

1 **Assessing the relative importance of methodological decisions in classifications of**  
2 **vegetation data**

3 Attila Lengyel<sup>1\*</sup> and János Podani<sup>2,3</sup>

4 <sup>1</sup>MTA Centre for Ecological Research, Institute of Ecology and Botany, Alkotmány u. 2-4., H-  
5 2163 Vácrátót, Hungary

6 <sup>2</sup>Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd  
7 University, Pázmány P. s. 1/C, H-1117 Budapest, Hungary

8 <sup>3</sup>Ecology Research Group of the Hungarian Academy of Sciences, Pázmány P. s. 1/C, H-  
9 1117 Budapest, Hungary. E-mail: podani@ludens.elte.hu

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

11

12 This is the pre-peer reviewed version of the following article: **Lengyel, A., Podani, J. (2015),**  
13 **Assessing the relative importance of methodological decisions in classifications of**  
14 **vegetation data. Journal of Vegetation Science 26: 804–815, doi: 10.1111/jvs.12268,**  
15 which has been published in final form at  
16 <http://onlinelibrary.wiley.com/doi/10.1111/jvs.12268/full>.

17

18 **Abstract**

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

25 **Location:** Rock grassland in Hungary.

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

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

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

46

#### 47 **Keywords**

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

#### 50 **Abbreviations**

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

#### 53 **Running head**

54 Relative importance of methodological decisions

55

#### 56 **Introduction**

57 Classification of vegetation has long been the primary research objective in phytosociology  
58 and still represents an integral part of vegetation science in general (Whittaker 1973; Mucina  
59 1997; Peet & Roberts 2013). It provides a firm reference basis for syntaxonomy, similarly to  
60 the classification of living organisms in biological systematics (or taxonomy). Scientific

61 communication would be impossible without a common basis for recognizing, separating,  
62 describing, naming and mapping plant communities, that is, without classification of  
63 vegetation units. In addition to the fact that syntaxonomy is conditioned upon taxonomy,  
64 there is a fundamental difference between these two fields of biology. Whereas the basic  
65 observational units of classification in conventional systematics are natural entities  
66 (individuals), community classification requires the use of – more or less – arbitrarily  
67 delineated tracts from the vegetation continuum. Therefore, one is faced with a multitude of  
68 methodological choices that have to be made in the real *topographical* space (Podani 1984),  
69 that is, in the field. These include the appropriate selection of sampling criteria, that is, plot  
70 size, shape, number and arrangement (Kenkel et al. 1989). Syntaxonomy and taxonomy  
71 share only the problem of *conceptual* and *methodological* decisions which concern the  
72 variables to describe the study objects, measurement scale, resemblance coefficient and  
73 clustering algorithm to be used during data processing (Podani 1989). Tradition, fashion,  
74 practicability, comparability with others' results, availability of software and similar, more or  
75 less subjective considerations may guide the user in this methodological jungle.  
76 Nevertheless, since no absolute and universally valid criteria are available, all decisions  
77 remain unavoidably arbitrary in every step of the study. An important philosophical  
78 conclusion is that any attempt to find and define unique classifications in vegetation science  
79 will be illusory – which does not mean that the effect of sampling and analysis upon the  
80 results should be disregarded in community analysis.

81 The importance of such methodological choices in multivariate analysis has long been  
82 recognized by several authors (plot size: Kenkel & Podani 1991; Otýpková & Chytrý 2006;  
83 Dengler et al. 2009; measurement scale: Jensen 1978; van der Maarel 1979; Wilson 2012;  
84 resemblance measure: Green 1980; Hajdu 1981; Wolda 1981; Hubálek 1982; clustering  
85 method: Milligan & Cooper 1987; Belbin & McDonald 1993; Dale 1995; Lötter et al. 2013;  
86 cluster number: Milligan 1996; Aho et al. 2008; Tichý et al. 2010). It is fair to say, however,  
87 that vegetation classifications are not equally influenced by the above-mentioned factors, and  
88 that differences are always case-dependent. In this regard, the evaluation of the relative  
89 importance of decisions influencing the classifications may be extremely helpful. Ecological  
90 interpretation of results is greatly enhanced, for example, if we learn that switching from  
91 abundances to presence-absence data is more critical than either changing the plot size or  
92 selecting among various clustering algorithms. In order to draw such conclusions, we need  
93 comparative studies that allow quantifying the amount of variance in the results attributable  
94 to a particular factor changed. One such approach was suggested earlier by Podani (1989) in  
95 which classification results of the same objects, each obtained by a given combination of  
96 choices related to sampling and data analysis, were mapped into an ordination. Then, each

97 ordination axis was identified by a given factor and the order of importance of these factors  
98 was determined based on the percentages of variance explained by the associated  
99 ordination dimensions. However, this method has limited applicability, because there is no  
100 guarantee that axes can be unambiguously identified with any of the factors modified.  
101 Furthermore, that approach required the use of all possible combinations of factors, which is  
102 a strong methodological limitation. A more general procedure is necessary which is able to  
103 partition total variation in the results into components which have one to one correspondence  
104 with the modified factors.

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

115

## 116 **Materials and Methods**

### 117 ***Data sets***

#### 118 *Actual community data*

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

#### 128 *Simulated spatial series data*

129 Artificial data matrices were generated for 80 virtual quadrats containing up to 100 species.  
130 For each quadrat, a probability of occurrence for each species was generated based on the  
131 lognormal distribution (mean = 2, SD = 2 on the ln scale). A predefined number of plant  
132 individuals were distributed over the species based on these probabilities. The total number  
133 of individuals in the sample unit was used as a proxy for plot size, assuming that these two  
134 are proportional to each other. Applied virtual 'plot sizes' were 25, 100, 225, 400, 625, 900,  
135 1225, 1600 individuals. Individuals were assigned to species such that those occurring in the  
136 smallest 'quadrat' were retained in all larger quadrats, thus providing a nested species  
137 composition similarly to the actual grassland data. In summary, simulated spatial series data  
138 were stored in a three-dimensional matrix with 80 locations, 100 species and 8 plot sizes.

### 139 *Methodological decisions*

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

### 153 *Resampling and the matter of plot sizes*

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

164 modelling experiments, the mean and the standard deviation of quadrat sizes are used as  
165 explanatory variables.

#### 166 *Data transformation and resemblance measures*

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

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

169 in which  $x_{ij}$  is the relative percentage cover value for species  $i$  in quadrat  $j$  ranging from 0 to  
170 1, and  $c$  is a parameter falling in the range  $[-\infty, \infty]$  such that  $c=0$  is not allowed. This  
171 procedure allows for weighting abundances differently by adjusting the  $c$  parameter. In cases  
172 with high positive  $c$ , transformed data approximate the presence/absence situation, thus  
173 giving more weight to less abundant species. Large negative values of  $c$  lead to  
174 overweighting the dominant species. If  $c$  is very close to 0, the relative abundance  
175 differences of species remain practically unaffected. However, in real situations data  
176 transformation is rather used for downweighting dominant species, therefore, we made  
177 separate 'full-range' analyses and 'narrow-range' analyses by changing the value of  $c$  within  
178  $[-16; +16]$  or  $(0; +16]$ . Note that  $c$  must not equal 0.

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

#### 186 *Classification algorithm*

187 A hierarchical classification was obtained from the dissimilarity matrix by agglomerative  
188 clustering. The fusion algorithm was the beta-flexible method because it allows for  
189 reproducing classifications of different grouping mechanisms by adjusting its  $\beta$  parameter  
190 within the interval  $[-1; 1]$  (Lance & Williams 1967; see also Podani 2000). Values of  $\beta$  close to  
191 1 tend to emphasise a chained group structure (similarly to the single link or nearest  
192 neighbour method), while negative  $\beta$  values lead to increased grouping tendency (as  
193 observed for complete link or farthest neighbour algorithms). In each trial, the value of  $\beta$  was  
194 chosen randomly from -1 to 1 ('full range'). However, in practice 'group-forming' methods are  
195 preferred, therefore  $\beta$  values were drawn from  $[-1; 0]$  for 'narrow-range' analyses. The cluster

196 level (simulating the case of an 'optimal non-hierarchical classification') was randomly  
197 chosen between 2 and 8. The hierarchical classification was 'cut' at this level and hereafter  
198 only this non-hierarchical clustering was used.

### 199 **Data analysis**

200 The 200 trials of the randomization resulted in 200 classifications of the same spatial series.  
201 From each classification, an incidence matrix,  $\mathbf{C}$ , was calculated in which  $c_{ij}$  is 1 if objects  $i$   
202 and  $j$  in the same cluster and 0 otherwise. Euclidean distances were calculated between all  
203 pairs of incidence matrices. This method is also called 'PAIRBONDS' (Arabie & Boorman  
204 1973; Podani 2000). These distances were then summarized into another distance matrix  
205 based on which principal coordinates analysis was computed. In the resulting ordination all  
206 points correspond to a non-hierarchical classification. Then, the following explanatory  
207 variables were fitted to the ordination diagram: mean and standard deviation of plot sizes,  
208 resemblance measure,  $c$  of Clymo's transformation,  $\beta$  of the flexible classification and the  
209 number of clusters. Trend surfaces of numerical variables were fitted onto the scatter plots  
210 by generalized additive models, while average scores were calculated for the resemblance  
211 measures. The relative importance of the explanatory variables was tested by constrained  
212 ordination: the Euclidean distances obtained earlier were subjected to a distance-based  
213 redundancy analysis (db-RDA, Legendre & Anderson 1999). When mean plot size, Clymo's  $c$   
214 and  $\beta$  were scaled on full-range, their squared terms were also included in the model as  
215 explanatory variables. Low (<2) values of generalized variance inflation factors (GVIF, Fox &  
216 Monette 1992) indicated negligible collinearity between model terms. The models were  
217 evaluated by comparing  $F$  ratios of the model terms vs. residual variation, by calculating  
218 adjusted  $R$ -squared measures and by visual observation of fitted explanatory variables on  
219 the PCoA diagrams. During the evaluation of db-RDA models, predictors with  $F$  ratios with a  
220 type I error rate of  $P < 0.01$  were considered significant.

221 Our variation partitioning approach relies on the basic assumption that db-RDA models can  
222 properly explain the variation among classifications attributed to the different methodological  
223 decisions. In order to validate our modelling technique, we applied a simulation test. The  
224 above described simulation analysis with narrow-range variables, starting from the sample  
225 selection and ending at calculation of explained variances was repeated many times.  
226 However, instead of the fully random parametrization of the six variables representing  
227 methodological decisions, some of them were 'fixed', i.e. they were given zero variance. For  
228 example, if plot size was fixed, only plots of the same size were selected from each location  
229 in all of the 200 classifications that were entered in each db-RDA. Of course, in such cases,  
230 the fixed variable was not included as an explanatory variable of the db-RDA, since it had no  
231 variation. The number of fixed variables was increased from zero to five in six steps and for

232 each number of fixed variables, 100 trials were performed. Then, average explained  
233 variation, unexplained and total variation were plotted against the number of fixed variables.  
234 We expected that explained variation would decrease with increases in the number of fixed  
235 variables because reducing the possible outcomes of methodological decisions should also  
236 reduce the variation among classification they account for. If unexplained variation also  
237 decreased with the increased number of fixed variables, we could conclude that variation  
238 caused by methodological decisions was not properly explained by the db-RDA model. On  
239 the contrary, approximately constant unexplained variation obtained for different numbers of  
240 fixed variables would mean that independently from the methodological decisions and the  
241 explanatory variables, there is a certain amount of inherent variation in the compositional  
242 data.

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

248

## 249 Results

250 Distance-based RDA models of simulated and grassland data sets explained different  
251 proportions of the total variation among classifications. The adjusted  $R^2$  values were higher  
252 for the simulated data sets (full-range: 0.466, narrow-range: 0.258) than the grassland data  
253 (full-range: 0.260, narrow-range: 0.157). In the model of the simulated data set with full-range  
254 variables flexible  $\beta$  ( $F=121.388$ ), cluster level ( $F=26.437$ ), mean plot size ( $F=6.592$ ), Clymo's  
255  $c$  ( $F=5.827$ ) and SD of plot sizes ( $F=3.455$ ) proved to have a significant effect at  $p<0.01$   
256 (Table 1). Mean plot size, Clymo's  $c$ , flexible  $\beta$  and cluster number showed a good fit on the  
257 first two dimensions of the PCoA ordination ( $P=4.1e-11$ ,  $P=9.2e-7$ ,  $P=3.4e-88$  and  $P=1.19e-  
258 14$ , respectively; Fig. 1). Values of flexible  $\beta$  changed gradually along the first PCoA axis  
259 with increasing  $\beta$  values in the positive direction, while mean plot size and Clymo's  $c$  showed  
260 a gradient along the second axis. A non-linear pattern was found for cluster number.  
261 Centroids of classifications with different resemblance measures fell close to each other.

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



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

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

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

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

290

## 291 **Discussion & Conclusions**

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

298 The order of importance of the predictors was not the same in all experiments, while some  
299 general trends did appear. Flexible  $\beta$ , cluster level, Clymo's  $c$  and mean plot size were

300 always among the significant model terms, and in many cases they were given the highest  
301 rank. Obvious interpretation is that decisions about clustering process, including the chaining  
302 algorithm and the number of clusters, influence most strongly the outcome of numerical  
303 classification of compositional data. Nevertheless, the other variables were also critical at  
304 least in one of the four scenarios.

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

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

332 The resemblance measure showed weaker effect than plot size and data transformation,  
333 however, it was still significant in the narrow-range analysis of the grassland data set, and it  
334 was near the pre-set significance level in the full-range trial of the same data. The matter of

335 choosing among resemblance measures is more deeply investigated compared to other  
336 methodological decisions, and many papers highlight the differences of the available indices  
337 (Campbell 1978; Legendre & De Cáceres 2013). Without questioning that different  
338 resemblance measures can be appropriate for specific purposes, and the choice between  
339 them had to be taken carefully, our results suggest that the importance of this decision may  
340 be over-emphasized in comparison with other decisions. Thus, we consider the importance  
341 of the resemblance measure as a good reference to assess the significance of the other  
342 explanatory variables. Nevertheless, it must be noted that we employed only four indices that  
343 are very popular among vegetation ecologists.

344 The  $\beta$  parameter of the flexible clustering was the most significant predictor in three cases.  
345 Its value with full range was more influential than with narrow range, which clearly indicates  
346 that decision on the classification method is most critical between chain-forming ( $\beta > 0$ ) and  
347 group-forming ( $\beta < 0$ ) methods, while differences within group-forming algorithms are not that  
348 substantial. This difference is the most striking with the grassland data, for which its effect is  
349 dropped from the 1<sup>st</sup> to the 3<sup>rd</sup> most important model term if compared to the full-range  
350 scenario. In recent works of numerical syntaxonomy (for example, Havlová 2006; Knollová &  
351 Chytrý 2004), of the distance-based methods chain-forming algorithms have received much  
352 fewer applications than group-forming ones which include the flexible method with negative  $\beta$   
353 values applied here. Much more widespread is Ward's agglomerative method (more  
354 precisely, incremental sum of squares) which also has a preference for spherical group  
355 shapes. The good performance of flexible method with  $\beta = -0.25$  and the Ward's method was  
356 also indicated by Lötter et al. (2013) but one is warned that groups show up apparently  
357 clearly in the resulting dendrograms even if in fact they do not exist in the data (Podani  
358 2000). Another very popular hierarchical method is TWINSpan (Hill 1979; Roleček et al.  
359 2009), however, its weaknesses are pointed out in several papers (Belbin & McDonald 1993;  
360 Dufrene & Legendre 1997; Lötter et al. 2013). The significant effect of clustering algorithm  
361 implies that during the comparison and revision of existing vegetation classifications the  
362 applied clustering methods should be taken into account carefully. Large differences  
363 between classifications of the same vegetation units of a certain area can be attributed to the  
364 different methods used, and therefore comparison of classification prepared by different  
365 algorithms may even be meaningless.

366 Cluster level was the second most significant model term in three of the four scenarios and  
367 the most important one for the grassland data set with narrow-range variables. In  
368 classification studies, the number of clusters is usually determined by an expert-based, i.e. a  
369 rather subjective method (but see Botta-Dukát et al. 2005 or Illyés et al. 2007). Cluster  
370 validation, including the choice of the optimal 'cut level', is the most data-specific decision

371 among those we studied here, therefore the only general recommendation that we could  
372 stress is to investigate and to use quantitative measures for this purpose instead of  
373 subjective assessment (for example, Milligan 1996; Aho et al. 2009; Tichý et al. 2010; Tichý  
374 et al. 2011). The validation tools are so numerous that their comparative study focusing on  
375 specific requirement for numerical syntaxonomy would be timely.

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

395 The second assumption was that the db-RDA model captured relevant information on  
396 variation among classifications. The first part of db-RDA was PCoA known to preserve the  
397 original distance structure of the input matrix. Then, the PCoA axes, as transformed variables  
398 of between-classification distances, were related to the explanatory variables (i.e. the  
399 methodological decisions) by usual RDA method. At this step, even patterns that are non-  
400 linear functions of the explanatory variables are decomposed into separate components for  
401 which the explanatory variables can be linearly related. To account for eventual non-linear  
402 relationships that cannot be revealed by this procedure, we included squared terms into the  
403 models and the distribution of the explanatory variables over the first two PCoA axes were  
404 also mapped by a flexible fitting method (GAM). These trend surfaces revealed that cluster  
405 number can show a non-linear pattern along the first two axes. However, this pattern can  
406 likely to be accounted for by db-RDA because cluster number came out as a highly

407 significant predictor in all cases. In our analysis to validate the appropriateness of our  
408 modelling approach, we found that the amount of unexplained variation of our models is not  
409 related to the number of fixed and randomized variables, that is, it is independent from the  
410 methodological decisions. This suggests that the variation caused by the random  
411 parametrization of the classifications is satisfactorily explained by the db-RDA models.  
412 Therefore, we do not suspect a significant amount of unexplained variation due to non-linear  
413 effects or interactions among methodological decisions. The unexplained variation may have  
414 several different origins. The most trivial reason is that the data set has a certain degree of  
415 robustness which explains low sensitivity to methodological changes. Robustness is  
416 obviously higher for the grassland data set that contains biologically interpretable patterns.  
417 Nevertheless, it is also present in the simulated data set since randomized data do not lack  
418 variation completely but this variation is comparable to what is expected by chance. Another  
419 possible source is the individual 'fate' of plots in the analysis. Two classifications can be  
420 identically parameterized in terms of the selected plot sizes but the sample to be analysed  
421 can still differ because it is not fixed which plot size should be selected from a certain  
422 location.

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

427

## 428 **Acknowledgements**

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

434

## 435 **References**

436 Aho, K., Roberts, D. W. & Weaver, T. 2008. Using geometric and non-geometric internal  
437 evaluators to compare eight vegetation classification methods. *Journal of Vegetation Science*  
438 19: 549–562.

- 439 Arabie, P. & Boorman, S. A. 1973. Multidimensional scaling of measures of distance  
440 between partitions. *Journal of Mathematical Psychology* 10: 148–203.
- 441 Belbin, L. & McDonald, C. 1993. Comparing three classification strategies for use in ecology.  
442 *Journal of Vegetation Science* 4: 341–348.
- 443 Botta-Dukát, Z., Chytrý, M., Hájková, P. & Havlová, M. 2005. Vegetation of lowland wet  
444 meadows along a climatic continentality gradient in Central Europe. *Preslia* 77: 89–111.
- 445 Campbell, B.M. 1978. Similarity coefficients for classifying relevés. *Vegetatio* 37: 101–109.
- 446 Chytrý, M. & Otýpková, Z. 2003. Plot sizes used for phytosociological sampling of European  
447 vegetation. *Journal of Vegetation Science* 14: 563–570.
- 448 Cramér, H. 1946. *The elements of probability theory and some of its applications*. Wiley, New  
449 York, US.
- 450 Dale, M.B. 1995. Evaluating classification strategies. *Journal of Vegetation Science* 6: 437–  
451 440.
- 452 Dengler, J., Jansen, F., Glockler, F., Peet, R.K., De Cáceres, M., Chytrý, M., Ewald, J.  
453 Oldeland, J., Lopez-Gonzalez, G., Finckh, M., Mucina, L., Rodwell, J.S., Schaminée, J.H.J. &  
454 Spencer, N. 2011. The global index of vegetation-plot databases 1 (GIVD): a new resource  
455 for vegetation science. *Journal of Vegetation Science* 22: 582–597.
- 456 Dengler, J., Löbel, S. & Dolnik, C. 2009. Species constancy depends on plot size – a  
457 problem for vegetation classification and how it can be solved. *Journal of Vegetation Science*  
458 20: 754–766.
- 459 Dufrene, M. & Legendre, P. 1997. Species assemblages and indicator species: the need for  
460 a flexible asymmetrical approach. *Ecological Monographs* 67: 345–366.
- 461 Fox, J. & Monette, G. 1992. Generalized collinearity diagnostics. *Journal of the American*  
462 *Statistical Association* 87: 178–183.
- 463 Goodman, L. & Kruskal, W. 1954. Measures of association for cross classifications. *Journal*  
464 *of the American Statistical Association* 49: 732–764.
- 465 Green, R.H. 1980. Multivariate approaches in ecology: The assessment of ecological  
466 similarity. *Annual Review of Ecology and Systematics* 11: 1–14.
- 467 Hajdu, L.J. 1981. Graphical comparison of resemblance measures in  
468 phytosociology. *Vegetatio* 48: 47–59.

- 469 Havlová, M. 2006. Syntaxonomical revision of the Molinion meadows in the Czech Republic.  
470 *Preslia* 78: 87–101.
- 471 Hill, M.O. 1979. *TWINSPAN. A FORTRAN program for arranging multivariate data in an*  
472 *ordered two-way table by classification of the individuals and attributes.* Cornell University,  
473 Ithaca, NY, US.
- 474 Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-  
475 absence) data: an evaluation. *Biological Reviews* 57: 669–689.
- 476 Illyés, E., Chytrý, M., Botta-Dukát, Z., Jandt, U., Škodová, I., Janišová, M., Willner, W. &  
477 Hájek, O. 2007. Semi-dry grasslands along a climatic gradient across Central Europe:  
478 Vegetation classification with validation. *Journal of Vegetation Science* 18: 835–846.
- 479 Jensen, S. 1978. Influences of transformation of cover values on classification and ordination  
480 of lake vegetation. *Vegetatio* 37: 19–31.
- 481 Kenkel, N., P. Juhász-Nagy & J. Podani. 1989. On sampling procedures in population and  
482 community ecology. *Vegetatio* 83:195-207.
- 483 Kenkel, N. C. & Podani, J. 1991. Plot size and estimation efficiency in plant community  
484 studies. *Journal of Vegetation Science* 2: 539–544.
- 485 Knollová, I. & Chytrý, M. 2004. Oak-hornbeam forests of the Czech Republic: geographical  
486 and ecological approaches to vegetation classification. *Preslia* 76: 291–311.
- 487 Knollová, I., Chytrý, M., Tichý, L. & Hajek, O. 2005. Stratified resampling of phytosociological  
488 databases: some strategies for obtaining more representative data sets for classification  
489 studies. *Journal of Vegetation Science* 16: 479–486.
- 490 Lance, G.N & Williams, W.T. 1967. A general theory of classificatory sorting strategies. I.  
491 Hierarchical systems. *Computer Journal* 9: 373–380.
- 492 Legendre, P, & De Cáceres, M. 2013. Beta diversity as the variance of community data:  
493 dissimilarity coefficients and partitioning. *Ecology Letters* 16: 951–963.
- 494 Legendre, P. & Anderson, M.J. 1999. Distance-based redundancy analysis: testing  
495 multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69:  
496 1–24.
- 497 Lengyel, A., Chytrý, M. & Tichý, L. 2011. Heterogeneity-constrained random resampling of  
498 phytosociological databases. *Journal of Vegetation Science* 22: 175–183.

- 499 Lötter, M.C., Mucina, L. & Witkowski, E.T.F. 2013. The classification conundrum: species  
500 fidelity as leading criterion in search of a rigorous method to classify a complex forest data  
501 set. *Community Ecology* 14: 121–132.
- 502 Maarel, E. van der. 1979. Transformation of cover-abundance values in phytosociology and  
503 its effects on community similarity. *Vegetatio* 39: 97–114.
- 504 Maarel, E. van der. 2005. *Vegetation ecology*. Blackwell Science, Oxford, UK.
- 505 Milligan, G.W. 1996. Clustering validation: Results and implications for applied analyses. In:  
506 Arabie, P., Hubert, L.J. & de Soete, G. (eds.) *Clustering and classification*, pp 341–375.  
507 World Scientific, River Edge, US.
- 508 Milligan, G.W. & Cooper, M.C. 1987. Methodology review: Clustering methods. *Applied*  
509 *Psychological Measurement* 11: 329–354.
- 510 Møller, A. & Jennions, M.D. 2002. How much variance can be explained by ecologists and  
511 evolutionary biologists? *Oecologia* 132: 492–500.
- 512 Mucina, L. 1997. Classification of vegetation: Past, present, and future. *Journal of Vegetation*  
513 *Science* 8: 751–760.
- 514 Otýpková, Z. & Chytrý, M. 2006. Effects of plot size on the ordination of vegetation samples.  
515 *Journal of Vegetation Science* 17: 465–472.
- 516 Peet, R.K. & Roberts, D.W. 2013. Classification of natural and semi-natural vegetation In:  
517 van der Maarel, E. & Franklin, J. *Vegetation ecology*, Second edition. pp 28–70. Wiley, NY,  
518 US.
- 519 Podani, J. 1984. Analysis of mapped and simulated vegetation patterns by means of  
520 computerized sampling techniques. *Acta Botanica Hungarica* 30: 419–441.
- 521 Podani, J. 1989. Comparison of ordinations and classifications of vegetation data. *Vegetatio*  
522 83: 111–128.
- 523 Podani, J. 1998. Numerikus cönológiai vizsgálatok a Sas-hegy (Budai hg.)  
524 dolomitzsiklagyepjeiben (A complex numerical analysis of dolomite rock grasslands of the  
525 Sas-hegy Nature Reserve, Budapest, Hungary. In Hungarian with English summary) In:  
526 Csontos P. (ed.), *Sziklagyeppek szünbotanikai kutatása*. Scientia, Budapest. pp. 213–229.
- 527 Podani, J. 2000. *Introduction to the exploration of multivariate biological data*. Backhuys  
528 Publishers, Leiden, NL



- 529 Reed, R.A., Peet, R.K., Palmer, M.W. & White, P.S. 1993. Scale dependence of vegetation-  
530 environment correlations: A case study of a North Carolina Piedmont woodland. *Journal of*  
531 *Vegetation Science* 4: 329–340.
- 532 Roleček, J., Tichý, L., Zelený, D. & Chytrý, M. 2009. Modified TWINSpan classification in  
533 which the hierarchy respects cluster heterogeneity. *Journal of Vegetation Science* 20: 596–  
534 602.
- 535 Tichý, L., Chytrý, M. & Smarda, P. 2011. Evaluating the stability of the classification of  
536 community data. *Ecography* 34: 807–813.
- 537 Tichý, L., Chytrý, M., Hájek, M., Talbot, S.S. & Botta-Dukát, Z. 2010. OptimClass: Using  
538 species-to-cluster fidelity to determine the optimal partition in classification of ecological  
539 communities. *Journal of Vegetation Science* 21: 287–299.
- 540 Whittaker, R.H. 1973. *Ordination and classification of vegetation*. Junk, The Hague, NL.
- 541 Wilson, B.J. 2012. Species presence/absence sometimes represents a plant community as  
542 well as species abundances do, or better. *Journal of Vegetation Science* 23: 1013–1023.
- 543 Wolda, H. 1981. Similarity indices, sample size and diversity. *Oecologia* 50: 296–302.
- 544

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

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

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

548

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

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

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

552

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

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

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

556

557

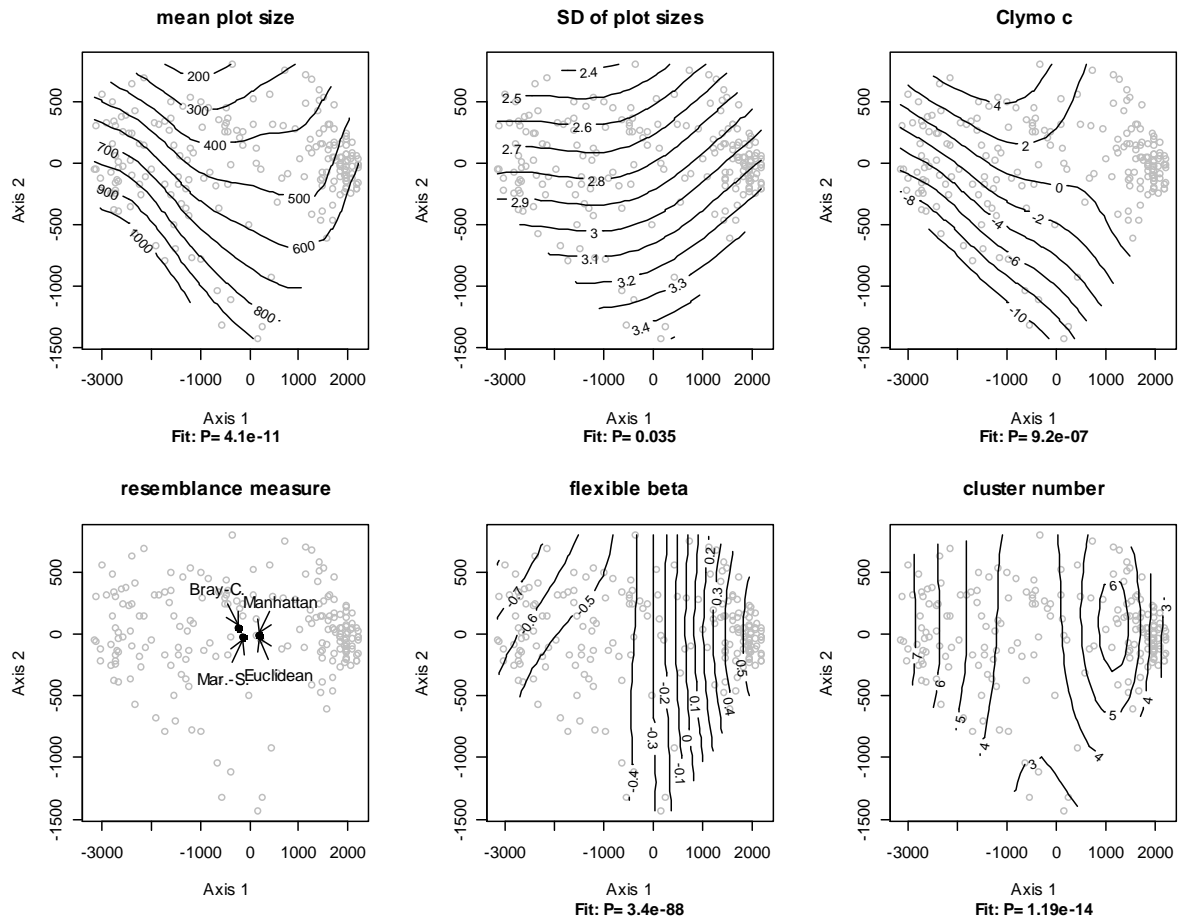
558 **Table 4.** Predictors of the db-RDA model for the grassland data set on narrow ranges of the  
 559 variables. P-values are based on 199 permutations.

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

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

561

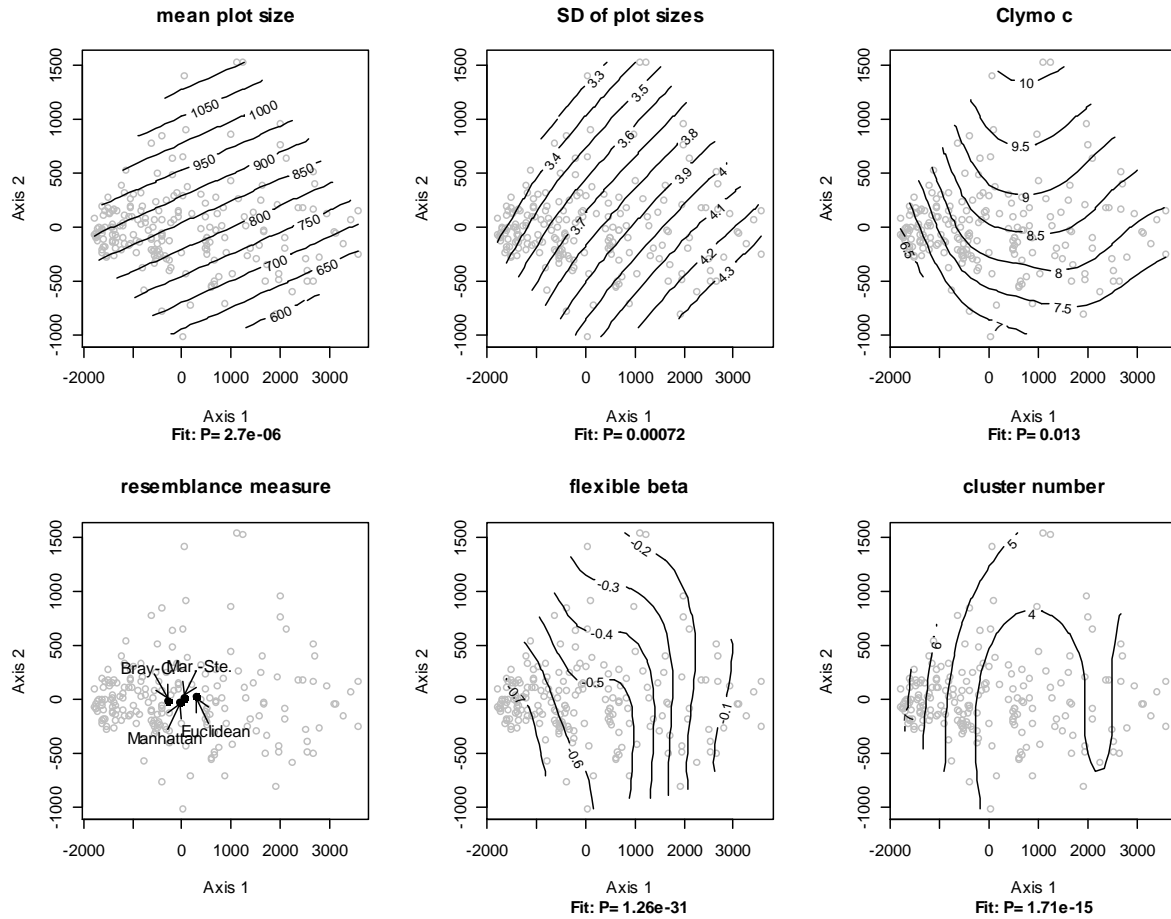
562 **Figure 1.** Principal coordinates analysis of classifications of the simulated data sets with the  
 563 full ranges of predictor variables. Continuous variables are fitted as trend surfaces *a*  
 564 *posteriori* by GAM, factor variables are fitted by averaging of object scores on the two  
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



566

567

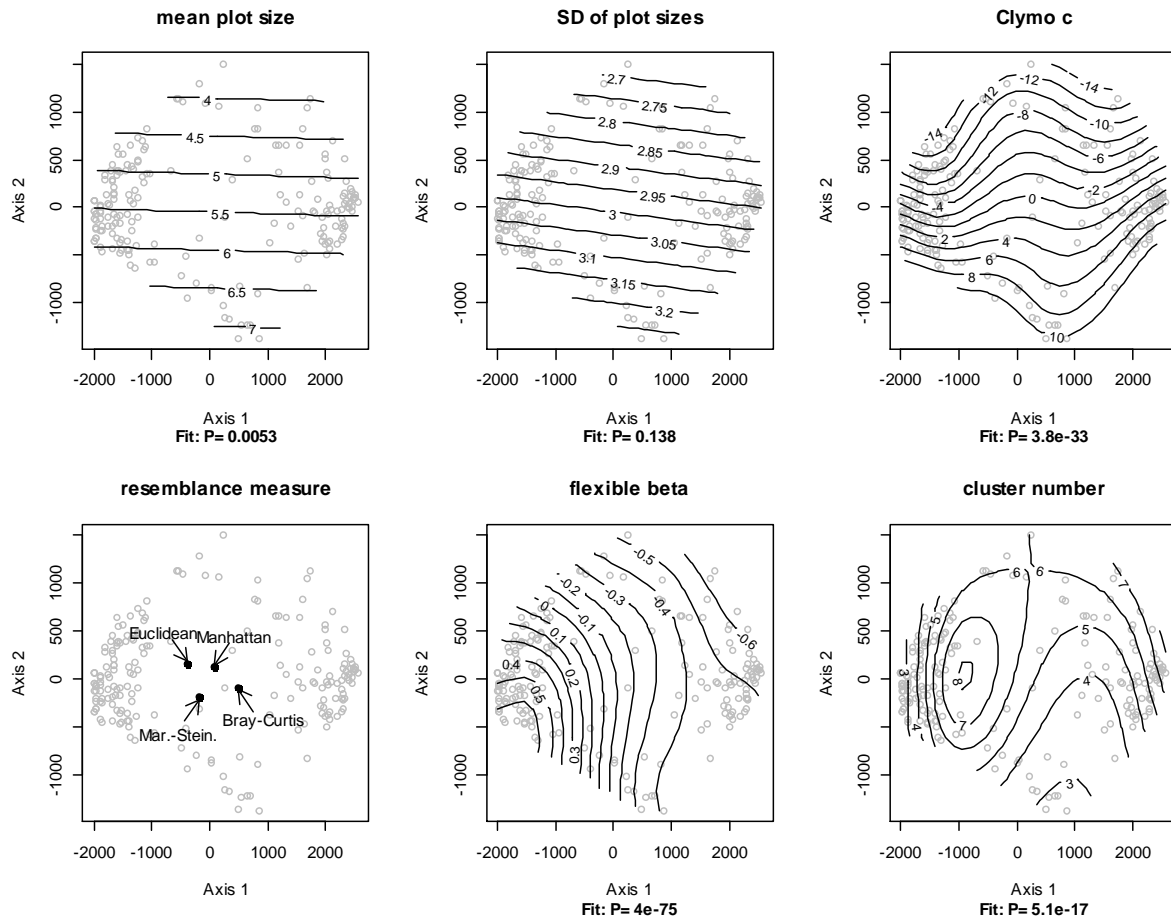
568 **Figure 2.** Principal coordinates analysis of classifications of the simulated data sets with the  
 569 narrow ranges of variables. Continuous variables are fitted as trend surfaces *a posteriori* by  
 570 GAM, factor variables are fitted by averaging of object scores on the two ordination axes.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



572

573

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



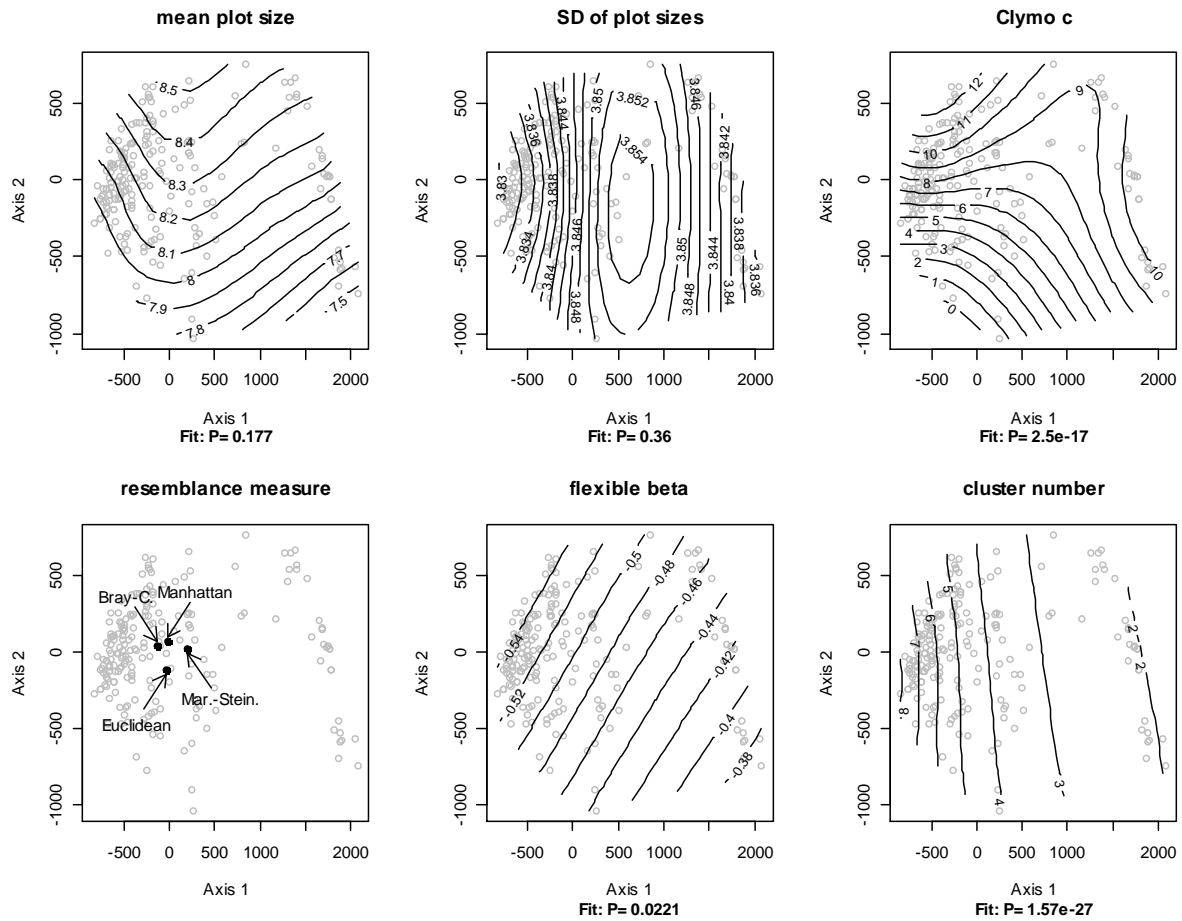
578

579

580



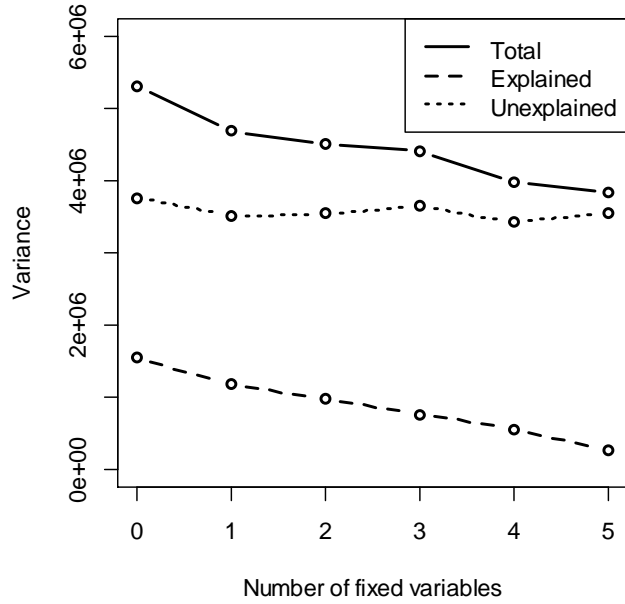
581 **Figure 4.** Principal coordinates analysis of classifications of the grassland data sets with the  
 582 narrow ranges of variables. Continuous variables are fitted as trend surfaces *a posteriori* by  
 583 GAM, factor variables are fitted by averaging of object scores on the two ordination axes.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



585

586

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