

## IS SAMPLING SUBJECTIVITY A DISTORTING FACTOR IN SURVEYS FOR VEGETATION DIVERSITY?

**Radim Hédl**

*Institute of Botany, Academy of Sciences of the Czech Republic, Department of Ecology, Poříčí 3b, 603 00 Brno, Czech Republic; e-mail hedl@ibot.cas.cz*

**Abstract:** Sampling design has long been a key methodology issue of vegetation science. There are still unsolved problems, amongst which the influence of the positioning (spatial arrangement) of plots remains a basic one. Random sampling is considered the only statistically valid option, along with its amendment by stratifying. However, huge amounts of plot-sample data originated from using subjective plot positioning, involving a great deal of the author's judgement, experience and bias. Field studies examining the role of subjectivity within a larger area (a landscape) are lacking. In the presented case-study, datasets sampled using computer-based geographically stratified random (R) and field-based subjective (S) designs are compared. In 1999–2000, 120 R and 244 S plots were spread across some 270 km<sup>2</sup>. Their properties are compared regarding the differences in vegetation variability by ordination, and diversity by rarefaction. Although the differences are not conspicuous, both datasets partly contain different vegetation types and the S dataset contains at least some species-richer plots. The number of R plots had been clearly underrated while the expected species diversity values were nearly reached by the S. More R than S plots would probably be needed to reach the expected diversity maximum.

**Keywords:** Beech forest, Czech Republic, Positioning of plots, Randomness and subjectivity, Sampling design, Vegetation diversity

### INTRODUCTION

Objective or subjective, mathematically accurate or intuitive, vegetation science can have either quality and employ both hemispheres of human's perception. It has evolved from the early subjective (e.g., KERNER VON MARILAU 1863) to the present-day mathematicized science (e.g., DALE & WEBB 1975, PLOTKIN & MULLER-LANDAU 2002, ACCAD & NEIL 2006). Phytosociology in terms of the Braun-Blanquet approach (VAN DER MAAREL 1975, EWALD 2003) is one of the most influential fields of vegetation science. Its usual aim is descriptive sampling, specifically the classification of vegetation. Along with vegetation science in general, sophisticated analytical tools have become so axiomatic that the common user does not think of their validity any more. This can have troublesome consequences, as pointed out most recently by PODANI (2006) and LÁJER (2007).

Nonetheless, before analysis data first have to be sampled in the field (BOX & FUJIWARA 2005: 107), which constitutes the most critical point of phytosociology – the traditionally subjective selection of samples (phytosociological relevés) directly in the field (EWALD 2003). This step relies on the researcher's personal experience, anticipation, intuition – and

affinity to this or that part of vegetation variability (HÉDL 2005). This results in an unpredictable bias within the datasets, which makes analyses of diversity based on relevés skewed (CHYTRÝ 2001).

### **Why does a sampling design matter?**

Descriptive sampling (EBERHARDT & THOMAS 1991) of vegetation using species-plots, e.g. relevés, can be applied for parameter estimation, such as that of cover or diversity (KENKEL et al. 1989). To obtain realistic estimates, the data is presumed to be a true sample, meaning a random draw from the whole (COCHRAN 1977). For a plot-sample, this means a random spatial arrangement of mutually independent plots (PODANI 1984, KENKEL et al. 1989, MCCUNE & GRACE 2002). An ecologically sound type of sampling design is a restricted random, often called stratified random (GREIG-SMITH 1983, PODANI 1984), usually with spatial or ecological restriction parameters. This type of design has however rarely been applied in phytosociological surveys (e.g. GRABHERR et al. 2003).

Comparisons of two or more spatial arrangements of species-plots for both the real (e.g. GOEDICKEMEIER et al. 1997, GIMARET-CARPENTIER et al. 1998) and virtual datasets (HIRZEL & GUIBAN 2002) are very rare; two unpublished theses are by HÉDL (2001) and LVONČÍK (2006). Surprisingly, it is probable that no published comparative study includes the most common design – the subjective sample arrangement.

Vegetation variability and diversity as derived from datasets with different amounts of subjective intervention are the subject of this paper. Two types of positioning of phytosociological relevés were jointly applied and their properties statistically analyzed. They were computer-based “objective” with randomized positioning of plots and field observation-based “subjective” sampling designs. It is assumed that the “objective” design represents real proportions between the species and the values of vegetation diversity. In contrary, the “subjective” design would probably be influenced by the author’s preferences and the unconscious tendency of sampling species-rich sites.

## **METHODS**

### **Sampled vegetation**

The study object is a beech- and mixed-beech (*Fagus sylvatica* L.) forest of the Rychlebské hory Mts., extending over about 270 km<sup>2</sup> in the north-east of the Czech Republic. Most of the area is forested at present; although it consists mainly of Norway spruce and mixed-spruce plantations, beech accounts for about 20%. Beech-dominated forest communities would probably constitute the natural vegetation in this area (NEUHÄUSLOVÁ et al. 1998), potentially dominating it except for the edges of ecological gradients.

### **Field sampling**

Two field samplings were applied independently from each other: geographically stratified random systematic (R) and subjective (S). The purpose of both was to record the diversity of the entire beechwood vegetation within the territory. The extent of the beechwood was mapped in the field by L. Bureš and R. Hédli, early in the 1990s. They used a set of 13 map

sheets scaled 1 : 10,000. The entire area occupied by the beechwood covered 2475 hectares (horizontal projection) spreading throughout the territory, elevations ranging from 350 to 1050 m a.s.l., occupying all aspects and slopes of 0 to 40 degrees.

For the R, a systematic plot design was established dividing each map using a regular  $0.5 \times 0.5$  cm square grid. The cells at least half-occupied with beechwood (numbering 9900) were subjected to a random selection of 130 cells. The selection procedure was not completely random; a geographical stratification was applied using 13 maps for strata. The number of cells selected for each map was proportional to the extent of the beechwood on the area covered by it. The procedure was carried out using a simple computer programme, NVS by P. Václavěk. The 130 R-cells were identified in the field as  $50 \times 50$  m squares.

Field sampling was executed from August to September 1999 and 2000. I subsequently visited the R-cells, roughly according to their geographical distribution. In the assumed centre of each square cell, I recorded one phytosociological relevé (see below). Ten plot centres fell outside the beech forest, resulting in  $N_R = 120$ . Walking for days and weeks in the forests, I worked like a classical phytosociologist in parallel. I sampled the beechwoods subjectively placing the S-plots, aiming at representing the vegetation variability as completely as possible. I tried to avoid any preference towards particular types of beechwood, while also including specific small sites. The resulting number of plots was  $N_S = 244$ .

Species-plots are the classical phytosociological relevés. All species present within a plot were recorded assessing the quantity of each species using a refined Braun-Blanquet abundance-dominance scale (degrees r, +, 1a, 1b, 2a, 2b, 2c, 3a, 3b, 4 and 5). Tree, shrub, herb and tree seedling layers were distinguished (moss layer usually covering 1–5%, was neglected). Relevés were squares sized 200–400 m<sup>2</sup>, the size being a function of vegetation heterogeneity.

### Data treatment and analyses

All 364 relevés were stored in Turboveg for Windows database programme (HENNEKENS & SCHAMINÉE 2001) and consequently edited using JUICE 6.4 programme (TICHÝ & HOLT 2006), keeping the four vegetation layers separate.

(1) The first type of analysis focuses on the interrelations between plots using a detrended correspondence analysis (DCA; MCCUNE & GRACE 2002) provided by the CANOCO programme package (TER BRAAK & ŠMILAUER 1998). Two separate parallel DCAs were performed:  $DCA_R$  for 120 R-plots and  $DCA_S$  for 244 S-plots. Prior to the analyses, abundance-dominance values of species were logarithmically transformed from percentages, obtaining values ranging from 1 to 2.94. Detrending was done using 26 segments (default), no downweighing of rare species was applied. The scores of the plots at the first and second ordination axes were jointly plotted in a DCA scatterplot. Plant species with scores falling into sectors with a clear prevalence of R or S plots were identified but not plotted.

(2) The second type of analysis concerns diversity measures using rarefaction methods (e.g., GOTELLI & COLWELL 2001) by the computer programme EstimateS (COLWELL 2005). Each dataset was re-sampled at random, the number of plots gradually increasing from 1 to  $N_R$  or  $N_S$  with 100 random draws at each step, randomization with replacement of samples. The number of species was estimated using the Mao Tau estimator (COLWELL et al. 2004),

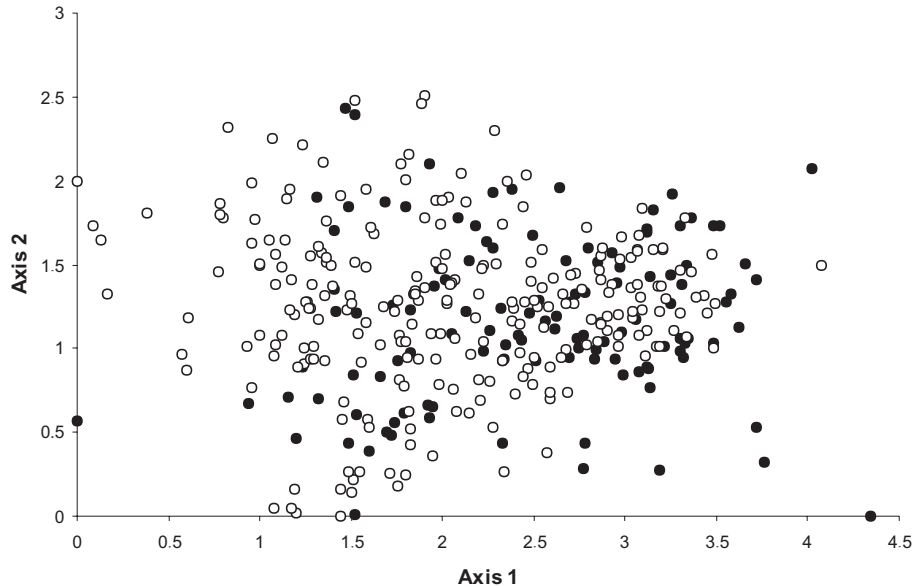


Fig. 1. Joint diagram of two parallel detrended correspondence analyses;  $DCA_R$  of R ("random") plots, full circles, and  $DCA_S$  of S ("subjective") plots, open circles. Positions at the first and second ordination axes are plotted, units are SD. The central part comprises an overlap of both groups, while the upper left-hand part is covered with S-plots and the bottom right-hand part with R-plots. For interpretation see text.

supplied with a mean observed species number. The sample-based rarefaction curves were jointly plotted for R and S datasets. Two kinds of scaling of  $x$ -axis were done, using either plots or individuals (i.e., every species record in the relevés); the former shows the species density, the latter, the species richness (see GOTELLI & COLWELL 2001).

## RESULTS

### Detrended correspondence analysis

The DCA diagram (Fig. 1) shows patterns composed of the R and S plots. The datasets overlap only partly, leaving the left-hand section of the diagram occupied with S plots (with one R exception), and the bottom-right part with R plots. Judging from the species with scores confined to these extreme parts, there is a diagonal gradient from acidic, species-poor and probably partly high-altitude beechwoods (*Luzulo-Fagion*) towards nutrient- and species-rich forests with lime and sycamore in the canopy: the ravine forests of deeply incised valleys (*Tilio-Acerion*). The former (S-margin) is characterized by *Avenella flexuosa*, *Vaccinium myrtillus*, *Calamagrostis epigejos*, *Luzula campestris*, *Abies alba* (tree layer), and *Sorbus aucuparia* (seedlings). The latter (R-margin) comprises more than twenty species including *Cardamine impatiens*, *Stachys sylvatica*, *Geranium robertianum*, *Euphorbia dulcis*, *Pulmonaria obscura*, *Galeopsis* spp., *Fraxinus excelsior* (tree layer and seedlings), *Tilia platyphyllos* (tree layer), and *Acer pseudoplatanus* (tree layer).

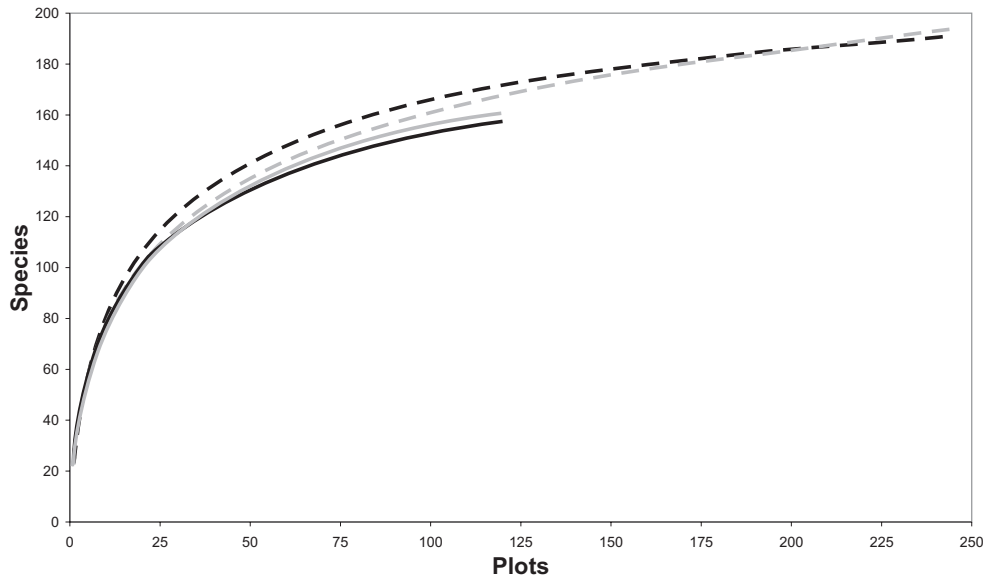


Fig. 2. Rarefaction curves for number of species (vertical axis) related to number of plots (horizontal axis) increasing from 1 to 120 for the R-dataset (solid lines) or 244 for the S-dataset (dashed lines). Black lines denote values of the Mao Tau species number estimator, grey lines are means of the observed species numbers (from 100 random draws). Number of species should be read as species density because related to the area of relevés constituting the dataset (both increasing). Note the higher species density increase for the S-dataset, slowly approaching the asymptote.

### Diversity

Rarefaction curves are presented in Fig. 2. Two black curves show the increasing values of Mao Tau diversity estimator. Two black lines show the values of Mao Tau diversity estimator as a function of aggregated plot size which is the variable controlling species density. This indicates that the “subjective” S-dataset contains an unknown number of relatively species-richer plots (the number cannot be established from the diagram). They contribute to a higher estimate of species density compared to the relatively species-poorer plots of the R-dataset. Fig. 3 shows this pattern explicitly; differences between species number gradually increase; there are relatively species-rich plots in the S-dataset that increasingly fall into the selection.

A clear difference between the datasets is the final species number, reaching 157 and 189 for R and S datasets, respectively. Neither of the curves reach the asymptote yet, but the S curve most probably would not exceed 200 species – the estimated total species number of the vegetation sampled. More R than S plots would be needed to reach the maximum values.

Taking plain observed mean values of species numbers into account (Fig. 2, grey lines), however, the difference is not as apparent as when using the estimator. When re-scaling the  $x$ -axis for individuals instead of plots, the pattern of curves is almost exactly the same (diagram not shown). This means that concerning species density (scaling for plots) and species richness (scaling for individuals) the sampling designs perform equally.

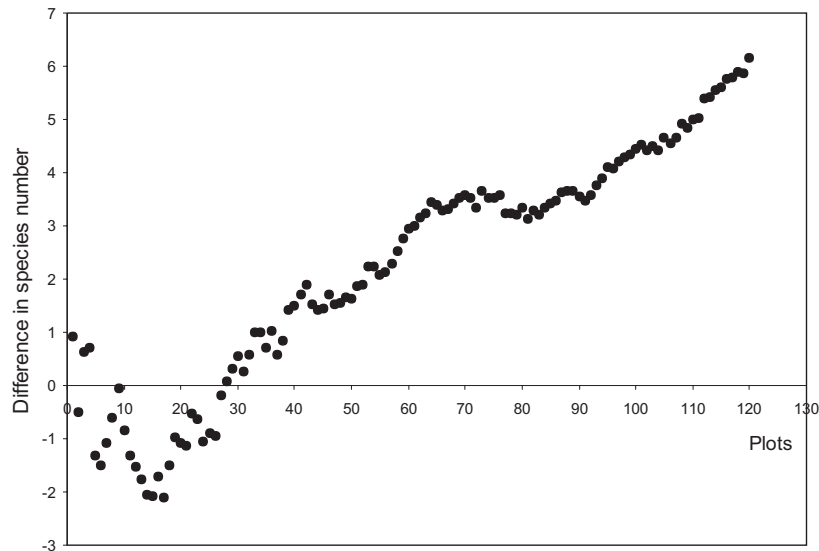


Fig. 3. Difference in observed mean species number, 100 random draws, between the S-dataset and R-dataset (compare with Fig. 2, grey lines). Number of plots is limited with  $N_R = 120$ . After 30–40 plots, the difference gradually increases meaning that the S-dataset contains relatively species-richer plots than the R-dataset.

## CONCLUSIONS

It was demonstrated that subjectivity in positioning of the plots within a vegetation survey can influence the properties of the result. The sample obtained is relatively more diverse than the real vegetation, containing a higher proportion of species-richer plots. It is hard or even impossible not to under- or oversample the particular vegetation types when selecting subjectively in the field, which is a characteristic drawback of phytosociology and similar types of ecological surveys and censuses. Consciously applied statistically-based survey designs (e.g., JONSSON & MOEN 1998, SCHMIDTLEIN 2003) should provide repeatable and more reliable results.

Subjective field selection, however, is probably the best strategy for capturing diverse sites (local “hot-spots”), which are usually of a very small extent considering the overall poor environments of the temperate landscapes. It was also shown that an *a priori* (non-field) assessment of the diversity and consequently the appropriate number of plots can easily fail. In the presented case, the number of plots was underrated when positioning plots at random.

These conclusions, however, can be strongly case-specific. Firstly, beechwoods are a relatively homogeneous sort of vegetation and it is not difficult to reach similar results with any arrangement of plots (which, I think, was not quite the case here). If the entire vegetation at the landscape level were included, the differences between the designs would presumably be higher. Secondly, objectivity of sampling is influenced by the experiences of the researcher(s) and overall circumstances. This, however, should be a subject of a psychological (BRUNSWIK 1955, DHAMI et al. 2004) rather than an ecological study.

**Acknowledgements:** I am much indebted to Tomáš Herben for his incentive ideas, and to two anonymous reviewers for their comments that helped to improve the first version of the manuscript substantially. I would like to thank Simon Botten and Péter Szabó for language correction. This publication was supported by the institutional long-term research plan, grant No. AV0Z60050516 of the Academy of Sciences of the Czech Republic.

## REFERENCES

- ACCAD A. & NEIL D.T. (2006): Modelling pre-clearing vegetation distribution using GIS-integrated statistical, ecological and data models: A case study from the wet tropics of Northeastern Australia. *Ecol. Modelling* 198: 85–100.
- BOX E.O. & FUJIWARA K. (2005): Vegetation types and their broad-scale distribution. In: VAN DER MAAREL E. (ed.), *Vegetation ecology*, Blackwell Science Ltd., pp.106–128.
- BRUNSWIK E. (1955): Representative design and probabilistic theory in a functional psychology. *Psychol. Rev.* 62: 193–217.
- CHYTRÝ M. (2001): Phytosociological data give biased estimates of species richness. *J. Veg. Sci.* 12: 439–444.
- COCHRAN W.G. (1977): *Sampling techniques*. J. Wiley, New York.
- COLWELL R.K. (2005): *EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5*. User's Guide and application published at: <http://purl.oclc.org/estimates>.
- COLWELL R.K., MAO C.X. & CHANG J. (2004): Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 85: 2717–2727.
- DALE M.B. & WEBB L.J. (1975): Numerical methods for the establishment of associations. *Pl. Ecol.* 30: 78–88.
- DHAMI M.K., HERTWIG R. & HOFFRAGE U. (2004): The role of representative design in an ecological approach to cognition. *Psychol. Bull.* 130: 959–988.
- EBERHARDT L.L. & THOMAS J.M. (1991): Designing environmental field studies. *Ecol. Monogr.* 61: 53–73.
- EWALD J. (2003): A critique for phytosociology. *J. Veg. Sci.* 14: 291–296.
- 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.
- GOTELLI N.J. & COLWELL R.K. (2001): Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4: 379–391.
- GRABHERR G., REITER K. & WILLNER W. (2003): Towards objectivity in vegetation classification: the example of the Austrian forests. *Pl. Ecol.* 169: 21–34.
- GREIG-SMITH P. (1983): *Quantitative plant ecology*. Blackwell, Oxford, London.
- HÉDL R. (2001): *Bučiny Rychlebských hor (Beechwoods of the Rychlebské hory Mts.)*. Diploma thesis, Charles University, Prague.
- HÉDL R. (2005): Co popisuje fytoocenologie? (What is described by phytosociology?). *Zprávy Čes. Bot. Společ.* 40: 301–314.
- HENNEKENS S.M. & SCHAMINÉE J.H.J. (2001): Turboveg, a comprehensive database 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. Modelling* 157: 331–341.
- JONSSON B.G. & MOEN J. (1998): Patterns in species associations in plant communities: the importance of scale. *J. Veg. Sci.* 9: 327–332.
- KENKEL N.C., JUHÁSZ-NAGY P. & PODANI J. (1989): On sampling procedures in population and community ecology. *Vegetatio* 83: 195–207.
- KERNER VON MARILAUN A. (1863): *Das Pflanzenleben der Donauländer*. Innsbruck.
- LÁJER K. (2007): Statistical tests as inappropriate tools for data analysis performed on non-random samples of plant communities. *Folia Geobot.* 42: 115–122.
- LVONČÍK S. (2006): *Lesní vegetace Slezské pahorkatiny a Ostravské pánve (Forest vegetation of the Slezská Uplands and Ostrava Basin)*. Diploma thesis, Masaryk University, Brno.
- MCCUNE B. & GRACE J.B. (2002): *Analysis of ecological communities*. MJM Press, Glenden Beach, Oregon.

- NEUHÄUSLOVÁ Z., BLAŽKOVÁ D., GRULICH V., HUSOVÁ M., CHYTRÝ M., JENÍK J., JIRÁSEK J., KOLBEK J., KROPÁČ Z., LOŽEK V., MORAVEC J., PRACH K., RYBNÍČEK K., RYBNÍČKOVÁ E. & SÁDLO J. (1998): *Mapa potenciální přirozené vegetace České republiky (Map of potential natural vegetation of the Czech Republic)*. Academia, Praha.
- PLOTKIN J.B. & MULLER-LANDAU H.C. (2002): Sampling the species composition of a landscape. *Ecology* 83: 3344–3356.
- PODANI J. (1984): Spatial processes in the analysis of vegetation: Theory and review. *Acta Bot. Hung.* 30: 75–118.
- PODANI J. (2006): Braun-Blanquet's legacy and data analysis in vegetation science. *J. Veg. Sci.* 17: 113–117.
- SCHMIDTLEIN S. (2003): Raster-based detection of vegetation patterns at landscape scale levels. *Phytocoenologia* 33: 603–621.
- TER BRAAK C.J.F. & ŠMILAUER P. (1998): *CANOCO reference manual and user's guide to Canoco for Windows: Software for canonical community ordination (version 4)*. Microcomputer Power, Ithaca, NY.
- TICHÝ L. & HOLT J. (2006): *JUICE, program for management, analysis and classification of ecological data*. Program manual, <http://www.sci.muni.cz/botany/juice/>
- VAN DER MAAREL E. (1975): The Braun-Blanquet approach in perspective. *Pl. Ecol.* 30: 213–219.

Received 9 October 2006, revision received and accepted 22 January 2007