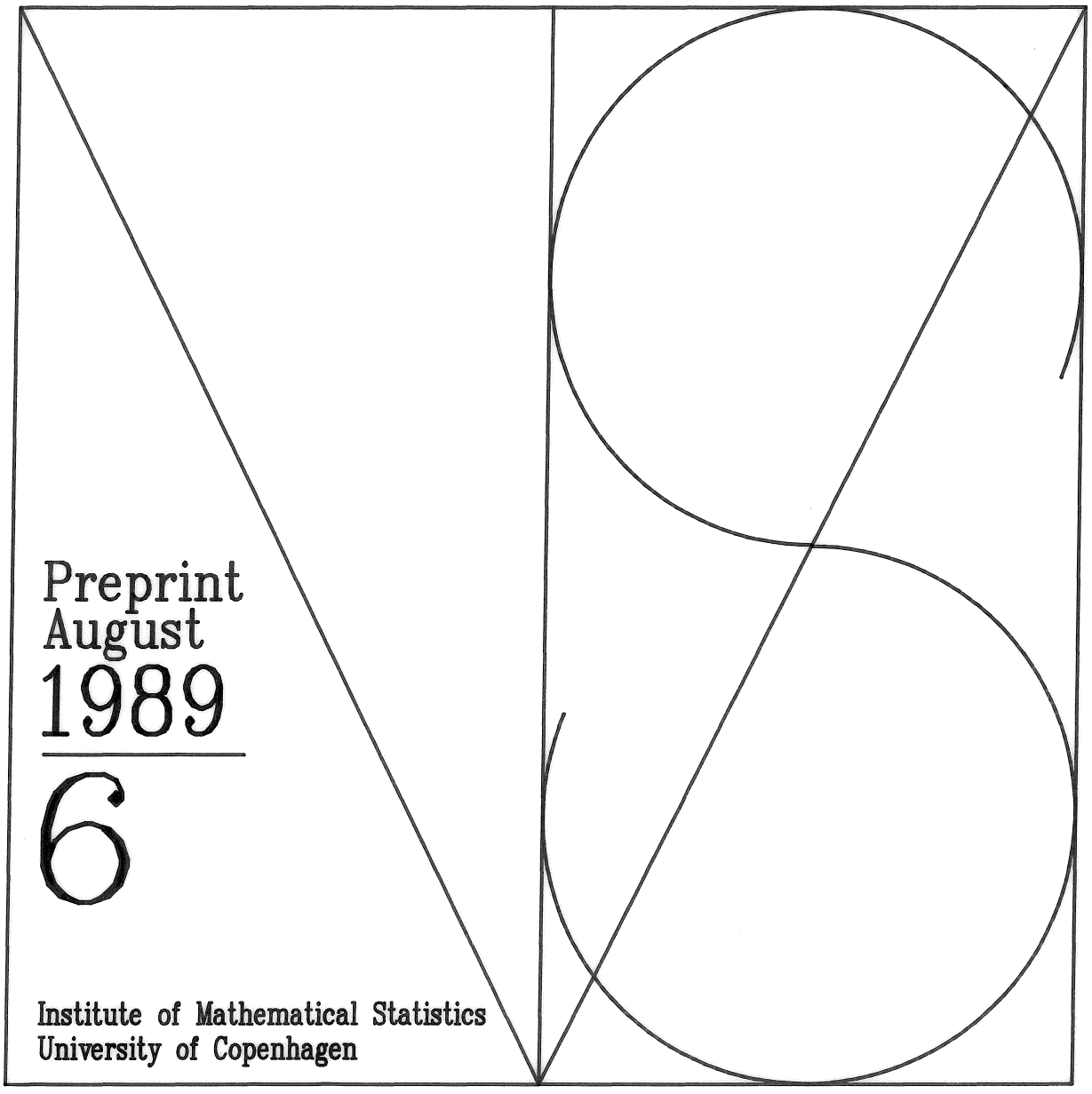
Steven K. Thompson

Adaptive Cluster Sampling: Designs with Primary and Secondary Units



Preprint
August
1989

6

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

Steven K. Thompson*

ADAPTIVE CLUSTER SAMPLING:

DESIGNS WITH PRIMARY AND SECONDARY UNITS

Preprint 1989 No. 6

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

August 1989

*University of Alaska Fairbanks, Alaska.

Adaptive Cluster Sampling: Designs with Primary and Secondary Units

Steven K. Thompson

Department of Mathematical Sciences,
University of Alaska Fairbanks,
Fairbanks, Alaska 99775, U.S.A.

SUMMARY

In adaptive cluster sampling designs, an initial probability sample is selected and, whenever the observed value of the variable of interest satisfies a given condition, units in the neighborhood of that observation are added to the sample. In this paper, the initial design is selected in terms of primary units, while subsequent sampling is in terms of secondary units. Such initial designs include systematic sampling, strip sampling, and other forms of classical cluster sampling. But because of the subsequent addition to the sample of secondary units in the neighborhood of any (secondary) unit that satisfies the condition of interest, the final "clusters" of units obtained through the procedure may be quite different in shape from the initial primary units. The methods described in this paper apply to such sampling situations as whale surveys in which the research vessel temporarily leaves the selected transect to close in on sighted whales, surveys of rare bird species in which initial observations are made at systematically selected sites and additional observations are made in the vicinity of any site at which sufficiently high abundance is observed, and aerial walrus surveys in which the aircraft searches to either side of the preselected transect line whenever a congregation of animals is encountered. Because conventional estimators of the population mean and total are biased with such a procedure, estimators which are unbiased under the adaptive designs are presented in this paper. Variance formulae and unbiased estimators of variance are also given. The designs are illustrated using a point pattern representing locations of individuals or objects in a spatially aggregated population; for such a population, the adaptive designs can be substantially more efficient than their conventional counterparts.

1. Introduction

Adaptive cluster sampling designs are those in which, following an initial probability selection of units, additional units are added to the sample in the neighborhood of any selected unit for which the observed value of the variable of interest ("y-value") satisfies a specified condition. Such designs are in marked contrast to conventional sampling designs, in which the probabilities for selecting samples do not depend on any population y-values.

Key words: Adaptive sampling; Animal abundance estimation; Cluster sampling; Ecological sampling; Systematic sampling.
July 31, 1989

The purpose of the adaptive strategies is to take advantage of population characteristics to obtain more precise estimates of population parameters with a given amount of effort. For example, many populations of animals and plants have aggregation tendencies due to such factors as schooling, flocking, dispersal patterns, and environmental patchiness. Often, the location and shape of the aggregations can not be predicted before a survey so that traditional means of increasing precision such as stratification are not sufficient. For such populations, adaptive sampling strategies may provide a way to dramatically increase the effectiveness of sampling effort.

Adaptive cluster sampling designs in which the initial sample is selected by simple random sampling, with or without replacement, are considered in Thompson (1989). Adaptive designs in which the sample size of a simple random sample within primary units or strata depends on initial observations within those primary units or strata are discussed in Francis (1984) and Kremers (1987). Adaptive strategies in which the sample size depends instead on observed values in neighboring primary units or strata are presented in Thompson and Ramsey (1983) and Thompson (1988). Seber (1986) and Cormack (1988) discuss the importance of adaptive sampling methods for ecological sampling.

In this paper, adaptive cluster sampling designs are considered in which the initial sample is selected in terms of primary units and subsequent additions to the sample are in terms of secondary units. For example, in an aerial survey of walruses or polar bears or in a ship survey of whales sighted by their spouts, the strip observed in each selected transect forms a primary unit. If, whenever animals are sighted, the area to the side of the transect is searched—with still further searching if additional animals are sighted while on this search—the searching pattern defines neighborhoods of secondary units added to the sample. In surveys of bird and fish species, the selection of sites at which to make observations is often done systematically and a single systematic selection forms a primary unit. If additional observations are made in the neighborhood of any site at which abundance is observed, the subsequent observations would not in general follow the initial systematic pattern. With such survey situations, one can think of the study region as partitioned into secondary units representing all possible sites at which observations may be made, while the primary units from which the initial sample is selected consist of clusters—such as long, thin strips or systematic arrangements—of the secondary units.

2. Designs

For the adaptive cluster sampling designs considered in this paper, the

population is composed of N primary units. Each primary unit contains M secondary units (which may be refered to simply as *units*). The MN units of the population are denoted u_{ij}, for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. Associated with the j-th secondary unit of the ith primary unit is a variable of interest y_{ij}. The object of inference is estimation of the population mean $\mu = (MN)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}$ or, equivalently, of the population total $\tau = MN\mu$.

For every (secondary) unit of the population, a collection of units called the *neighborhood* of that unit is defined. The neighborhood of unit u_{ij} includes unit u_{ij}, and, if unit u_{ij} belongs to the neighborhood of unit $u_{i'j'}$, then unit $u_{i'j'}$ belongs to the neighborhood of unit u_{ij}. In applications, the neighborhood of a unit will typically be defined as a contiguous set of surrounding units or a systematic pattern of surrounding units.

The unit u_{ij} is said to satisfy the *condition of interest* if the associated value y_{ij} is in a specified set C. For problems in the estimation of animal abundance, the condition may commonly be defined so that a unit satisfies the condition if its y-value equals or exceeds some constant c.

In the adaptive cluster sampling designs of this paper, an initial sample of n_1 primary units is selected by simple random sampling without replacement. Whenever the observed value of a (secondary) unit in the sample satisfies the condition of interest, all units in its neighborhood are added to the sample. If in turn any of these subsequently added units satisfies the condition, the units of its neighborhood are also added to the sample, so that finally the sample contains every unit in the neighborhood of any sample unit satisfying the condition.

A population with a given set of y-values can be uniquely partitioned into K sets called *networks* so that whenever a unit u_{ij} satisfying the condition is in the neighborhood of unit $u_{i'j'}$ also satisfying the condition, then units u_{ij} and $u_{i'j'}$ belong to the same network. Thus, if an initially selected primary unit intersects a given network, every unit in that network will be included in the sample. A unit that does not satisfy the condition belongs to a network consisting just of itself.

A unit u_{ij} which does not satisfy the condition will be included in the sample either if the primary unit which includes it is initially selected or if any primary unit of the initial selection intersects the network of one or more units satisfying the condition in the neighborhood of unit u_{ij}.

Note that while neighborhoods are defined by such relationships as physical proximity and do not depend on the y-values of the population, networks do depend on the population y-values, corresponding roughly to the natural aggregations of animals, plants, or other individuals in the

population.

If each initial primary unit consists of a set of units evenly spaced in some arrangement throughout the population, the initial sample will be termed a *systematic* initial sample. The initial primary units will be called *strips* if each initial primary unit consists of a row of units arranged in a straight line.

The draw-by-draw selection probability p_{ij} for unit u_{ij} is the probability in any initial draw of selecting any one of the primary units that intersects the network containing unit u_{ij} or, if unit u_{ij} does not satisfy the condition, selecting a primary unit that intersects the network of any unit satisfying the condition in the neighborhood of unit u_{ij}. That is,

$$p_{ij} = \frac{m_{ij} + a_{ij}}{N},$$

where m_{ij} is the number of primary units that intersect the network containing unit u_{ij}, and a_{ij} is the number of primary units which do not intersect the network of unit u_{ij} but intersect the network of one or more units satisfying the condition in the neighborhood of unit u_{ij}. For a unit satisfying the condition, $a_{ij} = 0$, while for a unit not satisfying the condition, $m_{ij} = 1$.

The probability α_{ij} that unit u_{ij} is included in the sample is the probability that one or more primary units of the initial sample intersects the network that includes unit u_{ij} or intersects a network of which unit u_{ij} is an edge unit. That is,

$$\alpha_{ij} = 1 - \binom{N - m_{ij} - a_{ij}}{n_1} \Big/ \binom{N}{n_1}.$$

The expected sample size, that is, the expected number of distinct secondary units in the final sample, is the sum of the inclusion probabilities (Godambe, 1955; and see Cassel, Särndal, and Wretman, 1977, p.11), so that the expected sample size ν expressed in terms of the equivalent number of primary units in the final sample is

$$E(\nu) = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_{ij}.$$

3. Estimators

With the adaptive cluster sampling designs described in this paper, standard estimators of the population mean and total are biased. With

spatially aggregated populations, for example, if additional units are added to the sample whenever high abundance is observed, the final sample tends to contain units with higher than average abundance, and the sample mean will overestimate the population mean. If, on the other hand, the estimator is formed by averaging first all y-values associated with the selection of a primary unit—that is, the units of the primary unit together with all units adaptively added to the sample as a result of initial selection of that primary unit—the mean of these averages may tend to underestimate the population mean, due to the fact that, whenever units with higher than average y-values are selected, additional sampling commences until low values are obtained, while when units with low values are selected, no such compensatory procedure commences.

In this section, therefore, estimators are given which are unbiased with the adaptive cluster sampling designs of this paper. Since these estimators are in fact design-unbiased, the unbiasedness does not depend on any assumptions about the population itself.

3.1 The Initial Sample Mean

One way to obtain an unbiased estimator is to ignore all units adaptively added to the sample and use the sample mean t_1 of the initial sample:

$$t_1 = \frac{1}{Mn_1} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}.$$

This estimator does not make use of the observations adaptively added to the sample. It is of interest in this paper because it offers the basis for nonadaptive alternatives with which the adaptive strategies may be compared.

From classical results on simple random sampling without replacement, t_1 is unbiased and has variance

$$var(t_1) = \frac{N - n_1}{M^2 N n_1} \sigma_1^2,$$

where

$$\sigma_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(\sum_{j=1}^{M} y_{ij} - M\mu \right)^2.$$

An unbiased estimator of variance is

$$\widehat{var}(t_1) = \frac{N - n_1}{M^2 N n_1} s_1^2,$$

where

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\sum_{j=1}^{M} y_{ij} - M t_1 \right)^2 .$$

An unbiased estimate of variance is, of course, not available for systematic samples with only one starting point (though see Wolters, 1984, for methods useful in practice).

3.2 An Estimator Based on Partial Selection Probabilities

It is also possible to obtain unbiased estimators which do make use of observations in addition to those initially selected. Estimators such as the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) and the "multiplicity" estimator (cf. Sirken, 1970, 1972) achieve unbiasedness by dividing each observation by its selection probability and multiplying by the number of times the unit was selected. With the adaptive cluster sampling designs of this paper, however, not all of the selection probabilities as given in Section 2 can be determined from the sample data. When a unit not satisfying the condition appears in the sample, one may not know whether its selection probability is influenced by the presence of units in its neighborhood which do satisfy the condition. The unbiased estimator of this section therefore depends only on aspects of the selection probilities which are known.

For the estimator of this section and the next, it will be convenient to relabel the variables in terms of the networks of the population rather than in terms of the individual units. Let the K networks of the population be labelled $1, \ldots, K$, and let y_i denote the total of the y-values in the i-th network. Define the indicator variable I_{ki} to be 1 if the k-th primary unit intersects the i-th network and 0 otherwise. Let x_i be the number of primary units which intersect the i-th network, that is, $x_i = \sum_{k=1}^{N} I_{ki}$. Consider the estimator

$$t_2 = \frac{1}{M n_1} \sum_{k=1}^{n_1} \sum_{i=1}^{K} \frac{y_i I_{ki}}{x_i}.$$

Note that variables only for those networks that are intersected by selected primary units enter into the above expression.

A network y-value is utilized in the estimator as many times as there are primary units in the initial sample that intersect it. Some observations in the data—associated with units not satisfying the condition and not included in the initial sample—are not utilized at all in the estimator.

The actual selection probability for network i is related to x_i but may also depend, in a manner not known from the data at hand, on other networks.

It is shown in Appendix A that t_2 is an unbiased estimator of the population mean with variance

$$var(t_2) = \frac{N - n_1}{M^2 N n_1} \sigma_2^2,$$

where

$$\sigma_2^2 = \frac{1}{N-1} \sum_{k=1}^{N} \left(\sum_{i=1}^{K} \frac{y_i I_{ki}}{x_i} - M\mu \right)^2 .$$

An unbiased estimator of the variance of y_2 is given by

$$\widehat{var}(t_2) = \frac{N - n_1}{M^2 N n_1} s_2^2,$$

where

$$s_2^2 = \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} \left(\sum_{i=1}^{K} \frac{y_i I_{ki}}{x_i} - M t_2 \right)^2 .$$

3.3 An Estimator Based on Partial Inclusion Probabilities

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) estimator achieves unbiasedness by dividing the y-value for each unit in the sample by the probability that the unit is included in the sample. With the adaptive cluster sampling designs of this paper, not all of these inclusion probabilities, as given in Section 2, are known from the sample data. In particular, the constants a_{ij} as defined in Section 2 may not be known because the sample may not reveal units that satisfy the condition in the neighborhoods of sample units that do not satisfy the condition. In this section, an unbiased estimator is given based on the partial knowledge of inclusion probabilities obtainable from the data.

Let π_i denote the probability that one or more of the primary units which intersects network i is included in the initial sample. With the adaptive cluster sampling designs of this paper, this probability is given by

$$\pi_i = 1 - \binom{N - x_i}{n_1} \Big/ \binom{N}{n_1} .$$

Let π_{ij} denote the probability that one or more of the primary units which intersect both networks i and j is included in the initial sample. With the designs of this paper,

$$\pi_{ij} = 1 - \left[\binom{N - x_i}{n_1} + \binom{N - x_j}{n_1} - \binom{N - x_i - x_j + x_{ij}}{n_1} \right] \Big/ \binom{N}{n_1},$$

where x_{ij} denotes the number of primary units which intersect both networks i and j. It is emphasized that the π's are not the actual network inclusion probabilities, but are computable from the sample data.

Define the indicator variable z_i to be 1 if one or more of the primary units which intersect network i are included in the initial sample and 0 otherwise. Consider the estimator t_3 given by

$$t_3 = \frac{1}{MN} \sum_{i=1}^{K} \frac{y_i z_i}{\pi_i},$$

so that the summation is over the distinct networks in the sample which intersect one or more primary units of the initial sample. The weight an observation receives in the estimator does not depend, as it does with t_2, on the number of intersecting primary units selected, as long as at least one of them is included in the initial sample. Also, some observations in the data may receive zero weight.

The estimator t_3 is unbiased for the population mean (see Appendix B) and has variance

$$var(t_3) = \frac{1}{M^2 N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right),$$

with the convention that $\pi_{ii} = \pi_i$.

An unbiased estimator of this variance is given by

$$\widehat{var}(t_3) = \frac{1}{M^2 N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{y_i y_j z_i z_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right),$$

provided that none of the joint probabilities π_{ij} are zero.

3.4 Improvement of the Estimators with the Rao-Blackwell Method

Since each of the above unbiased estimators is not necessarily a function of the minimal sufficient statistic, it is in principle possible that each

could be improved by the Rao-Blackwell method of taking its conditional expectation given the sufficient statistic. The minimal sufficient statistic for the finite population sampling situation is the unordered set of distinct, labelled observations (Basu, 1969), that is, the y-values together with the identities of the units with which they are associated, without regard to order of selection or number of times selected. Given a set of data, one could therefore compute the average of the estimator obtained over all possible reselections from the data of n_1 distinct primary units which would give rise to exactly the same set of data.

While the method can offer practical improvements in estimators for such designs as adaptive cluster sampling with initial simple random samples (Thompson, 1989) and certain designs in which sample sizes within strata are based on initial observations (Kremers, 1987), the method does not appear to be of great importance for most situations in which the designs considered in this paper would be applied. Because of the very specific patterns of primary and secondary units typical of the samples obtained, one would not often expect to find other initial primary unit selections that would give rise to exactly the same value of the minimal sufficient statistic.

4. Example

Sample calculations for the adaptive cluster sampling strategies of this paper will be illustrated with two types of designs. In one, the primary units consist of long, thin strips. The other has an initial systematic sampling design, with starting points chosen at random in a four-by-four square and the positions repeated throughout the study area. The neighborhood of a unit is defined to consist of itself together with all adjacent (sharing a full edge) units. Thus, for a unit not on the boundary of the study region, the neighborhood consists of five units in a cross shape.

The square study region of 400 units is depicted in Figures 1 and 2. The locations of individuals or objects in the study region were produced with a realization of a Poisson cluster process (cf. Diggle, 1985), with three parent locations selected at random and Poisson (mean=100) numbers of offspring distributed about each parent with a bivariate Gaussian distribution (with standard deviation 0.03 in the unit square). The object of sampling is to estimate the number of objects in the study region (the correct answer: 326) or, equivalently, the mean number per unit (0.815). The 400 population y-values for the example are listed in Appendix C.

A unit is considered to satisfy the condition if it contains at least one individual of the population, so that any time a selected unit contains one

or more individuals, the remaining units in its neighborhood are added to the sample. Figure 1 shows the sample obtained as the result of the initial selection of five of the strips. Figure 2 shows the sample obtained with an initial selection of two systematic starting points. Sample calculations will be carried out for these illustrated samples.

In the initial strip plot sample, the first (leftmost) of the primary units in the sample intersects two collections of units which satisfy the condition, leading to the additional clusters of units added to the sample. The network of units satisfying the condition within the uppermost of these clusters has total y-value 106; it can be determined from the sample that this network intersects four of the primary units. The lower cluster has total y-value 105 and also intersects four primary units. All other observations in the data are zero. The term $\sum_{k=1}^{n_1} y_i I_{ki}/m_i$ associated with the first sample primary unit is $(106/4) + (105/4) = 52.75$. For the second sample primary unit the term is 105/4=26.25, and, for the other three primary units in the sample, the term is zero. The estimate t_2 is $(1/20)(1/5)(52.75 + 26.25 + 0 + 0 + 0) = .79$. The variance estimate is $\widehat{var}(t_2) = (20 - 5)/[(20^2)(20)(5)](555.85625) = .2084$, in which 555.85625 is the sample variance of the values 52.75, 26.25, 0, 0, and 0.

The intersection probability for each of the two non-zero networks in the sample is $1 - \binom{20-4}{5}/\binom{20}{5} = .7183$, and the estimator t_3 is $(1/20)(1/20)[(106/.7183) + (105/.7183)] = .7344$. Since two primary units intersect both networks, the joint intersection probability for the two sample networks is $1 - [\binom{20-4}{5} + \binom{20-4}{5} - \binom{20-4-4+2}{5}]/\binom{20}{5} = .5657$. The sample estimate of variance is $\widehat{var}(t_3) = (1/20^2)(1/20^2)\{(106^2/.7183)[(1/.7183) - 1] + (105^2/.7183)[(1/.7183) - 1] + 2(106)(105/.5676)[(.5676/.7183^2) - 1]\} = .09963$.

For the sample with the initial systematic design, the first primary unit (based on the starting position in the third column of the second row) intersects the central left network, with y-total 106, while the second primary unit (starting position in fourth column of third row) intersects both that network and the top right network, which has total y-value 115. By dividing the study area into four-by-four squares, it can be determined that 10 primary units (that is, 10 of the 16 possible systematic samples) intersect the top right network and 13 intersect the central left network, while nine intersect both. For this sample, $t_2 = (1/25)(1/2)[8.1538 + 19.6538] = .5562$, and $\widehat{var}(t_2) = [(16 - 2)/(25^2)(16)(2)](66.125) = .0463$. The intersection probability for the top network is .875 and for the left network is .975, and their joint intersection probability is .8583. For this sample, $t_3 = (1/25)(1/16)(115/.875 +$

$106/.975) = .6004$, and $\widehat{var}(t_3) = (1/25^2)(1/16^2)\{(115^2/.875)[(1/.875) - 1] + (106^2/.975)[(1/.975) - 1] + 2(115)(106/.8583)[(.8583/((.875)(.975)) - 1] = .01684$.

The actual variances for each of the unbiased estimators are given in Table 1 for the design with the initial strips and in Table 2 for the initially systematic design. In addition to the estimators t_1, t_2, and t_3, the variance has been computed for the sample mean t_1^* of a simple random sample of primary units with sample size equal to the expected sample size under the adaptive designs. The variance of t_1^* is computed using the formula of Section 3.1 with sample size $E(\nu)$, even if fractional, in place of n_1. Thus, t_1^* offers one way to compare the adaptive strategies with nonadaptive counterparts of equivalent sample size. Sample sizes in the table are expressed in terms of primary units. One primary unit consists of 20 secondary units in the strip design and 25 secondary units in the systematic design. The tables give variances obtained with initial sample sizes ranging from one up to a sampling fraction of one-half.

With the initial strip design, the adaptive strategies with an initial sample size of one primary unit are slightly more efficient than the comparable nonadaptive strategy for the example population. The relative advantage of the adaptive strategies increases with increasing initial sample size, and also the efficiency of t_3 relative to t_2 increases. With an initial sample size of 10 (initial sampling fraction of one-half), the adaptive cluster sampling strategy adds less than one primary unit to the expected sample size [$E(\nu) = 10.76$], but is almost five times as efficient as the equivalent nonadaptive strategy [$var(t_1^*)/var(t_3) = .04917/.01008 = 4.88$].

With the initial systematic sampling design, the adaptive strategies are dramatically more efficient than their nonadaptive counterparts for the example population. Also, comparing Tables 1 and 2, one sees that even with conventional systematic strategies (t_1 and t_1^*), variances are considerable lower than with the conventional strategies using strips. (This result would be expected due to the positive, monotonically decreasing covariance density function of the Poisson cluster process—see for example, Matérn, 1988 and Thompson and Ramsey, 1987.) The efficiency of the adaptive strategy with t_3 relative to the comparable nonadaptive systematic strategy with t_1^* ranges from 152% for a single initial systematic sample (.12852/.08441=1.52) to infinity—the adaptive strategy has zero variance for initial selections of more than six, as the intersection probability for each of the three networks in the population becomes one with such a design.

ACKNOWLEDGEMENTS

This research was supported by National Science Foundation Grant DMS-8705812. The paper was written while the author was a sabbatical visitor at the Institute of Mathematical Statistics, University of Copenhagen, Denmark.

REFERENCES

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā, Series A* **31**, 441-454.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling,* New York: Wiley.

Cormack, R.M. (1988). Statistical challenges in the environmental sciences: A personal view. *Journal of the Royal Statistical Society, Series A* **151**, 201-210.

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns.* New York: Academic Press.

Francis, R.I.C.C. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research* **18**, 59-71.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B* **17**, 269-278.

Hansen, M.M. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* **14**, 333-362.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

Kremers, W.K. (1987). Adaptive sampling to account for unknown variability among strata. Preprint No. 128. Institut für Mathematik, Universitat Augsburg, Federal Republic of Germany.

Matérn,B. (1986). *Spatial Variation*, 2nd edition. Berlin: Springer.

Seber, G.A.F. (1986). A review of estimating animal abundance. *Biometrics* **42**, 267-292.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association* **63**, 257-266.

Sirken, M.G. (1972). Variance components of multiplicity estimators. *Biometrics* **28**, 869-873.

Thompson, S.K. (1988). Adaptive sampling. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 784-786.

Thompson, S.K. (1989). Adaptive cluster sampling. Preprint no. 5, Institute of Mathematical Statistics, University of Copenhagen.

Thompson, S.K. and Ramsey, F.L. (1983). Adaptive sampling of animal populations. Technical Report 82. Dept. of Statistics, Oregon State University, Corvallis.

Thompson, S.K. and Ramsey, F.L.(1987). Detectability functions in observing spatial point processes. *Biometrics* **43**, 355-362.

Wolter, K.M. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association* **79**, 781-790.

Appendix A

The estimator t_2 can be written

$$t_2 = \frac{1}{Mn_1} \sum_{i=1}^{K} \frac{y_i r_i}{x_i},$$

where the random variable r_i denotes the number of primary units in the initial sample that intersect the i-th network of the population. Under the design, r_i has a hypergeometric distribution with expected value $n_1 x_i / N$. The expected value of t_2 is thus

$$E(t_2) = \frac{1}{Mn_1} \sum_{i=1}^{K} \frac{y_i n_1 x_i}{x_i N} = \frac{1}{MN} \sum_{i=1}^{K} y_i = \mu,$$

so t_2 is an unbiased estimator of the population mean.

An easy way to obtain the variance of t_2 is to associate with primary unit i of the population the variable $w_i = (1/M) \sum_{i=1}^{K} y_i I_{ki} / x_i$, for $i = 1, \ldots, N$, using the notation of Section 3.2. Then t_2 is the sample mean of the w_i for a simple random sample of size n_1. The formulas for the variance and the estimator of variance for t_2 follow readily.

Appendix B

The estimator t_3 can be written

$$t_3 = \frac{1}{MN} \sum_{i=1}^{K} \frac{y_i z_i}{\pi_i},$$

where the random variable z_i equals 1 if one or more primary units of the initial sample intersect network i, and z_i equals 0 otherwise. Under the design, z_i is a Bernoulli random variable with expected value π_i. The expected value of t_3 is thus

$$E(t_3) = \frac{1}{MN} \sum_{i=1}^{K} \frac{y_i \pi_i}{\pi_i} = \frac{1}{MN} \sum_{i=1}^{K} y_i = \mu,$$

so that t_3 is an unbiased estimator of the population mean.

The variance of t_3 can be written

$$vat(t_3) = \frac{1}{M^2 N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{y_i y_j}{\pi_i \pi_j} cov(z_i, z_j).$$

Since $var(z_i) = \pi_i(1 - \pi_i) = \pi_i - \pi_i^2$, and $cov(z_i, z_j) = \pi_{ij} - \pi_i \pi_j$ for $i \neq j$,

$$var(t_3) = \frac{1}{M^2 N^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j),$$

using the convention that $\pi_{ii} = \pi$.

The expression in Section 3.3 for $\widehat{var}(t_3)$ is unbiased for $var(t_3)$ because $E(z_i z_j) = \pi_{ij}$.

Appendix C

The 400 y-values for the population in the example of Section 4 are listed below:

$$
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 14 & 25 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 22 & 38 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 6 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 11 & 26 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 22 & 19 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 17 & 26 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 10 & 26 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 5 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

Table 1
Variances with initial long, thin strip plots

n_1	$E[\nu]$	$var(t_1)$	$var(t_1^*)$	$var(t_2)$	$var(t_3)$
1	1.57	1.30628	0.80706	0.79253	0.79253
1	1.57	1.30628	0.80706	0.79253	0.79253
3	4.35	0.38959	0.24758	0.23637	0.19944
4	5.58	0.27501	0.17749	0.16685	0.12651
5	6.74	0.20625	0.13530	0.12514	0.08378
6	7.82	0.16042	0.10702	0.09733	0.05636
7	8.85	0.12768	0.08666	0.07746	0.03788
8	9.82	0.10313	0.07123	0.06257	0.02510
9	10.76	0.08403	0.05907	0.05098	0.01621
10	11.66	0.06875	0.04917	0.04171	0.01008

Table 2
Variances with initial systematic samples

n_1	$E[\nu]$	$var(t_1)$	$var(t_1^*)$	$var(t_2)$	$var(t_3)$
1	2.98	0.44078	0.12825	0.08441	0.08441
2	4.36	0.20570	0.07846	0.03939	0.01684
3	5.31	0.12734	0.05919	0.02439	0.00363
4	6.15	0.08816	0.04701	0.01688	0.00072
5	6.98	0.06465	0.03798	0.01238	0.00011
6	7.80	0.04898	0.03089	0.00938	0.00001
7	8.62	0.03778	0.02516	0.00724	0.00000
8	9.44	0.02939	0.02042	0.00563	0.00000

FIGURES

Figure 1. Adaptive cluster sample with initial selection of five strip plots.

Figure 2. Adaptive cluster sample with initial selection of two systematic samples.

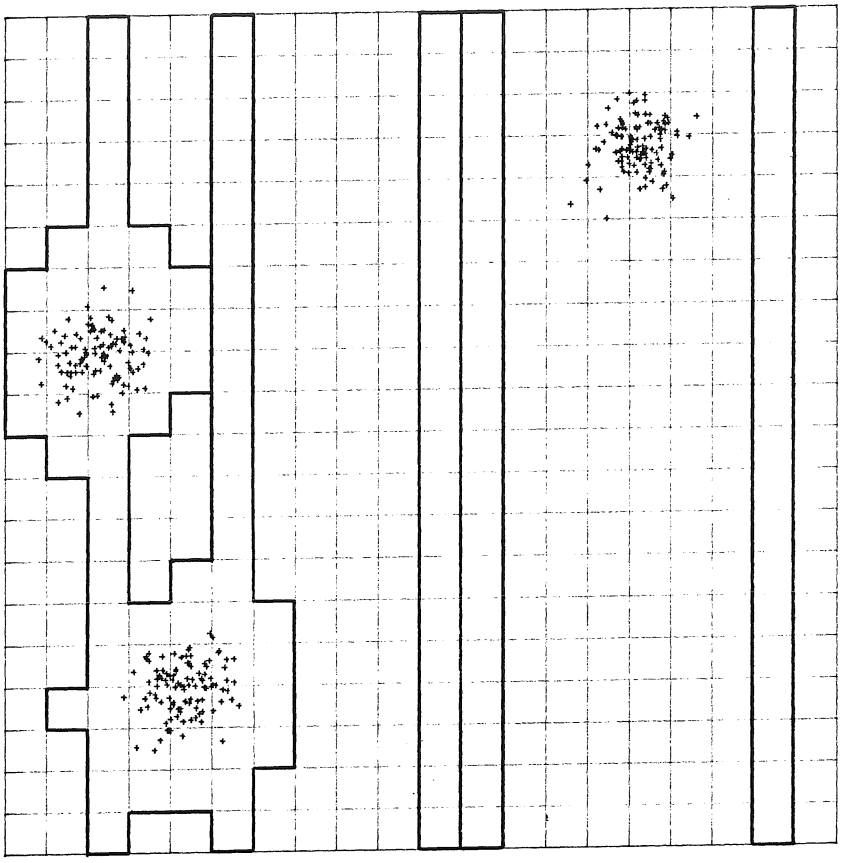


Figure 1

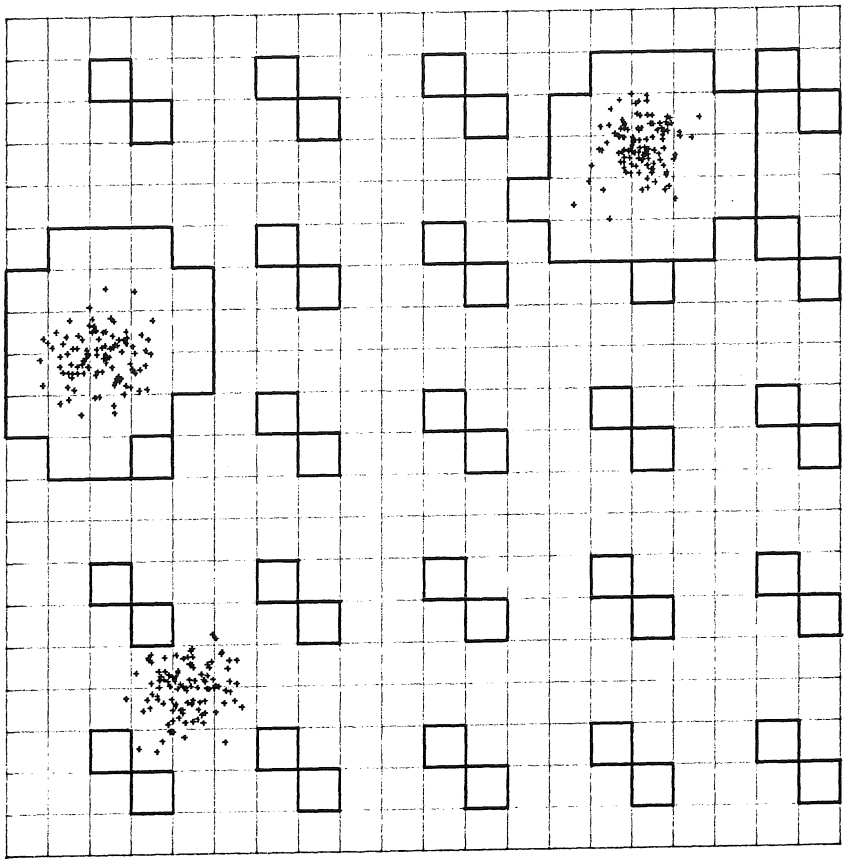


Figure 2

PREPRINTS 1988

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33.

No. 1 Jacobsen, Martin: Discrete Exponential Families: Deciding when the Maximum Likelihood Estimator Exists and Is Unique.

No. 2 Johansen, Søren and Juselius, Katarina: Hypothesis Testing for Cointegration Vectors - with an Application to the Demand for Money in Denmark and Finland.

No. 3 Jensen, Søren Tolver, Johansen, Søren and Lauritzen, Steffen L.: An Algorithm for Maximizing a Likelihood Function.

No. 4 Bertelsen, Aksel: On Non-Null Distributions Connected with Testing that a Real Normal Distribution Is Complex.

No. 5 Tjur, Tue: Statistical Tables for Personal Computer Users.

No. 6 Tjur, Tue: A New Upper Bound for the Efficiency of a Block Design.

No. 7 Bunzel, Henning, Høst, Viggo and Johansen, Søren: Some Simple Non-Parametric Tests for Misspecification of Regression Models Using Sign Changes of Residuals.

No. 8 Brøns, Hans and Jensen, Søren Tolver: Maximum Likelihood Estimation in the Negative Binomial Distribution.

No. 9 Andersson, S.A. and Perlman, M.D.: Lattice Models for Conditional Independence in a Multivariate Normal Distribution.

PREPRINTS 1989

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK, TELEPHONE + 45 1 35 31 33 .

No. 1 Bertelsen, Aksel: Asymptotic Expansion of a Complex Hypergeometric Function.

No. 2 Davidsen, Michael and Jacobsen, Martin: Weak Convergence of Twosided Stochastic Integrals, with an Application to Models for Left Truncated Survival Data.

No. 3 Johansen, Søren: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.

No. 4 Johansen, Søren and Juselius, Katarina: The Full Information Maximum Likelihood Procedure for Inference on Cointegration - with Applications.

No. 5 Thompson, Steven K.: Adaptive Cluster Sampling.

No. 6 Thompson, Steven K.: Adaptive Cluster Sampling: Designs with Primary and Secondary Units.