Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2004: Innovative Methods for Surveying Difficult-to-reach Populations

2004

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2004
Innovative Methods for Surveying Difficult-to-reach Populations

ADAPTIVE WEB SAMPLING FOR DIFFICULT-TO-REACH POPULATIONS

Steven K. Thompson[1]

This talk describes adaptive sampling designs in which, at any point in the sampling, the next unit or set of units is with high probability selected from a distribution that depends on the values of variables of interest in an active set of units already selected. With lower probability, the next selection is made from a distribution not dependent on those values. The active set may consist of the entire current sample, or only the most recently selected unit, or a wide range of other possibilities. Design-unbiased estimation with such designs is based on a combination of initial and conditional selection probabilities, and these preliminary estimators are improved using the Rao-Blackwell method. Markov chain resampling estimators are used for larger sample sizes. Network- and spatially-based application of the designs to a hidden human population at risk for HIV/AIDS and a wintering waterfowl survey are evaluated. The new designs can give efficiency gains over comparable conventional designs is some situations and, in comparison with other adaptive and link-tracing sampling methods, the present class of strategies has advantages in flexibility regarding adaptive criteria and breadth and depth of sample coverage, ease of implementation, control of sample sizes, and the availability of robust if computationally intense design-based estimators.

1. INTRODUCTION

This article describes classes of sampling designs in which, at any point during the sampling, the selection of the next unit or wave of units to include in the sample is with high probability selected with a distribution that depends on values of variables of interest in an ``active set'' of units already selected. With lower probability, the next unit is selected from a distribution not depending on values of variables of interest. The active set may consist of the entire
current sample, the most recently selected unit, the several most recently selected units, a wave of previously selected units that changes only at fixed intervals, a set of units within a certain spatial distance from the most recently selected unit, and many other possibilities.

An example of a design of this type in a spatial setting is illustrated in Figure 1. A population of rare, clustered point objects, representing for example animals or plants in a spatial study region, is shown at the upper left. The study region is divided into 100 units, a sample of which is to be selected for the purpose of estimating the total or mean number of point objects or other characteristics of the population. A population graph associated with this spatial population is constructed by representing each unit as a node (a circle in the depiction at the upper right). If a unit contains an ``interesting'' value, in this case having one or more of the point objects, an arc or arrow is drawn from it to each of its neighboring units, of which there are up to four. The graph is a directed one because there is no arc back from a unit at the edge of an aggregation, having no point-objects.

A sample of 20 nodes, shown at the lower left of Figure 1 with order of selection indicated, is selected from the graph using the following procedure. First, an initial sample of 12 units is selected by random sampling without replacement. It turns out that none of the initial sample units contain any of the objects of interest, but, since the total sample size of 20 has been decided on in advance, random sampling without replacement continues for another step. The 13th unit selected does contain some of the objects, and so has four links out from it. With 90% probability, one of these links will be selected at random and followed, while with 10% probability another unit will be selected at random from the units in the population not already selected. As it happens, it is determined to select a link, and the one chosen leads us to the 14th selected unit, which also contains objects of interest. Now, with 90%

[1] Steven K. Thompson, Department of Statistics, Pennsylvania State University, University Park, PA 16801 USA;
Address for correspondence, 2004-5 academic year: 132 Mesa Vista Street, Santa Fe, NM 87501 USA
Address from 1 September 2005: Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6 Canada

probability we can select a link from any of the six links, originating at the 13th and 14th units, that lead out from the current sample. This selection could be made at random but, because the 14th unit contains a lot more of the objects than does the 13th unit, we will use a design variation instead that selects the next link with probability proportional to the value of the originating node. This gives higher probability to following a link from node 14, and indeed that is what leads us to node 15. Continuing with this design, the remaining five units are selected. The active set in this example consists of the entire current sample, and selections depend adaptively on both the node values and the link-indicator values associated with this set.

Figure 2 shows a larger spatial population, with more and larger clusters, and illustrates some of the variety possible affecting depth and breadth of sample coverage. The sample shown in the bottom left of the figure was selected by independently selecting 16 starting units at random, with an active set sample of size 5 resulting from each of these starts. From each start, with probability 0.9 a link from that unit is selected at random to select the next unit for the sample, while with probability 0.1 a unit from outside the current sample is selected completely at random. With this design, the depth to which one follows links is limited to a maximum of five steps, and the total sample size is fixed at 80. For the sample shown in the bottom left of the figure, a single starting unit was selected at random, and an adaptive sample of 80 units was selected in sequence from there. At each point in the sampling, the entire current sample serves as the active set from which with probability 0.9 a link out is selected at random and followed.

Figure 3 shows a pure graph population, representing for example a hidden human population and its social network relations. Colors represent node characteristics, with the dark color indicating presence of the characteristic while a light color indicates its absence. Three types of samples are shown, each started from a single randomly selected node. The sample in the upper right was selected as a random walk (Lovasz 1993) modified to be without replacement and to allow for the possibility of a random jump depending on node value. At any point in the sampling, if the current node is a positive one (dark) then with probability 0.99 one of the links leading from it to an unsampled unit is selected at random and followed to add the next unit, while with probability 0.01, or if there are no more links to follow from the current unit, the next unit is selected at random from the units not already selected. If the present node is a zero-valued one (light), the probabilities are 0.9 of following a link and 0.1 of taking a random jump. The active set for the walk consists of only the most recently selected unit. In the lower two designs the whole current sample serves as the active set, and with probability 0.9 a link out from that set is selected. At the lower left, links out are selected at random. At the lower right, links are weighted so that only links from nodes with the characteristic of interest are selected.

2. ADAPTIVE WEB DESIGNS BASED ON ACTIVE SETS

An adaptive web design takes place in several steps. First, an initial sample s0 is selected by some design p0. At the kth step after the initial sample, selection of the next part of the sample sk depends on values associated with a current active set ak t,hat is, a subset or subsequence of the sample so far selected, together with any associated variables of interest. Thus, for $k \geq 1$, the selection distribution at step k is pk(sk | yak, wak), where ak is a subset or subsequence of the current sample.

One way to implement such a design is to select the next unit from a mixture distribution, so that with probability d the next unit is selected using a distribution based on the unit values or graph structure of the active set and with probability 1-d it is selected using a distribution based on the frame or spatial structure of the population. For example, with probability d one of the links from the active set is selected at random and the unit to which it connects is added to the sample, while with probability 1-d a new unit is selected completely at random from the population, or from those units not already in the sample. The probability d may itself depend on the values in the active set.

So far as the adaptive nature of selections is concerned, selections can be made unit-by-unit or in waves. Selection can be said to occur in waves if the active set remains constant for several selections in a row, so that a whole group of selections are based on a given active set. Snowball-type designs, for example, typically occur in waves, with a whole set of links selected from the previous wave of units or from all the units selected so far. Random walk related designs, on the other hand, involve unit-by-unit selection, with the active set consisting solely of the most recently selected unit.

Estimation with these designs can be done using either design-based or model-based methods. Efficient design-unbiased estimators are available by using a preliminary estimator, based on the initial sample or using conditional

probabilities along sample paths, and improving that estimator using the Rao-Blackwell method. Since the computations for the Rao-Blackwell estimators become prohibitive unless sample sizes are quite small, a Markov Chain Monte Carlo resampling method can be used which produces samples from the conditional distribution given the minimal sufficient statistic. For model-based estimation, Bayes methods provide the most practical approach. Except with the most simple stochastic graph or spatial models and for all but the most simple designs, the Bayes methods also require Markov chain Monte Carlo methods, to produce samples from the posterior distribution given the data. In addition to Gibbs or other steps to estimate parameters of the model, a data augmentation step is usually involved to produce a full graph or spatial population surrounding the sample.

3. DISCUSSION

The sampling strategies described in this article have a number of advantages relative to other adaptive and link-tracing strategies described in the literature, as well as providing efficiency gains in some situations over conventional sampling with the same sample size. In comparison with ordinary adaptive cluster sampling (Thompson 1990, Thompson and Seber 1996) and with some types of network or multiplicity sampling (Birnbaum and Sirken 1965), one advantage is that no connected component is required to to be sampled completely.

Relative to some of the standard snowball designs in graphs (Frank 1977a,b, 1978a,b, 1979, Frank and Snijders 1994), an advantage of the present designs is that sample size can be fixed in advance. Also in contrast to some of the standard methods for network and snowball designs, the strategies discussed in this paper are applicable in directed graph as well as undirected graph situations. A potential advantage over ordinary adaptive cluster sampling is that, instead of requiring a single fixed criterion, such as a y value in excess of a specified constant, for adaptively adding new units, the probability of adaptive additions in any region can continuously depend on unit values or link weights. On the other hand, adaptive cluster sampling offers some efficient design-unbiased estimators that can readily be computed by hand, whereas the estimators described in this paper tend to be computationally demanding.

In relation to optimal model-based sampling strategies (Zacks 1969, Chao and Thompson 1999), which can roughly be characterized as adaptively placing new units in proximity to high-valued or ``interesting'' observations while at the same time striving to spread them out to cover the study region, the proposed designs, while not optimal under any one model, approximate some of these characteristics while being much simpler to implement and avoiding the dependence on model-based assumptions. In relation to likelihood and Bayes inference methods for sampling in graphs or networks (Thompson and Frank 2000, Chow and Thompson 2003), the present designs should be seen as complimentary, providing likelihood-ignorable selection procedures to make the model-based inferences straightforward and valid, while at the same time offering the alternative of stand-alone design-based inference methods that are robust to departures from model-based assumptions.

ACKNOWLEDGMENTS

Support for this work has been provided by the National Center for Health Statistics and the National Science Foundation (DMS-0406229). The Visiting Faculty Program of the Statistical Sciences Group at Los Alamos National Laboratory provided time for carrying out some of this work.

REFERENCES

Birnbaum, Z.W., and Sirken, M.G. (1965), “Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates”, *Vital and Health Statistics*, Ser. 2, No. 11, Washington: Government Printing Office.

Chao, T-C., and Thompson, S.K. (2001), “Optimal adaptive selection of sampling sites”, *Environmetrics*, 12, pp. 517-538.

Chow, M. and Thompson, S.K. (2003), “Estimation with link-tracing sampling designs – a Bayesian approach”, *Survey Methodology*, 29, no. 2, pp. 197-205.

Frank, O. (1977a), “Survey sampling in graphs”, *Journal of Statistical Planning and Inference*, 1, pp. 235-264.

Frank, O. (1977b), “Estimation of graph totals”, *Scandinavian Journal of Statistics*, 4, pp. 81-89.

Frank, O. (1978a), “Estimating the number of connected components in a graph by using a sampled subgraph”, *Scandinavian Journal of Statistics*, 5, pp. 177-188.

Frank, O. (1978b), “Sampling and estimation in large social networks”, *Social Networks*, 1, pp. 91-101.

Frank, O. (1979), “Estimation of population totals by use of snowball samples”, *In Perspectives on Social Network Research*, edited by P.W. Holland and S. Leinhardt, New York: Academic Press, pp. 319-347.

Frank, O., and Snijders, T. (1994), “Estimating the size of hidden populations using snowball sampling”, *Journal of Official Statistics*, 10, pp. 53-67.

Lovasz, L. (1993), “Random walks on graphs: A survey”, In Miklos, D., Sos, D., and Szoni, T., eds., Combinatorics, Paul Erdos is Eighty, Vol. 2, pp. 1-46. Janos Bolyai Mathematical Society, Keszthely, Hungary.

Thompson, S.K. (1990), “Adaptive cluster sampling”, *Journal of the American Statistical Association*, 85, pp. 1050-1059.

Thompson, S., and Frank, O. (2000), “Model-based estimation with link-tracing sampling designs”, *Survey Methodology*, 26, no. 1, pp. 87-98.

Thompson, S.K., and Seber, G.A.F. (1996), *Adaptive Sampling*, New York: Wiley (Wiley Series in Probability and Mathematical Statistics).

Zacks, S. (1969), “Bayes sequential designs of fixed size samples from finite populations”, Journal of the American Statistical Association, 64, pp. 1342-1349.

Figure 1.

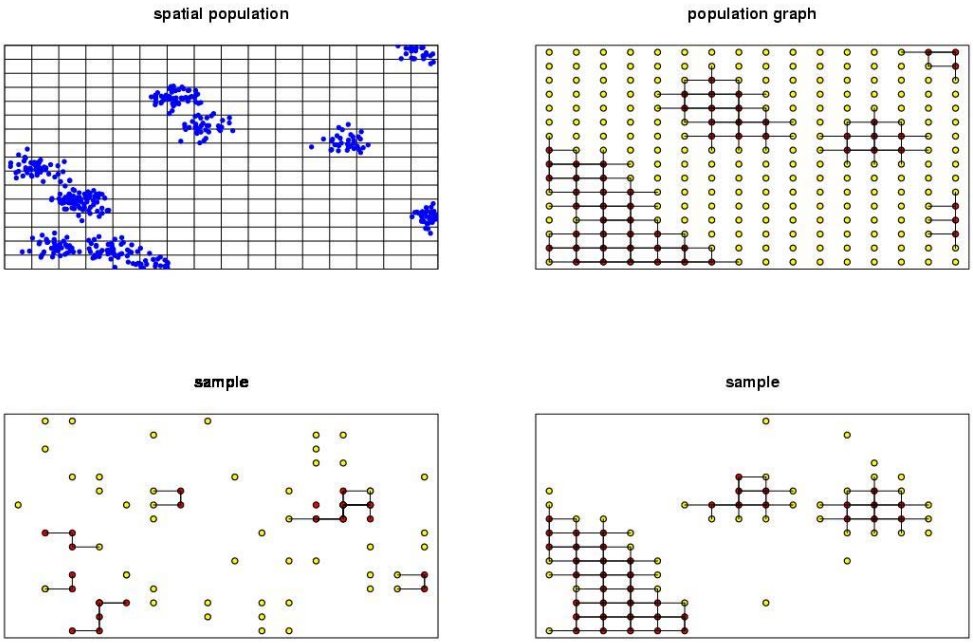

Figure 2.

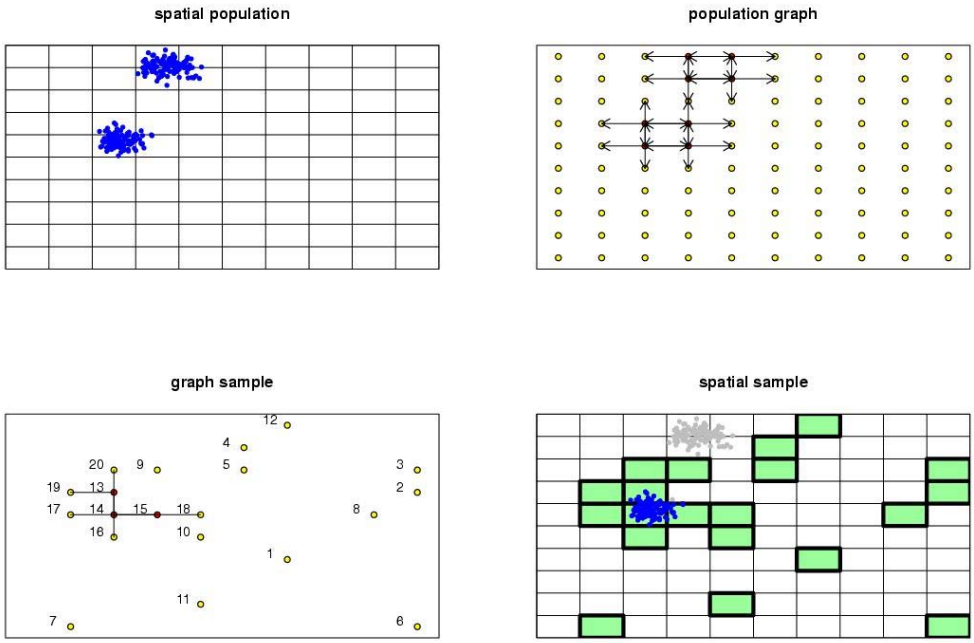

Figure 3.