ENVIRONMETRICS
Environmetrics 2003; **14**: 593–610 (DOI: 10.1002/env.606)

Variance estimation for spatially balanced samples of environmental resources

Don L. Stevens, Jr[1]*,† and Anthony R. Olsen[2]

[1]*Department of Statistics, Oregon State University, Corvallis, OR 97331-4501, U.S.A.*
[2]*U. S. Environmental Protection Agency, NHEERL Western Ecology Division, 200 S.W. 35th Street, Corvallis, OR 97333, U.S.A.*

SUMMARY

The spatial distribution of a natural resource is an important consideration in designing an efficient survey or monitoring program for the resource. We review a unified strategy for designing probability samples of discrete, finite resource populations, such as lakes within some geographical region; linear populations, such as a stream network in a drainage basin, and continuous, two-dimensional populations, such as forests. The strategy can be viewed as a generalization of spatial stratification. In this article, we develop a local neighborhood variance estimator based on that perspective, and examine its behavior via simulation. The simulations indicate that the local neighborhood estimator is unbiased and stable. The Horvitz–Thompson variance estimator based on assuming independent random sampling (IRS) may be two times the magnitude of the local neighborhood estimate. An example using data from a generalized random-tessellation stratified design on the Oahe Reservoir resulted in local variance estimates being 22 to 58 percent smaller than Horvitz–Thompson IRS variance estimates. Variables with stronger spatial patterns had greater reductions in variance, as expected. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: spatial sampling; Horvitz–Thompson; environmental monitoring

1. INTRODUCTION

Environmental studies invariably involve populations distributed over space. Traditionally, such studies tended to focus on relatively small and well-delimited systems. However, some of the environmental issues that we face today, such as global warming, long-range transport of atmospheric pollutants, or habitat alteration, are not localized. Understanding and quantifying the extent of symptoms of wide-spread concerns requires large-scale study efforts, which in turn needs environmental sampling techniques and methodology that are formulated to address regional, continental, and global environmental issues. Stehman and Overton (1994) give an overview of some statistical issues associated with environmental sampling and monitoring, and Gilbert (1987) has an extensive discussion of sampling methods for monitoring environmental pollution.

One of the more prominent features of many environmental populations is the arrangement of the population units throughout space. Nearby units interact with one another, and tend to be influenced by

*Correspondence to: Don L. Stevens, Statistics Department, 44 Kidder Hall, Oregon State University, Corvallis, OR. 97331, U.S.A.
†E-mail: stevens@stat.orst.edu
Contract/grant sponsor: U.S. Environmental Protection Agency/Dynamac International; contract/grant number: 68-C6-0005
Contract/grant sponsor: U.S. Environmental Protection Agency/Oregon State University; contract/grant number: CR82-9096-01

Received 12 March 2002
Accepted 26 August 2002

the same set of natural and anthropogenic factors. For example, neighboring trees in a forest interact by competing for energy and nutrients, and are influenced by the same set of physical and meteorological conditions, the same level of air- or water-borne pollutants, and the same set of landscape disturbances. Sampling designs that capitalize on this spatial aspect of environmental populations tend to be more efficient than simple random sampling (Dunn and Harrison, 1993).

There are several basic paradigms for incorporating the spatial aspect of an environmental population into a sample. Area sampling partitions the domain of the population into polygons, which can be treated either as strata or as population units themselves. Systematic sampling (Cochran, 1946; Cochran, 1977; Madow, 1949) using a regular grid is often applied, as are several variants that perturb the strict alignment (Bellhouse, 1977; Dunn and Harrison, 1993; Olea, 1984). Along the same lines, Munholland and Borkowski (1996) have used a Latin square with a single additional independent sample to achieve a spatially balanced sample. Breidt (1995) used a Markov process to generate a one-unit-per-stratum spatially distributed sample. A third approach is to use space to order a list frame of the population, then use the order of the list to structure the sample, say by defining strata as successive segments of the ordered list, or by systematic random sampling. For example, Saalfeld (1991) drew on graph theory to define a tree that leads to a spatially articulated list frame, and the National Agricultural Statistics Service has used serpentine strips (Cotter and Nealon, 1987) to order their primary sample units within a state. A related idea that originated in geography is the General Balanced Ternary (GBT) spatial addressing scheme (Gibson and Lucas, 1982). The concept behind a GBT address is similar to the concept of space-filling curves, such as first constructed by Peano (1890), or the Hilbert curve (Simmons, 1963). Wolter and Harter (1990) have used a construction similar to Peano's to construct a 'Peano key' to maintain the spatial dispersion of a sample as the underlying population experiences births or deaths. Saalfeld (1991) has also used the Peano key to maintain spatial dispersion of a sample.

We have synthesized several of these concepts to create a very powerful and flexible technique for selecting a spatially well-distributed probability sample. The technique is based on creating a function that maps 2-dimensional space into 1-dimensional space, thereby defining an ordered spatial address. We require that the function be quadrant-recursive (Mark, 1990), that is, that the image of any subquadrant be an interval. The quadrant-recursive property ensures that some 2-dimensional proximity relationships are preserved under the function. A restricted randomization, called hierarchical randomization (HR), is used to randomly order the spatial addresses. Systematic sampling along the randomly ordered address sequence is analogous to sampling a random tessellation of 2-dimensional space, and results in a spatially well-balanced random sample. We call the resulting design a Generalized Random Tessellation Stratified (GRTS) design. Details of the design are discussed in Stevens (1997), Stevens and Olsen (1999), Stevens and Olsen (2000), and Stevens and Olsen.[1] In this article, we provide an abbreviated description of the design, note some of it properties, and develop a variance estimator that is easily computable, approximately unbiased, and stable.

2. GENERALIZED RANDOM TESSELLATION STRATIFIED DESIGN

The GRTS design is developed as if we were selecting points in a continuous, two-dimensional target population. However, it works equally well for obtaining a spatially well-distributed sample of a finite

[1]Stevens DL Jr, Olsen AR. 2002. in review. Spatially-balanced sampling of natural resources in the presence of frame imperfections.

population consisting of discrete units with known spatial locations or a linear, continuous population embedded in 2-space, e.g. a stream network. In these two cases, let the domain be a 2-dimensional region containing the population. The hierarchically randomized, quadrant-recursive function in the design application assigns a random address to every one of the (uncountably infinite) points in the domain. Thus, every unit in the finite population will be assigned a random address, which can be used to induce a random order of the population. Similarly, every point in a linear network will be mapped onto a random point, in effect stringing the points of the network out onto a line in random order. In all three cases, systematic sampling along the random order will result in the corresponding sample units or points being well-distributed over the population domain.

Stevens (1997) derived inclusion and joint inclusion functions for several grid-based designs that were precursors to GRTS designs, and share some of their properties. The designs are all generalizations of the Random Tessellation Stratified (RTS) design (Overton and Stehman, 1993; Olea, 1984; Dalenius *et al.*, 1961). The RTS design selects random points in space via a 2-step process. First, a regular tessellation coherent with a regular grid is randomly located over the domain to be sampled, and second, a random point is selected within each random tessellation cell. The RTS design is a variation on a systematic design that avoids the alignment problems that can occur with a completely regular systematic design. Like a systematic design, an RTS design does not allow variable probability spatial sampling. Stevens (1997) introduced the Multiple-Density, Nested, Random-Tessellation Stratified (MD-NRTS) design to provide for variable spatial sampling intensity. The geometric concept underlying the MD-NRTS was the notion of coherent intensification of a grid: adding points to a regular grid in such a way as to result in a denser regular grid with similarly shaped but smaller tessellation cells.

We can view a quadrant-recursive function as being defined by the limit of successive intensifications of a grid covering the unit square, where a grid cell is divided into four sub-cells, each of which is subsequently divided into four sub-sub-cells, and so on. If we were to carry this recursion to the limit, and pair grid points with addresses based on the order in which the divisions were carried out, with each digit of the address representing a step in the subdivision, then we obtain a quadrant-recursive function. For example, suppose we begin with a point at $(1, 1)$, and replace it with four points $p_0 = (1/2, 1/2)$, $p_1 = (1/2,\ 1)$, $p_2 = (1,\ 1/2)$, and $p_3 = (1,\ 1)$. The next step of the recursion replaces each of the four points $\mathrm{p}_0, \ldots, \mathrm{p}_3$ with $\{\mathrm{p}_\mathrm{i} - \{(1,\ 1),\ (0, 1),\ (1, 0),\ (0, 0)\}/2^2\}$. Thus the point $\mathrm{p}_1 = (1/2,\ 1)$ is replaced with the four points $p_{10} = (1/4,\ 3/4), p_{11} = (1/4,\ 1), p_{12} = (1/2,\ 3/4)$ and $p_{13} = (1/2,\ 1)$. Figure 1(a) shows the first four points (larger dots), and the successor points to p_1

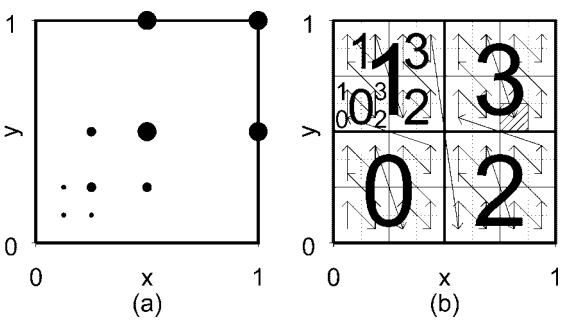

Figure 1. Schematic diagram of quadrant-recursive partition of unit square. (a) Sequence of successor points, and (b) Addresses induced by the partitioning and a path following the order of addressing

(smaller dots). In general, the nth step replaces each of the 4^n points $p_{i_1 i_2 \ldots i_n}$ with $\{p_{i_1 i_2 \ldots i_n} - \{(1,1), (0,1), (1,0), (0,0)\}/2^{n+1}\}$.

A spatially referenced address can be constructed following the pattern of the partitioning, with each new partition adding a digit position to the address. Thus, in the above example, the first group of four points are assigned the addresses '0', '1', '2' and '3', with '3' being the original point at (1, 1). The successor points to '2' receive the addresses '20', '21', '22' and '23', and so forth. If subquadrants are associated with the point in their upper-right corner, then the addresses induce a linear ordering of the sub-quadrants. Moreover, if we carry the process to the limit, and treat the resulting address as digits in a base-4 fraction, e.g. '$20131\ldots$' as the base 4 number $(0.20131\ldots)_4$, then the correspondence between grid point and address is a quadrant-recursive function.

Figure 1(b) shows the first four levels of the quadrant-recursive partitioning of the unit square with the associated addresses. Thus, for example, the address of the cross-hatched subquadrant is, as a base 4 fraction, $(0.320)_4$. If we were to carry the recursive-partitioning to the limit, every point in the subquadrant would be assigned an address beginning with $(0.320)_4$, and so would be in the interval $(0.320, 0.321)_4 = (56/64, 57/64)_{10}$.

The line connecting subquadrants in Figure 1(b) follows the same pattern within every subquadrant, that is, the subquadrants are linked together in the order lower left, upper left, lower right, upper right. A permutation of that order would still yield a quadrant-recursive function; however, the resulting addressing sequence would be different. In fact, a different permutation could be chosen for every partition of every subquadrant and the resulting mapping would still be quadrant recursive. If the permutations are chosen at random and independently from the set of all possible permutations, we call the resulting random address sequence a *hierarchical randomization* of the original sequence obtained using the order lower left, upper left, lower right, upper right within every subquadrant. If the process is carried to the limit, the result is a 1–1, onto, quadrant-recursive, hierarchically randomized function f that maps the unit square to the unit interval.

The next step in the sample selection is to induce a measure on the unit interval corresponding to the inclusion probability function on the population domain. Since f is 1–1 and onto, f^{-1} is well-defined (in fact, both f and f^{-1} are measurable functions). We define the induced measure by assigning to each interval of the form $(0, x]$ the total of the inclusion probability of the set $B(x) = f^{-1}([0, x])$. In the case of a finite population, the total is just the sum of the inclusion probabilities of all units in $B(x)$, i.e. $\sum_{u_i \in B(x)} \pi_i$, where π_i is the inclusion probability for population unit u_i. If the population is an infinite continuum, e.g. a linear or 2-dimensional extensive resource, then the total is $\int_{B(x)} \pi(s)\, \mathrm{d}\phi(s)$ where $\pi(s)$ is the inclusion density and $\phi(s)$ is a measure such that $\phi(B(x))$ gives the amount (length or area) of the resource in $B(x)$. In any case, we can define a distribution function $F(x)$ so that $F(x)$ is the total inclusion probability over $B(x)$. We have then that $F(0) = 0$ and $F(1) = M =$ expected sample size.

A systematic sample with a random start and unit selection interval will, on the average, locate M points in the interval $(0, M]$. Because F is increasing, F^{-1} maps the selected points onto points in the unit interval. We then use f^{-1} to map the points in the unit interval back to the population domain, thereby defining the sample. A schematic of the process is given in Figure 2. Because of the recursive construction of f, systematic sampling along the randomly ordered spatial address is analogous to sampling a random tessellation of 2-dimensional space, and results in a spatially well-balanced random sample. The construction of the functions f and F ensures that (i) the sample will be well-distributed over the population domain, and (ii) it will have the desired inclusion probability. Details of the sampling method and the construction of f, f^{-1}, F and F^{-1} are given in Stevens and Olsen.

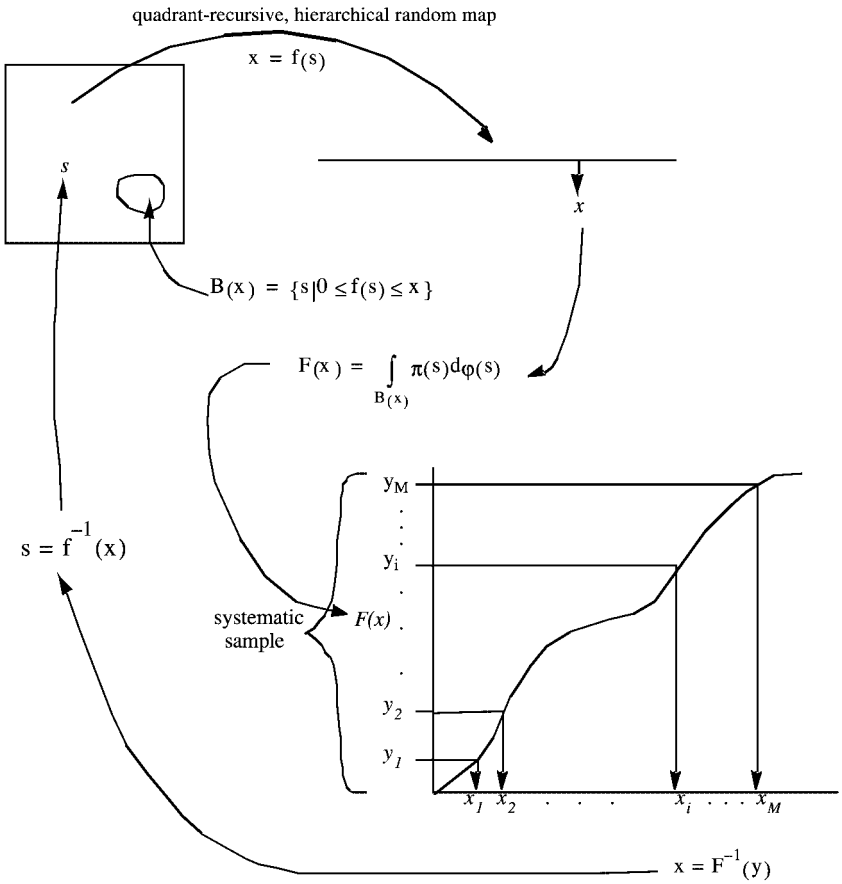

Figure 2. Schematic diagram of sample selection using the GRTS algorithm

3. ESTIMATION

The basic theoretical tool for population estimation using complex, variable probability sampling designs is the Horvitz–Thompson Theorem (Horvitz and Thompson, 1952), which is stated here in its continuous form (Cordy, 1993):

Continuous Horvitz–Thompson Theorem: Let $s_1, s_2, \ldots, s_n$ be a sample selected from a universe U according to a design with inclusion function $\pi(s)$ and joint inclusion function $\pi(s,t)$, with $\pi(s) > 0$ almost everywhere on U. Let $R \subset U$, and let $z(s)$ be a real-valued integrable function defined on R. An unbiased estimator of $\int_R z(s)\,\mathrm{d}s = z_\mathrm{T}$ is given by

$$\hat{z}_\mathrm{T} = \sum_{i=1}^{n} \frac{I_R(s_i) z(s_i)}{\pi(s_i)}$$

with variance (Horvitz and Thompson, 1952)

$$V_{\mathrm{HT}}(\hat{z}_{\mathrm{T}}) = \int_R \frac{z^2(s)}{\pi(s)}\mathrm{d}s + \int_R \int_R \left[\frac{\pi(s,\, t) - \pi(s)\pi(t)}{\pi(s)\pi(t)}\right] z(s)z(t)\ \mathrm{d}t\ \mathrm{d}s$$

or, equivalently (Yates and Grundy, 1953),

$$V_{\mathrm{YG}}(\hat{z}_{\mathrm{T}}) = \frac{1}{2}\int_U \int_U [\pi(s)\pi(t) - \pi(s,\, t)]\left[\frac{z(s)I_R(s)}{\pi(s)} - \frac{z(t)I_R(t)}{\pi(t)}\right]^2 \mathrm{d}t\ \mathrm{d}s$$

Corresponding estimators of variance are

$$\hat{V}_{\mathrm{HT}}(\hat{z}_{\mathrm{T}}) = \sum_{s_i \in R} \frac{z^2(s_i)}{\pi^2(s_i)} + \sum_{s_i \in R} \sum_{\substack{s_j \in R \\ j \neq i}} \left[\frac{\pi(s_i,\, s_j) - \pi(s_i)\pi(s_j)}{\pi(s_i,\, s_j)\pi(s_i)\pi(s_j)}\right] z(s_i)z(s_j)$$

and

$$\hat{V}_{\mathrm{YG}}(\hat{z}_{\mathrm{T}}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i,\, s_j)}{\pi(s_i,\, s_j)}\right]\left[\frac{z(s_i)I_R(s_i)}{\pi(s_i)} - \frac{z(s_j)I_R(s_j)}{\pi(s_j)}\right]^2$$

Both variance estimators are unbiased, provided $\pi(s,\, t) > 0$ almost everywhere in U.

The spatially balanced designs obtained by the composition of random grid placement, hierarchical randomization of a quadrant-recursive address, and systematic sampling have spatial distribution properties that are very similar to a simple RTS design at the same spatial resolution. This has been established by extensive simulation with a variety of populations.

Let C be a polygon congruent to the tessellation cells, let $C(0)$ be the cell enclosing the (non-random) origin, and $C(\mathrm{s})$ be $C(0)$ translated to the point s, that is, $C(s) = \{t \,|\, t - s \in C(0)\}$ Following Stevens (1997), the inclusion functions for the RTS design are

$$\pi(s) = \frac{1}{|C(s)|} = \frac{1}{|C|}$$

and

$$\pi(s,\, t) = \pi(s)\pi(t)\left\{1 - \frac{|C(s) \cap C(t)|}{|C|}\right\}$$

where $|C|$ denotes the area of C. The GRTS design has joint inclusion functions that are non-zero almost everywhere, and we have accurate, easily computable approximations for them, so that, in theory, the Horvitz–Thompson (HT) and Yates–Grundy (YG) estimators are applicable for variance estimation. However, in practice, the case is not so straightforward. The HT variance estimator has an unfortunate tendency to yield negative estimates. The YG estimator is guaranteed to be positive for the GRTS design, but tends to be unstable. The difficulty stems from the fact that the joint inclusion density appears in the denominator of the estimators. Like the RTS design, the GRTS design guarantees that $\pi(s,\, t) > 0$ for s, but $\pi(s,\, t) \to 0$ as $s \to t$. Our experience has shown that most

applications of the GRTS design with a modest number (30+) of sample points result in one or more point pairs with small values of $\pi(s, t)$. The corresponding terms in the variance estimators tend to be large in absolute value and to dominate the value of the estimators, leading to their unstable behavior. This seems to be a problem especially for variable probability samples where the inclusion density is discontinuous, since then the values of $z(\cdot)/\pi(\cdot)$ can be substantially different even for nearby points. Since many of the designs we envision will have discontinuous inclusion densities (e.g. at regional boundaries, or stream confluences), the HT and YG variance estimators are unsuitable.

A stable variance estimator can be obtained by treating the sample as if it arose from independent random sampling (IRS), where the n points are selected independently from an arbitrary density $f(s)$ over U. This results in an estimator analogous to the 'simplified' estimator given by Särndal *et al.* (1992), or the pps-wr estimator v_{10} given by Wolter (1985). For an IRS design, the inclusion density is $\pi(s) = nf(s)$, and the pairwise inclusion density is $\pi_{IRS}(s, t) = n(n-1)f(s)f(t) = (n-1)\ \pi(s)\ \pi(t)/n$. We know the true inclusion density for our design, and we obtain an approximate variance by replacing the true pairwise inclusion density with the IRS expression. When we do that, the HT variance estimator for $\hat{Z}_{\mathrm{T}}$ reduces to

$$\hat{V}_{\mathrm{IRS}}(\hat{Z}_{\mathrm{T}}) = \sum_{s_i \in R}\left(\frac{z(s_i)}{\pi(s_i)}\right)^2 - \frac{1}{n-1}\sum_{\substack{s_i, s_j \in R \\ i \neq j}}\left(\frac{z(s_i)}{\pi(s_i)}\right)\left(\frac{z(s_j)}{\pi(s_j)}\right) = \frac{n}{n-1}\sum_{s_i \in R}\left(\frac{z(s_i)}{\pi(s_i)} - \overline{\left(\frac{z}{\pi}\right)}\right)^2 = nV_{\mathrm{SRS}}(z/\pi)$$

where $V_{\mathrm{SRS}}(z/\pi)$ is the usual estimator of the population variance for a simple random sample (SRS) design applied to $z(s_i)/\pi(s_i)$.

The IRS estimator accounts for the non-constant inclusion density, but does not account for the spatially constrained nature of the GRTS design. If the response has some spatial pattern, at least to the extent that the responses for two points close together tend to be more similar than the responses at two points far apart, then the GRTS design will lead to more precise estimates than independent random sampling with the same inclusion function. Thus, the IRS estimator will be conservative, i.e. it will tend to overstate the variance.

Dunn and Harrison (1993) compared the IRS variance estimator to one based on post-stratification. The sample points from a systematic sample and a 1-per-stratum stratified random sample from two land-use data sets were formed into artificial strata by grouping adjacent points into strata of size 2. Near the boundary, strata with three points were allowed. Their results indicated that both the IRS and the post-stratification estimator tended to overestimate variance, although the post-stratification estimator was less biased than the IRS approximation.

Several authors (Yates, 1981; Wolter, 1985; Overton and Stehman, 1993) have considered a class of estimators based on contrasts. The general form of these estimators is

$$\hat{V}_{\mathrm{Ctr}}(\hat{Z}_{\mathrm{T}}) = \sum_i w_i y_i^2$$

where y_i is a contrast of the form $y_i = \sum_k c_{ik} z(s_k)$ with $\sum_k c_{ik} = 0$. For an RTS design, Overton and Stehman also considered a 'smoothed' contrast-based estimator of the form

$$\hat{V}_{\mathrm{SMO}}(\hat{Z}_{\mathrm{T}}) = \sum_i w_i (z_i - z_i^*)^2$$

where z_i^*, called the 'smoothed value' for data point z_i, is taken as a weighted mean of a point plus its nearest neighbors in the tessellation. Thus, for points near the edges of the population, fewer neighbors are used in calculating z_i^*.

The variance estimator we are proposing here is a contrast-based estimator that bears some resemblance to the Overton and Stehman smoothed estimator. We replace the single contrast $(z_i - z_i^*)^2$ with an average of several contrasts over data from a local neighborhood that is the construct in the GRTS design analogous to a tessellation cell and its nearest neighbors in the RTS design. Some of the justification for this approach is the observation that the selection from unit intervals on the line corresponds to selection from a random tessellation of the population domain, i.e. a random stratification. If we let $\tilde{B}$ denote the random event that determines the stratification, then the GRTS design, conditional on $\tilde{B}$, is a 1-sample-per-stratum spatially stratified sample. Recall that $\hat{Z}_T = \sum_{s_i \in R} z(s_i)/\pi(s_i)$, where $z(s_i)$ is a sample from the ith random stratum. Since the selections within strata are conditionally independent of one another, $E[\hat{Z}_T \mid \tilde{B}] = Z_T$, so that $V(\hat{Z}_T) = E[V(\hat{Z}_T \mid \tilde{B})] + V(E[\hat{Z}_T \mid \tilde{B}]) = E[V(\hat{Z}_T \mid \tilde{B})] = \sum_{s_i \in R} E\big[V\{z(s_i)/\pi(s_i)\} \mid \tilde{B}\big)\big]$. We form the neighborhood variance estimator by approximating $E\big[V\{z(s_i)/\pi(s_i)\} \mid \tilde{B}\big)\big]$ by averaging several contrasts over a local neighborhood $D(s_i)$.

The choice of a neighborhood is motivated by the following considerations. For a GRTS design, the joint inclusion function $\pi(s, t)$ is well-approximated by a function of the form $\pi(s, t) = \pi(s)\pi(t)\{1 - h(s, t)\}$, where $h(s, t)$ has the properties: $h(s, t) = h(t, s)$, $h(s, s) = 1$, $0 \le h(s, t) \le 1$, $h(s, s + \Delta s) \to 0$ as $|\Delta s|$ increases, and $h(s, s + \Delta s) = 0$ for $|\Delta s|$ greater than some constant. Stevens (1997) has shown this analytically for several variations on the basic RTS design, and we have investigated more complex applications via simulation. For $s \in R$, let $D(s)$ be the neighborhood of s where $h(s, t)$ is positive, i.e. let $D(s) = \{t \in R \mid h(s, t) > 0\}$. For t outside $D(s)$, the pairwise inclusion density factors $\pi(s, t) = \pi(s)\pi(t)$, an independence-like condition, so that $D(s)$ can be thought of as a neighborhood of influence for a sample point at s. It follows that $\pi(s, t) - \pi(s)\pi(t) = 0$, $t \notin D(s)$. Applying this relation in the YG variance gives

$$V_{YG}(\hat{Z}_T) = \frac{1}{2}\int_U \int_{D(s)} \big[\pi(s)\pi(t) - \pi(s, t)\big]\left[\frac{z(s)I_R(s)}{\pi(s)} - \frac{z(t)I_R(t)}{\pi(t)}\right]^2 \mathrm{d}t\,\mathrm{d}s$$

that is, only point pairs (s, t), with $t \in D(s)$, contribute to the variance. Thus, neighborhoods corresponding to the $D(s)$ are a natural choice on which to base a local estimate of variance.

In an equi-probable RTS design, the neighborhoods $D(s)$ are easy to determine. If the RTS is based on a tessellation with cells congruent to a polygon C, then $D(s)$ is a polygon similar to C but with four times the area. Moreover, in this case, the expected number of sample points falling in $D(s)$ is four. For the GRTS design, the case is not so straightforward. For example, a non-constant inclusion probability density distorts the shape of the $D(s)$. Even so, the expected number of samples falling in $D(s)$ is still four. We use this characteristic to define the local neighborhoods used in the estimator.

The neighborhoods $D(s_i)$ are developed by initially including the point s_i itself plus the next three nearest neighbors for each point. Thus, the minimum number of points in any $D(s_i)$ is four. Including more points tends to increase the local variance, since the variance is integrated over a larger portion of the population. Including fewer points tends to increase the variability of the local estimate. The neighborhoods are then adjusted by adding to $D(s_i)$ any points s_j such that $s_i \in D(s_j)$. This ensures that $s_j \in D(s_i) \Leftrightarrow s_i \in D(s_j)$, reflecting the requirement that $h(s, t) = h(t, s)$. The neighborhood total is calculated as $\bar{z}_D(s_i) = \sum_{s_j \in D(s_i)} w_{ij}[z(s_j)/\pi(s_j)]$. The weights w_{ij} are selected using the following criteria:

1. The weight w_{ij} should vary inversely as $\pi(s_j)$ and decrease as the distance between s_i and s_j increases.
2. $\sum_i w_{ij} = \sum_j w_{ij} = 1$, so that the neighborhood totals are averages over the neighborhoods, and the sum of the neighborhood totals is equal to the estimated overall total.

The weights are developed by first assigning a value that decreases as the rank of the distance between s_j and s_i among the points in $D(s_i)$ increases and is inversely proportional to $\pi(s_j)$. The formula for this first step is

$$w_{ij}^* = \frac{1 - (\text{rank}(s_j) - 1)/\text{count}(D(s_i))}{\pi(s_j)}$$

For example, if $D(s_1)$ contained five points, the points would be ranked 1 through 5 in order of their distance from s_1. Of course, s_1 receives rank 1, since it is the closest point to itself. The other four points would be ranked in terms of increasing distance from s_1. If all of the points have the same inclusion density, say $\pi(s_j) \equiv \pi$, then the point with rank 4 would get weight $[1 - ((4-1)/5)]/\pi = (2/5)/\pi$. The weights are normalized to satisfy each column total constraint by setting $\tilde{w}_{ij} = w_{ij}^* / \sum_{s_k \in D(s_i)} w_{ik}^*$. There is no unique way to satisfy both constraints in criterion 2, so we select the set of weights w_{ij} that minimizes $\sum_{i,j} (w_{ij} - \tilde{w}_{ij})^2$ while satisfying criterion 2. We solve this constrained minimization problem using Lagrange multipliers. The unconstrained minimization is then

$$\min_{w_{ij}, \lambda_k, \gamma_l} \sum_{i,j} \left(w_{ij} - \tilde{w}_{ij}\right)^2 + \sum_k \lambda_k \left(\sum_j w_{kj} - 1 \right) + \sum_l \gamma_l \left(\sum_i w_{il} - 1 \right)$$

The w_{ij} are easily eliminated from the set of linear equations obtained by setting derivatives to 0. The resulting set of equations in λ_k and λ_l are singular, and we use the Moore–Penrose generalized inverse (Rao and Mitra, 1971) to solve for $\hat{\lambda}_k$ and $\hat{\gamma}_l$. The minimizing set of weights is

$$w_{ij} = w_{ij}^* + \frac{\hat{\lambda}_i + \hat{\gamma}_j}{2}$$

The neighborhood-based variance estimator is then

$$\hat{V}_{\text{NBH}}(\hat{Z})_{\text{T}}) = \sum_{s_i \in R} \sum_{s_j \in D(s_i)} w_{ij} \left(\frac{z(s_j)}{\pi(s_j)} - \bar{z}_{D(s_i)} \right)^2 = \sum_{s_i \in R} \sum_{s_j \in D(s_i)} w_{ij} \left(\frac{z(s_j)}{\pi(s_j)} - \sum_{s_k \in D(s_i)} w_{ik} \frac{z(s_k)}{\pi(s_k)} \right)^2$$

We note that, by the using the symmetry of $h(s, t)$, the estimator can be rewritten as

$$\hat{V}_{\text{NBH}}(\hat{Z}_{\text{T}}) = \sum_{s_j \in R} \sum_{s_i \in D(s_j)} w_{ij} \left(\frac{z(s_j)}{\pi(s_j)} - \bar{z}_{D(s_i)} \right)^2$$

Because of the constraint $\sum_i w_{ij} = 1$, the term $\sum_{s_i \in D(s_j)} w_{ij} \left([z(s_j)/\pi(s_j)] - \bar{z}_{D(s_i)} \right)^2$ can be regarded as the average of several estimates of variance, each taking the mean over a somewhat different region corresponding to a different random tessellation. Thus, we interpret the term $\sum_{s_i \in D(s_j)} w_{ij} \left([z(s_j)/\pi(s_j)] - \bar{z}_{D(s_i)} \right)^2$ as an approximation to $\text{E}\left[V\left([z(s_j)/\pi(s_j)] \mid \tilde{B} \right) \right]$.

The use of a local neighborhood and decreasing weights as a function of distance from a point is similar to approaches used in geostatistics (Cressie, 1991). In kriging, variance estimates at a point are based on the variogram, which is a function of the distance between points, and is usually estimated from the data. The estimation process is developed assuming a stochastic process for the generation of the data. It is natural to use properties of that stochastic process to develop the variance estimator. In our case, the local neighborhood variance estimation process is developed to incorporate properties of the survey design.

4. SIMULATION STUDY

We have verified the performance of the estimator on a variety of real and constructed populations. Here we show results for a finite, a linear, and an extensive population. The finite and extensive populations are artificial; they were constructed to have features that seemed to cause the HR and YG variance estimators to be particularly unstable. The linear population consists of real stream traces from the upper portion of a watershed. The designs in each case incorporate variable probability.

For the three population types, we assigned population response values by picking (x, y) coordinates, and associating a z value by interpolating on a surface. This procedure ensured some degree of spatial association in the responses. We used two different surfaces (a 'smooth' and a 'rough' surface), shown in Figure 3, to define the response values. The surfaces themselves were defined by specifying values at a 101×101 matrix of xy-coordinates, and using the interpolation algorithm to specify the response at arbitrary coordinates.

4.1. Finite population simulation results

We generated a finite population consisting of 1000 units. The locations of the units were picked to achieve a population with wide variation in spatial density, to have voids and to have areas of densely packed population elements. Figure 4 shows the spatial pattern of the finite population used in the simulations. Variable probability was introduced by assigning, at random, 750 units with a relative weight of 1, 200 units with a weight of 2, and 50 units with a weight of 4.

We selected 1000 samples of size 50 from the population using the GRTS design. For each sample, we calculated the estimate of Z_T, and the IRS and NBH estimators of variance. We estimated the true variance by the variance of the 1000 estimates of Z_T. We calculated confidence interval coverage using normal theory confidence intervals, and the known total of the response.

Figure 5 shows a histogram of the 1000 estimates of the NBH variance of Z_T using response surface 1, the smoother surface. A vertical bar is drawn at the location of the true variance. Confidence interval coverages are also displayed in the figure. The estimator is approximately centered on the true value, although it appears to be biased slightly high. The confidence interval coverages are also slightly high, but close to nominal. Since the variance holds little intrinsic interest, other than as a means to calculate confidence, the agreement of the coverage with the nominal is important. Figure 6 shows the same 1000 samples as in Figure 5, but with response values from Surface 2. The results are very similar to the first case, with the roughness of the surface reflected in higher variance estimates. The results for both surfaces are summarized in Table 1. Note that the estimator based on the IRS approximation is too large by a factor of about 2; furthermore, it is also more variable than the neighborhood-based estimator.

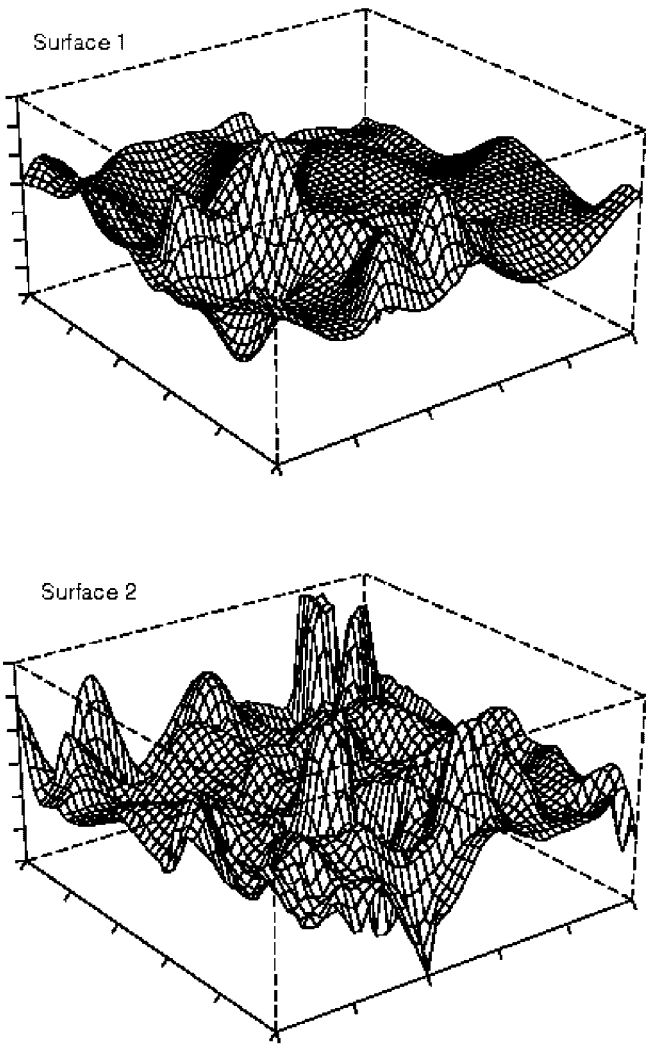

Figure 3. Surfaces used to assign population values

4.2. Linear population simulation results

We used the stream traces from a real stream network to test the estimator on a linear population. We assigned weights by Strahler order, using a weight function equal to the stream order. This weighting scheme is similar to one often used by the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP)(Herlihy, *et al.*, 2000; Larsen, *et al.*, 1991; Stevens, 1994). Stream networks tend to have much of their total length in smaller, lower-order reaches, e.g. in the headwater reaches. Placing greater weight on the higher-order reaches puts more sample points in the streams that are more likely to contain fish. The test population is shown in Figure 6. The results, shown in Figure 7, are very similar to the results to the finite population cases: the estimator was approximately unbiased, and confidence interval coverage was near nominal.

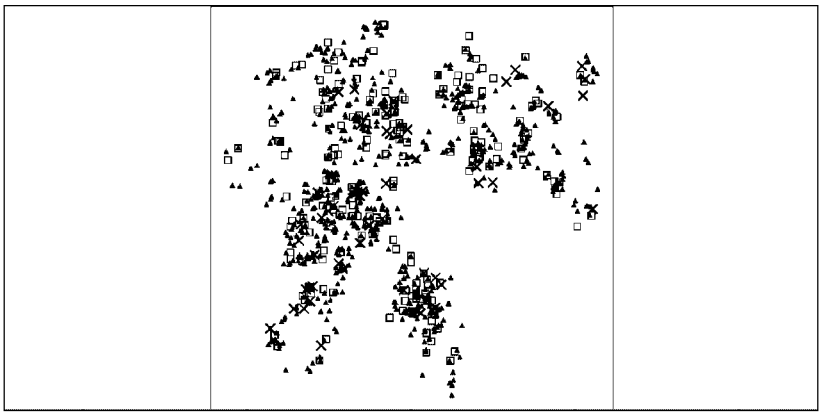

Figure 4. Locations of finite population used in simulation. Weight 1 units are represented by Δ, weight 2 units by $\square$, and weight 4 units by $\times$

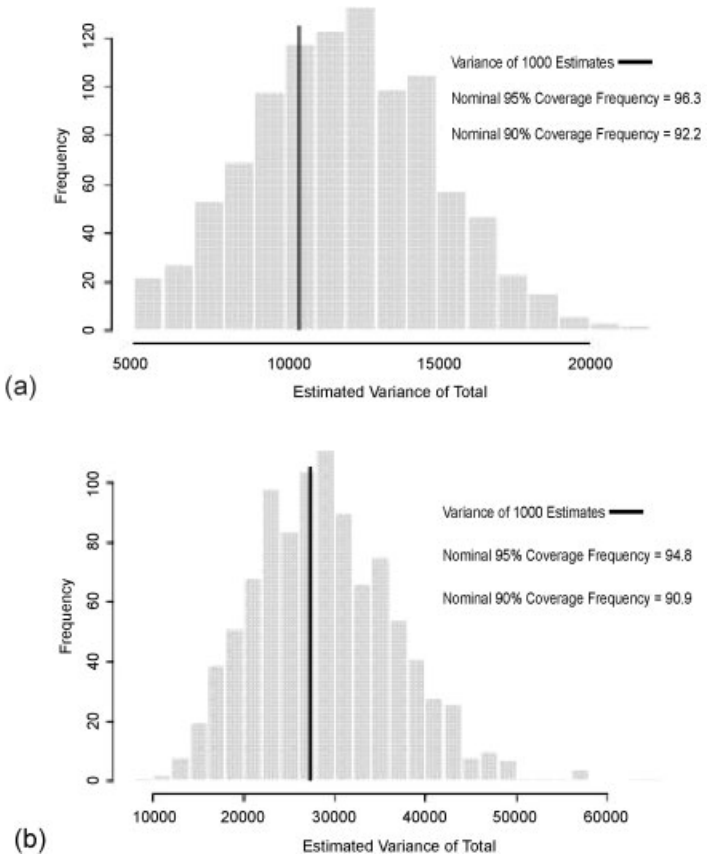

Figure 5. Histogram showing results of 1000 GRTS samples of size 50 from finite population using (a) Surface 1 and (b) Surface 2

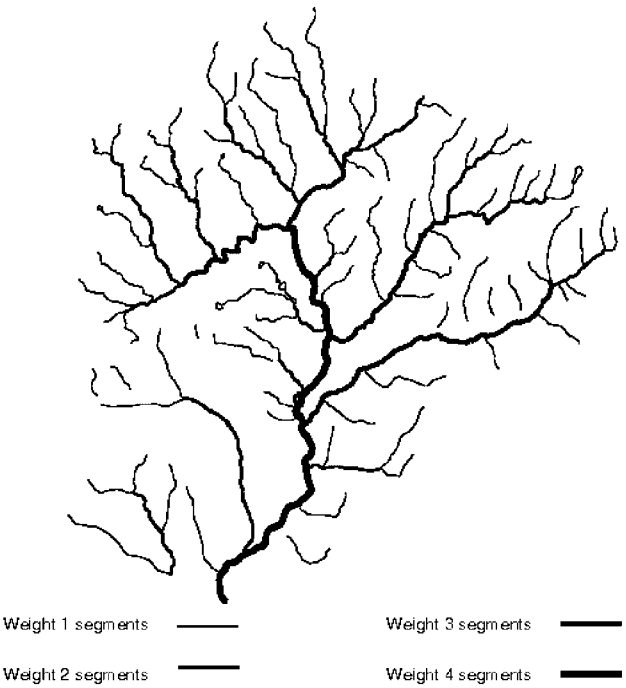

Figure 6. Linear network used in simulation study

4.3. Extensive population simulation results

The final test population was an artificial extensive population, which consisted of an irregular region split into two unequal-sized subregions. In this case, variable probability was introduced by specifying the sample sizes required in each region, with the larger sample specified for the smaller region. This implicitly defines the inclusion density function, constant over subregion, with the relative intensity of Region 2 to Region 1 equal to 2.66. The test population is displayed in Figure 8, with a histogram of the results for Surface 1 in Figure 9. The results for Surface 2 were quite similar.

The behavior of the NBH estimator is compared to the IRS estimator in Figure 10, where the histograms of each estimator are shown, drawn to the same scale. As for the finite population case, the IRS estimator is biased high, by a factor of about 2. Also, the NBH estimator is more stable than the IRS estimator, consistent with the finite population results.

Table 1. Summary statistics for 1000 samples from a finite population. V_{1000} is the variance of the 1000 estimates of the total; $\hat{V}_{\mathrm{NBH}}$ and $\hat{V}_{\mathrm{IRS}}$ are the means of the neighborhood- and IRS-based estimators, respectively. The columns headed 'SD' are the standard deviations of the variance estimators

Surface	V_{1000}	$\hat{V}_{\mathrm{NBH}}$	$SD_{\hat{V}_{\mathrm{NBH}}}$	$\hat{V}_{\mathrm{IRS}}$	$SD_{\hat{V}_{\mathrm{IRS}}}$
1	10472	12768	3526	21678	5632
2	27325	27899	8367	53513	15197

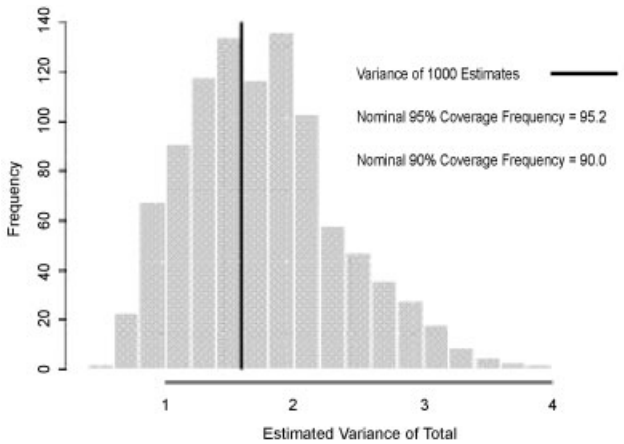

Figure 7. Histogram of estimated variance for 1000 GRTS samples of size 50 from a linear population

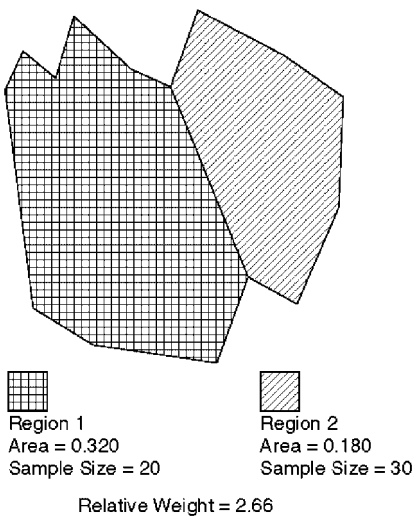

Figure 8. Two subregions of extensive population used in simulation study, showing their areas and sample sizes

5. EXAMPLE APPLICATION: SAMPLING THE OAHE RESERVOIR ON THE MISSOURI RIVER

The USEPA is completing a research study to develop ecological indicators for large rivers and associated reservoirs (Bolgrien *et al.*[2]). As part of this study, we developed a GRTS survey design for the Oahe Reservoir on the Missouri River. The Oahe reservoir is 1287.5 km^2 in area and approximately 325 km in length, extending from North Dakota to South Dakota. An extensive frame development

[2]Bolgrien DW, Angradi TR, Corry TD, Scharold JV, Schweiger EW, Kelly JR. 2002. A whole-lake water quality survey of a large reservoir of the Missouri River using a spatially-balanced probabilistic design. (in prep).

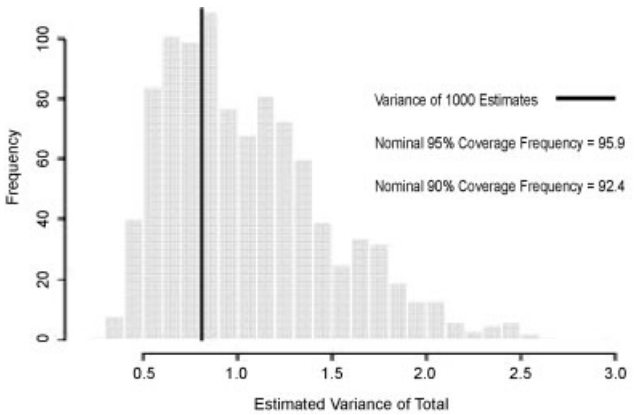

Figure 9. Histogram showing results of 1000 GRTS samples of size 50 from extensive population using Surface 1

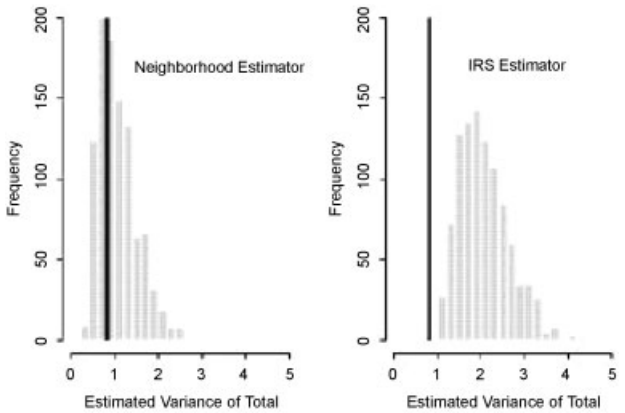

Figure 10. Comparison of neighborhood and IRS variance estimators. The histogram for the neighborhood estimator is the same as for Figure 8, scaled to match the IRS histogram

effort was used to delineate the reservoir boundaries. The National Elevation Database (NED) was used to create reservoir polygons. Polygons were edited and attributed using available navigation charts, topographic maps and local knowledge. Insufficient bathymetric data were available to create an explicit depth criteria but a variety of sources were used to eliminate upland extents of some bays to avoid probable dry, inaccessible or otherwise non-target areas. The shoreline was defined as the 510 m mean sea level contour 20 m above the normal pool elevation (490 m). The many bays on the reservoir were of particular interest, so they were individually delineated and categorized as small (0.5 to 1.45 km^2), medium (1.45 to 3.32 km^2) and large (3.32 to 8.0 km^2) bays. The remainder of the reservoir was designated as open water. The desired number of samples was 12 in each bay category and 15 in open water, for a total sample size of 51. The areas were 22.17 km^2, 50.56 km^2, 289.49 km^2 and 925.27 km^2 for small, medium, large and open water, respectively.

We defined a GRTS survey design with four density categories based on bays and open water. In addition to the desired sample size of 51 sites, an over sample of another 51 sites was included for a

Table 2. Lake Oahe population estimates from GRTS design. $\hat{V}_{NBH}$ and $\hat{V}_{IRS}$ are the mean of the neighborhood- and IRS-based estimators, respectively. Percent reduction is compared to IRS-based variance estimate

Response	Estimated mean	$\hat{V}_{IRS}$	$\hat{V}_{NBH}$	Percentage reduction
Chlorophyll-a (μg/L)	2.83	0.192	0.081	58.1
Temperature (°C)	23.37	0.077	0.035	55.0
Secchi depth (m)	2.61	0.121	0.073	22.5

total sample size of 102. After applying reverse hierarchical ordering, the base sample of 51 sites was taken from every other sample point on the ordered line, with the over sample being the remaining 51 sites. Initial weights (reciprocal of the inclusion probability) were 1.85, 4.24, 23.81 and 61.90 km^2 for small, medium, large and open water, respectively. While conducting fieldwork, four sites were found not to be in the reservoir target population due to frame inaccuracy. The sites were replaced by the first four sites on the over sample list, for a total of 55 sites included in the final sample of 4 non-target, 34 bay and 17 open water sites. We adjusted the weights to account for the use of the four over sample sites. Simple ratios of frame area to sum of weights within open water and all bays were applied to the initial weights. As expected, all non-target sites were located in bays, leading to the use of these two ratios.

Three response variables measured at each location were chlorophyll-a, water temperature at 2 m depth, and secchi depth (a water transparency measure). Table 2 gives the estimated population means, IRS Horvitz–Thompson variance estimates and neighborhood variance estimates. The neighborhood variance estimate ranges from 22.5 to 58.1 per cent reduction compared to the IRS variance estimate. Figure 11 displays the geographic spatial pattern for the observed data. In each case definite spatial patterns exist in the data. The neighborhood variance estimator takes advantage of the pattern, resulting in significantly lower variance estimates.

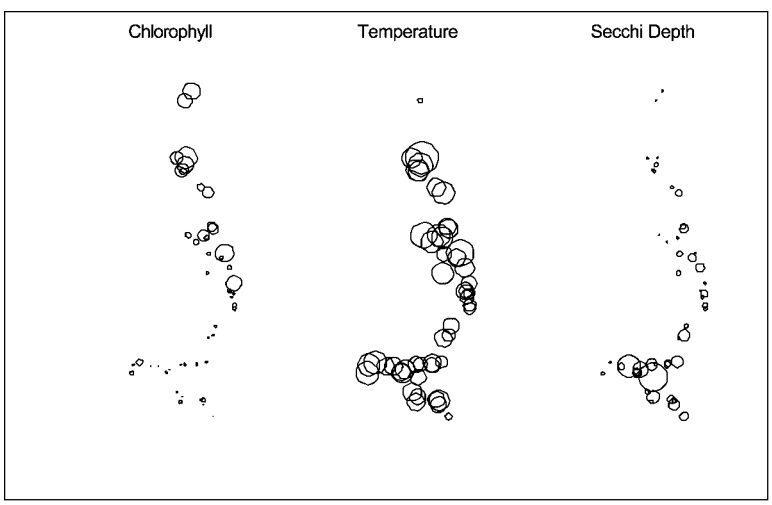

Figure 11. Spatial geographic pattern of relative magnitudes for chlorophyll, temperature and secchi depth in Oahe Reservoir. Symbol size is proportional to response magnitude

6. DISCUSSION

The neighborhood-based variance estimator performed quite well on all the test cases simulated. The estimator is approximately unbiased and, most importantly, the confidence interval coverage is very close to nominal. The distribution of the estimator tends to be somewhat skewed to the right, as is to be expected for a variance estimator, but does appear to be stable.

Although we have not explored the possibility, we expect that the estimator would also work very well for a randomly placed systematic design, or, indeed, for any design with an element of spatial balance, e.g. Breidt's (1995) Markov-chain design. The estimator draws on the smoothness of the local response surface; if the response has no spatial structure, then each local estimate of the total estimates the overall total, and thus, each local variance estimates the overall variance. Because of the constraints on the weighting function, the NBH variance estimator can be viewed as an average of several estimates of the same quantity. If, however, the local smoothness varies, then this will be reflected in higher variance estimates.

ACKNOWLEDGEMENTS

We wish to thank an anonymous reviewer for pointing out the Dunn and Harrison reference. The example data for the Oahe Reservoir were graciously provided by Dave Bolgrien, Billy Schweiger and Ted Angradi, who initiated the study and cooperated on its design. The research described in this article has been funded by the U.S. Environmental Protection Agency. This document has been prepared at the EPA National Health and Environmental Effects Research Laboratory, Western Ecology Division, in Corvallis, Oregon, through Contract 68-C6-0005 to Dynamac International, Inc., and Cooperative Agreement CR82-9096-01 to Oregon State University. It has been subjected to Agency review and approved for publication. The conclusions and opinions are solely those of the authors and are not necessarily the views of the Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES

Bellhouse DR. 1977. Some optimal designs for sampling in two dimensions. *Biometrika* **64**: 605–611.

Breidt FJ. 1995. Markov chain designs for one-per-stratum sampling. *Survey Methodology* **21**(1): 63–70.

Cochran WG. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics* **17**: 164–177.

Cochran WG. 1977. *Sampling Techniques*, 3rd edn. Wiley: New York.

Cordy C. 1993. An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. *Probability and Statistics Letters* **18**: 353–362.

Cotter J, Nealon J. 1987. *Area Frame Design for Agricultural Surveys*, U.S. Department of Agriculture, National Agricultural Statistics Service, Research and Applications Division, Area Frame Section.

Cressie N. 1991. *Statistics for Spatial Data*. Wiley: New York.

Dalenius T, Hájek J, Zubrzycki S. 1961. On plane sampling and related geometrical problems. *Proceedings of the 4th Berkeley Symposium on Probability and Mathematical Statistics* **1**: 125–150.

Dunn R, Harrison AR. 1993. Two-dimensional systematic sampling of land-use. *Applied Statistics* **42**: 585–601.

Gibson L, Lucas D. 1982. Spatial data processing using balanced ternary. *Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing*. IEEE Computer Society Press: Silver Springs, MD.

Gilbert RO. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold: New York; 320.

Herlihy AT, Larsen DP, Paulsen SG, Urquhart NS, Rosenbaum BJ. 2000. Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP mid-Atlantic pilot study. *Environmental Monitoring and Assessment* **63**(1): 95–113.

Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**: 663–685.

Larsen DP, Stevens DL Jr, Selle AR, Paulsen SG. 1991. Environmental Monitoring and Assessment Program, EMAP-Surface Waters: a northeast lakes pilot. *Lake and Reservoir Management* **7**(1): 1–11.

Madow WG. 1949. On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20**: 333–354.
Mark DM. 1990. Neighbor-based properties of some orderings of two-dimensional space. *Geographical Analysis* **2**: 145–157.
Munholland PL, Borkowski JJ. 1996. Simple Latin square sampling +1: a spatial design using quadrats. *Biometrics* **52**: 125–136.
Olea RA. 1984. Sampling design optimization for spatial functions. *Mathematical Geology* **16**: 369–392.
Overton WS, Stehman SV. 1993. Properties of designs for sampling continuous spatial resources from a triangular grid. *Communications in Statistics Part A: Theory and Methods* **22**: 2641–2660.
Peano G. 1890. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen* **36**: 157–160.
Rao CR, Mitra SK. 1971. *Generalized Inverse of Matrices and its Applications.* Wiley: New York.
Saalfeld A. 1991. Construction of spatially articulated list frames for household surveys. In *Proceedings of Statistics Canada Symposium 91, Spatial Issues in Statistics.* Statistics Canada: Ottawa, Ontario, Canada; 41–53.
Särndal C, Swensen B, Wretman J. 1992. *Model Assisted Survey Sampling.* Springer-Verlag: New York; 421–423.
Simmons GF. 1963. *Introduction to Topology and Modern Analysis*. McGraw-Hill: New York.
Stehman SV, Overton WS. 1994. Environmental sampling and monitoring. In *Handbook of Statistics*, Vol. 12, pp. 263–305, Patil GP, Rao CR (eds). Elsevier Science: Amsterdam, The Netherlands.
Stevens DL Jr. 1994. Implementation of a national environmental monitoring program. *Journal of Environmental Management* **42**: 1–29.
Stevens DL Jr. 1997. Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics* **8**: 167–195.
Stevens DL Jr, Olsen AR. 1999. Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics* **4**: 415–428.
Stevens DL Jr, Olsen AR. 2000. Spatially-restricted random sampling designs for design-based and model-based estimation. In *Accuracy 2000: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Delft University Press: The Netherlands; 609–616.
Wolter K. 1985. *Introduction to Variance Estimation.* Springer-Verlag. New York: 287.
Wolter KM, Harter RM. 1990. Sample maintenance based on Peano keys. In *Proceedings of the 1989 International Symposium: Analysis of Data in Time.* Statistics Canada: Ottawa, Ontario, Canada; 21–31.
Yates F. 1981. *Sampling Methods for Censuses and Surveys*, 4th edn. Griffin: London.
Yates F, Grundy PM. 1953. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* **B15**: 253–261.

