Teaching and Learning Sampling Using Computer Simulation and Expert Systems

Robin M Reich - Fort Collins, Colorado, USA
Loukas G Arvanitis - Gainesville, Florida, USA

1. Introduction

Sampling is the process of collecting, analysing, and interpreting data in order to test hypotheses and provide needed information for intelligent decisions. Sampling is an attractive alternative to complete enumeration. Instead of measuring or recording all members of a population of interest, one concentrates on a representative proportion, objectively selected, to estimate desired characteristics of the target population.

Sampling involves a number of critical decisions which affect the reliability of the estimates. For example, what factors influence the sample size? Does a specific design meet cost and precision criteria? How does the size, shape, and allocation of individual sampling units affect the efficiency of estimators? Under what conditions does stratified random sampling, for example, yield more precise results than simple random sampling? These types of questions must be addressed prior to selecting the most appropriate sampling method for a given situation.

Sukhatme and Sukhatme (1970) define relative efficiency (RE) between two procedures, A and B, as follows:

$$RE = \frac{1/V_B}{1/V_A} = \frac{V_A}{V_B} \quad (1)$$

where V_A and V_B are the variances that result from each one of the two procedures. If t_1 and t_2 are two estimated of a population parameter, t_1 is more efficient than t_2 if its mean square error (MSE) is less than that of t_2. The estimator t_1 is more precise than t_2 if $V(t_1) < V(t_2)$. Thus, the information supplied by t_1 is measured by the inverse of its MSE. For unbiased estimators, the efficiency and the precision are the same, since the MSE equals the variance plus the square of the bias.

If $C(t_1)$ and $C(t_2)$ denote the cost of two sampling procedures that yield estimators t_1 and t_2 respectively, then the cost-efficiency (CE) of t_1 compared to that of t_2 is:

$$CE(t_1/t_2) = V(t_2)C(t_2)/V(t_1)C(t_1). \quad (2)$$

This relationship determines the ratio of the amount of information per unit cost in the two cases (Murthy, 1967).

A given sampling design is efficient if it provides the best possible precision for a fixed cost, or when it provides a fixed level of precision at the lowest possible cost. Both conditions must be considered when dealing with the constraints of a particular survey problem. For example, if one anticipates that the cost of sampling exceeds the available budget, the precision of the estimate may be reduced due to the smaller sample size. However, the user may also explore other designs capable of providing the desired level of precision within a given budget.

Although these concepts are simple and straightforward for statisticians, student comprehension of cost-efficiency in sampling is troublesome. This is particularly the case in forestry, ecology, soils, range, and other plant sciences. Part of the problem is the fact that the elementary sampling units in these populations have an area associated with them such as circles, squares, rectangles, strips, etc. Thus, for a desired precision or cost, the same efficiency may be achieved by a large number of combinations of the number of observations (sample size) and the area of the elementary sampling units. Alternative sampling designs capable of providing the required precision of an estimator without exceeding the budget, may also be explored.

To enhance student comprehension of basic sampling concepts, the authors have developed a Forest Sampling Simulator (FOSS) for microcomputers (Arvanitis and Reich, 1989). It has been well-documented by numerous studies that computer simulation fosters understanding of complex systems by permitting students to manipulate individual parts and observe the effects of their action on the rest of the model (Heerman, 1988). This system is described in Section 2 below. In Section 3 its possible augmentation by an expert system is described, which would automatically generate the most appropriate sampling strategy for a particular situation, given the appropriate input information.

2. The FOSS system

FOSS is an interactive microcomputer software program written in BASICA for IBM™ PCs or other fully compatible units. The program consists of three main components: (a) POPULATION, (b) PLOT CONFIGURATION, and (c) SAMPLING DESIGN. The first generates a realisation of a particular type of forest; the second selects the shape and character of the sampling unit; and the third specifies the type of sampling design to be used.

2.1 *Population*

Users have the option of generating three spatial patterns of trees or other objects

whose positions, specified by x-y rectangular coordinates, are fixed on a plane. The three spatial patterns are *random*, *aggregated* and *regular* (or uniform). For a given number of trees in a population, *random* patterns are created by a pseudo-random generator whose starting point is determined randomly by the internal computer clock (Figure 1).

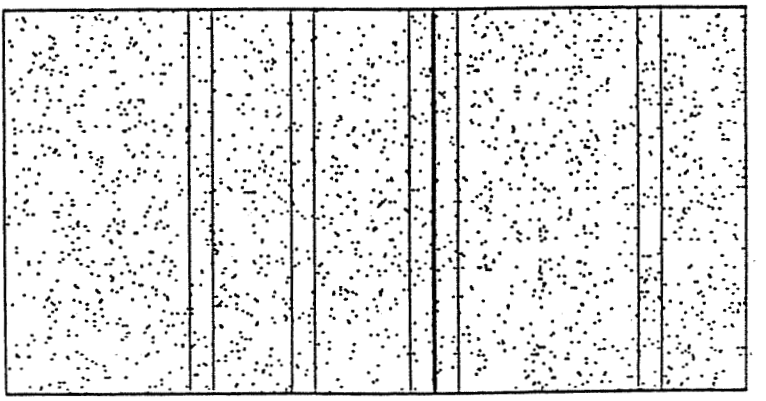

FIGURE 1

Random spatial pattern of 2000 tree centres within an area of 3.48ha. An example of strip sampling with replacement. Strips are approximately 7m in width.

Aggregated patterns (Figure 2) are simulated in two sequential steps. First, a user-specified number of cluster centres is located at random within the boundaries of the population. Second, the average number of trees per cluster and the size of the clusters are determined by a probability (P) and a scaling factor (S).

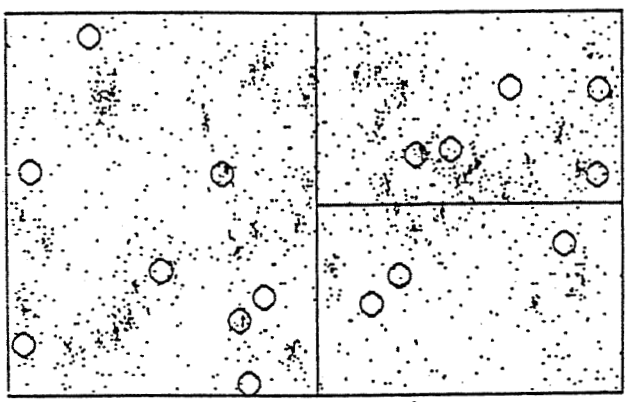

FIGURE 2

Aggregated spatial pattern of 2000 tree centres within an area of 3.48ha. Number of clusters 50; scaling factor .02; probability of aggregation 0.8. An example of stratified random sampling with Neyman allocation. Circular plots are approximately $51m^2$ in size.

The parameter (P) denotes the probability that a given tree will be located within randomly selected clusters. For each tree in the population, a pseudo-random number is generated between 0 and 1. If the number is greater than P, the tree is randomly located within the boundaries of the forest. If the number is less than P, the tree is randomly assigned to a cluster. Its distance from the cluster centre (D) is determined as follows (Reich, 1980):

(3) $$D = -\ln(RND)/S$$

where RND is a random number between 0 and 1, S is a scaling factor ($.01 < S < 0.2$), and ln is the natural logarithm.

The azimuth of an individual tree from the vertex of the cluster centre is determined by a random angle between 0 and 360 degrees. Tree distances extending beyond the population boundaries are continued along the same azimuth on the opposite side of the forest.

Regular or plantation-type forests or other plant populations are simulated by defining the distance between individual trees within and between rows (Figure 3).

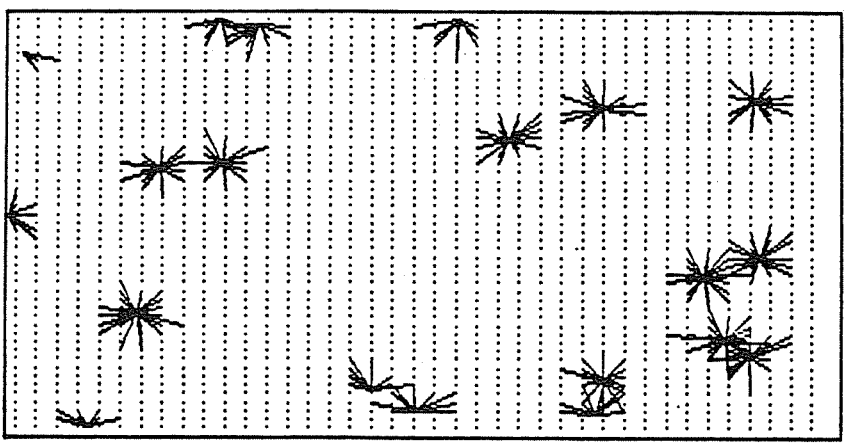

FIGURE 3

Plantation-type spatial pattern of 2000 tree centres in a forest of 3.48ha. Distances between rows 12 units; distance between trees in a row 6 units. Twenty random pps point samples (basal area factor $1.1m^2$). Lines radiating out from centre of points indicate distances to trees included in the sample.

Gradient: On command, FOSS generates an East-West and/or North-South gradient of plant locations (Figure 4) by transforming the original x-y coordinates using an exponential function.

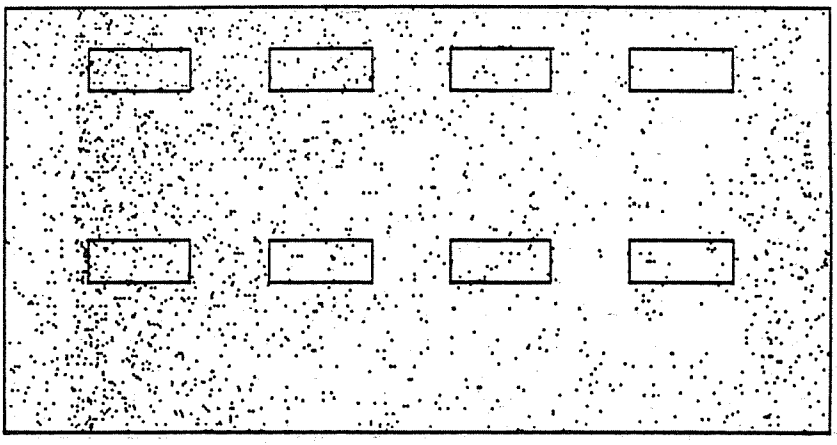

FIGURE 4

Spatial pattern of 2000 tree centres within an area of 3.48ha. An East-West gradient of value 1 is present. Eight rectangular plots of approximately 405m^2 each are systematically arranged.

With all these options which can be handled by FOSS, an infinitely large number of forests or other plant populations can be generated to depict very closely real-world settings and test the efficiency of various sampling designs. For each simulated forest, the user has the flexibility of selecting either the normal or the Weibull distribution for the population variable of interest.

2.2 *Plot configuration*

FOSS handles six different shapes of elementary sampling units: square, rectangle, circle, point, line, and strip. With the exception of point sampling, once a specific shape is selected, the user has the option of choosing different sizes of elementary sampling units. This applies also to lines (transects) where different lengths may be user-specified. Transect sampling is very popular among foresters, wildlife specialists, geologists, ecologists, soil specialists, and other scientists.

2.3 *Sampling Designs*

FOSS handles nine sampling designs: simple random with replacement (Figure 1), stratified random (Figure 2), systematic (Figure 4), list (pps), 3-P (sampling with probability proportional to prediction), double sampling, two-stage, quadrats, and distance sampling (Figure 5).

The students can select any of the populations, plot configurations and sampling designs described above. Each trial yields summary tables with sample statistics and true population parameters. Comparison of standard errors and confidence intervals are easy to perform in order to compute relative efficiencies among sampling designs. Students may repeat the process as many times as deemed appropriate to acquire a firm grasp of sampling variability.

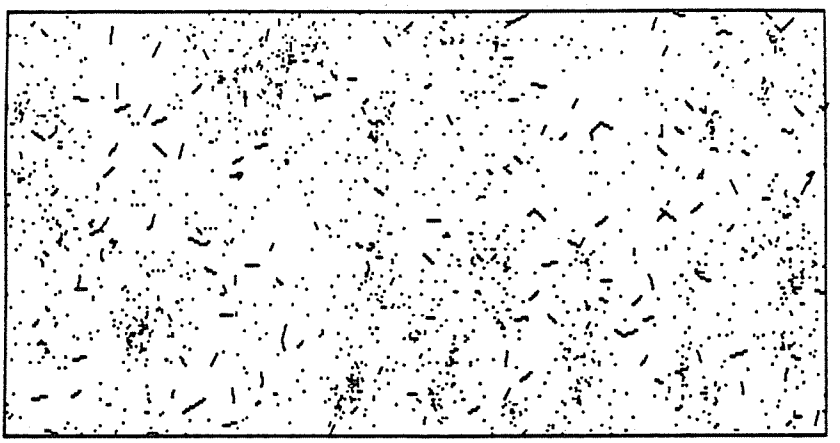

FIGURE 5

Aggregated spatial pattern of 2000 tree centres within an area of 3.48ha. Number of cluster 100; scaling factor .01; probability of aggregation 0.9. An example of distance sampling using 150 point-to-tree and tree-to-tree distances. Lines denote distances from points and tree centres included in the sample.

3. A conceptual expert simulation system for sampling designs

FOSS has been very successful in stimulating students to think independently and evaluate the outcome of their decisions. If one accepts the notion that the primary responsibility of a teacher is to assist students in advancing from memorising to independent thinking, then programs like FOSS provide a valuable contribution to the learning process. Through FOSS, users with relatively weak quantitative backgrounds begin to comprehend basic concepts of statistics and sampling. A logical extension of FOSS would be an intelligent procedure, a knowledge base reasoning system, to assist students in deciding which one of the many designs available to them is the best to apply in a specific case. Such a system should take into consideration the characteristics of the population, the available resources, and other constraints such as time, trained personnel, and priorities. To accomplish this task effectively without an intelligent computer-aided system would be cumbersome.

3.1 *Knowledge base system*

A knowledge base system would assist students in selecting a particular sampling design from among a number of alternatives, which would balance the cost and precision requirements in each specific case. As a result, students' comprehension of sampling efficiency and their ability to implement cost-effective inventories when elements are fixed in space (trees, lesser plants, and other items of interest).

During the past decade, a new decision-making tool, referred to as artificial intelligence (AI), has emerged. This new science involves programming computers to

imitate human behaviour that requires intelligence to make rational decisions. One remarkable and promising technology within AI involves knowledge base systems, often referred to as expert systems (ES). An ES is a logical program that enables a computer to mimic an expert in helping humans diagnose problems and select among alternative actions (Barrett and Jones, 1989). ES are developed using other computer programs, capable of manipulating symbols and numbers, thus enabling the computer to represent knowledge and logic symbolically. An ES can be used to teach non-experts the problem-solving approaches of experts (Holt, 1989).

In addition, simulation models like FOSS can provide critical input to ES to improve decisions when an expert's knowledge is limited (Beck and Jones, 1989). For example, in selecting the optimal plot area and shape, one has to have prior information on the structure of the population such as spatial distribution, number of elements, and amount of understory vegetation. Also, information is needed on the relationship between sampling variability and plot area. Thus, it seems reasonable to link simulation models with ES to provide reliable estimates of the above relationships essential to the decision-making process.

If information is available from a previous survey or pre-sampling, FOSS could be employed to simulate the relationship between plot size and sampling variability. This information may be subsequently used by the ES to identify the optimal plot area and sample size that minimises total survey time (measurement plus travel time) for a pre-determined level of precision. Because of the flexibility of FOSS, one can also approximate how this relationship changes with different plot configurations (square, rectangle, circle, strip, etc.) and spatial patterns such as random, aggregated, or uniform (Reich, 1980).

If one is contemplating whether to use, say, double sampling with regression instead of simple random sampling, it is important that a strong linear relationship exists between the auxiliary variable (x) and the variable of interest (y). The absence of such a relationship may affect the precision of the survey. Thus, if computers can identify feasible solutions under such conditions, human intelligence is being imitated and enhanced.

A characteristic of most ES is that incomplete and uncertain information can be used. These programs often pursue an alternative line of reasoning when information is unknown or suggest less confidence in the answer provided. To overcome this problem, uncertainty on the part of the user may be represented in the form of numerical certainty factors (Spiegelhalter, 1986). Such factors commonly range from -1 to 1, where -1 represents complete confidence that the information is false, and 1 implies complete confidence that it is true. Numerical certainty factors are measures of confidence, not statistical probabilities.

If the user is not confident that a strong linear relationship exists between, say, the auxiliary variable (x) and the variable of interest (y), one would be better off using simple random sampling. Thus, this type of information may be used to rank alternative sampling designs, based on the level of certainty of the information being provided to the ES.

3.2 *Brief description of the expert simulation system*

An expert simulation system usually has six main counterparts (Hayes-Roth et

al., 1983): language processor, input, knowledg base, design storage, design evaluation, and design refinement.

Language processor: An ES is usually based on a problem-oriented language, such as Level Five™ (Level Five, 1988), which relies on a user-friendly and fixed vocabulary to, (a) conveniently input and display information for the user, and (b) select and apply appropriate design-based rules using input from the user.

Input: The ES needs specific information before it can identify an appropriate sampling design. Such information falls into three categories: (a) characteristics of the population (i.e. spatial distribution of trees, number of trees per unit area, dimension of individual trees, density and height of the understory vegetation, etc.); (b) objectives and constraints (precision, costs, time) associated with the sampling design; and (c) knowledge gained from the previous surveys that may aid in selecting the most appropriate sampling design.

A major problem frequently encountered in designing an efficient survey is the lack of reliable information on the structure and variability of the population of interest. In this case, the system may use default stored in the knowledge base, or information obtained from a simulation model, such as FOSS (Arvanitis and Reich, 1989). This approach will approximate missing information, which may subsequently be improved as needed in the design phase.

Knowledge base: The knowledge base is the heart of the system. It uses rule-based formulae to capture human expert knowledge and experience which is stored for retrieval. Knowledge base rules are divided into two groups: design default rules and design change rules. The design default rules are used to develop the initial survey. The design change rules are used to modify the sampling method to ensure that it satisfies the stated objectives.

Design storage: Design specifications can be stored to the user would be able to modify or refine the sampling method at a later time.

Design evaluation: The initial sampling procedure may not satisfy all the constraints (precision, costs, and time) imposed by the user. In this phase, the ES would evaluate and rank applicable sampling designs to determine their conformity with objectives. If a sampling design does not meet the specifications, the system can make appropriate modifications using the design change rules stored in the knowledge base.

Design refinement: Using the knowledge base design rules, one will be able to modify the sampling method. In addition, changes would be possible on some of the constraints, previously imposed on the design, if this is deemed appropriate.

3.3 *Advantages of an expert simulation system*

The main value of an expert simulation system, such as Figure 6, is that students will learn to play an active role in designing efficient forest and other related surveys. Such a system also allows students to explore a wide variety of alternative solutions that would otherwise be impossible to accomplish in the real world. Other advantages include: (a) linking theoretical understanding with practical applications; (b) discovering new functional relationships among variables of interest; (c) building student-controlled simulations that employ theoretical concepts taught in a course in natural resource sampling; and (d) develop computer skills that will be applicable to their professional careers. It is anticipated that an expert system will foster critical

thinking and enhance problem-solving ability of students.

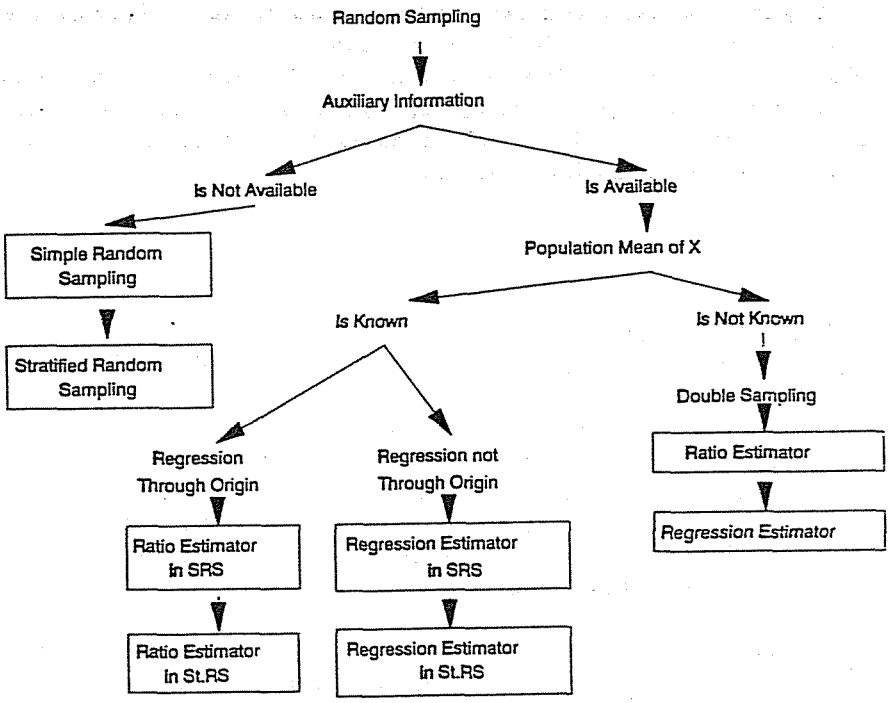

FIGURE 6
Flow chart of design base rules for selecting the "best" sampling design

References

Arvanitis, L G and Reich, R M (1989) Sampling simulation with a microcomputer. *COENOSES* 4(2), 73-80.

Barrett, J R and Jones, D D (eds) (1989) *Knowledge Engineering in Agriculture.* American Association of Agricultural Engineers, St Joseph, MI.

Beck, H W and Jones, J W (1989) Simulation and artificial intelligence. In: J R Barrett and D D Jones (eds) *Knowledge Engineering in Agriculture.* American Association of Agricultural Engineers, St Joseph, MI, 117-135.

Hayes-Roth, F, Waterman, D A and Lenat, D B (1983) *Building Expert Systems.* Addison-Wesley Publishing Co Inc, Reading, MA.

Heerman, B (1988) *Teaching and Learning with Computers.* Jossey-Bass Publishers, San Francisco.

Holt, D A (1989) The growing potential of expert systems in agriculture. In: J R Barrett and D D Jones (eds) *Knowledge Engineering in Agriculture.* American Association of Agricultural Engineers, St Joseph, MI, 1-11.

Level Five Research Inc (1988) *Level Five.* Fifth Avenue, Indiatlantic, FL.

Murthy, M N (1967) *Sampling Theory and Methods.* Statistical Publishing Society,

Calcutta.

Reich, R M (1980) *An Evaluation of Distance Sampling and Ratio Estimation : A Case Study in Slash Pine Plantations Infected by Fusiform Rust.* Master of Science Thesis, University of Florida, Gainesville, 139pp.

Spiegelhalter (1986) Uncertainty in expert systems. In: W Gale (ed) *Artificial Intelligence and Statistics.* Addison-Wesley Publishing Co Ltd, Reading, MA, 17-255.

Sukhatme, P V and Sukhatme, B V (1977) *Sampling Theory of Surveys with Applications.* Asia Publishing House, New Delhi.