Journal of Vegetation Science 16: 479-486, 2005
© IAVS; Opulus Press Uppsala.

Stratified resampling of phytosociological databases: some strategies for obtaining more representative data sets for classification studies

Knollová, Ilona*; Chytrý, Milan[1]; Tichý, Lubomír[2] & Hájek, Ondřej[3]

Department of Botany, Masaryk University, Kotlářská 2, CZ-611 37 Brno, Czech Republic;
E-mail [1]chytry@sci.muni.cz; [2]tichy@sci.muni.cz; [3]ohajek@sci.muni.cz
** Corresponding author; Fax +420 541211214; E-mail ikuzel@sci.muni.cz*

Abstract

Question: The heterogeneous origin of the data in large phytosociological databases may seriously influence the results of their analysis. Therefore we propose some strategies for stratified resampling of such databases, which may improve the representativeness of the data. We also explore the effects of different resampling options on vegetation classification.

Methods: We used 6050 plot samples (relevés) of mesic grasslands from the Czech Republic. We stratified this database using (1) geographical stratification in a grid; (2) habitat stratification created by an overlay of digital maps in GIS; (3) habitat stratification with strata defined by traditional phytosociological associations; (4) habitat stratification by numerical classification and (5) habitat stratification by Ellenberg indicator values. Each time we resampled the database, taking equal numbers of relevés per stratum. We then carried out cluster analyses for the resampled data sets and compared the resulting classifications using a newly developed procedure.

Results: Random resampling of the initial data set and geographically stratified resampling resulted in similar classifications. By contrast, classifications of the resampled data sets that were based on habitat stratifications (2–5) differed from each other and from the initial data set. Stratification 2 resulted in classifications that strongly reflected environmental factors with a coarse grain of spatial heterogeneity (e.g. macroclimate), whereas stratification 5 resulted in classifications emphasizing fine-grained factors (e.g. soil nutrient status). Stratification 3 led to the most deviating results, possibly due to the subjective nature of the traditional phytosociological classifications.

Conclusions: Stratified resampling may increase the representativeness of phytosociological data sets, but different types of stratification may result in different classifications. No single resampling strategy is optimal or superior: the appropriate stratification method must be selected according to the objectives of specific studies.

Keywords: Ellenberg indicator values; Geographical stratification; GIS; Habitat stratification; Phytosociological data; Plant community; Relevé; Sampling design; Survey.

Abbreviations: ASS = Phytosociological association; ELL = Ellenberg indicator values; GEO = Geographical stratification; GIS = Geographical information system; NUM = Numerical classification; RAN = Random resampling.

Introduction

Vegetation surveys and associated analyses of habitat types provide basic information for decision making in nature conservation, environmental management and landscape planning (Haila & Margules 1996). Correct estimates of biodiversity or natural resource quality of an area are critically dependent on the sampling design of such surveys. Appropriate sampling designs, however, may differ from the classical principles of probability sampling (Cochran 1977). While probability sampling is mainly concerned with estimating parameters of the target statistical populations (Gregoire 1998), the surveys aimed at vegetation or habitat typification focus on sampling maximum variation, which is not achieved with simple random or systematic arrangement of sampling units over the study area. Such sampling schemes tend to yield high numbers of replicates of common habitats whilst rare habitats may be under-represented, or even missing (Gillison & Brewer 1985; Kenkel et al. 1989; Bunce et al. 1996).

As a specific case of maximum variation sampling (Patton 1990), modern biological surveys of large areas increasingly use environmentally stratified sampling designs (Goedickemeier et al. 1997; Gimaret-Carpentier et al. 1998; Olsen et al. 1999; Yoccoz et al. 2001) or different kinds of adaptive sampling strategies (Thompson & Seber 1996; Stein & Ettema 2003). Strata are usually defined on the basis of environmental variables that have been shown in previous studies to influence species composition. If there is sufficient information on the spatial pattern of such variables, e.g. digital climatic or geological maps, appropriate strata for sampling can be created by overlaying these maps in geographical information systems (GIS). Sampling designs that use random arrangement of relevés within such strata are assumed to maximize the between-relevé variation and to give the same chance of common and rare habitat types being sampled. Examples of large area vegetation surveys based on environmentally stratified sampling plans include British

Countryside Survey (Bunce et al. 1996; Smart et al. 2003), the Austrian forest survey (Grabherr et al. 2003), or regional vegetation surveys in the Rocky Mountains (Stohlgren et al. 1997) and New South Wales (Cawsey et al. 2002). To reduce survey costs, the gradient directed transect or gradsect method was proposed (Gillison & Brewer 1985; Austin & Heyligers 1989; Wessels et al. 1998), in which only a few transects are sampled, located along the steepest environmental gradients in the area, instead of sampling the entire area.

Vegetation surveys of large areas are expensive and time consuming (Olsen et al. 1999), even if cost-effective methods such as gradsect are used. For this reason, the required information is often extracted from existing data (Haila & Margules 1996). In recent years, many electronic phytosociological databases have been compiled, mainly in Europe, but also elsewhere (Mucina et al. 2000; Ewald 2001; Hennekens & Schaminée 2001; Wiser et al. 2001; Chytrý & Rafajová 2003). Ewald (2001) estimated that ca. one million relevés had already been computerized and stored in electronic databases worldwide. However, relevés contained in the databases are usually heterogeneous, as they originate from various sources and were sampled for different purposes, often without any sampling plans. They are usually biased towards sites that are easily accessible or of special interest (e.g. nature reserves) and towards habitats that are attractive to the researchers (e.g. those containing rare species or high species richness). In contrast, some geographical areas or habitats may be under-represented.

Therefore, we propose that the use of existing data sets should be preceded by stratified resampling that would increase their representativeness. Such resampling cannot remedy the lack of data from some areas or habitats (these can only be obtained through additional field sampling), but it can improve the quality of the data sets by removing redundancy due to oversampling of some areas or habitats.

There are two basic approaches to resampling of the existing databases: (1) geographical stratification with strata corresponding to *a priori* delimited land areas such as quadrats in a geographical grid; (2) habitat stratification based on partitioning of environmental space into relatively homogeneous strata. There are several options for achieving habitat stratification. One possibility, similar to the planning of a new survey, is defining strata using a GIS overlay of a few digital maps of selected environmental variables that are likely to be the most relevant to plant distribution. The accuracy of GIS data, however, depends on the level of detail of the original maps, which is often coarser than the fine scale habitat mosaics that control plant distribution patterns. Unlike digital maps of large areas, species composition records in databases reflect this fine-scale environmental variation. Therefore, a promising option in the context of database resampling is habitat stratification based directly on the species composition of the available relevés. This can be achieved in several ways:

1. Use of traditional phytosociological associations as strata. In areas with a phytosociological tradition, such as many parts of Europe, relevés in the databases are often assigned to associations, which are based on specific floristic composition and implicitly contain expert knowledge of local habitats.
2. Construction of numerical classifications of all relevés in the database and use of the resulting clusters as strata.
3. Application of species indicator values, such as Ellenberg indicator values in central Europe (Ellenberg et al. 1992), to stratify relevés by relative values of estimated environmental variables.

The purpose of this paper is to compare different types of stratified resampling of phytosociological databases, with respect to the subsequent classifications of real data.

Material and Methods

Initial non-stratified data set

As an initial non-stratified data set, we used 6050 relevés of meadows and mesic pastures from the Czech Republic, stored in the National Phytosociological Database (Chytrý & Rafajová 2003) in the TURBOVEG program (Hennekens & Schaminée 2001). Only relevés with an expert assignment (as taken from the original sources) to the phytosociological class of European meadows, *Molinio-Arrhenatheretea*, were included in this data set. The plot size of relevés in phytosociological databases varies considerably (Chytrý & Otýpková 2003), and we tried to limit this variation by selecting only those from plots of 4 - 100 m^2. Each relevé contained a species list with cover estimates on the Braun-Blanquet or Domin scales (Kent & Coker 1992). To facilitate simultaneous use of all relevés, cover values from original scales were transformed into percentages. Cryptogam records were deleted, as they were missing in several relevés and the quality of their recording varied. Data handling and stratified resampling were conducted with a modified version of the JUICE 6.3 program (Tichý 2002).

Stratified resampling of the initial data set

The initial data set was stratified in the following ways:

Geographical stratification (GEO)

GEO divided the country into 9834 quadrats of a geographical grid of 2.5 longitudinal and 1.5 latitudinal minutes, i.e. ca. 3 km × 2.8 km. 1626 of these quadrats contained at least one relevé.

Habitat stratification by overlay of digital maps (GIS)

GIS defined strata by combining three digital maps: mean annual temperature, mean annual precipitation and potential natural vegetation (Neuhäuslová et al. 1997). The latter was used as a representative of the joint effect of bedrock, soil type and soil moisture. We used seven categories for temperature (< 4, 4-5, 5-6, 6-7, 7-8, 8-9, > 9°C), six categories for precipitation (< 500, 500-600, 600-700, 700-900, 900-1200, >1200 mm) and 17 units of potential natural vegetation as defined at the higher hierarchical level of the map. The overlay of these categories yielded 229 strata of which 118 contained at least one relevé.

Habitat stratification by associations (ASS)

ASS defined strata by assigning relevés to phytosociological associations corresponding to the standard phytosociological classification of the Czech Republic (Moravec et al. 1995). According to the expert assignments by the researchers who sampled the plots in the field, there were 56 associations of meadows in the initial data set.

Habitat stratification by numerical classification (NUM)

NUM defined strata as clusters that resulted from numerical classification of the initial non-stratified data set. Cluster analysis from the PC-ORD 4 package (McCune & Mefford 1999) was used. Percentage cover data were square-root transformed, distances between all pairs of relevés were computed as a relative Euclidean (chord) distance and the β-flexible method with $\beta = -0.3$ was used to create a dendrogram. Based on the resulting dendrogram, the relevés were divided into 60 groups.

Habitat stratification by Ellenberg indicator values (ELL)

ELL defined strata as subdivisions of six dimensional environmental space whose axes were Ellenberg indicator values for light, temperature, continentality, moisture, soil reaction and nutrients. For most vascular plant species of Central Europe, Ellenberg et al. (1992) tabulated ordinal values that expressed their relationships to these six factors. By calculating mean species values within each relevé, we obtained a single value for each relevé and each factor. Relevés were then stratified into three strata for each factor, corresponding to low, medium and high levels of the factor, so that each stratum contained an equal number of relevés. For the combined two-way stratification with two factors, we would obtain nine strata; in our case with six factors, we obtained $3^6 = 729$ strata, of which 573 contained at least one relevé.

Finally, we also created stratifications by combining the geographical stratification with different variants of the habitat stratifications (GIS-GEO, ASS-GEO, NUM-GEO, ELL-GEO). In all cases, habitat stratification was carried out first, with a subsequent geographical stratification within the habitat based strata.

From each of the nine stratified data sets, we resampled 1210 relevés, i.e. 20% of the initial data set. Resampling was random within strata and aimed at obtaining as even a number of relevés within strata as possible. First, one relevé was randomly selected from each stratum that contained at least one relevé. Then, another relevé was randomly selected from each stratum that contained at least two relevés, etc. When the number of resampled relevés approached the limit of 1210, random selection of the next relevé was only done in randomly selected strata and the process stopped when 1210 relevés were resampled. For the sake of comparison, we also made a random resampling (RAN) of 1210 relevés from the initial data set. Each resampling was repeated ten times, which yielded 100 resampled data sets.

Classification

Each of the resampled data sets was classified by cluster analysis. The options were the same as in the habitat stratification by numerical classification (see above). For each resampled data set we took eight clusters at the highest level of the dendrogram hierarchy. The results of each classification were summarized in a synoptic table, which contained eight columns with frequency occurrences for each species. The diagnostic value (fidelity) of each species for each column was calculated, using the phi coefficient of association (Sokal & Rohlf 1995), which quantified the degree of association between species and clusters (Chytrý et al. 2002). As the unequal numbers of relevés included in individual clusters resulted in higher Φ values for larger clusters, each of the eight clusters was virtually adjusted to the size of 151 relevés (i.e. 1/8 of the data set size), while holding the percentage occurrences of species within and outside the target clusters the same as in the original data set. This virtual equalization of the cluster sizes could result in high Φ values for some rare species that occurred in relevés of small clusters mainly by chance, even though their association with such clusters was not statistically significant because of their rarity. Therefore, we additionally tested the statistical significance of the concentration of each species in each cluster by Fisher's exact test, using actual cluster sizes. The Φ value was arbitrarily set to zero for all species whose occurrence concentration in a cluster was not significantly different from random at $P < 0.001$. Species with positive Φ values for particular clusters were then considered as diagnostic species of these clusters.

Comparison of classifications of resampled data sets

Classifications of different resampled data sets could not be compared by conventional methods, because these data sets only partly shared the same relevés. Therefore, we compared classifications through the similarity between sets of diagnostic species of particular columns of synoptic tables, assuming that clusters with similar diagnostic species represent corresponding vegetation types.

Euclidean distances between all pairs of columns of synoptic tables across all classified resampled data sets were calculated, based on phi coefficients. Only values of $\Phi > 0$ were considered. The resulting distance matrix had 800 rows and 800 columns (100 classified resampled data sets, each with eight columns). This matrix was further simplified into a 100 × 100 matrix of distances between the classifications of the resampled data sets. These distances were calculated in the following way:

The distance $d(X_i; Y)$ between the i-th column of the synoptic table of classified resampled data set X and all eight columns of the synoptic table of classified resampled data set Y was calculated as:

$$d\left(X_i;Y\right)=\frac{\left(n_Y-1\right)\cdot \min_j\left(ED_{ij}\right)}{\left(\sum_j ED_{ij}\right)-\min_j\left(ED_{ij}\right)} \tag{1}$$

where i and j denote columns of synoptic tables of classified resampled data sets X and Y, respectively; n_Y is the number of columns in classified resampled data set Y (eight columns in our case); ED_{ij} is Euclidean distance between phi coefficients in columns X_i and Y_j; $\min_j (ED_{ij})$ is the shortest of the distances between the column X_i and each of the columns of data set Y. Subsequently, the distance between classified resampled data sets X and Y was calculated by averaging distances for individual columns X_i of classified resampled data set X:

$$d\left(X;Y\right)=\left[\sum_i d\left(X_i;Y\right)\right]\Big/n_X \tag{2}$$

Such distance is an asymmetric measure, because the distances $d(X; Y)$ and $d(Y; X)$ differ. Therefore the same procedure was applied in the opposite direction:

$$d\left(Y_j;X\right)=\frac{\left(n_X-1\right)\cdot \min_i\left(ED_{ji}\right)}{\left(\sum_j ED_{ji}\right)-\min_i\left(ED_{ji}\right)} \tag{3}$$

$$d\left(Y;X\right)=\left[\sum_j d\left(Y_j;X\right)\right]\Big/n_Y \tag{4}$$

Finally the symmetric distance $D(X; Y)$ between the classified resampled data sets X and Y was calculated as the mean of $d(X; Y)$ and $d(Y; X)$:

$$D\left(X;Y\right)=\left[d\left(X;Y\right)+d\left(Y;X\right)\right]\Big/2 \tag{5}$$

These symmetric distances were calculated between all pairs of classified resampled data sets. They were arranged in a 100 × 100 distance, which was further simplified with principal coordinates analysis (PCoA), using the CANOCO 4.5 program (ter Braak & Šmilauer 2002).

Additionally, we compared classifications of the resampled data sets using Ellenberg indicator values. We calculated mean indicator values for each relevé and tested the null hypothesis of no difference in Ellenberg indicator values among clusters resulting from each classification (ANOVA from the STATISTICA 7 program; www.statsoft.com). The differences were significant ($P < 0.05$) for all Ellenberg indicator values in all data sets, but there were conspicuous differences in the values of the coefficient of determination (R^2), which specified the fit of the ANOVA model. The matrix of R^2-values for six Ellenberg indicator values and 100 resampled data sets was therefore analysed with principal components analysis (PCA on correlation matrix; CANOCO 4.5 package, ter Braak & Šmilauer 2002).

Results

The PCoA ordination diagram (Fig. 1) shows the similarity of classifications among differently resampled data sets. The randomly resampled, non-stratified data set (RAN) occupies a central position. From this position the other data sets diverge into two directions. One direction contains the geographical stratification (GEO), overlapping with all combinations of geographical and habitat stratifications. In the opposite direction there are habitat stratifications by overlay of digital maps (GIS) and by numerical classification (NUM). Isolated positions at marginal parts of the diagram are occupied by habitat stratification by associations (ASS, situated close to NUM) and habitat stratification by Ellenberg indicator values (ELL, situated close to GEO and combinations with GEO).

The PCA ordination diagram (Fig. 2) shows that resampling based on the geographical stratification (GEO) resulted in classifications that poorly reflected any ecological gradient, in particular the temperature gradient. Habitat stratification based on digital maps (GIS) led to classifications with clusters of contrasting temperature requirements but similar nutrient requirements, whereas the habitat stratification by Ellenberg indicator values (ELL) produced the opposite pattern. Clusters resulting from the classifications of the resampled data sets based on habitat stratification by associations (ASS) were differentiated mainly due to underlying temperature gradient and to some extent to soil reaction, continentality and light availability gradients. Classifications based on the other stratifications were placed in the middle of the ordination diagram; this indicates that these stratifications did not produce resampled data sets with extremely high or low heterogeneity with respect to some particular environmental gradient.

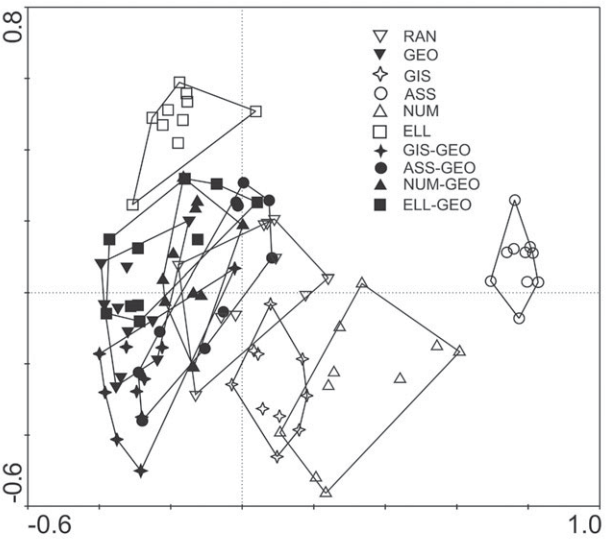

Fig. 1. Similarity of classifications resulting from differently resampled data sets, expressed through similarity of species composition among clusters of different classifications. Ordination was prepared using principal coordinates analysis (PCoA).

Discussion

Multivariate analyses of large phytosociological databases are increasingly used for reassessment of national or regional vegetation typologies (Rodwell 1990-2000; Schaminée et al. 1995-1999; Diekmann et al. 1999; Grabherr et al. 2003). Although the results of such analyses are often robust, most databases contain a non-representative sample of vegetation of the study area. This may occasionally result in incompatible classifications even for adjacent areas that are assumed to contain nearly identical vegetation types (Bruelheide & Chytrý 2000). With the recently emerging large databases (Ewald 2001), representativeness can be improved by techniques such as stratified resampling, which has been used in several recent studies (Ewald 2002, 2003; Chytrý et al. 2003; Kočí et al. 2003; Havlová et al. 2004, Lososová et al. 2004).

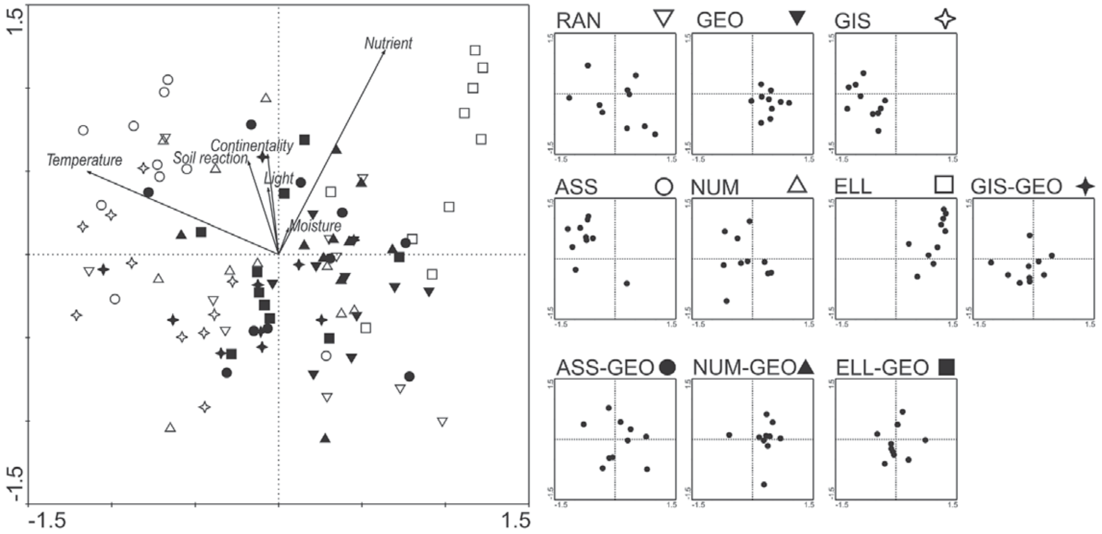

Fig. 2. Similarity of classifications resulting from differently resampled data sets, expressed through the between-cluster variation in Ellenberg indicator values. Ordination was prepared using principal components analysis (PCA). Small ordination diagrams show separate categories of the large diagram.

Our analyses have shown that the randomly resampled data sets (RAN) and the resampled data sets based on geographical stratification (GEO) or its combination with some of the habitat stratifications (GIS-GEO, ASS-GEO, NUM-GEO, ELL-GEO) tend to produce similar classifications (Fig. 1). Resampling based on the pure geographical stratification seems to produce classifications with poor ecological differentiation among clusters. In our case, the poorest differentiation with respect to temperature (Fig. 2) was probably due to the fact that ca. 50% of the country's area is within a narrow altitudinal range of 200 - 500 m (Chytrý & Rafajová 2003), and relevés from these altitudes predominate in the geographically resampled data sets. These data sets were also poorly differentiated with respect to soil reaction, reflecting the rarity of base-rich soils in the country, so the geographically stratified resampling increases the proportion of relevés from acidic soils and homogenizes the data set. The effect of geographical stratification, however, is dependent on habitat heterogeneity within the study area. Resampled data sets based on pure geographical stratification would probably better reflect environmental gradients in heterogeneous rather than homogeneous areas.

Unlike the geographical stratification, habitat stratifications led to classifications which were different both from each other and from the randomly resampled initial data set. The data sets resampled within strata based on the overlay of digital maps (GIS) produced classifications that were most similar to those of randomly resampled data sets (RAN) (Fig. 1), however, they strongly differed in temperature requirements of the resulting clusters (Fig. 2). This is not surprising, given that the mean annual temperature was used to define strata in GIS. However, it is striking that the clusters based on GIS stratification were poorly differentiated with respect to soil nutrient status. Obviously the scale of the digital map of potential natural vegetation was too coarse to reflect the fine-scale mosaic of habitats with different nutrient availability. The same result would probably be obtained if we used soil or bedrock maps, because the units of potential natural vegetation are often delineated according to the boundaries between different soil or bedrock types.

Habitat stratification by phytosociological associations (ASS) resulted in classifications that deviated the most from the others (Fig. 1). Their clusters strongly differed with respect to temperature and also to soil reaction (Fig. 2). This pattern may result from the intuitive judgement of the traditional phytosociologists, who perhaps tend to over-emphasize the variation along the altitudinal gradient (lowland vs montane communities) and the soil reaction gradient, although these two gradients may not be the most important for the species composition of the studies.

Habitat stratification by numerical classification (NUM) led to classifications that were intermediate between classifications based on GIS and ASS stratifications (Fig. 1). NUM is an iterative classification technique, in which the first classification is used to provide strata for resampling and the classification of the resampled data set is accepted as a final output. We believe that it has the potential to produce robust results, especially if more than two iterations are used, but further study is needed to explore its performance in more detail.

Habitat stratification by Ellenberg indicator values (ELL) led to different classifications than other habitat stratifications: they were more similar to the classifications based on geographical stratification (Fig. 1) and emphasized the differences in nutrient availability (Fig. 2). This suggests that ELL stratification is perhaps the most efficient for describing vegetation heterogeneity controlled by the environmental factors that form fine-scale mosaics. Downweighting of temperature differences due to this stratification reflects a narrow temperature range within the study area, which is further narrowed by the absence of the studied vegetation type (mesic grassland) in both the warmest and the coolest areas.

There seems to be no single, optimal resampling strategy, as there is no such strategy for field sampling. Hirzel & Guisan (2002) proposed a basic dichotomy between the equal-stratified sampling and the proportional-stratified sampling. The former uses a constant number of sampling sites in each habitat, while the latter selects the number of sites in each habitat proportionally to the habitat area. Our resamplings based on habitat stratifications approximated the equal-stratified sampling, combinations of the habitat and geographical stratification were closer to the proportional-stratified sampling, and the pure geographic stratification was similar to the combination of systematic and random sampling (random selection of relevés within systematically placed quadrats).

The most pronounced effects on the resulting classifications were obtained with pure habitat stratifications. Different kinds of habitat stratification, however, produced contrasting results. GIS stratification led to classifications that emphasized the differences in those environmental factors, which are variable on coarse spatial scales (temperature), while stratifications based on species composition (ASS, NUM, ELL) also reflected factors that tend to be variable on fine scales (e.g. nutrients). This result is also relevant for planning new field surveys, where environmental landscape stratification with GIS is increasingly used as a guide to site selection (Austin & Heyligers 1989; Stohlgren et al. 1997; Grabherr et al. 2003; Smart et al. 2003). Our results suggest that the data from GIS-planned surveys may over-estimate the heterogeneity due to coarse scale environmental factors. By contrast, factors that frequently change at small distances may be under-estimated, even though they may impose a strong control on plant communities, such as nutrients that depend on the position of the meadows in microtopography of river floodplains or on the input of artificial fertilizers in some places. If the heterogeneity due to fine scale factors is of interest, stratified resampling of the existing databases can yield, under certain circumstances, even more representative data sets than new field sampling according to some statistical plan.

In conclusion, stratified resampling can increase the representativeness of phytosociological databases, mainly by removing redundant relevés from oversampled areas or habitats. However, it can hardly provide a fully representative data set, because some areas or habitats are probably under-represented or missing already in the source database. Ideally, resampling should be combined with a gap analysis and subsequent additional field survey. In any case, specific strategies of both database resampling and field sampling should be selected carefully with respect to the aims of individual studies: different strategies may result in classifications, which reflect different environmental gradients.

Acknowledgements. We appreciated helpful comments on the previous version of this paper by Norman Kenkel and three anonymous referees. This study was funded from the projects GAČR 206/05/0020 and MSM 0021622416.

References

Austin, M.P. & Heyligers, P.C. 1989. Vegetation survey design for conservation: gradsect sampling of forests in Northeast New South Wales. *Biol. Conserv*. 50: 13-32.

Bruelheide, H. & Chytrý, M. 2000. Towards unification of national vegetation classifications: a comparison of two methods for analysis of large data sets. *J. Veg. Sci*. 11: 295-306.

Bunce, R.G.H., Barr, C.J., Clarke, R.T., Howard, D.C. & Lane, A.M.J. 1996. Land classification for strategic ecological survey. *J. Environ. Manage*. 47: 37-60.

Cawsey, E.M., Austin, M.P. & Baker, B.L. 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodivers. Conserv*. 11: 2239-2274.

Chytrý, M. & Otýpková, Z. 2003. Plot sizes used for phytosociological sampling of European vegetation. *J. Veg. Sci*. 14: 563-570.

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

Chytrý, M., Tichý, L., Holt, J. & Botta-Dukát, Z. 2002. Determination of diagnostic species with statistical fidelity measures. *J. Veg. Sci*. 13: 79-90.

Chytrý, M., Tichý, L. & Roleček, J. 2003. Local and regional patterns of species richness in Central European vegetation types along the pH/calcium gradient. *Folia Geobot*. 38: 429-442.

Cochran, W.G. 1977. *Sampling techniques*. 3rd ed. Wiley, New York, NY, US.

Diekmann, M., Eilertsen, O., Fremstad, E., Lawesson, J.E. & Aude, E. 1999. Beech forest communities in the Nordic countries – a multivariate analysis. *Plant Ecol*. 140: 203-220.

Ellenberg, H., Weber, H.E., Düll, R., Wirth, W., Werner, W. & Paulißen, D. 1992. *Zeigerwerte von Pflanzen in Mitteleuropa*. 2nd ed. *Scripta Geobot*. 18: 1-258.

Ewald, J. 2001. Der Beitrag pflanzensoziologischer Datenbanken zur vegetationsökologischen Forschung. *Ber. R. Tüxen Ges*. 13: 53-69.

Ewald, J. 2002. A probabilistic approach to estimating species pools from large compositional matrices. *J. Veg. Sci*. 13: 191-198.

Ewald, J. 2003. The sensitivity of Ellenberg indicator values to the completeness of vegetation relevés. *Basic Appl. Ecol*. 4: 507-513.

Gillison, A.N. & Brewer, K.R.W. 1985. The use of gradient directed transects or gradsects in natural resource survey. *J. Environ. Manage*. 20: 103-127.

Gimaret-Carpentier, C., Pélissier, R., Pascal, J.-P. & Houllier, F. 1998. Sampling strategies for the assessment of tree species diversity. *J. Veg. Sci*. 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.

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

Gregoire, T.G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. For. Res*. 28: 1429-1447.

Haila, Y. & Margules, C.R. 1996. Survey research in conservation biology. *Ecography* 19: 323-331.

Havlová, M., Chytrý, M. & Tichý, L. 2004. Diversity of hay meadows in the Czech Republic: major types and environmental gradients. *Phytocoenologia* 34: 551-567.

Hennekens, S.M. & Schaminée, J.H.J. 2001. TURBOVEG, a comprehensive data base management system for vegetation data. *J. Veg. Sci*. 12: 589-591.

Hirzel, A. & Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. *Ecol. Model*. 157: 331-341.

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

Kent, M. & Coker, P. 1992. *Vegetation description and analysis – a practical approach*. Wiley, New York, NY, US.

Kočí, M., Chytrý, M. & Tichý, L. 2003. Formalized reproduction of an expert-based phytosociological classification: A case study of subalpine tall-forb vegetation. *J. Veg. Sci.* 14: 601-610.

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. *J. Veg. Sci.* 15: 415-422.

McCune, B. & Mefford, M.J. 1999. *PC-ORD. Multivariate analysis of ecological data.* Version 4. MjM Software Design, Gleneden Beach, OR, US.

Moravec, J., Balátová-Tuláčková, E., Blažková, D., Hadač, E., Hejný, S., Husák, Š., Jeník, J., Kolbek, J., Krahulec, F., Kropáč, Z., Neuhäusl, R., Rybníček, K., Řehořek, V. & Vicherek, J. 1995. Rostlinná společenstva České republiky a jejich ohrožení (Red list of plant communities of the Czech Republic and their endangerment). 2nd ed. *Severočeskou Přír., Suppl.* 1995: 1-206.

Mucina, L., Bredenkamp, G.J., Hoare, D.B. & McDonald, D.J. 2000. A national vegetation database for South Africa. *S. Afr. J. Sci.* 96: 497-498.

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 the potential natural vegetation of the Czech Republic 1 : 500 000).* Institute of Botany, Průhonice, CZ.

Olsen, A.R., Sedransk, J., Edwards, D., Gotway, C.A., Liggett, W.A., Rathbun, S., Reckhow, K.H. & Young, L.J. 1999. Statistical issues for monitoring ecological and natural resources in the United States. *Environ. Monit. Assess.* 54: 1-45.

Patton, M.Q. 1990. *Qualitative evaluation and research methods.* 2nd ed. Sage, Newbury Park, CA, US.

Rodwell, J.S. (ed.) 1990-2000. *British plant communities.* Vols. 1-5. Cambridge University Press, Cambridge, UK.

Schaminée, J.H.J., Stortelder, A.H.F., Westhoff, V., Weeda, E.J. & Hommel, P.W.F.M. (eds.) 1995-1999. *De vegetatie van Nederland. Deel 1-5.* Opulus Press, Uppsala, SE.

Smart, S.M., Clarke, R.T., van de Poll, H.M., Robertson, E.J., Shield, E.R., Bunce, R.G.H. & Maskell, L.C. 2003. National-scale vegetation change across Britain; an analysis of sample-based surveillance data from the Countryside Surveys of 1990 and 1998. *J. Environ. Manage.* 67: 239-254.

Sokal, R.R. & Rohlf, F.J. 1995. *Biometry.* 3rd ed. Freeman, New York, NY, US.

Stein, A. & Ettema, C. 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agr. Ecosyst. Environ.* 94: 31-47.

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

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

Thompson, S.K. & Seber, G.A.F. 1996. *Adaptive sampling.* Wiley, New York, NY, US.

Tichý, L. 2002. JUICE, software for vegetation classification. *J. Veg. Sci.* 13: 451-453.

Wessels, K.J., van Jaarsveld, A.S., Grimbeek, J.D. & van der Linde, M.J. 1998. An evaluation of the gradsect biological survey method. *Biodiv. Conserv.* 7: 1093-1121.

Wiser, S.K., Bellingham, P.J. & Burrows, L.E. 2001. Managing biodiversity information: development of New Zealand's National Vegetation Survey databank. *N.Z.J. Ecol.* 25: 1-17.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. 2001. Monitoring of biological diversity in space and time. *Trends Ecol. Evol.* 16: 446-453.

Received 1 July 2004;
Accepted 26 July 2005.
Co-ordinating Editor: N. Kenkel.