New Methods for Sampling Sparse Populations

Anna Ringvall[1]

Abstract.—To improve surveys of sparse objects, methods that use auxiliary information have been suggested. Guided transect sampling uses prior information, e.g., from aerial photographs, for the layout of survey strips. Instead of being laid out straight, the strips will wind between potentially more interesting areas. 3P sampling (probability proportional to prediction) uses information not available prior to the survey (e.g., the quality of downed logs), for selection of substrates for species inventories. Then, the surveyor's judgments of the substrates' suitability for the species of interest are used as the base for selection. Initial studies have shown that these methods have a potential to improve the efficiency of surveys of sparse populations.

Introduction

Biodiversity-oriented surveys often need to survey objects that from, a sampling perspective, are sparse or rare, which is problematic because precision of estimates generally will be low unless large areas are covered. In traditional forest surveys, oriented towards variables of interest for timber production, systematically distributed plots have turned out to be a cost-efficient method. When sampling sparse objects, however, the walking time is a proportionally large part of the cost for the survey. In that case, plot-based methods tend to be inefficient because only a small area is covered by the sample and a relatively long time is spent traveling between plots. With a transect-based method such as a strip survey, a larger area is covered for the same walking distance, which should be more cost-efficient (Ståhl and Lämås 1998).

Due to the problems of sampling sparse objects, probability sampling methods are sometimes abandoned for purposive sampling. In these cases, satellite images, aerial photos, or different types of maps together with knowledge about species' preferences are used to select areas to survey. In the surveyed areas, surveyors use their knowledge of species' habitat and substrate preferences to search for the species of interest. Such use of auxiliary information is also useful in probability sampling, both in the design and in the estimation phase, to improve the precision of estimates (e.g., Thompson 1992).

This article presents an approach for sampling sparse objects in which the same type of information and knowledge that is used in purposive surveys is utilized but in a probability sampling context. The approach has two steps, which can be used together or independent of each other. In the first step, information from satellite images or aerial photos is used to improve the design of a strip survey. In the second step, the surveyor's knowledge of species' substrate preferences is used for a more efficient subsampling of substrates for species inventories. Step one is a new method called guided transect sampling (Ståhl *et al.* 2000) and step two is an existing method, 3P sampling (probability proportional to prediction), but with a new application (Ringvall and Kruys 2005). The idea behind the approach is to use probability sampling methods to replicate a skilled surveyor purposively seeking for the species of interest.

Guided Transect Sampling

In the first step, information that can be obtained prior to the survey is used. For example information obtained from satellite images or aerial photos can be available and used at different scales. For example, if information is available at a stand scale it can be used to stratify stands and sample the

[1] Assistant Professor, Swedish University of Agricultural Sciences, Department of Forest Resource Management and Geomatics, S-901 83 Umeå, Sweden. E-mail: Anna.Ringvall@resgeom.slu.se.

more interesting strata more frequently. If the information is available at a pixel level, it can be used in a plot-based design (at least theoretically) for stratification of pixels. For sparse objects, however, it is probably most efficient to use all time in the forest to actually survey, which means using a transect-based method. Guided transect sampling (GTS) is one way of using auxiliary information with a high spatial resolution in the design of transect-based surveys.

An overview of the method is given in figure 1. Before the survey, the whole area is divided into a grid cell system with cells of some certain size. For each cell, a covariate value is obtained with help of remote sensing. The covariate is some variable that is related to the object of interest. In the first stage, wide strips are selected in the area to be surveyed, for example, with simple random sampling. Inside each selected wide strip a transect is selected across the strip by successively selecting one grid cell in each column. The selection of this survey line can be made in different ways, but the general idea is that it is in some way based on the covariate values in each grid cell. Some straightforward alternatives will later be described.
A strip survey is then conducted along the selected line. It should be a continuous survey but it is so far approximated by a survey of the entire grid cells selected. Instead of being laid out straight, the strip will wind between the potentially more interesting areas.

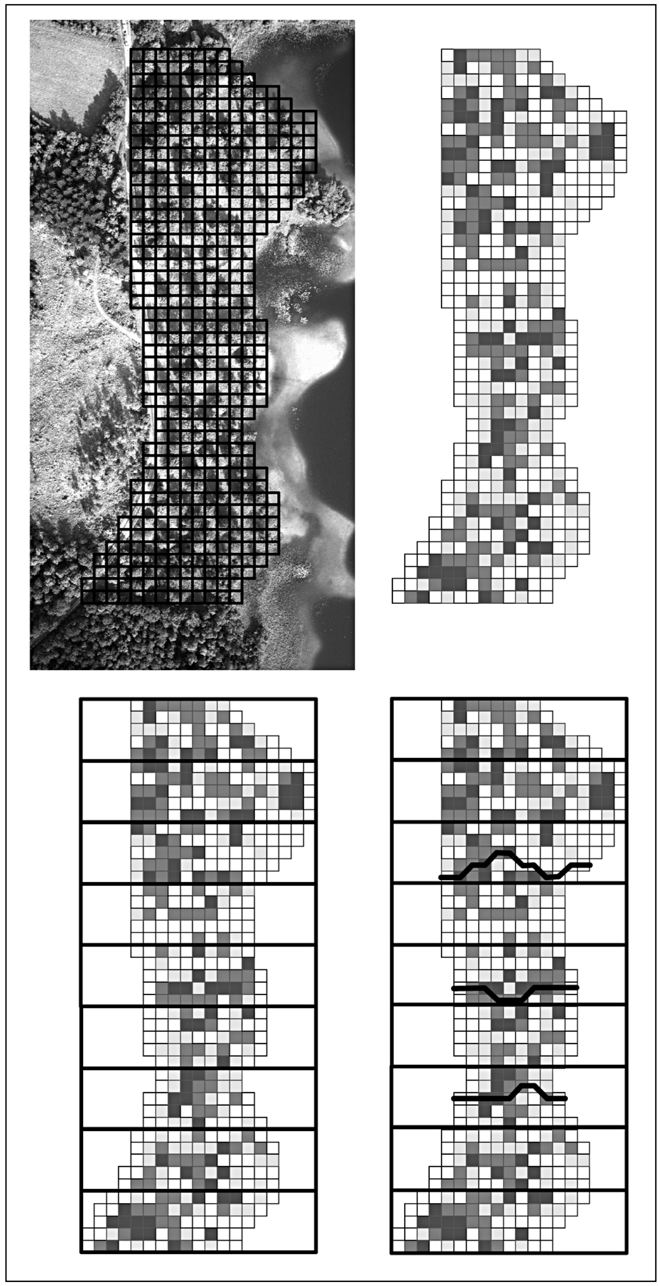
The total of the quantity of interest in each selected first stage strip can be estimated with a Horvitz-Thompson estimator as

$$\hat{Y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{\pi_{ij}}$$

Where:
n_i = the number of sampled grid cells in first stage strip *i*.
y_{ij} = the value of the variable of interest in grid cell *j*, first stage strip *i*.
π_{ij} = the inclusion probability of this grid cell.
The inclusion probabilities depend on how the second stage selection of the survey strip is carried out. The covariate information can be further used for estimation purposes by using a generalized ratio estimator

$$\hat{Y}_{iR} = X_i \frac{\hat{Y}_i}{\hat{X}_i}$$

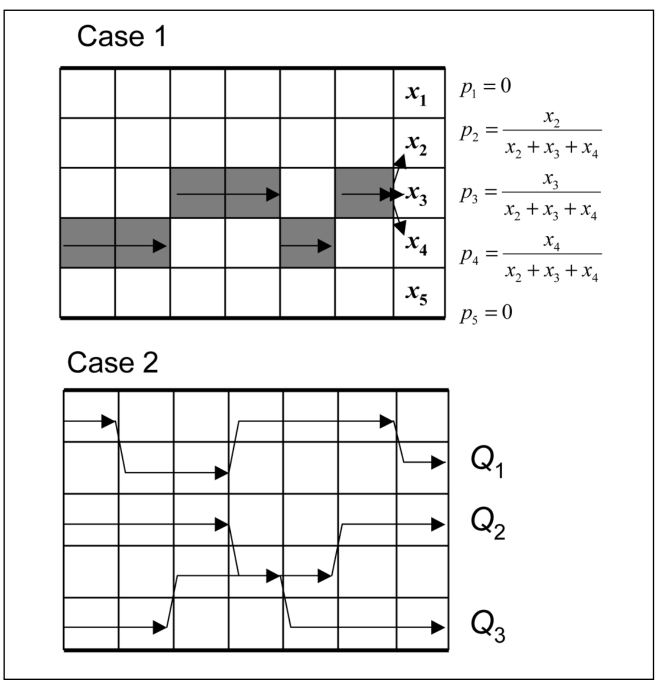
Figure 1.—*An overview of guided transect sampling. Prior to the survey, the area of interest is partitioned into a grid cell system (top left). For each grid cell a covariate value is assessed (top right). In the first stage, wide strips are randomly selected (bottom right), and in the second stage, survey strips are selected within the selected first stage trips by successively selecting one grid cell in each column based on the covariate data.*



where X_i is the total of covariate values in first stage strip i and $\hat{Y}_i$ and $\hat{X}_i$ are the Horvitz-Thompson estimator of the total of the variable of interest and of the covariate variable, respectively. How estimates are made for the whole area of study depends on how first stage strips are selected. For more details about estimation, variances, and derivation of inclusion probabilities see Ståhl *et al.* (2000).

Selecting the transect within the selected first stage strips can be done in different ways. This selection is referred to as the strategy for guidance since it can be seen as it is guided by the covariate information. Two straightforward alternatives for the strategy of guidance are shown in figure 2. With case 1, transitions are made to one of the neighboring cells in the next grid cell column. The selection of the grid cell to enter is made with probability proportional to the covariate values in the neighboring cells in the next column. The inclusion probability for a particular grid cell will depend both on the covariate value in the grid cell and its neighboring cells in the same column but also on the inclusion probabilities in the neighboring grid cells in the previous column so when calculating inclusion probabilities, recursive calculation will be needed. This strategy for guidance is rather short-sighted because at each step only the covariate information in the next column is used. Instead case 2 might be a better use of the covariate information. In this case, many transects are first simulated with transitions between grid cells made as in case 1; i.e., with probability proportional to the covariate values in the neighboring cells. For each simulated transect the covariate values in all grid cells passed are summed giving a sum Q for each transect. One whole transect is then selected with probability proportional to the sum of covariate, the Q-value. The inclusion probability for a particular grid cell will then be the sum of the Q-values for all transects that pass that particular grid cell divided by the sum of the Q-values for all transects simulated. With this strategy, transects that passes many interesting cells will have a higher probability of being selected. Case 1 in this article is equal to case 1 in Ståhl *et al.* (2000). Case 2 in this article is similar to case 3 in Ståhl *et al.* (2000), but with the difference that the probability of transition to a neighboring grid cell is proportional to the covariate value in the grid cell. Case 3 in Ståhl *et al.* (2000) is an equal probability of transition to all neighboring cells in the next column.

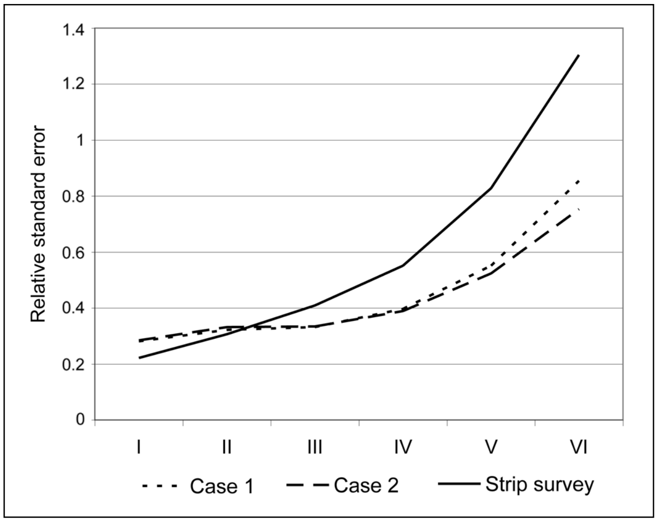
Figure 2.—*Two alternatives for the selection of survey strips in the second stage. With case 1, transitions between grid cells are made with probability proportional to the covariate values in the neighboring cells in the next grid-cells column. With case 2, one whole transect is selected with probability proportional to its Q-value, the sum of covariate values in all cells passed by the transect.*



Evaluation

The two described cases for the strategy of guidance were evaluated in simulated forest types. Although simulated they were created to resemble a real sampling situation. The forest types were created to resemble conifer forest with spots of deciduous trees and the object of interest considered to be some red-listed species connected to deciduous trees. GTS was compared with a standard strip survey with an equal area sampled with both methods. Comparisons were made in terms of standard errors of estimates. For details about the simulated forests and how comparisons were made, see Ståhl *et al.* 2000. The results of the comparison are shown in figure 3 for the case in which ratio estimators were applied, both in GTS and in the strip survey. The width of the first stage strips in GTS was five grid cells. When the object of interest is rather common, the methods perform equally well, but with a decreasing abundance

Figure 3.—*Comparison of the performance of two cases of guided transect sampling and a strip survey in six forest types. Ratio estimators were used both in guided transect sampling and in the strip survey. Forest types are ordered with a decreasing abundance to the right.*



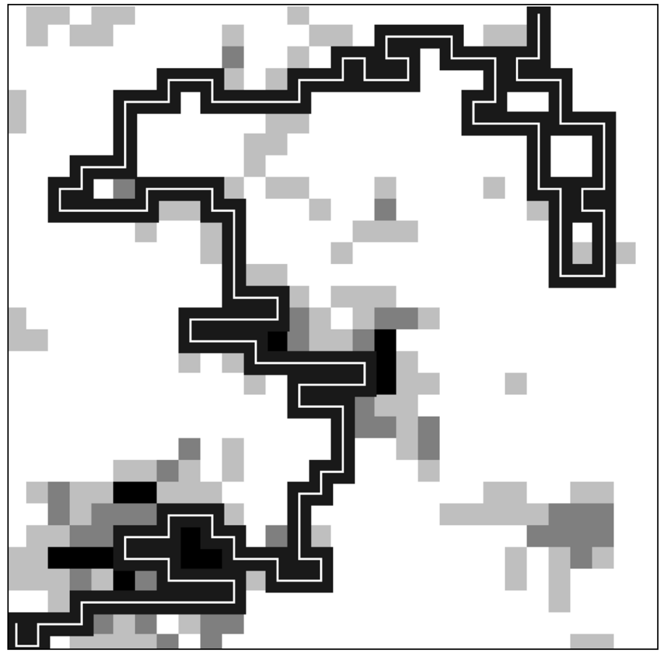
of the object of interest GTS is an improvement over a standard strip survey. The two cases considered were equivalent, so case 2 did not imply a better use of the available information as had been suggested.

Unrestricted GTS

With the strategies for the second stage guidance presented above, only one cell in each column is selected, which can be a restriction because interesting areas can only partly be covered. As an alternative, unrestricted GTS has newly been suggested (Ringvall *et al.* [in press]). With this variant of GTS, the first stage strips are skipped and transects are selected directly through the whole area (fig. 4).

With unrestricted GTS, a large amount of transects are first simulated according to certain restrictions, such as how many cells that must be passed and how transitions between cells are made. Besides giving a higher probability to cells with a higher covariate value, a higher probability can be given to continue in the same direction to avoid too many changes of direction. For each simulated transect, the values of covariate in cells passed

Figure 4.—*Example of a selected survey strip with unrestricted guided transect sampling, with which the first stage strips are skipped and transects are simulated directly through the whole area of interest.*



are summed, the *Q*-value, and finally one transect is selected with probability proportional to this *Q*-value (case 2). Hence, there is a higher probability of selecting a transect that passes a lot of interesting areas.

In the first evaluation of the performance of unrestricted GTS in simulated forest types, however, unrestricted GTS was only a rather limited improvement in comparison with the earlier suggested two-stage design (Ringvall and Ståhl 2007).

3P Subsampling

In the second step of the suggested approach, information that can only be obtained when at the sampling location (e.g., information about the quality of downed logs) is utilized. Whether a plot or strip survey, a species survey on all included substrates can be time consuming. Some sort of subsampling procedure might then be needed. In a purposive survey of specific species, the surveyors use their knowledge of species and substrate preferences to search for the species where they are likely to be found. In a probability sampling framework, 3P sampling has been suggested as a way to use this knowledge for a more efficient subsampling of substrates (Ringvall and Kruys 2005). In the original setup, the selection of sample trees is based on a quick measurement of the volume (e.g., Husch *et al.* 2003). In this context, the selection for sample trees is based on the surveyor's judgment of the probability for finding the species of interest on a given substrate on a scale 0–1. The probability that a substrate is selected for species survey is then proportional to this judge. The judge could also be for the substrates quality for a group of species surveyed on a scale 0–1. The result is that more time will be spent on the potentially more interesting substrates although all substrates have a probability of being selected.

Evaluation

The potential of using 3P sampling in this context was tested in a survey of needle lichens growing on the bark of old and coarse oak trees (Ringvall and Kruys 2005). To see the difference between different judges, three surveyors participated in the study. During the survey the probability of finding the species *Cyphelium inquinans* (Sm.) Trevis was judged by the surveyors. 3P subsampling was compared with a simple random subsampling in terms of standard errors of estimates. Comparisons were made both for estimates of the number of occurrences of the species *Cyphelium inquinans* (one occurrence was one tree with presence of the species) and for the number of occurrences of a group of nine species surveyed. The later case was a test for the ability of the judgment of the probability of finding a specific species to serve as a judgment for the substrate's suitability for a group of species. In one area, 3P sampling was a large improvement, while the improvement was more modest in the other (table 1). In the first area, a clear difference existed between the "good" and the "bad" trees. In the other area, it was more difficult to distinguish the good substrates. For details about the study and further results see Ringvall and Kruys (2005).

Conclusions

In this article, an approach for improving the efficiency of surveys of sparse objects by using the same type of information and knowledge used in purposive surveys has been presented. The evaluations of the methods in the approach have shown that they are promising. The evaluations made so far, however, are rather limited and mainly based on simulated forest types. Before the methods can be recommended and used in real applications, further evaluations are needed. Further, although the methods imply a considerable improvement there is still a risk that, with a reasonable sampling effort, the precision of estimate will be too poor.

Both practical and theoretical problems still need to be solved or considered. A theoretical problem is, for example, variance estimation in unrestricted GTS. Another problem is the availability of auxiliary information. For GTS to be useful, the auxiliary information much be easily available and at a low cost. *k*-Nearest Neighbor estimates providing for larger areas are a good example of such information. Also in these cases, however, the precision of estimates of sparse objects tends to be very low. For example, standard errors of estimates of volume of deciduous trees can sometimes be as high as more than 100 percent (e.g., Mäklelä and Pekkarinen 2001). The suggested methods are based on so called unequal probability sampling,

Table 1.—*The standard errors of a 3P subsampling with a ratio estimator given in percent of the standard errors of a simple random subsampling*[a]*, on an average for three surveyors, and, within parentheses, the range of their results.*

Variable	Area 1 (%)	Area 2 (%)
No. of occurrences of *Cyphelium inquinans*	46 (38–50)	84 (66–99)
No. of occurrences of nine surveyed lichens	58 (56–63)	88 (79–96)

[a] *(SE of 3P sampling/SE of simple random sampling)*100; values below 100 indicates that 3P sampling performed better than simple random sampling. Results are for a subsampling fraction of 0.5.*

and the risk of using such methods is well-known in sampling theory (e.g., Thompson 1992). If units have been assigned very low probability of inclusion while they actually have a high value of the variable of interest, the precision of estimates will be very poor. One way of preventing this is to not allow units to have too low probability of inclusion.

In contrast to these problems, the fast development in the Geographic Information System and field computer area and emerging techniques and improvements in the remote sensing area are promising for a possible implementation of these methods. The approach was described as it resembles the way a skilled surveyor walks through the forest, which might appeal to people who otherwise would prefer purposive methods.

Acknowledgments

The author thanks the co-authors of the paper on which this article is based: Professor Göran Ståhl, Associate Professor Tomas Lämås, and Dr. Nicholas Kruys. The work in these articles was financed by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning and Hildur and Sven Wingqvist's Foundation for Forest Research. Participation in the symposium was financed by Bröderna Edlunds Donationsfond and Karl-Erik Önnesjös stiftelse för vetenskaplig utveckling och forskning.

Literature Cited

Husch, B.; Beers, T.W.; Kershaw, J.A. 2003. Forest mensuration. 4th ed. New York: John Wiley. 456 p.

Mäkelä, H.; Pekkarinen, A. 2001. Estimation of timber volume at the sample plot level by means of image segmentation and Landsat TM imagery. Remote Sensing of Environment. 77: 66-75.

Ringvall, A.; Kruys, N. 2005. Sampling of sparse species with probability proportional to prediction. Environmental Monitoring and Assessment. 104: 131-146.

Ringvall, A.; Snäll, T.; Ekstrom, M.; Ståhl, G. [In press]. Unrestricted guided transect sampling for surveying sparse species. Canadian Journal of Forest Research.

Ståhl, G.; Lämås, T. 1998. Assessment of coarse woody debris: a comparison of probability sampling methods. In: Bachmann, P.; Köhl, M.; Päivinen, R., eds. Proceedings, conference on assessment of biodiversity for improved forest planning. EFI Proceeding No. 18. Kluwer, Dordrecht: 241-248.

Ståhl, G.; Ringvall, A.; Lämås, T. 2000. Guided transect sampling for assessing sparse populations. Forest Science. 46: 108-115.

Thompson, S.K. 1992. Sampling. New York: John Wiley. 343 p.