Guided Transect Sampling
- An Outline of the Principle

Göran Ståhl
Anna Ringvall
Tomas Lämås

Arbetsrapport 19 1997

SVERIGES LANTBRUKSUNIVERSITET
Institutionen för skoglig resurshushållning
och geomatik
S-901 83 UMEÅ
Tfn: 090-16 58 25 Fax: 090-14 19 15

ISSN 1401-1204
ISRN SLU-SRG-AR--19--SE

ABSTRACT

Guided transect sampling is primarily intended for the sampling of sparse, geographically scattered, populations for which there exist no list of the units. Basically, it consists of a two-stage design, using wide strips in the first stage and a subsampling procedure in each strip in the second stage. The subsampling is guided by prior information, e.g. in the form of remote sensing image data. Different strategies can be used for the guidance, resulting in different probabilities of inclusion of population units, and consequently in slightly different estimators.

The general principle for second stage subsampling guidance can be coupled with a number of methods for how the samples should be selected along the "guided route". Strip sampling, line transect sampling, adaptive cluster sampling, and plot sampling are examples of methods that can be used. However, in the theoretical set-up of the method, it is assumed that all objects in grid-cells passed by the survey transect are found. The grid-cells, covering the entire area under study, also contain the covariate data that are used for directing the sampling effort.

INTRODUCTION

In sampling, existing information about the population studied or about some covariate can be utilized in many ways to enhance the precision of the estimate of the parameter of interest. Techniques such as stratification and PPS use the information for the sample selection, while, e.g., ratio or regression estimators use the information for estimation purposes (e.g. Cochran 1977). The two ways of using prior information can be combined.

Sparse populations, for which there exist no list, generally pose substantial problems in sampling. This is the case in forestry, e.g. when studying elements of importance for biodiversity or a population of some threatened species. To be able to find enough sampling units for the precision to be acceptable, large areas must be covered. Generally, methods like line transect sampling (Burnham *et al.* 1980) or strip surveying (e.g. Lämås & Fries 1995) are used for the purpose, rather than the plot based methods of timber cruising. Still, inventories tend to be expensive and/or provide imprecise results.

In this paper, a method thought to be efficient for the sampling of sparse populations is presented. The method, *guided transect sampling*, is basically a two-stage design in which wide strips are laid out in the first stage. In the second stage, the subsampling within strips is guided by prior information, e.g. in the form of remote sensing image data. The method has some similarities with the covariate-directed sampling approach proposed by Patil *et al.* (1996).

THE METHOD

An overview of the method, in its basic form, is given in Figure 1 below. In the forest area delineated, strips too wide to be entirely surveyed are first randomly laid out. Secondly, a route for the subsampling within each strip is guided by prior information. The details of this guidance are described below.

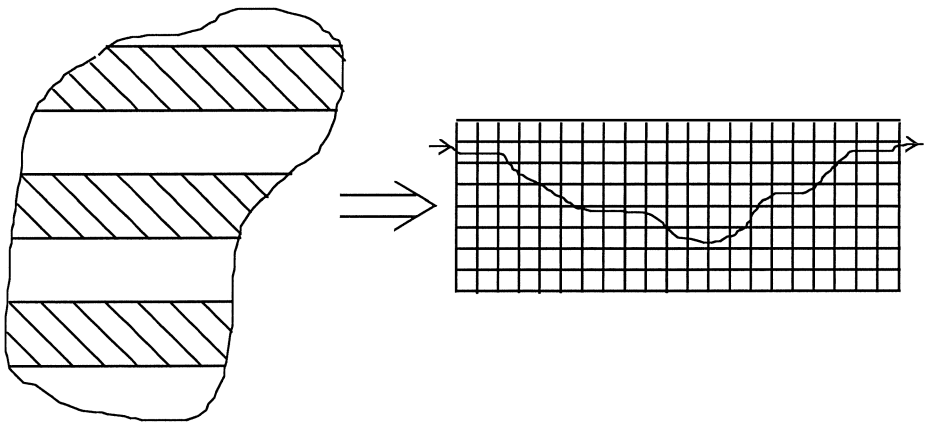

FIGURE 1. A general outline of guided transect sampling. A first stage sampling of wide strips (left) is followed by a second stage guided subsampling within each strip (right).

The entire area of interest is partitioned into grid cells of some suitable size, e.g. 20 by 20 meters (or possibly rectangular to simplify the field work). For each such cell, a covariate value is assessed prior to the sampling. E.g., the covariate could be the estimated volume of deciduous trees in case the population under study is known to prefer deciduous forests to coniferous forests. Such prior volume estimates can be obtained by, e.g., using satellite data and the kNN-method (e.g. Nilsson 1997).

In order to facilitate the theoretical description of the method, an assumption is made that all sampling units are detected and counted/measured once the surveyor enters the grid cell they are situated in. Also, the method relies on use of GPS, differential in real time, for the guidance of the surveyor through the forest. However, in simple cases it should also be possible to use a compass and a measuring tape.

The first stage strips are laid out randomly, with the restriction that they should match with the system of grid cells. The second stage is a subsampling of grid-cells along a survey line within each first stage strip. Many different strategies can be used for determining where the second-stage transect should be located. One basic idea is, however, that the grid-cells in some manner should be selected with probabilities proportional to their covariate values. Another basic idea is that the field-work should not be too complicated, implying that the survey transect should be some, more or less, connected curve from the beginning to the end of a strip. This is also the reason for introducing the first stage strips. Without them, the survey lines within the forest area would tend to be complicated. Finally, the idea is also to use some line-based inventory rather than plots in order to obtain a more efficient search for individuals of the sparse population. In this theoretical description of the method, a strip survey is approximated by a continuous survey of neighboring grid-cells. The surveyor is assumed to perform an entire search for objects in all grid-cells entered.

Conforming to all this, many different strategies for the subsampling within each strip can still be identified. Some straightforward possibilities are:

i) Random walk (Markovian) with the probability to enter a neighboring cell, in the direction of the survey line, given by the cell's covariate value (Figure 2a).

ii) As (i) but allowing the surveyor to step from a particular cell to any of the grid-cells in front. That is, "big steps" are allowed, since the surveyor in this case may go directly from one side of the strip to the other. The strip will no longer be connected (Figure 2b).

iii) Random simulation of entire transects through a strip (without considering the covariate data at this stage). Transition is only allowed to neighboring cells. A large number of transects are simulated. For each one, the sum of cell-wise covariate values is calculated. This sum, or some transformation of it, is used for selecting one particular transect by PPS (Figure 2c).

To make the method useful from a practical point of view, the grid-cells should generally be rectangular (very elongated in the direction of the strips) in order to avoid too much zigzagging for the surveyor. An alternative to this would be to assign higher probabilities for straight continuation than for changing to another row in the grid-cell system. However, in all figures in this theoretical desciption of the method, square grid-cells are used.

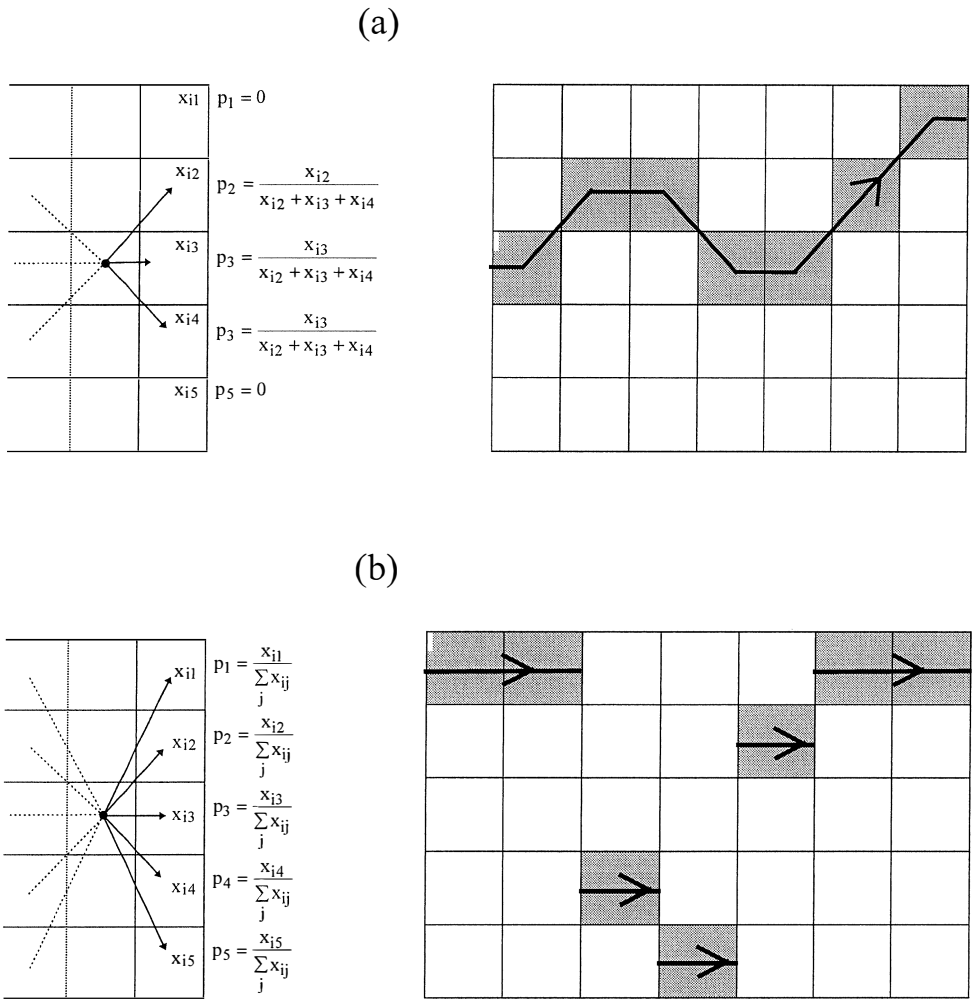

FIGURE 2. Different principles for guiding the subsampling. In (a) transition is only allowed to neighboring cells, in (b) transition is allowed to any onward cell, while in (c) entire transects are simulated. In (a) and (b), the probabilities of transition (the p-values) are determined from the covariate values (x-values) in the next stage, denoted i. In (c), entire transects are determined from the sum (Q-values) of covariates in grid-cells visited.

The choice of principle for the guidance will affect the precision of the method. The first principle (i) may appear appealing from a practical and computational point of view. The risk is, however, that this rather short-sighted approach will not be very efficient. The reason is outlined in Figure 3.

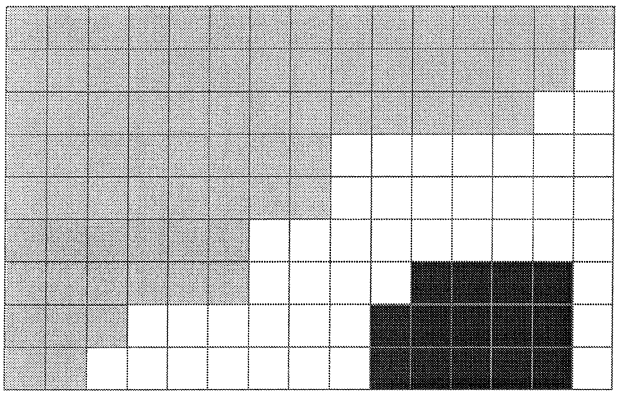

FIGURE 3. The risk of using guiding principle (i). The "hot-spot" in the lower right part of the strip will usually be missed due to the spatial pattern of covariate data.

Due to the spatial distribution of grid-cells with high covariate values, the surveyor will in this example very seldom find the potential "hot-spot" in the lower right part of the strip if guiding strategy (i) is used. This problem is avoided if strategy (ii) is used. A problem with this strategy is, however, that it is somewhat impractical in the field, since the surveyor will often be instructed to move from one side of the strip to the other. To enhance the strategy, the probabilities could be weighted by the distance one would have to move, but this will not be further discussed here. Another possibility would be to perform measurements also when moving "vertically".

Strategy (iii) is probably a good compromise. It will lead to "connected" transects, and problems as the one in Figure 3 will be avoided. The computational burden will, however, be substantial.

ESTIMATION

Whatever strategy is used for guiding the subsampling, the same general principle can be used for the estimation of population parameters such as the population total. The general principle is to use the Horvitz-Thompson (HT) estimator, by which a population total is always unbiasedly estimated as:

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

Here, y_i is the quantity of interest on the i^{th} unit sampled, π_i its probability of inclusion, and n is the number of objects in the sample. The probabilities of inclusion of different grid-cells, and thus the probabilities of inclusion of population units in these grid-cells, will vary

depending on what strategy is used for guiding the subsampling and also on the width and number of strips used. The derivation of the probabilities of inclusion of grid-cells will now be described for the three cases. Knowing these probabilities, the HT-estimator can be used for estimating a population total.

To derive the probabilities, the situation is first studied conditioned on the outcome of the first stage layout of strips. That is, the probability of inclusion of a grid-cell within a given strip is first studied. Then, the probability of inclusion of a grid-cell given the entire design is derived. The conditional probabilities of inclusion are called θ_{ij}, i being an index for grid-cell column ("stage") and j an index for grid-cell row ("state").

Case (i): To obtain the conditional probability of inclusion for a particular grid-cell in the first case, recursive calculations must be made starting from the beginning of the strip. The grid-cell to enter the strip in is selected by PPS among all possible cells in stage 1. The conditional probability of inclusion for state j in stage 1 is consequently, with x being the covariate value:

$$\theta_{1j} = \frac{x_{1j}}{\sum_k x_{1k}}$$

Looking next at the conditional probabilities of inclusion for grid-cells in stage 2, these depend on the probabilities of inclusion of cells in the first stage. The following recursive formula could be used from stage 2 onwards, to the end of the strip:

$$\theta_{ij} = \sum_{m \in N_{-1}} \theta_{(i-1)m} \cdot \frac{x_{ij}}{\sum_{k \in M} x_{ik}}$$

In this formula, N_{-1} is the set of neighbors to grid-cell ij in stage i-1 (usually consisting of three cells, but at the boundaries of the strip only two). Moreover M is the set of neighbors to the grid-cell (i-1)m in stage i, giving the possible transitions from cell (i-1)m onwards.

Case (ii): In this case, no recursive calculations are needed since the conditional probability of inclusion of a grid-cell depends only on the covariate values of the grid-cells in that particular stage. The probabilities are given by:

$$\theta_{ij} = \frac{x_{ij}}{\sum_k x_{ik}}$$

Case (iii): In this case entire transects are first simulated without considering covariate information. Next, covariate data for cells passed are summed, and perhaps transformed, giving a value Q_k for the entire transect k. A large number of transects are simulated and one of them selected PPS to the Q-values. The conditional probability of inclusion of grid-cell ij will in this case be:

$$\theta_{ij} = \frac{\sum_{m \in S} Q_m}{\sum_{k \in K} Q_k}$$

Here, S is the set of transects that pass grid-cell ij, and K is the set of all transects simulated.

So far, the conditional probabilities of inclusion of grid-cells have been derived. To arrive at unconditional probabilities, the first stage random layout of strips must also be considered.

Different approaches to unbiased estimation can be identified here. One possibility is to use the conditional probabilities to estimate the population "subtotal" within each strip and then to use standard techniques from strip sampling to calculate the overall population total for the forest area.

Another possibility would be to recognize that the conditional probability of inclusion of a grid-cell can be very different depending on what strip it happens to belong to. This is due to that the environment of the grid-cell, within the strip, affects how often a transect will pass it. Consequently, conditional probabilities should be calculated for a specific grid-cell for all possible ways in which a strip may be laid out ($\theta = 0$ if the grid-cell is outside the strip). Then, the unconditional probability of inclusion is the average of the conditional probabilities, if all potential strips are equally likely.

In case more than one strip is randomly laid out, the single strip probability of inclusion should be multiplied by the number of strips to obtain the expected number of inclusions of a grid-cell (which is used just as the probability of inclusion in the HT-estimator).

ASSESSING THE PRECISION

So far, no studies have been made concerning the properties of guided transect sampling, such as assessment of the precision of the method. One way to assess the precision would be to simulate different populations and assume a stochastic model for the relationship between population units and covariate data. Given such an outcome of a "forest", the formula for the true variance of the HT-estimator could be used to assess the precision in an "ideal" case. The variance of the HT-estimator is:

$$V(\hat{Y}) = \sum_{i=1}^{N} \frac{\pi_i (1-\pi_i)\, y_i^2}{\pi_i^2} + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)\, y_i\, y_j}{\pi_i\, \pi_j}$$

To calculate this variance for a given design and "forest", the individual probabilities of inclusion of each grid-cell should be calculated, as well as the joint probabilities of inclusion of each pair of grid-cells (π_{ij}). The y-values are the quantities of interest for the population units and N is the total number of individuals in the population. Calculation of π_i-values is made according to the principles outlined above. Calculation of π_{ij}-values can be made in the same general manner (looking at one pair of grid-cells at the time). For large grid-cell matrices, the computational effort required could be substantial.

DISCUSSION

The basic method outlined can be adjusted and extended in many ways. E.g., all cells along the transect need not be inventoried. In case the population is not too sparse, a subsample of grid-cells, or circular plots, along the transect would be more efficient.

For practical reasons, the approach outlined is not very appropriate. It would probably be very laborious to perform the inventory in grid-cells along the transect. To make the method more useful from a fieldwork point of view, strip sampling or line transect sampling would

probably be better alternatives. Moreover, the GPS-guidance cannot be expected to be very accurate. At least some ten meters deviation from the line must often be expected. Still, the guidance principle outlined could probably be used as an adequate approximation.

The cost of inventory could probably also be considered in a non-standard way. Including a digital elevation model, e.g., the cost of moving from one grid-cell to the other could be included and movements upwards or downwards in steep terrain be avoided.

Moreover, PPS could be used not only for selecting the route through strips, but also for selecting the strips. These could, e.g., be selected with probability proportional to their total covariate values.

Adaptive designs would also be a challenge. Adaptive cluster sampling (Thompson 1992) would be possible to include without too much difficulty. Adaptive designs with regard to how the probabilities of transition from one grid-cell to the next should be selected would be substantially more difficult, yet interesting, to address.

The covariate data should in many cases also be possible to use for estimation purposes in, e.g., ratio or regression estimators.

Finally, an alternative approach would be to let the surveyor choose his route through the first stage strips subjectively. Using intuition and knowledge about the objects or species searched for, the surveyor should select a route providing a high probability for finding units of the population. The GPS would record the route selected, and, possibly, the probability of inclusion of units could be estimated using a model for the relationship between covariate data and the subjective choice of route. A simpler way of estimation would be to define strata from the covariate information and use poststratification, although units in the strata are subjectively selected.

ACKNOWLEDGEMENTS

The authors thank Prof. G.P. Patil, Dr. W. Myers, Dr. C. Taillie, Dr. Z. Lou, Prof. B. Ranneby, and Ass. Prof. S. Holm for fruitful discussions.

LITERATURE

Burnham, K.P., Anderson, D.R. and Laake, J.L. 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monograph 72.

Cochran, W.G. 1977. Sampling techniques (3d ed.). John Wiley & Sons, New York.

Lämås, T. and Fries, C. 1995. An integrated forest inventory in a managed north Swedish forest landscape for estimating growing stock and coarse woody debris *in* The Monte Verità conference on forest survey designs, May 2-7 1994, Monte Verità, Switzerland, Köhl, M., Bachmann, P., Brassel, P. and Preto, G. (eds.). Swiss Federal Institute for Forest, Snow and Landscape Research, Birmensdorf.

Nilsson, M. 1997. Estimation of forest variables using satellite image data and airborne lidar. Swedish University of Agricultural Sciences, Silvestria 17.

Patil, G.P., Grigiletto, M. and Johnson, G. 1996. Using covariate-directed sampling of Emap hexagons to assess the statewide species richness of breeding birds in Pennsylvania. Pennsylvania State University, Dept. of Statistics. Technical report 95-1102.

Thompson, S. 1992. Sampling. John Wiley & Sons, New York.

Serien Arbetsrapporter utges i första hand för institutionens eget behov av viss dokumentation.

Författarna svarar själva för rapporternas vetenskapliga innehåll.

1995 1 Kempe, G. Hjälpmedel för bestämning av slutenhet i plant- och ungskog. ISRN SLU-SRG-AR--1--SE

2 Riksskogstaxeringen och Ståndortskarteringen vid regional miljöövervakning. - metoder för att förbättra upplösningen vid inventering i skogliga avrinningsområden. ISRN SLU-SRG-AR--2--SE.

3 Holmgren, P. & Thuresson, T. Skoglig planering på amerikanska västkusten - intryck från en studieresa till Oregon, Washington och British Columbia 1-14 augusti 1995. ISRN SLU-SRG-AR--3--SE.

4 Ståhl, G. The Transect Relascope - An Instrument for the Quantification of Coarse Woody Debris. ISRN SLU-SRG-AR--4--SE.

5 Törnquist, K. Ekologisk landskapsplanering i svenskt skogsbruk - hur började det?. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--5--SE.

1996 6 Persson, S. & Segner, U. Aspekter kring datakvaliténs betydelse för den kortsiktiga planeringen. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--6--SE.

7 Henriksson, L. The thinning quotient - a relevant description of a thinning? Gallringskvot - en tillförlitlig beskrivning av en gallring? Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--7--SE.

8 Ranvald, C. Sortimentsinriktad avverkning. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--8--SE.

9 Olofsson, C. Mångbruk i ett landskapsperspektiv - En fallstudie på MoDo Skog AB, Örnsköldsviks förvaltning. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--9--SE.

10 Andersson, H. Taper curve functions and quality estimation for Common Oak (Quercus Robur L.) in Sweden. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--10--SE.

11 Djurberg, H. Den skogliga informationens roll i ett kundanpassat virkesflöde. - En bakgrundsstudie samt simulering av inventeringsmetoders inverkan på noggrannhet i leveransprognoser till sågverk. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--11--SE.

12 Bredberg, J. Skattning av ålder och andra beståndsvariabler - en fallstudie baserad på MoDo:s indelningsrutiner. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--12--SE.

13 Gunnarsson, F. On the potential of Kriging for forest management planning. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--13--SE.

14 Holm, S. & Thuresson, T. samt jägm.studenter kurs 92/96. En analys av skogstillståndet samt några alternativa avverkningsberäkningar för en del av Östads säteri. ISRN SLU-SRG-AR--14--SE.

15 van Kerkvoorde, M. A sequential approach in mathematical programming to include spatial aspects of biodiversity in long range forest management planning. ISRN SLU-SRG-AR--15--SE.

16 Tormalm, K. Implementering av FSC-certifiering av mindre enskilda markägares skogsbruk. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN SLU-SRG-AR--16--SE.

1997 17 Engberg, M. Naturvärden i skog lämnad vid slutavverkning. - En inventering av upp till 35 år gamla föryngringsytor på Sundsvalls arbetsområde, SCA. Examensarbete i ämnet skogsuppskattning och skogsindelning. ISRN-SRG-AR--17--SE.

18 Christoffersson, P & Jonsson, P. Avdelningsfri inventering - tillvägagångssätt och tidsåtgång. ISRN-SRG-AR--18--SE.

19 Ståhl, G., Ringvall, A. & Lämås, T. Guided transect sampling - An outline of the principle. ISRN-SRG-AR--19--SE.