Guided Transect Sampling—A New Design Combining Prior Information and Field Surveying

Anna Ringvall, Göran Ståhl, and Tomas Lämås

Abstract.—Guided transect sampling is a two-stage sampling design in which prior information is used to guide the field survey in the second stage. In the first stage, broad strips are randomly selected and divided into grid-cells. For each cell a covariate value is estimated from remote sensing data, for example. The covariate is the basis for subsampling of a transect through the broad strip. The general idea of guided transect sampling can be combined with different transect-based field inventory methods. This paper describes guided transect sampling and provides some results from a theoretical evaluation.

Sparse populations, for which no list exists, generally pose substantial problems in sampling. This is the case in forestry, such as when studying elements of importance for biodiversity or a population of some indicator species. The plot-based methods generally used for timber cruising tend to give imprecise estimates for rare objects since only a small area is covered and a relatively long time is spent traveling between plots. For sparse objects, transect-based methods like strip surveying should be more cost-efficient because a larger area is covered (Lämås and Fries 1995). Line transect sampling (e.g., Buckland *et al.* 1993) has mainly been used for assessing wildlife populations, but is also a good alternative for sparse inanimate objects. Lately, adaptive cluster sampling (Thompson 1990, Roesch 1993) has been proposed as an efficient method for sparse and aggregated populations. Still, inventories can be expensive and/or provide imprecise results.

Prior information about the population studied can be used in many ways to enhance the precision of estimates. Techniques such as stratification and probability proportional to size (PPS) use the information for selecting the samples, while ratio and regression estimators, for example, use the information for estimation purposes.

This paper concerns a newly developed method (Ståhl *et al.* 1997) thought to increase the efficiency in transect-based sampling methods when the object of interest is sparse and geographically scattered. The method, guided transect sampling (GTS), is a two-stage probability sampling design. In the first stage, wide strips are laid out, and in the second stage, the subsampling of a transect inside the wide strip is guided by prior information, as in the form of remote sensing data. The aim of this paper is to give a general description of GTS and to provide some preliminary results regarding its efficiency.

THE METHOD

An overview of the method is provided in figure 1. In the target area delineated (left in fig. 1), broad strips are first randomly laid out. The strips are partitioned into grid-cells of some suitable size, e.g., 20 x 20 m. For each such cell, a covariate value is assessed prior to sampling. For example, the covariate could be the estimated volume of deciduous trees if the population studied is known to prefer deciduous trees to conifers in mixed stands. Such volume estimates can be obtained from satellite imagery or aerial photos (e.g., Tomppo 1986, Nilsson 1997).

Secondly, grid-cells are subsampled along a survey line across each first-stage strip (right in fig. 1). The subsampling is based on the covariate values in each grid-cell. The general idea of GTS can be combined with different kinds of transect-based inventory methods, such as strip surveying, line transect sampling, and adaptive cluster sampling with a strip as the initial unit. However, in the theoretical description of the method in this article, GTS is supposed to be combined with a strip survey

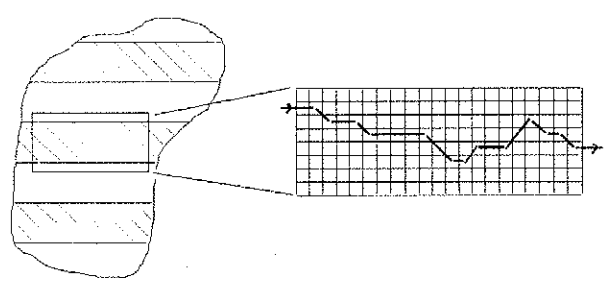


Figure 1.—*The general outline of GTS. A first-stage sampling of wide strips (left) is followed by a second-stage guided subsampling of a transect within each strip (right).*

Graduate Research Assistant, Professor in Forest Inventory, and Assistant Professor in Forest Management Planning, respectively, Department of Forest Resource Management and Geomatics, Swedish University of Agricultural Sciences Umeå, Sweden.

approximated as a survey of the entire grid-cells selected. In reality, the method also relies on the use of GPS, differential in real time, to guide the survey or along the predetermined transects.

Strategies for Guidance

Different strategies can be used for selecting the second-stage transect. Some straightforward possibilities are:

1. Random walk with transitions allowed only to neighboring cells in the next grid-cell column (fig. 2a). The decision about which cell to enter is made PPS to the neighboring cells' covariate values.

2. As strategy 1, but allowing the surveyor to step from a particular cell to any of the grid-cells in front. The second-stage transect will no longer be connected (fig. 2b).

3. Random simulation of entire transects through the first-stage strip without considering the covariate data. Transitions are allowed only to neighboring cells in the next grid-cell column as in strategy 1. However, here, there is an equal probability of transition to each of the neighboring cells since no covariate data are considered at this moment. Many transects are simulated, and for each transect the sum of covariate values in all grid-cells passed by that transect (the Q-value) is calculated. Finally, one transect is selected PPS to this sum of covariates (fig. 2c).

The choice of subsampling strategy will affect the simplicity of applying the method in the field as well as the precision of GTS. Theoretically, strategy 2 should be most efficient since cells with large covariate values always have a high probability of being selected. The simplicity of surveying along a continuous transect will, however, be lost. Strategy 1 gives connected transects, but it might be inefficient because interesting areas sometimes will have a very low probability of being selected. Strategy 3 could be regarded as a compromise between strategy 1 and 2. It will lead to connected transects, and the inefficiency that might occur with strategy 1 is reduced if enough transects are simulated.

ESTIMATION

Attention will first be given to conditional estimation within a primary unit. The general principle is to use the Horvitz-Thompson (HT) estimator (e.g., Cochran 1977), by which a first-stage strip total, Y_i, is unbiasedly estimated as:

$$\hat{Y}_i = \sum_{j=1}^{m_i} \frac{y_{ij}}{\pi_{ij}} \qquad (1)$$

Here, y_{ij} is the characteristic of interest in the jth grid-cell sampled in first-stage strip i, (π_{ij} the probability of including this grid-cell in the sample, and m_i the number of grid-cells sampled in first-stage strip i.

To arrive at an estimator for the entire area under study, the method for selecting the first-stage strips must be considered. If this selection is made by simple random sampling, assuming all primary units to have equal size, an estimator of the population total, $\hat{Y}_{tot}$, is obtained as:

$$\hat{Y}_{tot} = \frac{1}{n}\frac{A}{a}\sum_{i=1}^{n}\hat{Y}_i \qquad (2)$$

with A being the total area of the compartment, a the area of a first-stage strip, n the number of first-stage strips sampled, and $\hat{Y}_i$ the estimator of the population (sub-)total within first-stage strip i.

Inclusion Probabilities

The probabilities of inclusion of grid-cells will vary depending on what strategy is used for guiding the subsampling. Below, the derivation of inclusion probabilities will be described for the three cases previously discussed. For simplicity, a rectangular forest with m columns is assumed. All first-stage strips are assumed to have a width of r rows and to extend across the entire area under study, parallel with two of the sides in the rectangle. All derivations are made within a selected first-stage strip. In this section, notations for grid-cells are two-dimensional, indicating the row and the column.

Strategy 1: The grid-cell to enter the strip is selected PPS among all possible cells in column 1. The probability of inclusion of the grid-cell in row j, column 1, is then:

$$\pi_{j1} = \frac{x_{j1}}{\sum_{l=1}^{r} x_{l1}} \qquad (3)$$

with x_{j1} being the covariate value of the jth cell in the first column and r the number of rows in the first-stage strip. Note that all the x-values must be larger than zero for the strategy to be theoretically sound. If a grid-cell's covariate is zero initially, some small number must be added before the formula is applied.

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

$$\pi_{jk} = \sum_{l \in \Omega_{k-1}} \left(\pi_{l(k-1)} \cdot \frac{x_{jk}}{\sum_{s \in \psi_l} x_{sk}} \right) \qquad (4)$$

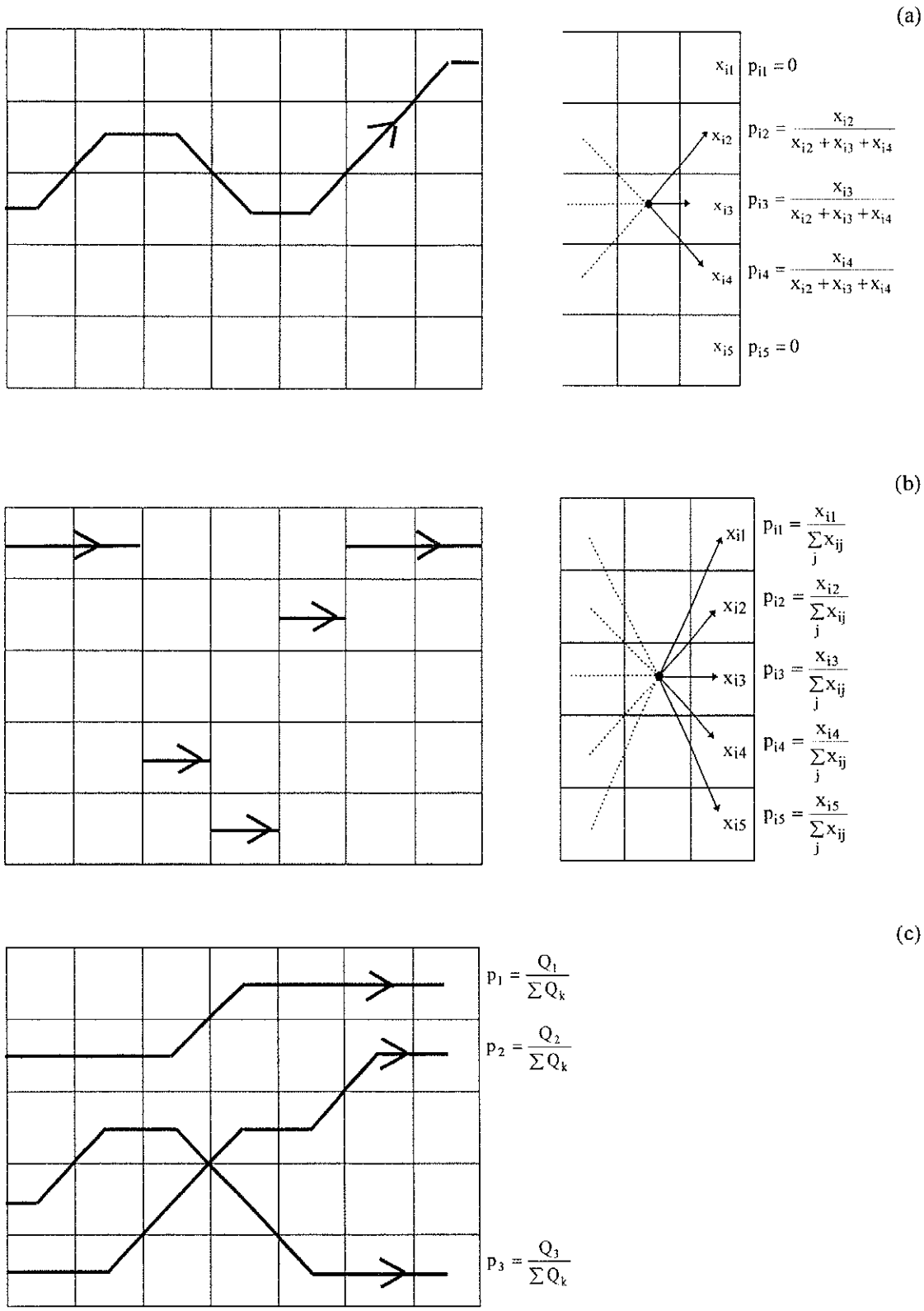


Figure 2.—*Different strategies for guiding the subsampling. In (a) transition is allowed only to neighboring cells, in (b) transition is allowed to any onward cell, and in (c) entire transects are simulated. In (a) and (b), the probabilities of transition to the onward cells (the p-values) are determined from the covariate values (x-values) in the cells in the onward column, here denoted i. In (c), an entire transect is selected PPS to the sum (Q-values) of covariates in grid-cells passed by the transect.*

In this formula, Ω_{k-1} is the set of neighbor cells in the previous column (k-1) of the grid-cell in row j, column k. The set usually consists of three cells, although at the upper and lower boundary of a first-stage strip, it consists of only two cells. Moreover, Ψ_l is the set of neighbors in front of the grid-cell in row l, column k-1, providing the possible transitions from that particular grid-cell onwards. Generally, this set also consists of three cells except at the borders of the first-stage strip.

Strategy 2: In this case the probability of inclusion of a grid-cell depends only on the covariate values of the grid-cells in that particular column. The probabilities are always given by:

$$\pi_{jk} = \frac{x_{jk}}{\sum_{l=1}^{r} x_{lk}} \tag{5}$$

As in strategy 1, all x-values must be larger than zero; otherwise, a small number must be added to all the x-values before applying the formula.

Strategy 3: Here, entire transects are first simulated without considering the covariate information. Next, covariate data for grid-cells passed by the transect are summed, resulting in a value Q_l for the entire transect l. Out of the large number of transects simulated, one is selected PPS to the Q-values. The probability of inclusion of a particular grid-cell (conditioned on the realization of transects) is:

$$\pi_{jk} = \frac{\sum_{s \in S} Q_s}{\sum_{l \in L} Q_l} \tag{6}$$

Here, S is the set of transects that pass the grid-cell in row j, column k, and L is the set of all transects simulated. For this formula to be theoretically sound, every grid-cell must be passed by at least one transect.

Precision of Estimates

The precision, in terms of the variance, can be derived from the general formula for conditional variances (e.g., Cochran 1977):

$$Var(\hat{Y}_{tot}) = Var\left[E(\hat{Y}_{tot}|\,S)\right] + E\left[Var(\hat{Y}_{tot}|\,S)\right] \tag{7}$$

where $\hat{Y}_{tot}$ is the estimator of the population total and S the set of first-stage strips in the sample.

Under the assumption of simple random sampling, without replacement, of n first-stage strips among N possible, formula (7) can be further elaborated to:

$$Var(\hat{Y}_{tot}) = \left(\frac{A}{a}\right)^2 \frac{1}{nN}\left[\left(\frac{N-n}{N-1}\right)\sum_{i=1}^{N}(Y_i - \bar{Y})^2 + \sum_{i=1}^{N} Var(\hat{Y}_i)\right] \tag{8}$$

Here, N is the total number of first-stage strips, and the rightmost variance terms are those from the second stage guided transect inventories. These are calculated with the formula for the variance of an HT estimator (e.g., Cochran 1977):

$$Var(\hat{Y}_i) = \sum_{j=1}^{M_i} \frac{(1-\pi_j)}{\pi_j} y_j^2 + \sum_{j=1}^{M_i}\sum_{k \neq j}^{M_i} \frac{(\pi_{jk} - \pi_j \pi_k)}{\pi_j \pi_k} y_j y_k \tag{9}$$

where π_j and π_k now is the inclusion probability of grid-cell j and k, respectively, π_{jk} the joint inclusion probability of the cells j and k, M_i the number of grid-cells in first-stage unit i, and y_j the quantity of interest in the jth cell. The joint probabilities of inclusion are calculated according to the same principles that are used for calculating the individual probabilities of inclusion. Focus is set on one grid-cell at the time, and the probability that this particular cell occurs in the sample simultaneously with another is calculated.

COMPARISON OF GTS VERSUS STRIP SURVEYING

To give an indication of the efficiency of GTS and the performance of the different strategies for subsampling, comparisons were made with a traditional strip survey in two fictitious "forests." The example was designed to resemble inventories of some scarce species in restricted areas of coniferous forests with patches of deciduous trees. Many red-listed species depend on deciduous trees (Berg *et al.* 1993), and the covariate was considered to be the volume of deciduous trees assessed from aerial photos or satellite imagery. The parameter of interest could then be the population total of some quantity, such as the total biomass of the species of interest. The two forests had different levels of scarcity of the covariate and the species of interest.

Simulation of Fictitious Forests

The two forests (Forest I and Forest II) consisted of rectangular matrices of 50 by 25 grid-cells. The quantity of the covariate and variable of interest was simulated in four steps:

1. First, each grid-cell was independently assigned either a covariate value equal to zero, or a covariate value greater than zero. The proportion of grid-cells with a nonzero covariate was 0.3 in Forest I and 0.1 in Forest II.
2. For grid-cells with a covariate value greater than zero, the value was independently simulated from a normal distribution with mean 50 and standard deviation 35, truncated to prevent values below zero.

3. The probability of occurrence of the species of interest in grid-cell j, P_j, was determined with the following (slightly modified) logistic model:

$$P_j = \frac{\exp(\alpha + \beta \cdot x_j) + P^*}{1 + \exp(\alpha + \beta \cdot x_j)} \qquad (10)$$

where x_j is the value of the covariate in grid-cell j, P^* the probability of occurrence of the species of interest when the covariate value is zero (0.03 in Forest I and 0.01 in Forest II), α and β parameters chosen to make P_j close to P^* when the value of the covariate is close to zero and P_j near 1 when the covariate is above 100 (α=-5, β=0.01). A uniform random number was generated, and if the random number was smaller than P_j, the species of interest occurred in grid-cell j.

4. For cells with occurrence of the species of interest, the biomass of the species, y_j, was modeled as:

$$y_j = \gamma \cdot x_j + \varepsilon_j \cdot x_j + \delta_j \qquad (11)$$

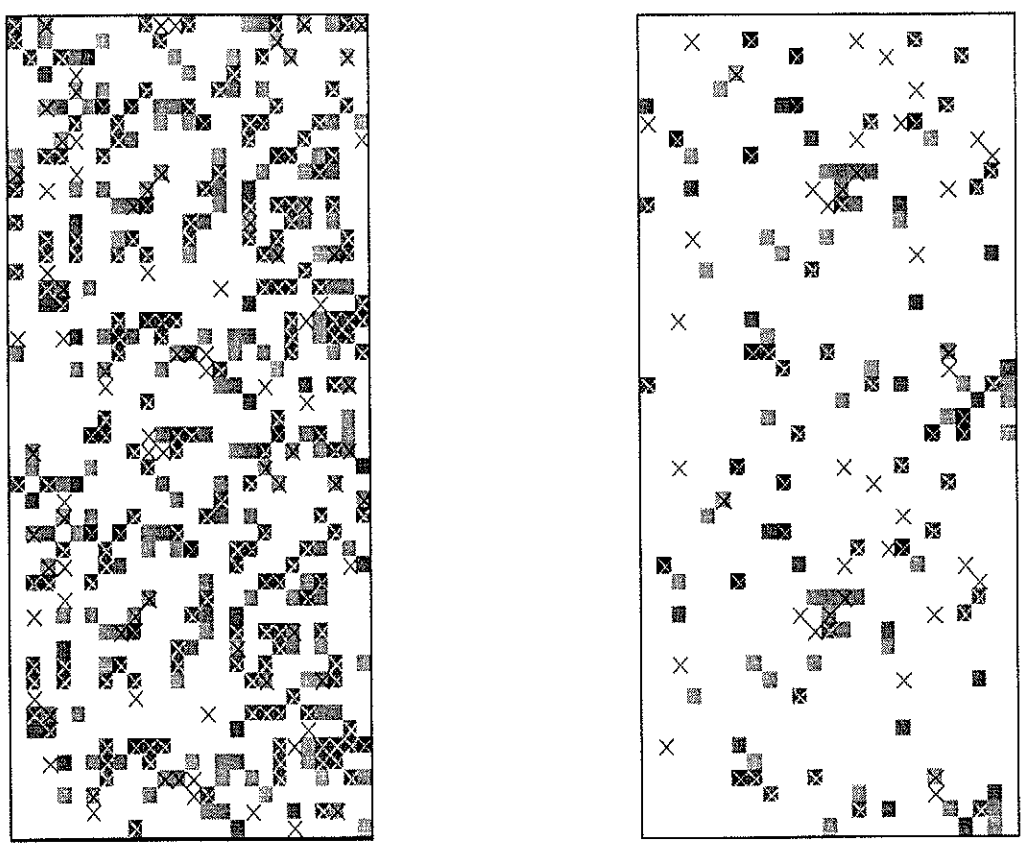
where x_j is the value of the covariate in grid-cell j, γ a scaling constant (γ=0.2), ε_j a normally distributed random variable (μ=0, σ^2=0.02), and δ_j a log-normally distributed random variable (μ=0, σ^2=1), added to assign a substantial random component to grid-cells where the value of the covariate was low.

Finally, a small constant was added to all grid-cells to give grid-cells with a covariate equal to zero an inclusion probability larger than zero.

Both forests were produced in 10 replicates to avoid conclusions based on extreme patterns in single outcomes. Examples of the two forest types are shown in figure 3. The mean values of the population total in the 10 replicates were 3164.3 (Forest I) and 1122.8 (Forest II), respectively.

Inventory Design

In this theoretical setup, it was assumed that entire grid-cells were selected and that the width of the strip or transect was equal to the size of the grid-cells so that



(I) (II)

Figure 3.—*Examples of the two simulated forests used in the study. Colored squares indicate grid-cells with a nonzero covariate. The darker the color, the higher the value of the covariate. The crosses indicate presence of the species of interest in a grid-cell.*

selected grid-cells were entirely surveyed. Considering this setup and the small forests used, there is a finite number of possible samples of strips and primary units that makes it possible to analytically calculate the estimators' true variances. No simulations are thus involved in the comparison of the methods' efficiency, only in generating the forests.

For GTS, variances were calculated for a design in which one first-stage strip was selected with simple random sampling. Within the first-stage strip, a transect was subsampled using each of the three strategies for guidance earlier described. For each possible first-stage strip, inclusion probabilities and joint inclusion probabilities were calculated for each grid-cell and each pair of grid-cells using formulas (3-6). Then, the variance of the estimator was calculated using formulas (8-9). Two different widths for the first-stage strips were tested, 5 and 10 rows of grid-cells, respectively. For the traditional strip survey, variances were calculated for a design where one strip was selected with simple random sampling. The width of the strip was one grid-cell, and thus equal numbers of grid-cells were sampled with both GTS and the standard strip survey. The variance for the estimator was calculated as:

$$Var(\hat{Y}_{tot}) = \left(\frac{A}{a_{ST}}\right)^2 \frac{1}{nN_{ST}} \frac{N_{ST}-n}{N_{ST}-1} \sum_{i=1}^{N_{ST}} \left(Y_{i(ST)} - \overline{Y}_{ST}\right)^2 \qquad (12)$$

where $\hat{Y}_{tot}$ is the estimator of the population total, A the area of the forest, a_{ST} the area of the strip, N_{ST} the number of possible strips, n the number of strips sampled, $Y_{i(ST)}$ the total in the ith strip, and $\overline{Y}_{ST}$ the average of strip totals.

Results

Results are shown in figure 4 as:

$$R = \frac{Var_{GTS}(\hat{Y}_{tot})}{Var_{ST}(\hat{Y}_{tot})} \qquad (13)$$

where Var_{GTS} and Var_{ST} are the mean values of variances obtained with GTS and the standard strip survey, respectively, in each of the 10 replicates of the two forest types.

As suspected, strategy 2 was always most efficient in terms of statistical efficiency. Since this strategy leads to disconnected transects and, consequently, time spent walking without observing eventual objects, the strategy will probably be less cost-efficient. A possible improvement would be to allow transitions between rows only each 5th or 10th column. For the two strategies leading to connected transects, strategy 3 seems to be the most reliable. The risk of poor efficiency using strategy 1 is especially evident using a first-stage strip with a width of 10 rows in Forest I. An improvement of strategy 1 would be to look several columns ahead, not just one, when

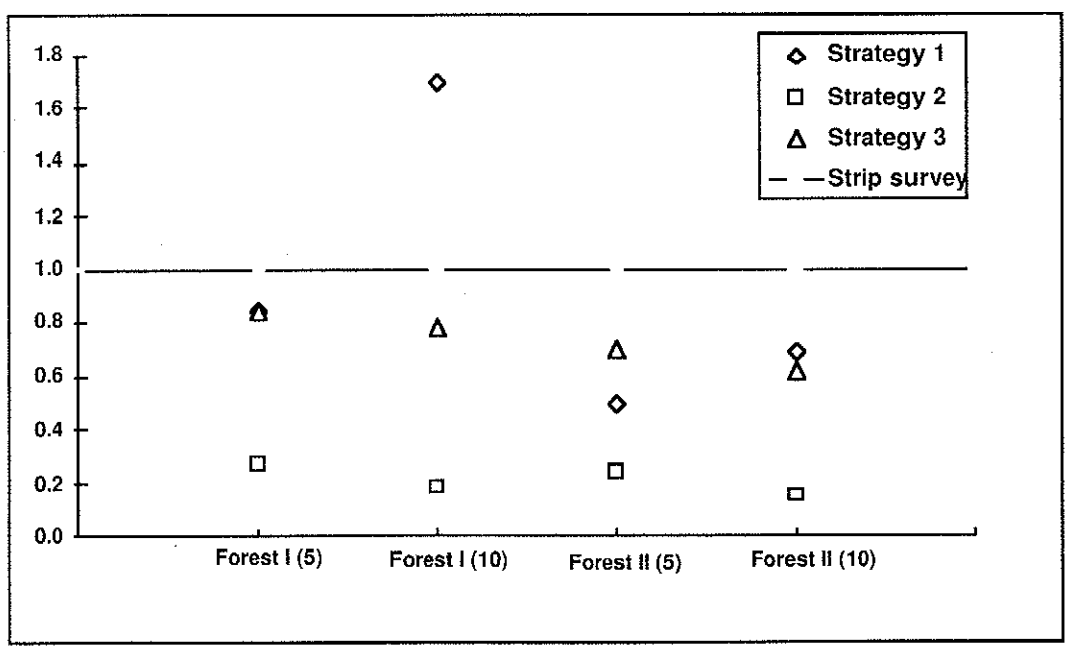


Figure 4.—*Comparison of guided transect sampling versus a standard strip survey, shown as the ratio between the variance of guided transect sampling and the variance of the strip survey. Results are shown for three different strategies of subsampling in two different forest types (I-II) and with two different widths of first-stage strips (5 and 10 rows of grid-cells).*

determining which transition to make. Both strategy 1 and strategy 3 seem to be relatively more efficient when the covariate and object of interest is less common.

DISCUSSION

This small study showed that GTS could improve the efficiency in a traditional strip survey for sampling sparse populations. However, the efficiency depends on the relationship between the variable of interest and the covariate, and when this relationship is expected to be poor, the method's potential is limited. Moreover, no costs were considered in this study, and although the same number of grid-cells was sampled using both GTS and strip surveying, the cost of performing GTS can generally be expected to be higher. Also, the availability and precision of the GPS guidance must be considered since deviations from the selected transects will affect the precision of the method.

The basic method presented can be developed and adjusted in many ways. The three strategies used for the second-stage guidance in this study are just some straightforward alternatives from a larger number of possibilities. Some improvements to the strategies developed for this study have already been discussed. Another more complicated, yet interesting, alternative would be to find an optimal strategy considering costs for entering different cells, and then for a given total cost, search for the transect giving the lowest variance, using some kind of optimizing routine.

In this study, GTS was combined with a strip survey, approximated as a survey of entire grid-cells selected. In practice, it would be desirable to survey strips along continuous lines. Then, only parts of grid-cells will be surveyed when changing rows in the grid-cell system, and inclusion probabilities will also have to be modified. Other interesting possibilities would be to combine GTS with line transect sampling or adaptive cluster sampling. This, however, would require further theoretical development. For objects or species that for some reason are difficult to survey using line-based methods, GTS could be combined with a sparse sample plot inventory.

Another possible way to improve the method would be to use the covariate data not only for selecting transects, but also for selecting the entire first-stage strips with PPS, and for estimation purposes.

ACKNOWLEDGMENT

The following people reviewed this manuscript: Jeffrey H. Gove, USDA Forest Service, Northeastern Research Station. Durham, NH, USA, and Sören Holm, Swedish University of Agricultural Sciences. Umeå, Sweden.

LITERATURE CITED

Berg, Å.; Ehnström, B.; Gustavsson, L.; Hallingbäck, T.; Jonsell, M.; Weslien, J. 1994. Threatened plant, animal and fungus species in Swedish forests: distribution and habitat associations. Conservation Biology. 8: 718-731.

Buckland, S.T.; Anderson, D.R.; Burnham, K.P.; Laake, J.L. 1993. Distance sampling. Estimating abundance of biological populations. London: Chapman & Hall. 446 p.

Cochran, W.G. 1977. Sampling techniques, 3d ed. New York, NY: John Wiley & Sons. 428 p.

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

Nilsson, M. 1997. Estimation of forest variables using satellite image data and airborne lidar. Silvestria 17. Umeå, Sweden: Swedish University of Agricultural Sciences. Ph.D. dissertation.

Roesch, F.A., Jr. 1993. Adaptive cluster sampling for forest inventories. Forest Science. 39: 655-669.

Ståhl, G.; Ringvall, A.; Lämås, T. 1997. Guided transect sampling. An outline of the principle. Working document 19. Umeå, Sweden: Department of Forest Resource Management and Geomatics. SLU. 9 p.

Thompson, S.K. 1990. Adaptive cluster sampling. Journal of American Statistical Association. 85: 1050-1059

Tomppo, E. 1986. Stand delineation and estimation of stand variables by means of satellite images. Research Notes 19. Helsinki, Finland: University of Helsinki, Department of Forest Mensuration and Management.