COENOSES 2(1): 9-18, 1987

COMPUTERIZED SAMPLING IN VEGETATION STUDIES

J. Podani
Research Institute for Ecology and Botany, Hungarian Academy of Sciences, Vácrátót, H-2163
and
Dept. Plant Taxonomy and Ecology, L. Eötvös University, Budapest, H-1083

Keywords: Computer, Simulation, Sampling, Pattern, Estimation, Computer, Programs, Classification, Plot size

Abstract: This paper discusses the utility of computerized sampling methods in vegetation studies. A brief review of published work and a presentation of new results demonstrate that computerized sampling may be beneficial in a wide variety of research areas, including the optimization of sampling design and the recognition of the scale of spatial pattern.

Motivation

A quantitative vegetation study may be divided into two main stages: sampling and data analysis. In the second stage, all ecologists seem to agree that computers have become indispensable tools (see the ever-increasing bulk of literature on the development and application of computer techniques in vegetation science). But what of the first step? Is it possible that computers could be of some help in the sampling stage as well? Is there a real need for computers at the very beginning of a vegetation survey? Is it at all conceivable that computerized sampling will ever be as routinely used as multivariate analytical methods, or will it be confined to some isolated fields of ecology? I hope that this brief review of published work and new results will bring us closer to the answers to these questions. However, before turning to the discussion of the possible utility of computerized sampling, it seems worthwhile to deal with the principles.

Principles of computerized sampling

A definition

It is appropriate at this stage to define the term «computerized sampling» in order to avoid any possible misunderstanding. The adjective «computerized» is meant to be more restrictive than «computer assisted». «Assistance» allows for the case, for example, when sample plots are laid down in the field according to random coordinates generated by a computer. This definition is not of concern here. Furthermore, sampled randomization tests and other Monte Carlo-type experiments that simulate data directly (*e.g.*, Swan 1970, Ek 1971, Gauch and Whittaker 1972, 1976, Ricklefs and Lau 1980, Carpenter and Chaney 1983, Lagonegro 1984, Prentice and Werger 1985, among many others) are also beyond the scope of this paper. I am concerned here with procedures which simulate field sampling. Thus, computerized sampling is defined as a completely automated process for taking simulated units from an appropriate sampling universe, which is housed within the memory of the computer. The simulation is completed with the recording and output of data requested.

Types of sampling universes

Depending on the level of vegetation description, two different types of sampling universe are conceivable. At the level of the population of one or more species, the distinguishability of individuals and species on the map will be important. This is not the case with community level maps, which express vegetation structure in terms of syntaxonomic units. In this paper, the first type, called pre-analytical, will be of primary concern. Since the second type is conditioned upon some analytical steps by which the syntaxonomic units are created, they are rarely if ever sampled.

The preparation of a sampling universe for computerized sampling

A fundamental question is the manner by which the sampling universe is stored in the computer core. Szöcs (1979) described a procedure, the photocomputational method, which illustrates the general idea. The first step is *fixation*; the real, primary picture of the vegetation is mapped using a camera or some other recording device. This map is subjected to *transformation* to convert it into a simplified geometric image that will be ready for *digitalization*, a process for numerically coding the sampling universe. The resulting set of «data» can then be used as input to a sampling simulator.

Types of transformation

The most crucial of the above steps is transformation, since it is at this point that the decision of how the actual vegetation map will simplify to a form suitable to digitalization is made. The method of transformation should be selected by considering some or all of the following: 1) the required level of resolution, 2) the size of individual plants relative to the size of the study area and plots, 3) the shape of plants, 4) the type of data to be recorded, and 5) the objective of the study.

In the simplest and commonest case, each individual plant is represented by a *point* whose position is described

using a Cartesian co-ordinate system. The two-dimensional map obtained in this way is applicable for recording abundance and presence/absence data but cannot be used in estimating cover. Such maps are efficient in the study of the spatial pattern of trees (*e.g.*, Diggle 1979, Bonnicksen and Stone 1981), since no additional information is conveyed on the spatial distribution of individuals by a more complex map also showing, say, the vertical projection of crowns. However, the representation of grasslands and similar vegetation types is much less accurate using point-scatter maps. One alternative, the representation of plant patches by dense point clusters, appears to be suboptimal. As an example, Podani (1984b) found that the simplification of a cover map into a scattergram, by replacing patches with point clusters, led to the underestimation of resemblance and information theory functions at small plot size, though this effect diminished as plot size increased.

Regular, two-dimensional figures (usually circles) are also used to approximate the vertical projection of plant parts (*e.g.*, basal area of trees, Arvanitis and O'Regan 1967, Sukwong *et al.* 1971). In such cases, the parameters required to describe the shape and size of individuals or patches render the processes of digitalization and sampling unit simulation much more difficult.

Transformation into *irregular* figures would require even greater sophistication and much more computer memory. Two possibilities for simplification deserve mentioning: the representation of vegetation as a mosaic of several phases (sensu Matérn 1979) or as small-celled grid images. To my knowledge, such maps have not yet been analyzed using computerized sampling. This is likely attributable to the difficulties associated with coding and programming.

Simulation of the sampling universe

The problems of fixation and transformation are avoided using computer-generated vegetation patterns. The simulation of vegetation structure allows the researcher to govern the pattern formation processes such that many properties of the resulting maps are known (Szöcs 1979). Reference is made here to a few works which may be consulted for Monte Carlo methods of pattern generation. Ripley (1979, 1981), Cliff and Ord (1981) and Diggle (1979, 1983) described models for stochastic point processes applicable to the generation of various types of spatial distribution of a single species. An algorithm to simulate random mosaics (Voronoi polygons or Dirichlet tessellations) of two or more species is presented in Green and Sibson (1978, see also Matérn 1979). Czárán (1984) suggested a procedure for the simulation of a point pattern of several species by considering competition as well as seed dispersal.

Simulation of the sampling procedure

Simulation of the sampling procedure yields a sample which is a subset of the sampling universe. Practically all sampling methods can be simulated, but the solution for any part of the universe, whether or not to include in a sampling unit, will not be equally easy for all plot shapes. For example, for circular plots it is sufficient to find the individuals that fall from the centre of the plot not farther than the specified radius. The search needs more application of co-ordinate geometry for randomly-oriented rectangles and even more calculations are required for elliptical sampling units. In any case, the selection or location of sampling units must involve a chance component.

As in the field, the user of a sampling simulator must specify several sampling characteristics. These include:

1) The type of sampling unit (random plant, random point, plot, line, etc.) and the type of data to be recorded (distance, abundance, presence/absence, cover, etc.);

2) The number of sampling units (sample size);

3) The arrangement of sampling units (*e.g.*, systematic, random, stratified random).

If plots are used, three more decisions must be made regarding;

4) the plot shape;

5) the plot size; and

6) the orientation of anizodiametric plots.

Some computer programs designed for sampling simulation are discussed in the Appendix.

Problems associated with computerized sampling

In addition to the problems of transformation, computerized plot sampling leads to two other difficulties that may introduce a bias into the results. The first problem revolves around the fact that usually no overlap is allowed between sample plots and the borders of the study area. Consequently, in the case of randomly-arranged plots, plants close to the border will have a smaller chance of being included in the sample than others; this is termed the «edge effect» (*e.g.*, O'Regan and Palley 1965, Wensel 1975, Ripley 1981). The edge effect is illustrated in Figure 1a, in which square units located parallel to the edge are used to simplify the illustration. Let s denote the side length of quadrats, x_i the distance of point i from the boundary, and p_i the probability that point i falls within a randomly located quadrat. For any point for which $x_i > s$, the p_i probability will be proportional to s^2. However, p_i will be lower for points with $x_i < s$, since all quadrats whose centroid is closer to the boundary than $s/2$ would cross it and would be deleted from the sample. In these cases, p_i is proportional to $s \cdot x_i$; that is, sampling intensity will decrease continuously towards the edge. This seems a serious disadvantage, though Podani (1984b) has suggested that edge effect should always be interpreted in accordance with the objective of the survey. Whenever the objective is the estimation of population variables and textural variables (*e.g.*, species/individual diversity), edge effect should be corrected for to avoid bias. This may be achieved using Wensel's (1975) method (a case of the toroidal edge correction of Ripley 1981), which complements incomplete sampling units by fragments taken from the opposite edge

of the study area (Fig. 1b). This implies that points on the opposite edge can be pooled. In this sense, the correction is absolutely meaningless if structural characteristics of species assemblages (*e.g.*, interspecific associations, resemblance, supraindividual diversity) and their dependence on sampling are of interest. In these cases, the problem of edge effect applies equally to simulated and field sampling, and therefore needs no correction (cf. Podani 1984b). An important consequence of the edge effect problem is that conclusions drawn from a simulated sample do not apply directly to the whole vegetation from which the map was taken. However, if the objective is the analysis of spatial pattern in a population, edge effects need to be corrected using special formulae (*e.g.*, Diggle 1979) or allowing a «guard» area around the sampled site (Ripley 1981, Galiano 1982).

The second question involves deciding whether or not to permit the overlapping of random plots for a given sample. Again, the answer is goal-dependent. If population variables are to be estimated, sampling with overlapping plots appears to be analogous to the random sampling of individual objects with replacement (Cochran 1977, p. 29). However, it is unclear if the same statistics directly apply to samples derived from plots. If the objective of the study is the analysis of vegetation structure, the overlap of plots produces no artifacts in the estimation of structural variables, as shown by Podani (1984b). Indeed, overlap among plots must be allowed in such cases, in order to capture as many species combinations as possible (see below).

Applications of computerized sampling

Simulation of sampling has served several purposes in ecology and related fields. Though these applications could be categorized in many different ways, I consider the underlying main objective of the study as the most important. Accordingly, the primary distinction here will be made between surveys aimed at the *estimation* of population variables and textural variables, and those analyzing *structural* properties of a single population (autophenetic pattern) or an assemblage of population (synphenetic pattern, sensu Juhász-Nagy 1984 p. 401).

Estimation

Having a full representation of the sampling universe in computer memory, the surveyor may readily determine the «true» values of population parameters and textural variables by complete enumeration. This knowledge makes possible comparative studies on the performance of different sampling techniques and statistical estimators.

Population variables

The history of estimation-oriented computerized sampling goes back to the early 1960's. To my knowledge, Palley and O'Regan (1961) first used a sampling simulator

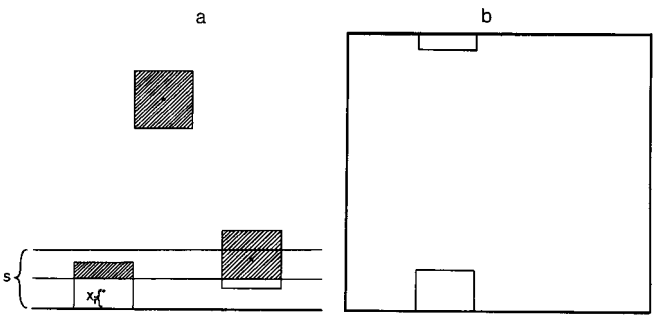

Fig. 1. a. Edge effect. Shaded areas are proportional to the probability that an individual in the plot center will be sampled (see text); b. Wensel's correction for edge effect.

program. They compared the precision of tree density, basal area and timber volume estimates obtained by two different *plot* sampling methods (Bitterlich point and line sampling) for various types of forests. In another paper (O'Regan and Palley 1965), the relationship between plot size and the variances of the above mentioned measurements was studied. Considering efficiency, plot sampling was found to be inferior to Bitterlich sampling. Arvanitis and O'Regan (1967) extended the study of optimality criteria to sample size as well as to the cost of measurements. With these considerations included, sampling by plots was found to perform better for all variables except basal area. These papers clearly showed how useful computerized sampling may be for selecting optimal sampling strategies in applied research. Predicting a promising future of this strategy, the authors went on to state that «Increase in storage capacity [of computers] and development of more accurate, rapid, and economical methods of constructing stem maps of large forests (by means of aerial photographs and electronic devices) would increase the efficiency of the computerized approach and would help to solve the difficult sampling problems...».

In forest science, computerized sampling has continued to play an important role. Wensel and John (1969) and Wensel (1975) were concerned with edge effect corrections in simulated sampling. Sukwong *et al.* (1971) returned to the old issue of comparing the precision of estimates from variable and fixed-radius plot samplig. Artificial populations with random and aggregated spatial pattern were analyzed. Their approach differed from the former one in that, instead of subjecting a whole map to sampling simulation, only fragments of forest around the centre of each plot were simulated to save computer time. O'Regan et al. (1973) determined the optimum number and size of random plots of estimating density at a fixed cost.

Pickford and Hazard (1978) described a program to simulate *line intersect* sampling. The effect of the number and length of lines on the precision of estimates of logging residue volume was analyzed.

Whenever the density of populations is examined using plant-to-plant or point-to-plant *distances*, the objectives of estimation and pattern analysis coincide. In such cases, density estimation is accompanied with inferences on the spatial pattern (cf. Diggle 1979, p. 119). For example, Batcheler (1973) derived empirical density estimators,

which incorporate an index of non-randomness, from simulated populations. Lamacraft *et al.* (1983) compared four distance-based density estimators, two sampling strategies, and actual and random plant patterns using simulated sampling. A general treatment of the distance-based density estimators is given by Warren and Batcheler (1979), which also includes many other references to simulation studies.

Textural variables

Sampling techniques and estimators for textural variables (*e.g.*, species number, species/individual diversity) for describing species assemblages may also be examined using simulated sampling. However, fully-automated comparisons based on maps appear to be lacking. Nevertheless, there are a few relevant papers which are worth mentioning here. Nosek (1976) used semi-automated sampling to analyze the performance of sampling methods in diversity studies. Kobayashi (1981) and Heltshe and Forrester (1985) reported on the comparison of diversity estimates using Monte Carlo data simulation. Finally, Hahn (1982) explored some possibilities for the correction of edge effects in diversity estimation; his suggestions deserve more attention in future simulation work.

Autophenetic pattern

Departures from randomness, and the scale of pattern are the primary concern when pattern analysis is performed at the population level. Clearly, a map showing the locations of all individuals provides more information regarding spatial pattern than field studies based on either quadrat or plotless sampling (cf. Diggle 1979). Advantages of using mapped patterns and simulated sampling include:

1) the ability to apply more sophisticated analytical techniques, *e.g.*, those based on all plant-to-plant distances within the study region (an example is in Galiano 1982);

2) the possibility of more exhaustive investigations of the same region through analysis using different techniques. This job could not be done in the field easily and the trampling of herbaceous plants is also avoided in this way;

3) the ability to assess the performance of various pattern analysis techniques. Goodall and West (1979) provide an example; they compared several indices of non-randomness by means of computerized sampling in artificial populations.

In many recent statistical investigations of mapped data, computerized sampling is a routinely used technique of data collection, without explicit references to it in the methodological part of papers. This is certainly the case for plotless techniques, in which the simulation of sampling consists merely of placing random points on the plane or of selecting random plants from a list.

The idea of using computerized sampling to study spatial pattern was first proposed in the early 1970's. Goodall and West (1972) outlined a method for pattern analysis based on low-level aerial photographs, whereas La France (1972) investigated the effect of plot shape on the variance/mean ratio along artificial gradients. Since then, mapped *point* patterns of forest trees have been of primary concern in many statistical studies. Diggle (1979) presented a detailed illustrative study of several actual maps and showed how powerful the nearest-neighbor techniques are if mapped patterns are analyzed. He gives many references for further reading. Bonnicksen and Stone (1981) simulated five different techniques, including nearest neighbor and quadrat methods, for the pattern analysis of tree classes (so the statistical population analyzed was in fact composed of several species). Recently, Franklin *et al.* (1985) have applied computer simulated sampling to the spectral analysis of digitalized distributional data for trees.

I am not aware of any report on the computerized sampling study of patterns other than point scatters. Perhaps Matérn's (1979) paper may give the necessary starting information towards this objective.

Synphenetic pattern

Studies of the spatial structure of communities, no matter what the level of resolution, recognize pattern of the synphenetic type. Classification and ordination of vegetation are obvious cases in point. Considering the wide interest in multivariate techniques, a relative imbalance will strike the eye of the reader of the forthcoming paragraphs. Only four published works are available in which the computer is used as exhaustively as possible for evaluating the dependence of structural studies on the underlying pattern and sampling.

Structural variables

Structural variables express specific information on vegetation pattern as a single value. An essential difference between textural and structural variables is the *scale dependence* of the latter, implying that the optimization of textural and structural variable estimates requires different solution. Specifically, whereas plot-size increases always result in more precise estimates of species diversity, etc., the same cannot be said of structural variables (cf. Podani 1984a).

Resemblance (Orlóci 1972), is an important structural variable upon which most multivariate analyses are based. Its dependence on plot size, although repeatedly emphasized by some authors (*e.g.*, Greig-Smith 1983), remains a neglected topic. A first step towards a better understanding of this problem may be to determine the expected resemblance of two non-overlapping random plots. This can be estimated by simulating many pairs of plots using mapped data. Podani (1984b) presented results of sampling experiments concerning the influence of plot size on four resemblance coefficients, in two mapped and two simulated communities. Two indices, which exclude the number of «negative» matches from the numerator, showed a monotonic increase with increasing plot size. Euclidean distance and the related simple matching coef-

ficient exhibited a peaked effect, indicating a plot size at which the sample shows maximum heterogeneity (*maximum area* in terms of expected distance). The effect of pattern was manifested in that the maximum area was always smaller in the random versions.

Computerized sampling may also be used to evaluate the properties of resemblance coefficients. For example, Lim and Khoo (1985) investigated the effect of sample size and species number on estimates of Gower's similarity coefficient. Test communities with varying species abundance relations were generated and subsequently sampled using simulated quadrats. The between-community similarity coefficients obtained from the sample were compared with their corresponding expectations. The estimates were biased when the expected similarity was either high or low. However, since plot size effects were not investigated, the conclusions drawn cannot be considered definitive.

A family of *information theory* functions, related to Shannon's entropy, was suggested by Juhász-Nagy (1976, 1984) and Juhász-Nagy and Podani (1983) for analyzing the scale of synphenetic pattern in the presence/absence case. Characterisic areas were defined in terms of the maxima and minima of these functions (florula diversity, local distinctiveness, associatum). In the simulated sampling study referred to above (Podani 1984b), these characteristic areas were identified and compared to the maximum area of expected distance. There was good agreement between the results of the two different approaches regarding maximum areas and departures from randomness. Estimation problems were also considered by analyzing the effect of plot shape, arrangement and sample size.

The relationship between sampling and one of the information theory characteristic functions, namely local distinctiveness, was examined by Williams *et al.* (1969). They suggested a variant, based on mapped tree localizations, of multiple nearest neighbor sampling for the elucidation of small-scale synphenetic pattern of forests. Each tree, together with its m nearest neighbors, is considered as a sampling unit. The local distinctiveness (or «information content» in the terminology of the authors) is calculated for the set of all sampling units. As it turned out, the change of local distinctiveness over m (clump size) also exhibited a peaked effect and was the function of the mean distance from the reference individuals.

Ordination

The effect of sampling and community pattern on ordinations was examined by La France (1972), using artificial communities and partly-automated sampling. The recognition of an underlying one-dimensional gradient was found to be highly affected by plot shape and orientation — perhaps, not a too surprising result. Small-scale pattern imposed on the overall pattern led to a more distinct group structure in the ordinations. However, the approach of La France, which examines the relationships between sampling, plant pattern and the results of multivariate analysis, has subsequently received very little attention.

Classification

Computerized sampling and pattern simulation have been almost completely neglected in classificatory studies. This is all the more surprising when one considers that cluster analysis is used as a standard tool for pattern recognition. The usefulness of the computerized sampling approach is demonstrated by Williams *et al.* (1969). The point clump sets (described above under *Structural variables*) at various values of m, as well as contagious quadrats of different sizes were input to cluster analysis. The resulting groups at the four-cluster level were allocated onto a field map in order to evaluate their ecological relevance and to illustrate the dependence of classification on the sampling procedure used. The results suggested the existence of a clear pattern within the study area, with the most clear-cut separation of groups at m=12. This was the clump size at which the local distinctiveness was maximum, a conclusion which apparently escaped the attention of the authors. Some resemblance between the classifications of larger plots and clumps of m=12 is also obvious, indicating that a search for maximum local distinctiveness may aid in finding the scale of synphenetic pattern.

Two new examples

The previous section showed earlier applications of computerized methods for evaluating the effect of sampling upon the results of multivariate analysis. That the possibilities are by no means exhausted will be demonstrated here using two examples. The first examines the effect of sample size upon alternative classifications of species, whereas the second is concerned with the recognition of scale in simulated communities.

Sample size dependence of interspecific relationships

In general, an increase of sample size results in greater precision of estimates of the true population values of the sampling universe. This is also true of some structural variables (*e.g.*, resemblance), for the following reasons. Assume that interspecific correlation or expected resemblance of plots is to be calculated for a given study region based on presence/absence data at a fixed plot size. Since the sampling universe is continuous in space, an infinite number of plots could be located within the area. However, the number of possible florulas (species combinations within plots) has an upper bound, which is determined by community pattern (Juhász-Nagy and Podani 1983). The probability of selecting a given florula is proportional to the area within which any plot would capture the same species combination. The size of such areas determines in a complex way the expected values of resemblance. Increases of sample size will lead to an increased representation of possible florulas, and their relative frequencies will reflect more and more closely their true relative proportions. This in turn implies a greater precision of resemblance structures. This situation would also hold for the case when abundance or cover are estimated.

Material and methods

Given the difficulty of examining the above problem using field data, it is apparent that the effect of sample size on resemblance and subsequent classifications may be best demonstrated with a computerized approach. A map showing the point pattern of the most abundant six species in a perennial sand steppe community (Kiskunság National Park, Hungary) was used for this purpose. The 2.4 m^2 area was mapped by Z. Szöcs using his photocomputational technique (Szöcs 1979). The study region is fairly homogeneous, with most species showing an aggregated spatial distribution.

The study region was sampled by the computer program ELSAM (see Appendix). A total of 2400 circular plots was taken. The plot size was selected so as to yield the maximum of florula diversity within the region. The resemblance matrix of the six species was calculated and cluster analysis performed for each of the following sample sizes: 5, 10, 35, 75, 150, 200, 300, 600, 1200 and 2400. The combinations of clustering algorithms and resemblance coefficients were as follows:

a) simple matching coefficient – furthest neighbor sorting;

b) Euclidean distance for binary data – incremental sum of squares agglomeration (SSA);

c) PHI coefficient – nearest neighbor sorting;

d) Euclidean distance using counts – SSA;

e) Euclidean distance using logtransformed counts – SSA; and

f) Euclidean distance using counts standardized by range – SSA.

The resemblance measures and clustering algorithms are described in Orlóci (1978). The computations were performed at the University of Western Ontario, London, using the cluster analysis program NCLAS (Podani 1984c).

Results

The six series of ten dendrograms are not illustrated here. Instead, the topology of the hierarchies that did not change after a certain sample size was reached (stable results) are presented (Fig. 2). This limit varied considerably with the data type and the standardization technique used. The results suggest that untransformed counts require the smallest sample size to yield an invariant classification (Fig. 2d). For presence/absence data, a larger sample size was necessary to obtain an invariant classification. At sample size 200, identical results were produced by methods a and b. The logarithmic transformation of counts gave a close approximation to the binary case and resulted in similar classification at the same number of plots (Fig. 2e). The PHI coefficient produced a quite different result, and a substantially larger sample size was necessary. This coefficient has been most widely used in measuring interspecific correlation, and the present results demonstrate that this choice is probably inappropriate unless a fairly large number of plots are enumerated. The classification based on counts standardized by range did not reach the invariance state even at a sample size of 2400.

These results show that the «optimum» sample size (*i.e.*, the smallest number of plots beyond which the classification is unchanged) is strongly dependent upon the choice of the data analytical technique. A more exhaustive treatment, including the simulation of the point pattern, is clearly required for a deep exploration of this problem.

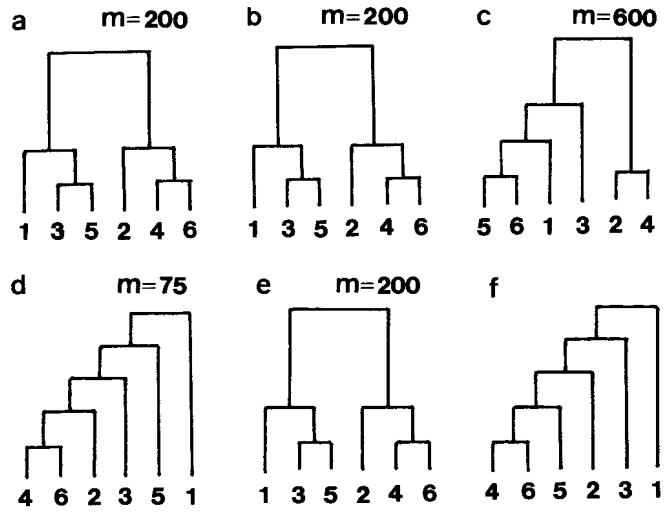

Fig. 2. The effect of sample size on the hierarchical classification of six species based on different measures of resemblance (see text). m is the sample size beyond which the topology did not change. m is not shown for dendrogram f since no stable result was obtained.

Community classification and plot size

This section examines the problem of scale dependence of community classification and its analysis. Although the literature abounds with proposals regarding the selection of plot size and sampling design for community classification, the problem remains to be investigated thoroughly. The complexity of the problem requires the aid of computer systems. Indeed, a completely computerized approach in which all steps, starting from community simulation and ending with cluster analysis, has much to be recommended since only in this way can the user control the factors which may seriously influence the results. These factors include:

A. Community properties
   1. Number of communities within the region
   2. Number of species
   3. Spatial distribution of individuals
   4. Species density
   5. Distribution of abundances
   6. Distinctness of communities regarding factors A.2-5
   7. Boundary type (sharp or fuzzy)

B. Sampling characteristics
   1. Plot size
   2. Plot shape
   3. Arrangement of plots
   4. Sample size

C. Decisions concerning data analysis
1. Data type
2. Data standardization
3. Resemblance coefficient
4. Structural variables
5. Clustering algorithm

This list is far from complete, but does indicate the complexity of the problem and the enormous amount of computation required before definitive conclusions regarding, say, the relationships between autophenetic and synphenetic pattern can be made. In this paper, a deliberately simple case is presented to illustrate the computational strategy. The influence of plot size on the recognition of two simulated community types is demonstrated, and the potential utility of some structural variables in plot size optimization is examined.

Materials and methods

Two artificial «communities» separated by a sharp boundary line were simulated within a study region of 10 by 24 units in size: community A on the left half, community B on the right. The number of species and the distribution of individuals were selected so as to yield a high overall similarity of the two communities. Twenty-eight of the species were common to both communities, being randomly dispersed over the entire study region. Four differential species were responsible for community differences, with two restricted to community A, the other two to community B; all were randomly dispersed. Species abundances ranged from 1-524 and approximated a lognormal distribution. The number of individuals of the differential species was 32, 64, 32 and 64, respectively; the total number of individuals was 2538.

Two sampling strategies were simulated. Sets of random circular plots, each with a sample size of 150, were taken using program ELSAM (see Appendix). The increasing series of plot radii was as follows: 0.2, 0.4, 0.5, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 2.0 and 2.5 units. To avoid the overlap of plots allowed by ELSAM, systematic sampling with square units was also performed using program SAMPROC (see Appendix), at the following quadrat sizes: 0.12, 0.64, 1.0, 1.8, 3.0, 5.0, 6.7, 8.7, 11.0 and 20.0 squared units. Because of the limited size of the study area, sample size decreased from 70 to 10 as plot size increased.

Structural variables, whose change will be evaluated in comparison with classification results, were calculated using the SYN-TAX II program package (Podani 1984c). The information theory functions mentioned above, plus dissociatum (Juhász-Nagy 1984), were computed using program INPRO2. To eliminate sample size effects, only the samples taken by ELSAM were used. Expected resemblance estimates were obtained by program EXPRES (see Appendix) for the same series of plot sizes.

In order to illustrate the effects of the existence of large-scale heterogeneity (i.e., distinct communities), the structural variables were also estimated from samples restricted to the separate community types. Sample sizes and plot sizes were identical with those used in the sampling of the entire study region.

The sample plots randomized over the whole region were classified using program NCLAS for each combination of plot size and arrangement. The clustering algorithm was sum of squares agglomeration based on a Euclidean distance matrix of presence/absence data.

Results

The dependence of structural variables on plot size, and the differences between the one and two-community cases, are illustrated in Figure 3. Arrows point to the maxima (or minima) of functions. The curves for community B are omitted since they were very similar to those found for community A.

The information theory functions (Fig. 3a) have no distinct extreme values; there are maximum or minimum intervals rather than definite peaks and deep throughs, particularly for florula diversity and dissociatum. Comparison with earlier findings suggests that this may be due to the presence of rare species (Podani 1984b). The divergence of results obtained from community A and from the entire region is apparent for local distinctiveness and associatum. In the single-community case, maxima are shifted to a smaller plot size and absolute values are lower, indicating that these functions are sensitive to the existence of large-scale heterogeneity. This sensitivity shows that local distinctiveness and associatum may be of greater utility in optimizing plot size prior to classification.

Expected values of the Sorensen and Russell-Rao indices increase monotonically with plot size (Fig. 3b). This inherent behavior is apparently little affected by actual plant pattern, rendering these functions inapplicable for recognizing the scale of synphenetic pattern (Podani 1984b). By contrast, the expectations for binary Euclidean distance and the simple matching index reach a maximum and a minimum respectively, indicating a plot size at which the sample is expected to be maximally heterogeneous. However, these extremes are not striking, since a wide range of plot sizes gave very similar estimate. As with local distinctiveness and associatum, the maximum area of expected distance is smaller for the one-community case.

The effect of plot size upon classifications is best illustrated by the allocation of results onto the map at the two-group level (Figs 4-5). In this way, we can see how the recognition of the two communities is affected by plot size. At small plot sizes, a mosaic structure is indicated. As plot size increases, the distinction of the two communities becomes more apparent. Optimal recognition of the two communities occurs at a plot size of 5.3 sq. units for random circles and at quadrat size of 11 sq. units for systematic sampling. Above this level, the plots begin to overlap the boundary line, obscuring recognition of the two communities. A plot size range 5.0-12.5 gives fairly «acceptable» results for both random and systematic sampling. However, there is a difference in the plot size which

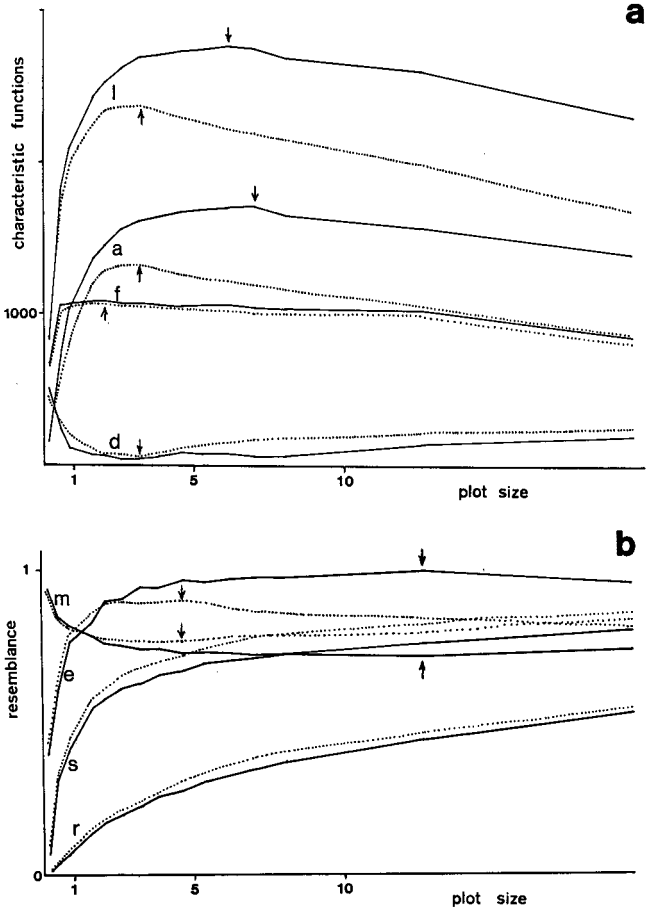

Fig. 3. The change of structural variables over plot size in the entire study area (solid lines), and in community A (dotted lines). a. Information theory functions: l: local distinctiveness, a: associatum, f: florula diversity, d: dissociatum. b. Expected resemblance: m: simple matching coefficient, e: normalized Euclidean distance, s: Sorensens index, r: Russell-Rao index.

gave the «best» classification. Plot size is smaller for the random design, since the overlap among plots facilitates «chaining» in the course of the clustering process, so that closely-positioned plots will tend to aggregate. In systematic sampling, the plots falling on the boundary line confuse the results at most plot sizes used.

Finally, the comparison of the two pattern recognition processes, (*i.e.*, the series of structural variables and the series of allocated classifications) deserves attention. The principal question is whether structural variables can be used for optimizing plot size in classification studies. Since systematic sampling yielded very small sample sizes at large plot sizes, this comparison can be made for the randomly arranged plots only.

Comparison of Figures 3 and 4 reveals that the «optimum» plot size is very close to the maximum area of local distinctiveness, and similar to that of associatum. This suggests that local distinctiveness, and perhaps associatum, may give a good approximation to the plot size for which the recognition of synphenetic pattern is optimal (see also the results of Williams *et al.* 1969, which were also obtained from overlapping units). At first sight, expected distance would appear less appropriate for this purpose; the peak is not so striking and the curve is fairly flattened within a considerable range of plot sizes. It is noted, however, that within this range, the recognition of communities was not seriously influenced by plot size. A practical implication of it is that after finding this range, other criteria (*e.g.*, cost) may dictate the selection of a particular plot size to be subsequently used.

Concluding remarks

This paper has reviewed and discussed the potential utility of computer simulated sampling in vegetation science. It is apparent from the many examples presented that great benefit may be derived from such an approach for the study of fundamental theoretical problems in vegetation sampling. Simulated sampling should receive more attention in the following research areas:

1) comparison of different sampling strategies;
2) the optimization of sampling design in accordance with the objective of the survey;
3) the joint application of a number of pattern analysis methods to the same segment of actual or artificial communities;
4) simultaneous analysis of principal properties of vegetation (population parameters, diversity, resemblance, etc.) in order to find the link between research areas which until now have been separated by wide methodological gaps; and
5) the study of the performance of pattern analysis methods, with emphasis on the relative impact of arbitrary decisions on the results.

For many, computerized sampling of simulated com-

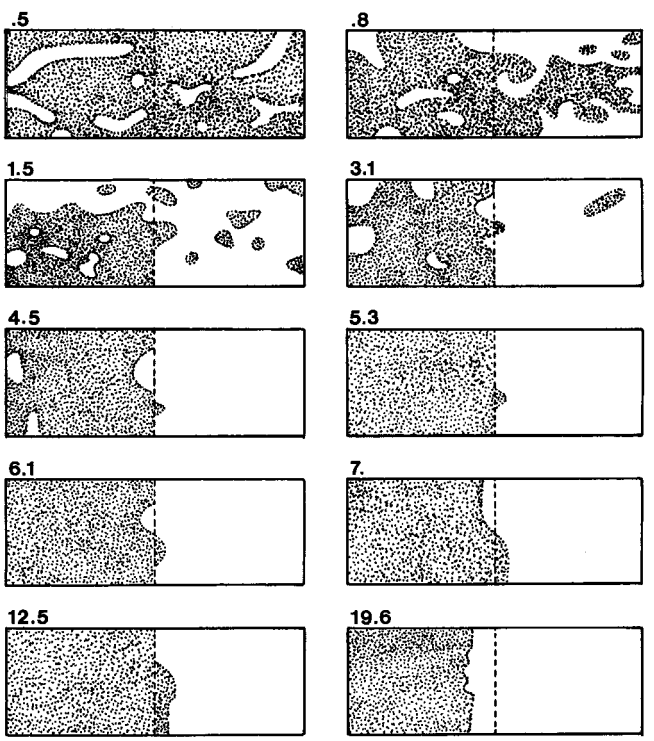

Fig. 4. The effect of plot size on the recognition of synphenetic pattern. The samples consist of random circles of area shown at top left of each map.

munities may seem to be the ultimate departure from real world, an «art for art's sake» approach. One should not forget, however, that just as in many other – and possibly more exact – fields of science, the simplifications inherent in simulation represent inevitable initial steps towards the understanding of the complexity of nature.

Acknowledgement. The author is grateful to N. C. Kenkel (University of Manitoba) whose critical comments and suggestions led to substantial improvements in the final version of this paper. I also thank P. Juhász-Nagy (L. Eötvös University), L. Orlóci (University of Western Ontario) and G. Copp (University of Lyon) for comments on early drafts of the manuscript.

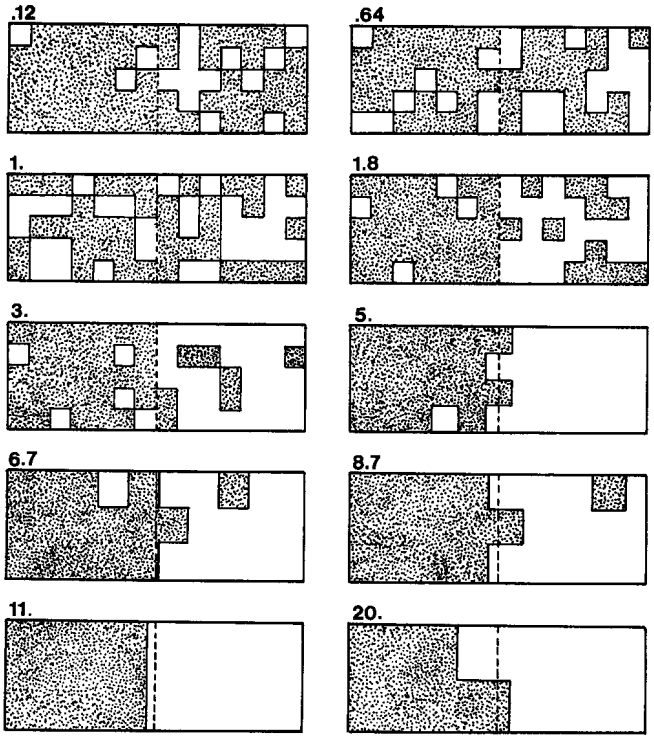

Fig. 5. The effect of plot size on the recognition of synphenetic pattern. Each sample consist of systematically arranged quadrats of area shown at top left of each map.

REFERENCES

ARVANITIS, L. G. and W. G. O'REGAN. 1967. Computer simulation and economic efficiency in forest sampling. Hilgardia 38:133-164.

BATCHELER, C. L. 1973. Estimating density and dispersion from truncated or unrestricted joint point-distance nearest neighbour distances. Proc. N.Z. Ecol. Soc. 20:131-147.

BONNICKSEN, T. M. and E. C. STONE. 1981. The giant Sequioa-mixed conifer forest community characterized through pattern analysis as a mosaic of aggregations. Forest Ecol. and Manage. 3:307-328.

CARPENTER, S. R. and J. E. CHANEY. 1983. Scale of spatial pattern: four methods compared. Vegetatio 53:153-160.

CLIFF, A. D. and J. K. Ord. 1981. *Spatial Processes.* Pion, London.

COCHRAN. W. G. 1977. *Sampling Techniques.* 3rd ed. Wiley, New York.

CORMACK, R. M. and J. K. Ord (eds). 1979. *Spatial and Temporal Analysis in Ecology.* Statistical Ecology vol. 8. International Co-operative Publishing House, Fairland, Maryland, USA.

CZÁRÁN, T. 1984. A simulation model for generating patterns of sessile populations. Abstracta Botanica 8:1-13.

DIGGLE, P. J. 1979. Statistical methods for spatial point patterns in ecology. In: R. M. Cormack and J. K. Ord (eds), pp. 95-150.

DIGGLE, P. J. 1983. *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

EK, A. R. 1971. A comparison of some estimators in forest sampling. Forest Sci. 17:2-13.

FRANKLIN, J., J. MICHAELSEN and A. H. STRAHLER. 1985. inSpatial analysis of density dependent pattern in coniferous forest stands. Vegetatio 64:29-36.

GALIANO, E. F. 1982. Pattern detection in plant populations through the analysis of plant-to-all-plants distances. Vegetatio 49:39-43.

GAUCH, H. G. and R. H. WHITTAKER. 1972. Coenocline simulation. Ecology 53:446-451.

GAUCH, H. G. and R. H. WHITTAKER. 1976. Simulation of community patterns. Vegetatio 33:13-16.

GOODALL, D. W. and N. E. WEST. 1972. An integrated set of computer programs for studying plant dispersion patterns. Paper read to meeting of Ecological Society of America, Minneapolis, Aug. 31, 1972. Abstract in Bull. Ecol. Soc. Amer. 53:44.

GOODALL, D. W. and N. E. WEST. 1979. A comparison of techniques for assessing dispersion patterns. Vegetatio 40:15-27.

GREEN, P. J. and R. SIBSON. 1978. Computing Dirichlet tessellations in the plane. Computer J. 21:168-173.

GREIG-SMITH, P. 1983. *Quantitative Plant Ecology*, 3rd ed. Butterworths, London.

HAHN, I. 1982. Einige Probleme der Probeentnahme bei der Schatzung der Anten- und Individuendiversität II. Eine mögliche Individuenzahlkorrektion. Botanikai Közlemények 69:59-70. (in Hungarian, with German summary).

HELTSHE, J. F. and N. E. FORRESTER. 1985. Statistical evaluation of the jackknife estimate of diversity when using quadrat samples. Ecology 66:107-111.

JUHÁSZ-NAGY, P. 1976. Spatial dependence of plant populations. I. Equivalence analysis (an outline for a new model). Acta Bot. Acad. Sci. Hung. 22:61-78.

JUHÁSZ-NAGY, P. 1984. Spatial dependence of plant populations. 2. A family of new models. Acta Bot. Hung. 30:363-402.

JUHÁSZ-NAGY, P. and J. PODANI. 1983. Information theory methods for the study of spatial processes and succession. Vegetatio 51:129-140.

KOBAYASHI, S. 1981. Diversity indices: relations to sample size and spatial distribution. Jap. J. Ecol. 31:231-236.

LAFRANCE, C. R. 1972. Sampling and ordination characteristics of computer-simulated individualistic communities. Ecology 53:387-397.

LAGONEGRO, M. 1984. SPAGHET: A coenocline simulator useful to calibrate software detectors. Stud. Geobot. 4:63-99.

LAMACRAFT, R. R., M. H. FRIEDEL and V. H. CHEWINGS. 1983. Comparison of distance based density estimates for some arid rangeland vegetation. Aust. J. Ecol. 8:181-187.

LIM, T. M. and H. W. KHOO. 1985. Sampling properties of Gower's general coefficient of similarity. Ecology 66:1682-1685.

MATÉRN, B. 1979. The analysis of ecological maps as mosaics. In: R. M. Cormack and J. K. Ord (eds), pp. 271-288.

NOSEK, J. N. 1976. Comparative analysis of some diversity functions under different conditions of sampling in sandy meadow. Acta Bot. Acad. Sci. Hung. 22:415-436.

O'REGAN, W. G. and M. N. PALLEY. 1965. A computer technique for the study of forest sampling methods. Forest Sci. 11:99-114.

O'REGAN, W. G., R. W. SEEGRIST and R. L. HUBBARD. 1973. Computer simulation and vegetation sampling. J. Wildlife Manage. 37:217-222.

ORLÓCI, L. 1972. On objective functions of phytosociological resemblance. Am. Midland Nat. 88:28-55.

ORLÓCI, L. 1978. *Multivariate Analysis in Vegetation Research.* 2nd ed. Junk, The Hague.

ORLÓCI, L., C. R. RAO and W. M. STITELER. (eds). 1979. *Multivariate Methods in Ecological Work.* Statistical Ecology vol. 7. International Co-operative Publishing House, Fairland, Maryland, USA.

PALLEY, M. N. and W. G. O'REGAN. 1961. A computer technique for the study of forest sampling methods. I. Point sampling compared with line sampling. Forest Sci. 7:282-294.

PICKFORD, S. G. and J. W. HAZARD. 1978. Simulation studies on line intersect sampling of forest residue. Forest Sci. 24:468-483.

PODANI, J. 1984a. Spatial processes in the analysis of vegetation: theory and review. Acta Bot. Hung. 30:75-118.

PODANI, J. 1984b. Analysis of mapped and simulated vegetation patterns by means of computerized sampling techniques. Acta Bot. Hung. 30:403-425.

PODANI, J. 1984c. SYN-TAX II. Computer programs for data analysis in ecology and systematics. Abstracta Botanica 8:73-94.

PRENTICE, I. C. and M. J. A. WERGER. 1985. Clump spacing in a desert dwarf scrub community. Vegetatio 63:133-139.

RICKLEFS, R. E. and M. LAV. 1980. Bias and dispersion of overlap indices: results of some Monte Carlo simulations. Ecology 61:1019-1024.

RIPLEY, B. D. 1979. Simulating spatial patterns: Dependent samples from a multivariate density. Appl. Stat. 28:109-112.

RIPLEY, B. D. 1981. *Spatial Statistics.* Wiley, New York.

STAUFFER, H. B. and G. D. NIGH. 1981. Available: A computer model which simulates quadrat sampling for tree density and spacing. Forest Sci. 27:31-32.

SUKWONG, S., W. E. FRAYER and E. W. MOGREN. 1971. Generalized comparisons of the precision of fixed-radius and variable-radius plots for basal-area estimates. Forest Sci. 17:263-271.

SWAN, J. M. A. 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology 51:89-102.

SZÖCS, Z. 1979. New computer-oriented methods for the study of natural and simulated vegetation structure. In: L. ORLÓCI et al. (eds), pp. 301-308.

WARREN, W. G. and C. L. BATCHELER. 1979. The density of spatial patterns: robust estimation through distance methods. In: R. M. Cormack and J. K. Ord (eds), pp. 247-270.

WENSEL, L. C. 1975. The treatment of boundary-line overlap in a forest sampling simulator. Hilgardia 43:143-159.

WENSEL, L. C. and H. H. JOHN. 1969. A statistical procedure for combining different types of sampling. Forest Sci. 15:307-317.

WILLIAMS, W. T., G. N. LANCE, L. J. WEBB, J. G. TRACEY and J. H. CONNELL. 1969. Studies in the numerical analysis of complex rain-forest communities. IV. A method for the elucidation of small-scale forest pattern. J. Ecol. 57:635-654.

APPENDIX

Computer programs for simulated sampling

The following programs, which are available under the FORTRAN package SYN-TAX II (Podani 1984c), are applicable to the study of point patterns within a rectangular study region:

ELSAM: random plots of circular, elliptical or rectangular shape with random or uniform orientation are taken. The output is a species by plots matrix of counts.

SAMPROC: systematic and restricted random sampling by circular or rectangular plots are simulated.

EXPRES: random pairs of plots are simulated to calculate the expectation of six resemblance coefficients.

The entire program package may be ordered from:

SISSAD, Viale Campi Elisi 62, Trieste, Italy.

Programs mentioned in the recent literature include:

INTRSCT: line intersect sampling simulator program (Pickford and Hazard 1978). The sampling universe is described by individual piece dimensions and spatial locations.

An unnamed FORTRAN program was written by Galiano (1982) for calculating plant-to-all-plant distances based on plant co-ordinates. The output includes a histogram of distances.

SAMPLE: An interactive FORTRAN program by Stauffer and Nigh (1981). It takes a sample from a simulated regular, random, or aggregated pattern of trees using circular plots placed randomly or systematically. The output consists of the estimate for density, the frequency distribution of the number of individuals per plot, the results of goodness of fit tests, and indices of spatial pattern. This program is extremely useful for instructive purposes. Inquiries about this program should be sent to: Research Branch, Ministry of Forests, 1450 Government Street, Victoria, B.C. Canada V8W 3E7.

Manuscript received: May 1986