

Jensen, S. 1978. Influences of transformation of cover values on classification and ordination of lake vegetation. *Vegetatio* 37: 19–31.

Kenkel, N., P. Juhász-Nagy & J. Podani. 1989. On sampling procedures in population and community ecology. *Vegetatio* 83:195-207.

Kenkel, N. C. & Podani, J. 1991. Plot size and estimation efficiency in plant community studies. *Journal of Vegetation Science* 2: 539–544.

Knollová, I. & Chytrý, M. 2004. Oak-hornbeam forests of the Czech Republic: geographical and ecological approaches to vegetation classification. *Preslia* 76: 291–311.

Knollová, I., Chytrý, M., Tichý, L. & Hajek, O. 2005. Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies. *Journal of Vegetation Science* 16: 479–486.

Lance, G.N & Williams, W.T. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems*. Computer Journal* 9: 373–380.

Legendre, P, & De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters* 16: 951–963.

Legendre, P. & Anderson, M.J. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69: 1–24.

Lengyel, A., Chytrý, M. & Tichý, L. 2011. Heterogeneity-constrained random resampling of phytosociological databases. *Journal of Vegetation Science* 22: 175–183.

Lötter, M.C., Mucina, L. & Witkowski, E.T.F. 2013. The classification conundrum: species fidelity as leading criterion in search of a rigorous method to classify a complex forest data set. *Community Ecology* 14: 121–132.

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

Maarel, E. van der. 2005. *Vegetation ecology*. Blackwell Science, Oxford, UK.

Milligan, G.W. 1996. Clustering validation: Results and implications for applied analyses. In: Arabie, P., Hubert, L.J. & de Soete, G. (eds.) *Clustering and classification*, pp 341–375. World Scientific, River Edge,US.

Milligan, G.W. & Cooper, M.C. 1987. Methodology review: Clustering methods. *Applied Psychological Measurement* 11: 329–354.

Møller, A. & Jennioins, M.D. 2002. How much variance can be explained by ecologists and evolutionary biologists? *Oecologia* 132: 492–500.

Mucina, L. 1997. Classification of vegetation: Past, present, and future. *Journal of Vegetation Science* 8: 751–760.

Otýpková, Z. & Chytrý, M. 2006. Effects of plot size on the ordination of vegetation samples. *Journal of Vegetation Science* 17: 465–472.

Peet, R.K. & Roberts. D.W. 2013. Classification of natural and semi-natural vegetation In: van der Maarel, E. & Franklin,J.*Vegetation ecology*, Second edition. pp 28–70. Wiley, NY, US.

Podani, J. 1984. Analysis of mapped and simulated vegetation patterns by means of computerized sampling techniques. *Acta Botanica Hungarica* 30: 419–441.

Podani, J. 1989. Comparison of ordinations and classifications of vegetation data. *Vegetatio* 83: 111–128.

Podani, J. 1998. Numerikus cönológiai vizsgálatok a Sas-hegy (Budai hg.) dolomitsziklagyepjeiben (A complex numerical analysis of dolomite rock grasslands of the Sas-hegy Nature Reserve, Budapest, Hungary. In Hungarian with English summary) In: Csontos P. (ed.), *Sziklagyepek szünbotanikai kutatása*. Scientia, Budapest. pp. 213–229.

Podani, J. 2000. *Introduction to the exploration of multivariate biological data*. Backhuys Publishers, Leiden, NL

Reed, R.A., Peet, R.K., Palmer, M.W. & White, P.S. 1993. Scale dependence of vegetation-environment correlations: A case study of a North Carolina Piedmont woodland. *Journal of Vegetation Science* 4: 329–340.

Roleček, J., Tichý, L., Zelený, D. & Chytrý, M. 2009. Modified TWINSPAN classification in which the hierarchy respects cluster heterogeneity. *Journal of Vegetation Science* 20: 596–602.

Tichý, L., Chytrý, M. & Smarda, P. 2011. Evaluating the stability of the classification of community data. *Ecography* 34: 807–813.

Tichý, L., Chytrý, M., Hájek, M., Talbot, S.S. & Botta-Dukát, Z. 2010. OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science* 21: 287–299.

Whittaker, R.H. 1973. *Ordination and classification of vegetation*. Junk, The Hague, NL.

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

Wolda, H. 1981. Similarity indices, sample size and diversity. *Oecologia* 50: 296–302.

Table 1. Predictors of the db-RDA model for the simulated data set on full ranges of the variables. P-values are based on 199 permutations.

	Df	Var%	F	P
flexible β	1	32.569	121.388	0.005
cluster level	1	7.093	26.437	0.005
mean plot size	1	1.769	6.592	0.005
Clymo's c	1	1.563	5.827	0.005
SD of plot sizes	1	0.927	3.455	0.005
resemblance measure	3	0.966	1.200	0.093
Clymo's c squared	1	0.324	1.208	0.150
flexible β squared	1	0.274	1.021	0.360
Residual	189	50.709	-	-
Total	199	100.000	-	-

R^2=0.493, R^2_{adj}=0.466

Table 2. Predictors of the db-RDA model for the simulated data set on narrow ranges of the variables. P-values are based on 199 permutations.

	Df	Var%	F	P
flexible β	1	13.614	36.524	0.005
cluster level	1	9.146	24.538	0.005
mean plot size	1	1.112	2.984	0.005
SD of plot sizes	1	0.956	2.564	0.005
Clymo's c	1	0.857	2.300	0.005
resemblance measure	3	1.195	1.068	0.265
Residual	191	71.193	-	-
Total	199	100.000	-	-

R^2=0.288, R^2_{adj}=0.258

Table 3. Predictors of the db-RDA model for the grassland data set on full ranges of the variables. P-values are based on 199 permutations.

	Df	Var%	F	P
flexible β	1	16.232	43.651	0.005
cluster level	1	3.668	9.865	0.005
Clymo's c	1	2.898	7.793	0.005
Clymo's c squared	1	1.368	3.678	0.005
mean plot size	1	0.820	2.206	0.005
resemblance measure	3	1.425	1.278	0.015
SD of plot sizes	1	0.489	1.314	0.036
flexible β squared	1	0.461	1.241	0.055
Residual	189	70.281	-	-
Total	199	100.000	-	-

R^2=0.297, R^2_{adj}=0.260

Table 4. Predictors of the db-RDA model for the grassland data set on narrow ranges of the variables. P-values are based on 199 permutations.

	Df	Var%	F	P
cluster level	1	12.011	28.336	0.005
Clymo's c	1	1.630	3.847	0.005
flexible β	1	1.204	2.841	0.005
mean plot size	1	1.135	2.678	0.005
resemblance measure	3	1.769	1.391	0.005
SD of plot sizes	1	0.443	1.045	0.300
Residual	191	80.958	-	-
Total	199	100.000	-	-

R^2=0.190, R^2_{adj}=0.157

Figure 1. Principal coordinates analysis of classifications of the simulated data sets with the full ranges of predictor variables. Continuous variables are fitted as trend surfaces *a posteriori* by GAM, factor variables are fitted by averaging of object scores on the two ordination axes. Axes 1 and 2 explain 62.5% and 2.6% of the total variation, respectively.

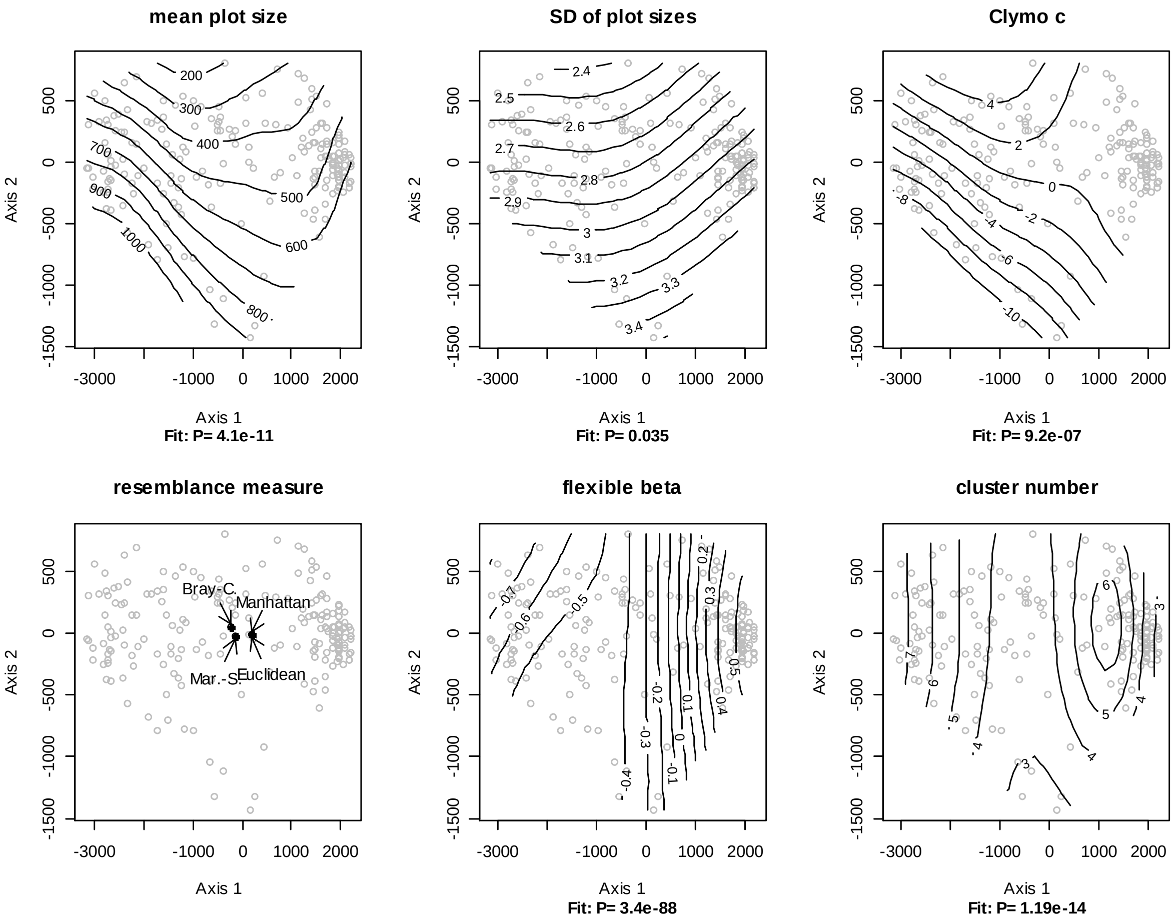


Figure 2. Principal coordinates analysis of classifications of the simulated data sets with the narrow ranges of variables. Continuous variables are fitted as trend surfaces *a posteriori* by GAM, factor variables are fitted by averaging of object scores on the two ordination axes. Axes 1 and 2 explain 38.6% and 2.7% of the total variation, respectively.

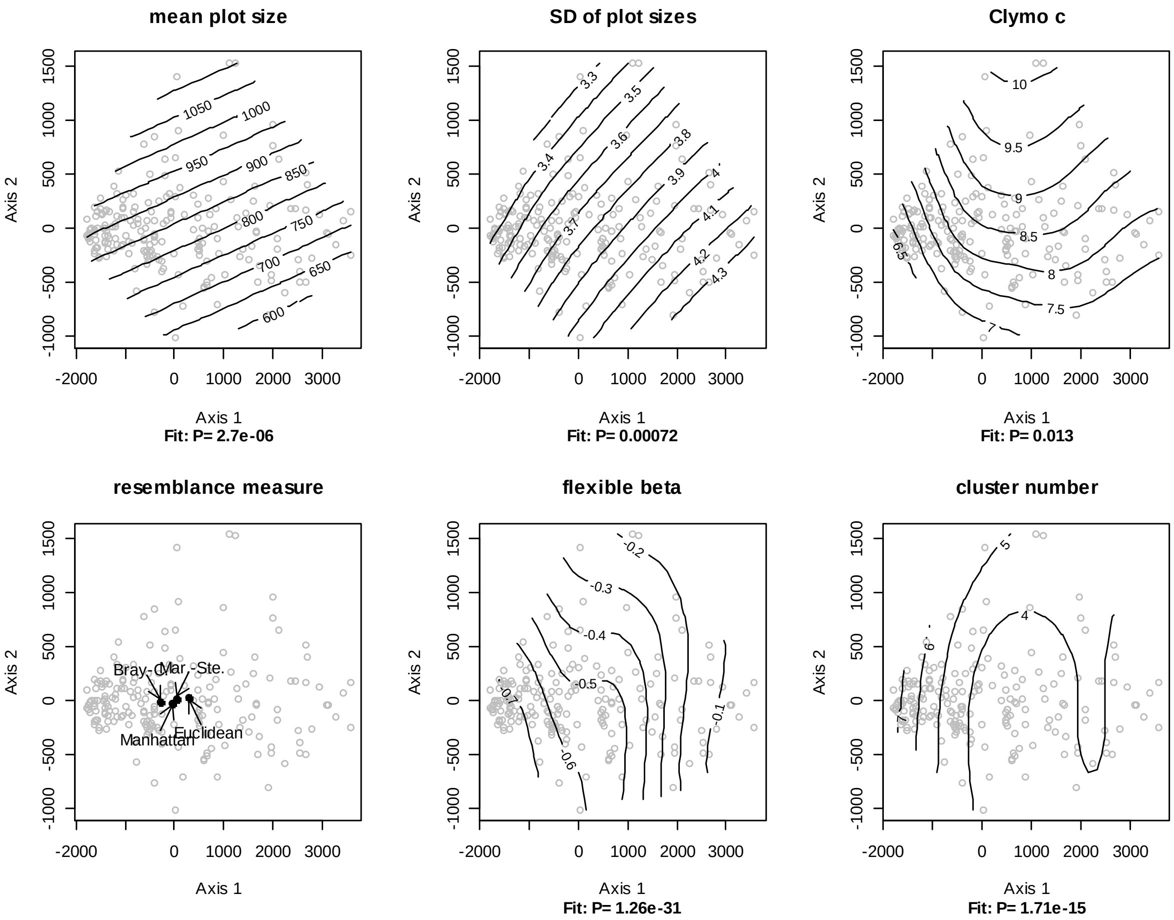


Figure 3. Principal coordinates analysis of classifications of the grassland data sets with the full ranges of variables. Continuous variables are fitted as trend surfaces *a posteriori* by GAM, factor variables are fitted by averaging of object scores on the two ordination axes. Axes 1 and 2 explain 29.7% and 2.8% of the total variation, respectively.

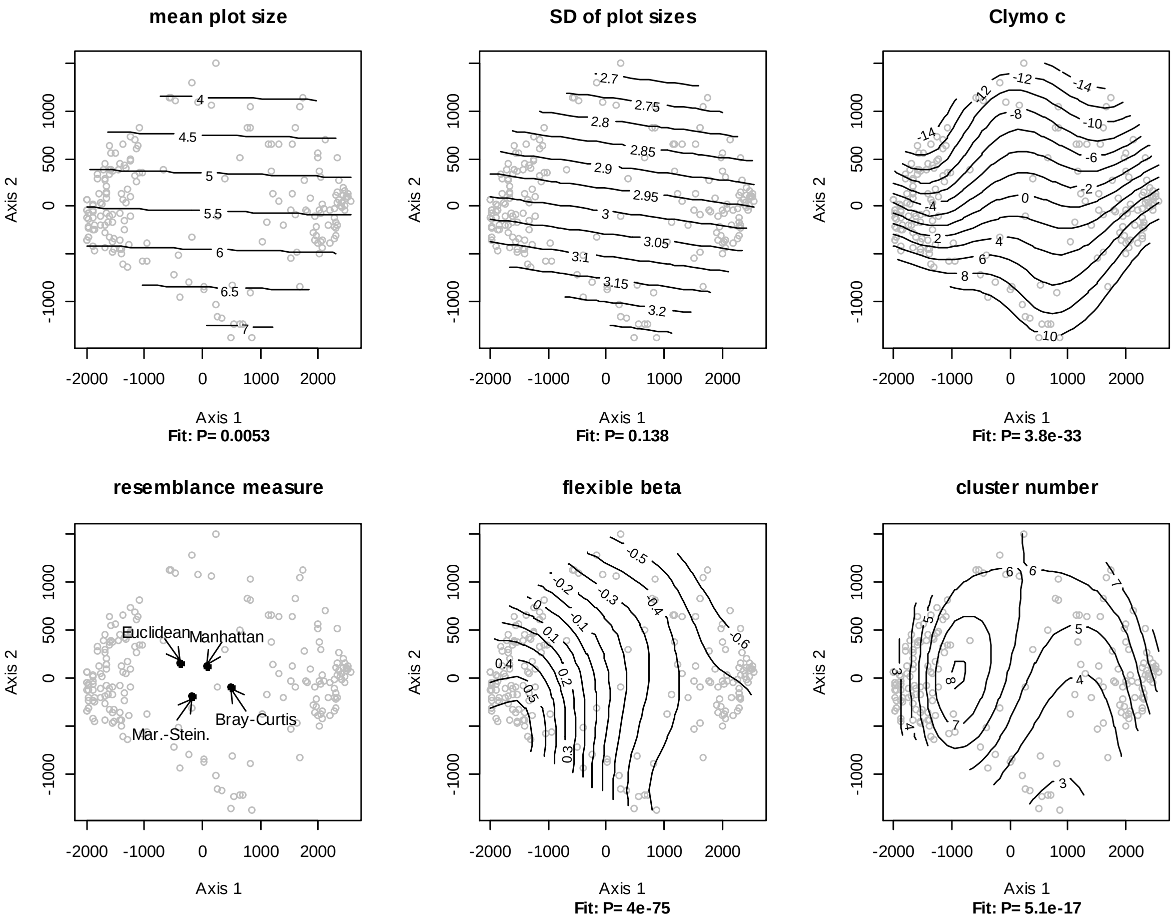
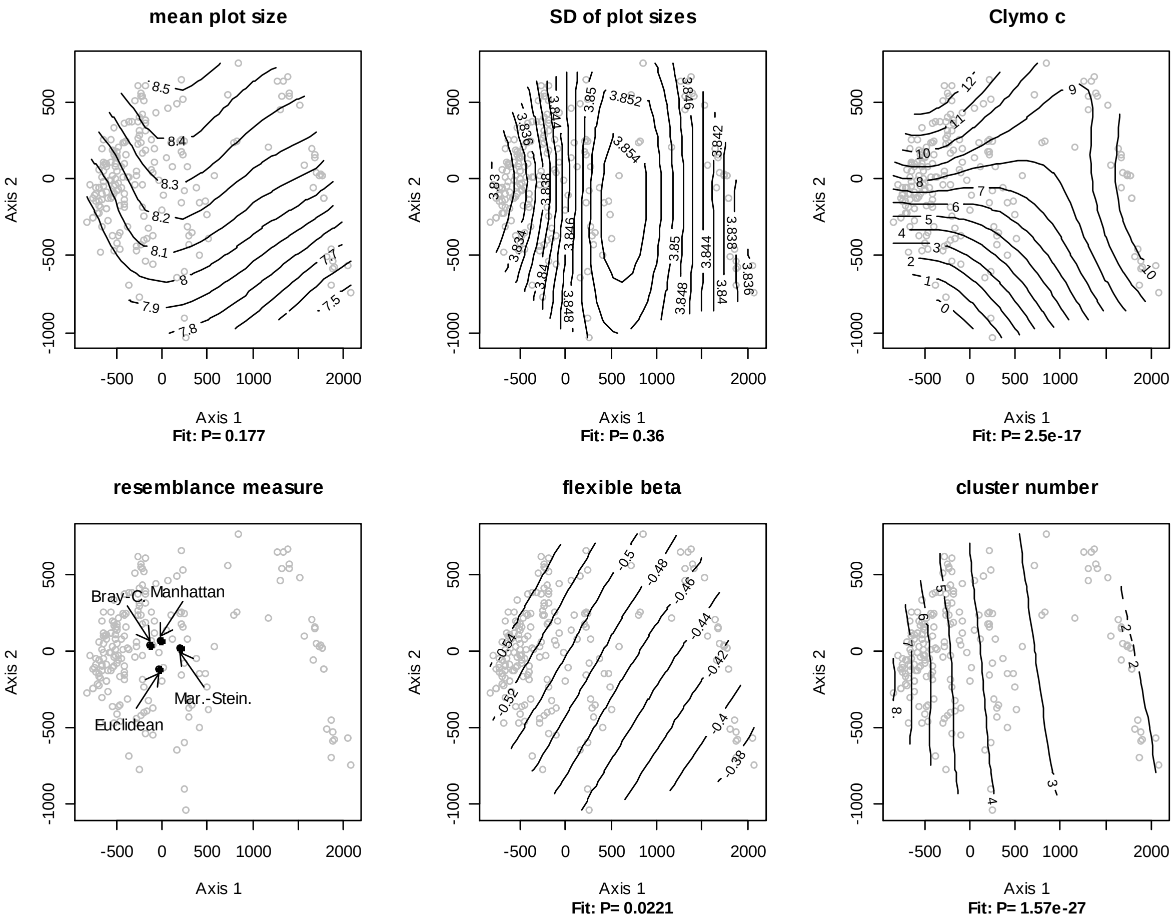
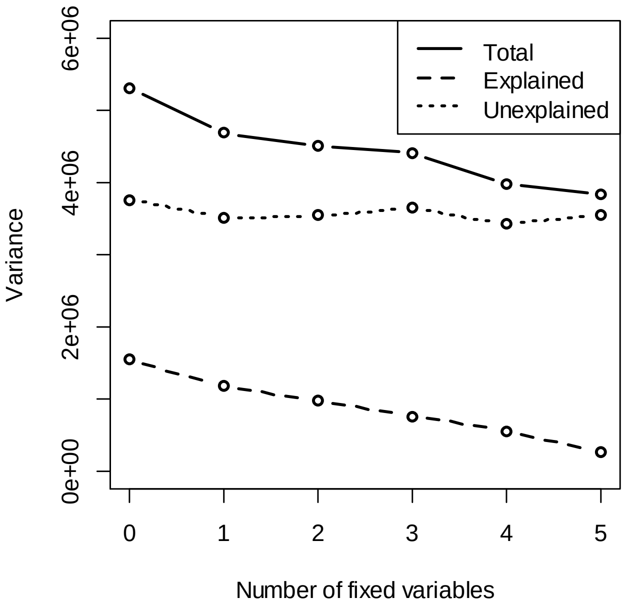


Figure 4. Principal coordinates analysis of classifications of the grassland data sets with the narrow ranges of variables. Continuous variables are fitted as trend surfaces *a posteriori* by GAM, factor variables are fitted by averaging of object scores on the two ordination axes. Axes 1 and 2 explain 18.5% and 3.4% of the total variation, respectively.



587 **Figure 5.** Relationship between average explained, unexplained and total variation and the
588 number of fixed variables out of the six variables in the simulation with narrow-range settings.



589