Journal of Vegetation Science **22** (2011) 281–291

Bias in vegetation databases? A comparison of stratified-random and preferential sampling

Dana Michalcová, Samuel Lvončík, Milan Chytrý & Ondřej Hájek

Keywords
Alpha diversity; Beta diversity; Endangered species; Invasive species; Neophytes; Ordination; Parameter estimation; Relevé; Sample-based rarefaction; Species number; Stratification; Subjectivity

Received 21 May 2010
Accepted 24 November 2010
Co-ordinating Editor: Ingolf Kühn

Michalcová, D. (corresponding author, danmich@sci.muni.cz), **Chytrý, M.** (chytry@sci.muni.cz) & **Hájek, O.** (ohajek@sci.muni.cz): Department of Botany and Zoology, Masaryk University, Kotlářská 2, CZ-61137 Brno, Czech Republic
Lvončík, S. (lvonca@seznam.cz): Department of Plant Biology, Faculty of Agronomy, Mendel University in Brno, Zemědělská 1, CZ-613 00 Brno, Czech Republic

Abstract

Aim: Vegetation plots collected since the early 20th century and stored in large vegetation databases are an important source of ecological information. These databases are used for analyses of vegetation diversity and estimation of vegetation parameters, however such analyses can be biased due to preferential sampling of the original data. In contrast, modern vegetation survey increasingly uses stratified-random instead of preferential sampling. To explore how these two sampling schemes affect vegetation analyses, we compare parameters of vegetation diversity based on preferentially sampled plots from a large vegetation database with those based on stratified-random sampling.

Location: Moravian Karst and Silesia, Czech Republic.

Methods: We compared two parallel analyses of forest vegetation, one based on preferentially sampled plots taken from a national vegetation database and the other on plots sampled in the field according to a stratified-random design. We repeated this comparison for two different regions in the Czech Republic. We focussed on vegetation properties commonly analysed using data from large vegetation databases, including alpha (within-plot) diversity, cover and participation of different species groups, such as endangered and alien species within plots, total species richness of data sets, beta diversity and ordination patterns.

Results: The preferentially sampled data sets obtained from the database contained more endangered species and had higher beta diversity, whereas estimates of alpha diversity and representation of alien species were not consistently different between preferentially and stratified-randomly sampled data sets. In ordinations, plots from the preferential samples tended to be more common at margins of plot scatters.

Conclusions: Vegetation data stored in large databases are influenced by researcher subjectivity in plot positioning, but we demonstrated that not all of their properties necessarily differ from data sets obtained by stratified-random sampling. This indicates the value of vegetation databases for use in biodiversity studies; however, some analyses based on these databases are clearly biased and their results must be interpreted with caution.

Introduction

Vegetation plots, or relevés, collected since the early 20th century and stored in large vegetation databases, are increasingly recognized as an important source of ecological information. There are more than 1 800 000 computerised vegetation plots in European databases alone (Schaminée et al. 2009). Although most of these plots were originally sampled for the purpose of local vegetation description and classification, currently they are often used for other types of analyses and meta-analyses, such as description of vegetation patterns using classification or ordination methods (e.g. Duckworth et al. 2000; van Kley & Schaminée 2003; Lososová et al. 2004; Illyés et al. 2007), local species richness or alpha diversity (Pyšek et al. 2005; Láníková & Lososová 2009; Lloret et al. 2009; Dupré et al.

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

2010), between-site variation or beta diversity (Ewald 2002; Lososová et al. 2004) or representation of alien plant species in vegetation types (e.g. Chytrý et al. 2005, 2008; Lososová & Simonová 2008; Vilà et al. 2010).

Most of the plots stored in vegetation databases were sampled preferentially (Podani 1984; Margules & Austin 1994; Lepš & Šmilauer 2007; Roleček et al. 2007), i.e. based on the researcher's subjective decision, without applying any clear formal criteria for plot selection (Goedickemeier et al. 1997). Like other sampling methods, preferential sampling has its advantages and disadvantages. If applied in a small area of a few hectares it can representatively record all vegetation types (Roleček et al. 2007), which seems sufficient for the purposes of local vegetation survey. However, analyses of subjectively positioned vegetation plots may lead to biased results due to different author's experience, anticipation, intuition or affinity to particular vegetation types (Olsen et al. 1999). In studies aimed at the estimation of vegetation parameters across landscapes, such as mean species richness or mean proportion of certain species groups (e.g. endangered or alien species), preferential sampling may lead to biased results due to violation of statistical assumptions of randomness and independence of observations (Kershaw & Looney 1985; Chiarucci 2007; Lájer 2007). Some authors (Chiarucci 2007; Lájer 2007) do not recommend statistical analyses of preferentially sampled data, while others argue for a compromise solution, suggesting the use of statistical tests but with cautious interpretation of their results (Botta-Dukát et al. 2007; Diekmann et al. 2007; Lepš & Šmilauer 2007; Økland 2007).

In studies of vegetation diversity, the weaknesses of preferential sampling cannot be simply solved by its replacement with random or systematic sampling, as sometimes recommended. Due to non-random distribution of vegetation types across landscapes, random or systematic sampling usually results in data sets with many plots from common vegetation types and few or no plots from rare types (Kenkel et al. 1989; Botta-Dukát et al. 2007; Lájer 2007; Wilson 2007). Therefore, in studies aiming at the description of maximum variation of vegetation across a landscape, stratified-random sampling seems to be a promising alternative to preferential sampling (Austin & Heyligers 1989; Margules & Austin 1994; Bunce et al. 1996). In this sampling method, a study area is stratified in advance according to different environmental variables that are thought to affect vegetation patterns, such as climate, altitude, topography, bedrock or soil types (Kenkel et al. 1989; Margules & Austin 1994; Michaelsen et al. 1994; Goedickemeier et al. 1997; Stohlgren et al. 1997b; Firbank et al. 2003), according to land cover or vegetation maps (Ortega et al. 2004; Acosta et al. 2005; Holt et al. 2008; Šamonil et al. 2009) or combinations of these. After landscape stratification, random sampling is applied within each stratum. In this design the statistical assumption of randomness is valid and non-explicit subjective decisions are replaced by formal, clearly described and repeatable criteria. Vegetation plot data sets obtained using this design are appropriate for statistical analyses and, at the same time, they may contain plots representing both common and rare vegetation types (Chiarucci 2007; Wilson 2007). Stratified-random design has been used in several vegetation survey studies (e.g. Coetzee et al. 1995; da Silva et al. 1997; Siebert et al. 2002; Grabherr et al. 2003), studies focussed on recording rare species (Margules & Austin 1994), estimation of species richness or diversity (Stohlgren et al. 1997a, c; Gimaret-Carpentier et al. 1998), assessment of alien plant distribution (Stohlgren et al. 1997c) or for monitoring programmes focussed on agriculture or natural resource conservation (Olsen et al. 1999).

In spite of the advantages and increasing popularity of stratified-random sampling, many current analyses of vegetation diversity are based on data sets extracted from large vegetation databases in order to save time and money needed for new field sampling. Such analyses are biased due to the preferential sampling of the original data, but the magnitude of this bias is unknown. Thus the reliability of these analyses is also unknown. It is therefore necessary to explore how vegetation analyses based on sets of preferentially sampled plots differ from those based on formalised sampling schemes. There are several studies comparing preferentially sampled data sets with those sampled randomly or systematically (e.g. Holeksa & Woźniak 2005; Botta-Dukát et al. 2007; Diekmann et al. 2007; Hédl 2007), but hardly any studies comparing data and analyses based on preferential versus stratified-random sampling.

The aim of this study is to compare vegetation analyses based on preferentially sampled plots with those obtained through stratified-random sampling. We compared two parallel analyses of forest vegetation, one based on a national vegetation database and the other on a new data set sampled in the field according to the stratified-random design. We repeated this comparison for two different regions in the Czech Republic. We focussed on vegetation properties that are commonly analysed based on large vegetation databases, including alpha (within-plot diversity), cover and proportional number of species belonging to different groups (e.g. endangered or alien species), total species richness of the sample, beta diversity and ordination patterns.

Methods

Study areas

The two study areas, Moravian Karst and the Czech part of Silesia, are situated in the eastern part of the Czech

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

Republic (Fig. 1). The study area of Moravian Karst was delimited as a rectangle with edge coordinates of 49°13′-49°26′N and 16°38′-16°48′W and included the karst area over Devonian limestone and its surroundings, with predominant granitoids, conglomerates and graywacke (Kodym et al. 1966). The mean annual precipitation is 550–700 mm and mean annual temperature is 6–9 °C (Vesecký et al. 1958). Deciduous and mixed forests cover approximately 40% of the area. Natural vegetation is deciduous forest belonging mainly to the phytosociological alliances *Fagion sylvaticae*, *Carpinion betuli*, *Tilio-Acerion*, *Quercion pubescenti-petraeae*, *Quercion petraeae* and *Alnion incanae* (Neuhäuslová et al. 1997). The second study area, Silesia, includes two phytogeographical districts of the national phytogeographical division (Skalický 1988): Silesian Hills (Slezská pahorkatina) and Ostrava Basin (Ostravská pánev). The main bedrock types are graywacke, clay shale, loess and loess loam (Kodym et al. 1966). The mean annual precipitation is 600–1000 mm and mean annual temperature is 6–9 °C (Vesecký et al. 1958). Deciduous and mixed forests cover approximately 20% of the study area. Common types of natural forest vegetation belong to the alliances *Fagion sylvaticae*, *Carpinion betuli*, *Genisto germanicae-Quercion*, *Alnion incanae*, *Alnion glutinosae* and *Salicion albae* (Neuhäuslová et al. 1997).

Stratified-random sampling

Forest vegetation sampling using the stratified-random design was conducted by Michalcová (2009) in the Moravian Karst in 2005–2008 and Lvončík (2006) in Silesia in 2004–2005. We focussed on deciduous forests, which correspond to natural vegetation in both areas, and mixed forests, i.e. deciduous forests with an admixture of planted conifers. The stratification of both areas was done before field sampling using the geographical information system ArcGIS (version 8.3; ESRI, Redlands, CA, USA; http://www.esri.com). Points placed on a regular grid with 0.1-km spacing in both longitudinal and latitudinal directions were generated for each of the two study areas. For each point, the following data were obtained from digital maps: (1) mean annual precipitation (Vesecký et al. 1958), (2) mean annual temperature (Vesecký et al. 1958), (3) slope inclination, (4) aspect, (5) presence/absence of alluvium, (6) presence/absence of limestone and (7) soil acidity on a five-point ordinal scale (estimation based on rock and soil type). Presence/absence of limestone was a better predictor of species composition than the available data on soil acidity, but significant areas of limestone were only found in the Moravian Karst; therefore limestone was used only in the Moravian Karst, while soil acidity was used only in Silesia. Each variable was divided into categories, which differed slightly in each area to accommodate specific features of the two landscapes (Table 1). For slope aspect, only two categories of north-facing (315-45°) and south-facing (135-225°) slopes were used to emphasize the difference in irradiation, whereas sites on slopes of other aspects were not considered. All grid points with the same combination of variable categories were assigned to the same stratum. Then, points located in other land-cover types than deciduous or mixed forest according to the CORINE land-cover data (Bossard et al. 2000) were excluded, and of the remaining points one from each stratum was selected at random. These points were found in the field using a GPS device (Garmin eTrex) and sampled using plots of 200 m^2 from June to September. A few plots found in the field that were in clear-cut or wind-thrown forests were not sampled, and other random points from the same stratum were sampled instead. The cover of all vascular plant species was estimated using a nine-degree Braun-Blanquet abundance–dominance scale (Westhoff & van der Maarel 1978). A total of 72 plots were recorded in the Moravian Karst and 150 plots in Silesia.



Fig. 1. The study areas within the Czech Republic.

Table 1. Variables used for environmental stratification of the study areas.

Environmental variable	Categories	
	Moravian Karst	Silesia
Mean annual precipitation (mm)	550–600, 601–650, 651–700	601–700, 701–800, 801–900, 901–1000
Mean annual temperature (°C)	6–7, 7.1–8, > 8	6–7, 7.1–8, > 8
Slope inclination (°)	0–4, 5–15, > 15	0–15, > 15
Slope aspect	North/south	North/south
Alluvium	Yes/no	Yes/no
Limestone	Yes/no	–
Soil acidity	–	Strongly acidic, moderately acidic, neutral, slightly alkaline

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

Preferential data sets

Corresponding sets of preferentially sampled vegetation plots were selected from the Czech National Phytosociological Database (Chytrý & Rafajová 2003) according to the following criteria: (1) they were from the areas of Moravian Karst or Silesia, as delimited above (Fig. 1); (2) they were from deciduous or mixed forests; (3) tree layer cover was higher than 40%; (4) plot size ranged from 100 to 300 m^2; (5) they were recorded after 1980; and (6) they were sampled after 15 May to eliminate vegetation plots with species occurring only in the early spring. There were 119 and 111 plots from the Moravian Karst and Silesia, respectively, that fulfilled criteria 1 and 2, and 73 and 61 plots that fulfilled all six criteria. Only the plots following all six of these criteria were used for the analysis.

Data analysis

Vegetation plots were stored in a TURBOVEG database (version 2.83; Hennekens & Schaminée 2001) and subsequently edited using the JUICE program (version 7.0; Tichý 2002). All records of bryophytes, lichens and woody plant seedlings and saplings were removed because of their inconsistent sampling in the plots from the database. Occurrences of the same species in the tree and shrub layers were merged and their covers summed on the basis of random overlap of their foliage (see Chytrý et al. 2005: 343 for details).

Parallel analyses and estimations of vegetation parameters were performed on preferential and stratified-random data sets from each of the two areas. The following properties were compared between the two data set types: (1) number of vascular plant species (without woody plant seedlings and saplings) per plot; (2) number, percentage cover and proportional number of endangered species per plot; (3) number, percentage cover and proportional number of invasive species per plot; (4) number, percentage cover and proportional number of neophytes per plot; (5) sample-based rarefaction curves, representing estimates of total number of all species, endangered species, invasive species and neophytes in the whole data sets; (6) beta diversity; and (7) patterns in ordination diagrams.

Endangered species were defined as those belonging to categories C1, C2 and C3 in the national list of endangered species (Holub & Procházka 2000), which correspond to the IUCN categories CR, EN and VU, respectively. Invasive species and neophytes (post-1500 aliens) followed the national list of Pyšek et al. (2002). The differences between species numbers, covers and proportions were tested using the non-parametric Mann-Whitney *U* test.

Beta diversity was calculated as multivariate dispersion (Anderson et al. 2006). This method quantifies beta diversity as the average dissimilarity (or distance) between individual observation units and their group centroid in a multivariate space using a selected dissimilarity measure. The differences in beta diversity between data sets based on the two sampling methods were tested using a multivariate test for homogeneity in dispersion. This test calculates the *F*-statistic to compare the average dissimilarity between observation units and their centroid. The *P*-value is then obtained by permuting least-square residuals (Anderson et al. 2006). This procedure was performed in the PRIMER program (version 6, PRIMER-E Ltd, Ivybridge, UK; http://www.primer-e.com/) using the PERMDISP routine (distance-based test for homogeneity of multivariate dispersions) with log-transformed percentage cover data, Euclidean distance and 999 permutations.

To estimate the total number of vascular plant species, endangered species and aliens (invasive species and neophytes) in each data set, sample-based rarefaction curves (Gotelli & Colwell 2001; Colwell et al. 2004) were computed using the EstimateS program (version 8.2.0; R. K. Colwell; http://purl.oclc.org/estimates). These curves enable comparisons of data sets at a comparable sampling effort, i.e. the same number of plots (Colwell et al. 2004). Each data set was resampled at random without replacement. The number of plots was gradually increased, with 50 random draws at each step. The number of all species and of species from different groups (endangered species, invasive species and neophytes) was estimated using the Mao Tau estimator including its 95% confidence intervals (Colwell et al. 2004). The rarefaction curves were created using the R program (version 2.8.0; R Development Core Team 2008, Vienna, AT; http://www.r-project.org).

Detrended correspondence analysis (DCA) was used to compare vegetation variability between the preferential and stratified-random data sets. For each of the two study areas, DCA ordination was computed jointly for all plots from both data sets, using the CANOCO software (version 4.5; ter Braak & Šmilauer 2002). Species cover values recorded on the Braun-Blanquet scale were transformed to mid-percentage values and log-transformed. Detrending by segments was applied. The scores for vegetation plots from either Moravian Karst or Silesia were plotted jointly on the DCA diagrams and visualised using the R program.

Results

The number of vascular plant species per plot was higher in the preferential data set than in the stratified-random data set for the Moravian Karst, but there was no

Journal of Vegetation Science
Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

difference between the two data sets for Silesia (Table 2). The number, percentage cover and proportional number of endangered species per plot was higher in the preferential data sets from both areas. A higher number, percentage cover and proportional number of invasive species and neophytes per plot was found in the preferential data set from Silesia. In the Moravian Karst, the figures for invasive species and neophytes did not differ between the two data sets, except for the proportional number of invasive species per plot, which was higher in the stratified-random data set (Table 2).

The rarefaction curves plotted for the preferential data sets of all vascular plant species (Fig. 2a and b) and endangered species (Fig. 2c and d) were higher than the curves plotted for the stratified-random data sets, i.e. the former estimated a higher number of species. The rarefaction curves for the invasive species (Fig. 2e and f) and neophytes (Fig. 2g and h) were not significantly different since their confidence intervals overlapped.

In both areas, the multivariate dispersion was significantly higher in the preferential data sets, indicating their higher beta diversity.

In ordination diagrams of detrended correspondence analysis (DCA) from the Moravian Karst (Fig. 3a) the first axis reflected a gradient of light and moisture, causing the change of forest vegetation from dry and open oak forests (*Quercion petraeae* and *Quercion pubescenti-petraeae*, left side), through mesic beech and hornbeam forests (*Fagion sylvaticae* and *Carpinion betuli*), to moist and shady ravine forests (*Tilio-Acerion*, right side). The second axis corresponded to a soil pH gradient, causing the change from basiphilous oak forests (*Quercion pubescenti-petraeae*, bottom) to more acidophilous oak forests (*Quercion petraeae*, top). The plots on the extreme top of the ordination diagram from Silesia (Fig. 3b) were from open forests with *Betula pendula* and *Sorbus aucuparia*, whereas the plots on the right side were from moist and open white willow alluvial forests (*Salicion albae*). The majority of plots (the scatter in the middle of the ordination diagram) were recorded in mesic beech and hornbeam forests (*Fagion sylvaticae* and *Carpinion betuli*) and ravine forests (*Tilio-Acerion*). In both ordination diagrams there was a clear tendency of plots from the stratified-random sampling to be concentrated in the central part of point scatters, and of the plots from preferential sampling to occur both in the central part and at the margins of point scatters. This indicates that preferentially sampled data sets contain more plots with unusual or extreme species composition.

Discussion

Our results indicate that the data sets sampled using the stratified-random and preferential methods differ in several respects, but that not all the studied properties are significantly affected by the sampling method. Some properties differed only in one of the two study areas, while some others did not differ in any of them.

A higher number of vascular plant species per plot in the preferentially sampled data set than in the stratified-randomly sampled data set was found in the Moravian Karst, but not in Silesia, where both data sets did not differ in local species richness (Table 2). In a comparison of preferential and systematic random samples of Polish forests, Holeksa & Woźniak (2005) found a nearly

Table 2. Comparison of vegetation parameters estimated from stratified-random and preferential data sets from the Moravian Karst and Silesia.

	Moravian Karst			Silesia		
	Preferential (mean ± SE)	Stratified-random (mean ± SE)	Mann-Whitney *U* test (*P*)	Preferential (mean ± SE)	Stratified-random (mean ± SE)	Mann-Whitney *U* test (*P*)
Vascular plant species						
Number per plot	31.0 ± 1.6	25.6 ± 1.5	< 0.05	20.4 ± 1.0	20.1 ± 0.8	n.s.
Endangered species						
Number per plot	1.0 ± 0.2	0.4 ± 0.1	< 0.01	0.3 ± 0.1	0.05 ± 0.02	< 0.001
Proportional number per plot (%)	2.9 ± 0.4	1.8 ± 0.5	< 0.01	1.3 ± 0.3	0.2 ± 0.1	< 0.001
Cover per plot (%)	5.5 ± 1.3	0.7 ± 0.1	< 0.001	2.0 ± 0.8	0.4 ± 0.3	< 0.001
Invasive species						
Number per plot	0.5 ± 0.1	0.6 ± 0.1	n.s.	1.5 ± 0.2	0.8 ± 0.1	< 0.001
Proportional number per plot (%)	1.5 ± 0.3	2.4 ± 0.4	< 0.05	8.3 ± 0.8	4.8 ± 0.4	< 0.001
Cover per plot (%)	1.6 ± 0.7	2.7 ± 1.2	n.s.	21.3 ± 3.6	4.4 ± 0.7	< 0.001
Neophytes						
Number per plot	0.4 ± 0.1	0.5 ± 0.1	n.s.	1.4 ± 0.1	0.9 ± 0.1	< 0.001
Proportional number per plot (%)	1.4 ± 0.3	2.1 ± 0.4	n.s.	7.7 ± 0.8	5.0 ± 0.4	< 0.01
Cover per plot (%)	1.5 ± 0.7	2.5 ± 1.3	n.s.	21.7 ± 3.6	4.6 ± 0.7	< 0.001
Beta diversity (multivariate dispersion)	7.3 ± 0.1	6.2 ± 0.1	< 0.001	7.4 ± 0.2	6.4 ± 0.1	< 0.001

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

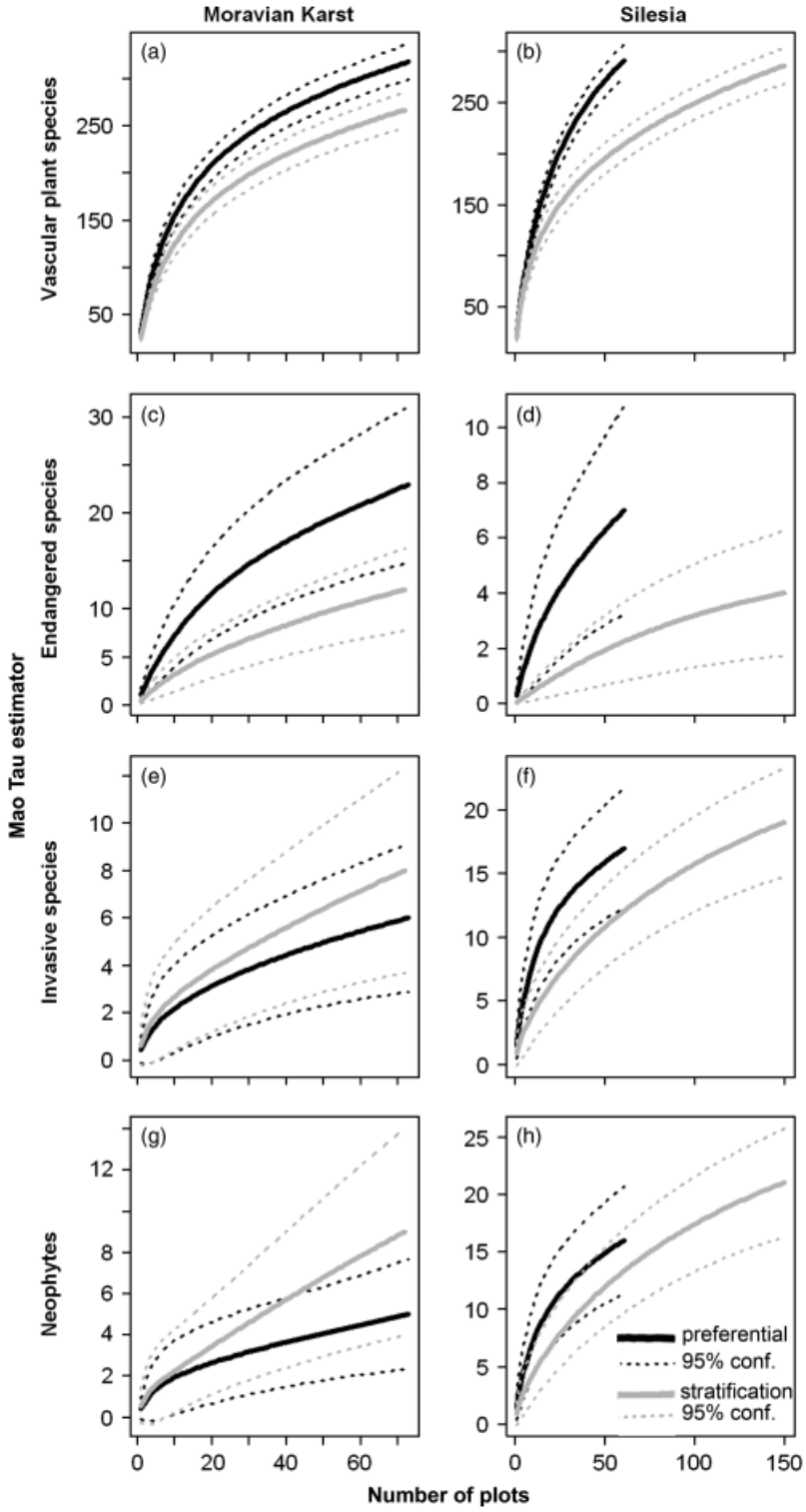

Fig. 2. Sample-based rarefaction curves for the total numbers of vascular plant species (**a**, **b**), endangered species (**c**, **d**), invasive species (**e**, **f**) and neophytes (**g**, **h**).

identical mean number of species per plot. Similarly, Diekmann et al. (2007) found no significant differences in species richness between preferential and random samples of German forests, except for plots sizes $< 4\,m^2$. In specific situations, preferentially sampled plots may have an even lower mean species richness than randomly

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

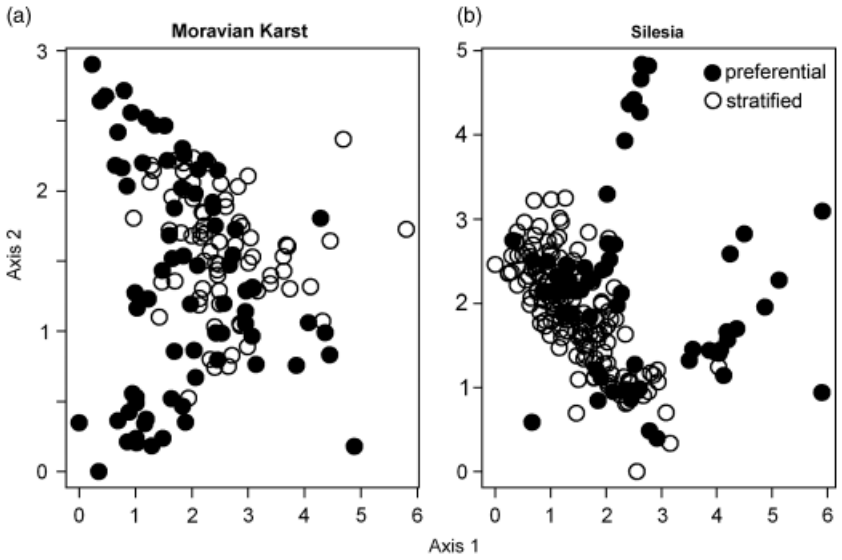

Fig. 3. Ordination diagrams of detrended correspondence analyses computed jointly for preferential and stratified-random data sets for each of the two study areas. Axes are scaled in standard deviation units.

sampled plots (Botta-Dukát et al. 2007). These results indicate that the estimates of species richness can be biased in preferentially sampled vegetation plots, but that this is often not the case. However, if plots from vegetation databases are to be used for the analyses of species richness patterns, only one single plot size or a narrow range of plot sizes should be selected to avoid the species–area effects (Dengler et al. 2009). In addition, plots with larger sizes than those commonly used in phytosociology should also be avoided, because such plots were often used by phytosociologists to sample species-poor vegetation, which leads to biased estimates of species richness (Chytrý 2001).

The over-estimation of the number of endangered species in preferentially sampled plots compared to the stratified-random sample, which was consistently found in both areas (Table 2), is not unexpected because researchers often tend to prefer sites that are valuable from a conservation viewpoint. Also, large vegetation databases contain plots obtained from studies focussed on particular endangered species, in which vegetation plots were recorded only at sites where this species occurred in order to show its relationships to habitats or vegetation types. Vegetation plots stored in large databases are therefore probably not an appropriate data source for estimating number, proportional number or percentage cover of endangered species, at least in Central European forests.

Comparisons of the number, proportional number and percentage cover of alien species (invasive species and neophytes) per plot yielded inconsistent results. In the Moravian Karst, where the representation of aliens in forest vegetation was generally lower, the estimates seem to be unaffected by the sampling method. In contrast, the preferential data set from Silesia contained more aliens per plot than the stratified-random data set from the same area. This rather surprising outcome is caused by the uneven spatial distribution of plots in the preferential data set. In both preferential data sets, plots were spatially more clustered than those in the stratified-random samples. In the Silesian data set, a large number of plots were from forests near the city of Ostrava where the concentration of aliens was high, probably due to more frequent disturbance and greater propagule pressure. These results indicate that estimates of the level of invasion from vegetation databases (e.g. Chytrý et al. 2008; Vilà et al. 2010) may be only weakly affected by researcher preference for sites that are not invaded, at least in forest vegetation. This result is contrary to our expectation that data sets from the database would contain fewer alien species than those sampled in the stratified-random way, because of the suspected tendency of phytosociologists to select weakly invaded or non-invaded sites while sampling natural forests in a preferential way.

Likewise, for the mean numbers of species per plot, the most striking differences in the total numbers of species in the entire data sets were found for the number of all vascular plant species (Fig. 2a and b) and endangered species (Fig. 2c and d). Rarefaction curves are higher if either alpha diversity (number of vascular plant species per plot) or beta diversity (between-plot differences in species composition) or both are higher. Our finding for the total number of vascular plant species agrees with that

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

of Hédl (2007), whose rarefaction curve from the preferential data set was also slightly higher than the curve computed from the geographically stratified-random data set. Similarly, Diekmann et al. (2007) found a higher species number in a data set with plots sampled preferentially in comparison with plots sampled randomly. Although these results indicate that preferentially sampled plots are biased in terms of parameter estimation (i.e. estimation of the mean local number of all or endangered species), preferential sampling is, in this case, the more efficient method for estimation of total species richness found in the study area.

Moreover, the patterns in ordination diagrams showed differences between the preferential and stratified-random data sets (Fig. 3). Several plots sampled preferentially tended to be positioned on the margins of the plot scatters, representing vegetation types with peculiar species composition, while the plots sampled stratified-randomly were concentrated in the middle-left part of the diagrams. (Note that a slightly different pattern was demonstrated by Roleček et al. (2007), who used the same but a less restricted data set; here we removed plots recorded before 1980 or recorded in early spring.) This indicates that the stratified-random data sets contained more homogeneous vegetation with a predominance of common species, whereas some plots from the preferential data sets were located in more extreme environmental conditions and thus contained more specialist species. Roleček et al. (2007) explained this pattern by the poor ability of the environmental stratification, based on commonly available digital maps, to identify those habitats that contain rare vegetation types. These vegetation types were clearly identified by preferential sampling, which in contrast tended to avoid sites with "average" species composition. Therefore, Roleček et al. (2007) recommended a two-step sampling scheme for the purpose of documentation of maximum variability of vegetation, beginning with stratified-random sampling, and followed by preferential sampling of rare sites with peculiar species composition that were not sampled according to the stratified-random scheme. This pattern also indicates that some vegetation types can be missed by stratified-random sampling due to the imperfections of environmental data sets used for stratification, such as their coarse resolution (Michaelsen et al. 1994; Holt et al. 2008), absence of data on the variables that are most relevant to plants and subjective selection of strata by the researchers (Knollová et al. 2005). Therefore, preferential sampling can be a useful addition to stratified-random sampling when new data are to be obtained in the field. Preferential sampling within strata or a combination of stratified-random sampling and "sequential sampling of unusual environments" (Palmer et al. 2002) can also provide a promising option. The adequacy of such a scheme can be assessed at the beginning or during sampling using some of the five methods proposed in Neldner et al. (1995), or the final data set can be stratified and resampled randomly within strata in order to avoid misinterpretations due to uneven spatio-temporal distribution (Knollová et al. 2005; Haveman & Janssen 2008). Nevertheless, when data from the databases are analysed, researchers should exercise caution for potentially poor representation or absence of some vegetation types.

Conclusions

We demonstrated that although vegetation data stored in large databases are influenced by researcher subjectivity in plot positioning, not all of their properties are necessarily biased; at least they may not differ significantly from the corresponding properties of data sets sampled in a stratified-random way. In our study, the preferentially sampled data sets obtained from a national phytosociological database over-estimated the local number of endangered species and beta diversity relative to estimations based on the stratified-random data sets. Nevertheless, estimates of the local number of all species and representation of alien species did not differ between the two data sets, at least in one of the two study areas. This result, together with some previous case studies that compared results from preferential sampling with those from random or systematic sampling, suggests that in some cases databases can provide correct estimates of vegetation parameters, although in other cases such estimates can be biased. Our comparison of positions of plots from both sampling techniques in ordination diagrams indicates that, for the description of diversity in species composition using ordination (and probably also classification), preferentially sampled data sets are still valuable because stratified-random sampling across a large area may lead to overlooking some rare vegetation types with unusual species composition that tend to be better represented in preferentially sampled data sets.

Large vegetation databases, which typically contain collections of preferentially sampled plots from different sources, will undoubtedly remain a valuable source of data for various analyses. The results based on these analyses may often be very close to results based on the data obtained using a formalized sampling design; however researchers should be cautious of potential pitfalls and use databases only for such analyses where the potential bias due to preferential sampling is limited.

Acknowledgements

We are grateful to Lubomír Tichý for performing the PERMANOVA analysis. This paper benefited from

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

valuable comments on a previous version made by Jörg Ewald, Florian Jansen and Petr Petřík. This study was supported by the Czech Science Foundation (GA ČR 206/09/0329), the Ministry of Education of the Czech Republic (MSM0021622416) and by the research project 526/09/H025.

References

Acosta, A., Carranza, M.L. & Izzi, C.F. 2005. Combining land cover mapping of coastal dunes with vegetation analysis. *Applied Vegetation Science* 8: 133–138.

Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. 2006. Multivariate dispersion as a measure of beta diversity. *Ecology Letters* 9: 683–693.

Austin, M.P. & Heyligers, P.C. 1989. Vegetation survey design for conservation: gradsect sampling of forests in North-eastern New South Wales. *Biological Conservation* 50: 13–32.

Bossard, M., Feranec, J. & Otahel, J. 2000. *CORINE land cover technical guide – addendum 2000*. European Environment Agency, Copenhagen, DK.

Botta-Dukát, Z., Kovács-Láng, E., Rédei, T., Kertész, M. & Garadnai, J. 2007. Statistical and biological consequences of preferential sampling in phytosociology: theoretical considerations and a case study. *Folia Geobotanica* 42: 141–152.

Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C. & Lane, A.M.J. 1996. Land classification for strategic ecological survey. *Journal of Environmental Management* 47: 37–60.

Chiarucci, A. 2007. To sample or not to sample? That is the question ... for the vegetation scientist. *Folia Geobotanica* 42: 209–216.

Chytrý, M. 2001. Phytosociological data give biased estimates of species richness. *Journal of Vegetation Science* 12: 439–444.

Chytrý, M. & Rafajová, M. 2003. Czech National Phytosociological Database: basic statistics of the available vegetation plot-data. *Preslia* 75: 1–15.

Chytrý, M., Pyšek, P., Tichý, L., Knollová, I. & Danihelka, J. 2005. Invasions by alien plants in the Czech Republic: a quantitative assessment across habitats. *Preslia* 77: 339–354.

Chytrý, M., Maskell, L.C., Pino, J., Pyšek, P., Vilà, M., Font, X. & Smart, S.M. 2008. Habitat invasions by alien plants: a quantitative comparison among Mediterranean, sub-continental and oceanic regions of Europe. *Journal of Applied Ecology* 45: 448–458.

Coetzee, J.P., Bredenkamp, G.J. & van Rooyen, N. 1995. Plant communities of the subhumid cool temperate mountain bushveld of the Pretoria–Witbank–Heidelberg area, South Africa. *South African Journal of Botany* 61: 114–122.

Colwell, R.K., Mao, C.X. & Chang, J. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 85: 2717–2727.

da Silva, J.A., Salomão, A.N., Gripp, A. & Leite, E.J. 1997. Phytosociological survey in Brazilian forest genetic reserve of Cacador. *Plant Ecology* 133: 1–11.

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.

Diekmann, M., Kühne, A. & Isermann, M. 2007. Random *vs* non-random sampling: effects on patterns of species abundance, species richness and vegetation–environment relationships. *Folia Geobotanica* 42: 179–190.

Duckworth, J.C., Bunce, R.G.H. & Malloch, A.J.C. 2000. Vegetation gradients in Atlantic Europe: the use of existing phytosociological data in preliminary investigations on the potential effects of climate change on British vegetation. *Global Ecology and Biogeography* 9: 187–199.

Dupré, C., Stevens, C.J., Ranke, T., Bleeker, A., Peppler-Lisbach, C., Gowing, D.J.G., Dise, N.B., Dorland, E., Bobbink, R. & Diekmann, M. 2010. Changes in species richness and composition in European acidic grasslands over the past 70 years: the contribution of accumulative atmospheric nitrogen deposition. *Global Change Biology* 16: 344–357.

Ewald, J. 2002. A probabilistic approach to estimating species pools from large compositional matrices. *Journal of Vegetation Science* 13: 191–198.

Firbank, L.G., Barr, C.J., Bunce, R.G.H., Furse, M.T., Haines-Young, R., Hornung, M., Howard, D.C., Sheail, J., Sier, A. & Smart, S.M. 2003. Assessing stock and change in land cover and biodiversity in GB: an introduction to countryside survey 2000. *Journal of Environmental Management* 67: 207–218.

Gimaret-Carpentier, C., Pélissier, R., Pascal, J.-P. & Houllier, F. 1998. Sampling strategies for the assessment of tree species diversity. *Journal of Vegetation Science* 9: 161–172.

Goedickemeier, I., Wildi, O. & Kienast, F. 1997. Sampling for vegetation survey: some properties of a GIS-based stratification compared to other statistical sampling methods. *Coenoses* 12: 43–50.

Gotelli, N.J. & Colwell, R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4: 379–391.

Grabherr, G., Reiter, K. & Willner, W. 2003. Towards objectivity in vegetation classification: the example of the Austrian forests. *Plant Ecology* 169: 21–34.

Haveman, R. & Janssen, J.A.M. 2008. The analysis of long-term changes in plant communities using large databases: the effect of stratified resampling. *Journal of Vegetation Science* 19: 355–362.

Hédl, R. 2007. Is sampling subjectivity a distorting factor in surveys for vegetation diversity? *Folia Geobotanica* 42: 191–198.

Hennekens, S.M. & Schaminée, J.H.J. 2001. TURBOVEG, a comprehensive database management system for vegetation data. *Journal of Vegetation Science* 12: 589–591.

Holeksa, J. & Woźniak, G. 2005. Biased vegetation patterns and detection of vegetation changes using a phytosociological databases. A case study in the forest of the Babia Góra

National Park (the West Carpathians, Poland). *Phytocoenologia* 35: 1–18.

Holt, E.A., McCune, B. & Neitlich, P. 2008. Spatial scale of GIS-derived categorical variables affects their ability to separate sites by community composition. *Applied Vegetation Science* 11: 421–430.

Holub, J. & Procházka, F. 2000. Red list of vascular plants of the Czech Republic – 2000. *Preslia* 72: 187–230.

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.

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

Kershaw, K.A. & Looney, J.H.H. 1985. *Quantitative and dynamic plant ecology*. Edward Arnold, London, UK.

Knollová, I., Chytrý, M., Tichý, L. & Hájek, 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.

Kodym, O., Patočka, B. & Cihelka, M. 1966. *Geologický atlas ČSSR (Geological atlas of Czechoslovakia)*. Ústřední geologický ústav, Praha, CZ.

Lájer, K. 2007. Statistical tests as inappropriate tools for data analysis performed on non-random samples of plant communities. *Folia Geobotanica* 42: 115–122.

Láníková, D. & Lososová, Z. 2009. Rocks and walls: natural *versus* secondary habitats. *Folia Geobotanica* 44: 263–280.

Lepš, J. & Šmilauer, P. 2007. Subjectively sampled vegetation data: don't throw out the baby with the bath water. *Folia Geobotanica* 42: 169–178.

Lloret, F., Solé, A., Pino, J., Vayreda, J., Font, X. & Terradas, J. 2009. Patterns of species impoverishment in managed forests of Catalonia (NE Spain). *Journal of Vegetation Science* 20: 675–685.

Lososová, Z. & Simonová, D. 2008. Changes during the 20th century in species composition of synanthropic vegetation in Moravia (Czech Republic). *Preslia* 80: 291–305.

Lososová, Z., Chytrý, M., Cimalová, Š., Kropáč, Z., Otýpková, Z., Pyšek, P. & Tichý, L. 2004. Weed vegetation of arable land in Central Europe: gradients of diversity and species composition. *Journal of Vegetation Science* 15: 415–422.

Lvončík, S. 2006. *Lesní vegetace Slezské pahorkatiny a Ostravské pánve (Forest vegetation of the Silesian Hills and Ostrava Basin)*. MSc. thesis, Masaryk University, Brno, CZ.

Margules, C.R. & Austin, M.P. 1994. Biological models for monitoring species decline: the construction and use of databases. *Philosophical Transactions of The Royal Society, B – Biological Sciences* 344: 69–75.

Michaelsen, J., Schimel, D.S., Friedl, M.A., Davis, F.W. & Dubayah, R.C. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science* 5: 673–686.

Michalcová, D. 2009. *Diverzita lesní vegetace Moravského krasu a přilehlého území (Diversity of forest vegetation of Moravian Karst and adjacent areas)*. MSc. thesis, Masaryk University, Brno, CZ.

Neldner, V.J., Crossley, D.C. & Cofinas, M. 1995. Using geographic information systems (GIS) to determine the adequacy of sampling in vegetation surveys. *Biological Conservation* 73: 1–17.

Neuhäuslová, Z., Moravec, J., Chytrý, M., Sádlo, J., Rybníček, K., Kolbek, J. & Jirásek, J. 1997. *Mapa potenciální přirozené vegetace České republiky 1: 500 000 (Map of potential natural vegetation of the Czech Republic 1: 500 000)*. Botanický ústav AV ČR, Průhonice, CZ.

Økland, R.H. 2007. Wise use of statistical tools in ecological field studies. *Folia Geobotanica* 42: 123–140.

Olsen, A.R., Sedransk, J., Edwards, D., Gotway, C.A., Liggert, W., Rathbun, S., Reckhow, K.H. & Young, L.J. 1999. Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment* 54: 1–45.

Ortega, M., Elena-Roselló, R. & García del Barrio, J.M. 2004. Estimation of plant diversity at landscape level: a methodological approach applied to three Spanish rural areas. *Environmental Monitoring and Assessment* 95: 97–116.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S. & Wohlgemuth, T. 2002. Quantitative tools for perfecting species lists. *Environmetrics* 13: 121–137.

Podani, J. 1984. Spatial processes in the analysis of vegetation: theory and review. *Acta Botanica Hungarica* 30: 75–118.

Pyšek, P., Jarošík, V., Kropáč, Z., Chytrý, M., Wild, J. & Tichý, L. 2005. Effects of abiotic factors on species richness and cover in Central European weed communities. *Agriculture, Ecosystems and Environment* 109: 1–8.

Pyšek, P., Sádlo, J. & Mandák, B. 2002. Catalogue of alien plants of the Czech Republic. *Preslia* 74: 97–186.

Roleček, J., Chytrý, M., Hájek, M., Lvončík, S. & Tichý, L. 2007. Sampling design in large-scale vegetation studies: do not sacrifice ecological thinking to statistical purism! *Folia Geobotanica* 42: 199–208.

Šamonil, P., Polesná, K. & Unar, P. 2009. Plant community variability within potential natural vegetation units: a case study from the Bohemian Karst. *Journal of Forest Science* 55: 485–501.

Schaminée, J.H.J., Hennekens, S.M., Chytrý, M. & Rodwell, J.S. 2009. Vegetation-plot data and databases in Europe: an overview. *Preslia* 81: 173–185.

Siebert, S.J., van Wyk, A.E., Bredenkamp, G.J. & du Plessis, F. 2002. The grasslands and wetlands of the Sekhukhuneland centre of plant endemism, South Africa. *Bothalia* 32: 211–231.

Skalický, V. 1988. Regionálně fytogeografické členění (regional phytogeographical division). In: Hejný, S. & Slavík, B. (eds.) *Květena České socialistické republiky 1 (Flora of the Czech Socialist Republic 1)*, pp. 103–121. Academia, Praha, CZ.

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science

Stohlgren, T.J., Chong, G.W., Kalkhan, M.A. & Schell, L.D. 1997a. Rapid assessment of plant diversity patterns: a methodology for landscapes. *Environmental Monitoring and Assessment* 48: 25–43.

Stohlgren, T.J., Chong, G.W., Kalkhan, M.A. & Schell, L.D. 1997b. Multiscale sampling of plant diversity: effects of minimum mapping unit size. *Ecological Applications* 7: 1064–1074.

Stohlgren, T.J., Coughenour, M.B., Chong, G.W., Binkley, D., Kalkhan, M.A., Schell, L.D., Buckley, D.J. & Berry, J.K. 1997c. Landscape analysis of plant diversity. *Landscape Ecology* 12: 155–170.

ter Braak, C.J.F. & Šmilauer, P. 2002. *CANOCO reference manual and CanoDraw for Windows user's guide. Software for Canonical Community Ordination (version 4.5)*. Biometris, Wageningen NL & České Budějovice, CZ.

Tichý, L. 2002. JUICE, software for vegetation classification. *Journal of Vegetation Science* 13: 451–453.

van Kley, J.E. & Schaminée, J.H.J. 2003. Large-scale ordination and gradient analysis of salt-marsh communities in the Netherlands in the light of the Dutch National Vegetation Classification. *Phytocoenologia* 33: 335–347.

Vesecký, A., Petrovič, Š., Briedoň, V. & Karský, V. (eds.) 1958. *Atlas podnebí Československé republiky (Climate atlas of the Czechoslovak Republic)*. Ústřední správa geodesie a kartografie, Praha, CZ.

Vilà, M., Pino, J., Montero, A. & Font, X. 2010. Are island plant communities more invaded than their mainland counterparts? *Journal of Vegetation Science* 21: 438–446.

Westhoff, V. & van der Maarel, E. 1978. The Braun-Blanquet approach. In: Whittaker, R.H. (ed.) *Classification of plant communities*, pp. 287–399. W. Junk, The Hague, NL.

Wilson, J.B. 2007. Priorities in statistics, the sensitive feet of elephants, and don't transform data. *Folia Geobotanica* 42: 161–167.

Doi: 10.1111/j.1654-1103.2010.01249.x © 2011 International Association for Vegetation Science