Variance Estimation for Subpopulation Parameters from Samples of Spatial Environmental Populations

Don L. Stevens, Jr., Thomas M. Kincaid, Dynamac Corporation
Don L. Stevens, Jr, Dynamac Corporation, 200 SW 35th Street, Corvallis, Oregon 97333

Key words: environmental sampling, domain estimation, continuous population sampling.

Abstract: A Randomized Tessellation Stratified (RTS) design selects a point sample from a spatial population by randomly locating a grid over a region containing the population domain and randomly selecting one point in each grid cell, retaining only those points that fall within the domain. In this paper, we investigate variance estimation for the RTS design for samples from arbitrary subpopulations, utilizing the continuous population form of the Horvitz-Thompson (HT) theorem. In particular, we explore the geometry of the variance near boundaries of geographically defined subpopulations, propose an easily computable, nearly unbiased modification of the Yates-Grundy (YG) variance estimator, and present simulation results to illustrate the modification, and contrast it with the HT estimator.

1. INTRODUCTION

This paper was presented at JSM '97 in the session "Surveying the Environment in the 21st Century: Advancing Theory and Addressing Institutional Constraints". We speak to this theme in our Introduction, and argue that the Multiple Density RTS design is ideally suited for satisfying institutional constraints on large-scale environmental monitoring. In the subsequent sections, we attempt to advance some of the theory and understanding of environmental sampling.

Many of the troublesome problems that beset the environment today, such as global warming, long-range transport of atmospheric pollutants, or habitat alteration, are not localized. Traditional environmental studies that focus on relatively small and well-delimited systems, such as the watershed of a second-order stream, may work well for understanding processes and quantifying rates, but cannot by themselves provide knowledge of status, condition, or change at regional or national scales. For example, an intensive study of the stream reaches in a watershed is of little use in estimating the number or proportion of stream-miles with healthy wild trout populations in the Pacific Northwest. Understanding and quantifying the extent of symptoms of widespread concerns requires broad-scale study efforts to address regional, continental, and global environmental issues. Thus, we see a need for environmental sampling and monitoring programs that span areas that range from a portion of a state to a continent or more. We believe the successful large-scale monitoring programs, i.e., the ones that are implemented and survive long enough to accomplish the program's goal and actually collect scientifically useful information, will share several characteristics. First, they will involve several federal, state, or local government agencies, industry, and a variety of special interest groups. In order to have political (and consequent financial) support, the programs will have to involve and accommodate all stakeholders. This leads to the second characteristic: the programs will have multiple objectives; will measure multiple resource types; and will measure multiple responses for each resource. Each interest group, agency, or industry will have its own agenda and needs, not always in concordance with one another. Finally, to ensure that the data collected does indeed satisfy program objectives, the programs will have a sound and rigorous statistical design that holds the various pieces together.

We note that an almost certain consequence of the first two characteristics is that a simple design approach, e.g., simple random sampling (SRS), will not suffice. There will be resources or sub-regions that demand special attention in the form of more intensive sampling, that is, more sample points per unit (length or area). The particular interest may stem from a scientific interest (the only place where a certain species occurs); stakeholder interest (a watershed supplying a town's drinking water); an environmental health issue (an area known to have toxic contamination); or a regulatory issue (permits for waste water discharge require sampling near the outfall). Whatever the source of the pressure, it is real, and the design must be able to accommodate it by allowing variable spatial density.

Sometimes, those special interest areas will be recognized at the time of the initial design, and can be accommodated at that time, e.g., by stratification. Occasionally (or often, depending on one's optimism), those "special" sub-populations will not be recognized at the time the sample is originally selected. Subpopulation analysis may be sufficient; however, we may also need to take additional samples in that subpopulation. Thus, we need a sampling design that can accommodate varying sample intensity, provide a means to selectively augment the sample after the fact, and allow arbitrary, post-design, specification of subpopulations for analysis.

We can foresee at least two distinct types of subpopulations being identified for post-design analysis. One we encounter frequently is a spatially defined subpopulation, e.g., phosphorus concentration of streams in the Texas blackland prairie ecoregion. The other arises in the multi-response nature of monitoring program, e.g., the subpopulation of all y-values defined by $x \in X_c$, where x and y are both sample measurements, and X_c is some criterion set. For example, we may want to split out the subpopulation of canopy density for forests growing on soils with pH < 7. Both types of subpopulation identify a spatial region occupied by the subpopulation, the first explicitly, the second implicitly. If we want our design to allow analysis for any arbitrarily defined subpopulation, we optimize our chances of having an acceptable sample size by making the original sample have an "even" spatial distribution. Now, any sample with a constant inclusion probability density has the property that the expected sample size in a region is proportional to the size of the region. In addition, we need the sample size variance to be small.

We will talk today about a design approach that we believe can be a positive aid in achieving successful large-scale, multi-objective, environmental sampling programs. We will describe the implementation of the design technique, and some of the practical aspects of the analysis of the resulting data.

2. Design Background

The general technique is described in detail in Stevens (1997), and we give only a short summary here. We cover the region with a randomly-located grid (we use a triangular grid) and locate one random point in each grid cell. This is the Randomized Tessellation Stratified (RTS) design (Bellhouse, 1977; Dalenius, Hájek, and Zubrzycki, 1961; Olea, 1984), and the two randomizations (grid location and point within cell) guarantee a non-zero joint inclusion probability for every pair of distinct points. The basic RTS design achieves even spatial spread, but does not provide the flexibility to vary spatial density. We gain that feature by utilizing nested point grids to extend the RTS design to a design with multiple spatial densities, called a Multiple-Density RTS (MD-RTS) design. Briefly, the technique depends on the observation that points can be added to a square or triangular point grid so that (1) the resulting collection of points remains a square or triangular grid, respectively, and (2) the original points form a subset of the higher density point grid. For example, Figure 1 shows the triangular grid resulting from a 7-fold increase in point density, with the original points denoted with "•" and the added points with "×". Each original point has been replaced by a cluster of 7 points, consisting of the original point plus 6 successor points.

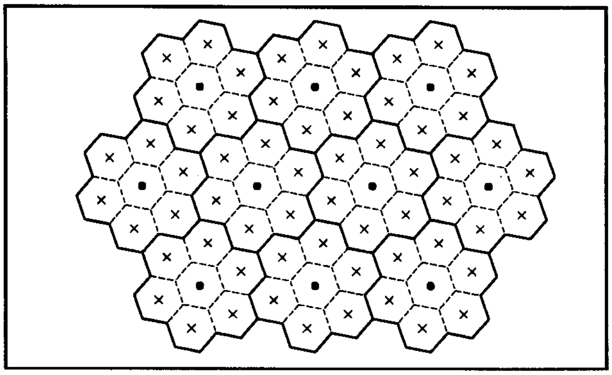

Figure 1. Triangular grid (•), 7-fold enhancement (×), with corresponding tessellations.

The natural tessellation of a grid is obtained by associating with each grid point the area closer to that point than to any other. For a triangular grid, the natural tessellation consists of hexagons centered on the grid points. This natural tessellation is not the only one, however. We can also create a tessellation for a particular grid by joining the natural tessellation polygons associated with the same point in an enhanced version of the grid. For example, if we join the seven hexagons of a point and its successors (the dashed lines in Figure 1) in a 7-fold enhancement of a triangular grid, we obtain a composite tessellation as shown by the solid lines in Figure 1.

These observations give us a means to vary the spatial density of a sample. We illustrate the ideas involved with a single example here. Suppose we have two regions, R_1 and R_2, and sample size requirements dictate a 7-fold increase in sample point density for R_2. Pick a universe that contains the union of R_1 and R_2, cover it with a triangular grid at a density suitable for R_1, and then create a 7-fold enhancement, keeping track of the original grid points and their successors. Let S_2 be an RTS design at the enhanced density. Let S_1 be a subsample selected by choosing one point at random from each group of seven associated with each original grid point. S_1 is an RTS samples relative to the composite tessellation. We then take our sample from R_i to be $R_i \cap S_i$, and the total sample to be $R_1 \cup R_2$.

3. Horvitz-Thompson Estimation for Continuous Populations

We consider the case of a response $z(s)$ defined on a region R that is a subset of a universe U, where our objective is to estimate the mean $\mu_z(R) = \int_R z(s)ds/|R|$ the distribution function $F_z(x) = \int_R I_{\{s|\, z(s) \le x\}}(s)ds/|R|$, where $|R|$ denotes the size (length, area, volume) of R and $I_A(x)$ is defined as $I_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$.

R may be defined and known at the time the sample is selected, but is often defined after the sample is drawn, and, in fact, we may need sample information to determine R. For example, suppose we sample bottom sediments in a large estuary, and want to analyze chemical concentration data for coarse-grained sediments separately from data on fine-grained sediments. We define R to be the subset of the estuary with coarse-grained sediments. We can unambiguously determine membership in R, but the boundaries of R are unknown when the sample was drawn, and in fact, may never be known completely. For a second example, R may be that portion of the estuary with depth less than 20 m, where we use an existing bathymetric map to determine depth. In this case, the complete boundary of R is available to us before we draw the sample.

A sampling design on U is specified by a joint distribution of n random variables, and a sample of a surface $z(s)$ over R is chosen by selecting a sample $S = \{s_1, s_2, ..., s_n\}$ of n locations from U with the distribution of S specified by a probability measure P_n on U^n. Assume that $f(s_1, s_2, ..., s_n)$, the joint probability density function (pdf) of the sample locations, $f_i(s)$, the marginal pdf of s_i, and $f_{ij}(s, t)$, the joint pdf of s_i and s_j, $i \neq j$, all exist. For $s \in U$, the inclusion density function is defined by $\pi(s) = \sum_{i=1}^{n} f_i(s)$, and the joint inclusion density function for $s, t \in U$ is defined by $\pi(s, t) = \sum_{i=1}^{n} \sum_{j \neq i}^{n} f_{ij}(s, t)$.

Horvitz and Thompson (1952) provided an estimator of the population total for variable-probability, without-replacement, finite-population sampling design, along with an expression for the variance of the estimated total and a related variance estimator. Alternative expressions for the variance and its estimator were provided by Yates and Grundy (1953) and Sen (1953).

Cordy (1993) showed that a version of the Horvitz-Thompson theorem holds when sampling from U when the inclusion density and pairwise inclusion density function are defined as above. The continuous version of the Horvitz-Thompson theorem provides an estimator of the total (integral) of z over R and associated variance estimators in terms of the functions $z(s)$, $\pi(s)$, and $\pi(s,t)$. An estimator of the mean is obtained by dividing estimated total by the size of R. As in the finite population case, the ratio estimator of the mean (also known as the Hájek estimator (Hájek, 1971; Thompson, 1992)), obtained by dividing by the estimated size of R, tends to be less variable and nearly unbiased. It is also well-suited to subpopulation estimation, as the size of R need not be known. The theorem is stated here for the continuous case; the finite population case is analogous. The continuous versions of both the Horvitz-Thompson variance (denoted HT) and the Yates-Grundy variance (denoted YG) are also given.

THEOREM: (Continuous Horvitz-Thompson): Let $s_1, s_2, ..., s_n$ be a sample selected from a universe U according to a design with inclusion function $\pi(s)$ and joint inclusion function $\pi(s, t)$, as defined above, 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 (approximately) unbiased estimator of $\mu_z(R)$ is given by $\hat{\mu}_z = \sum_{i=1}^{n} \frac{I_R(s_i) z(s_i)}{\pi(s_i)} \,/\, |\hat{R}|$, with variance

$$V_{HT}(\hat{\mu}_z) = \frac{1}{|\hat{R}|^2} \left\{ \int_R \frac{\Delta^2(s)}{\pi(s)} ds + \iint_{RR} \left[\frac{\pi(s, t) - \pi(s)\pi(t)}{\pi(s)\pi(t)} \right] \Delta(s)\Delta(t) dt ds \right\} \qquad (1)$$

or equivalently,

$$V_{YG}(\hat{\mu}_z) = \frac{1}{2|\hat{R}|^2} \iint_{UU} \left[\pi(s)\pi(t) - \pi(s, t) \right] \left[\frac{\Delta(s) I_R(s)}{\pi(s)} - \frac{\Delta(t) I_R(t)}{\pi(t)} \right]^2 dt ds \qquad (2)$$

where $\Delta(s) = z(s) - \hat{\mu}_z$ and $|\hat{R}| = \sum_{i=1}^{n} \frac{I_R(s_i)}{\pi(s_i)}$.

Corresponding estimators of variance are

$$\hat{V}_{HT}(\hat{\mu}_z) = \frac{1}{|\hat{R}|^2} \left\{ \sum_{s_i \in R} \frac{\Delta^2(s_i)}{\pi^2(s_i)} + \sum_{s_i \in R} \sum_{\substack{s_j \in R \\ j \neq i}} \left[\frac{\pi(s_i, s_j) - \pi(s_i)\pi(s_j)}{\pi(s_i, s_j)\pi(s_i)\pi(s_j)} \right] \Delta(s_i)\Delta(s_j) \right\} \qquad (3)$$

and

$$\hat{V}_{YG}(\hat{\mu}_z) = \frac{1}{|\hat{R}|^2} \left\{ \sum_{i=1}^{n} \sum_{j>i}^{n} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i, s_j)}{\pi(s_i, s_j)} \right] \left[\frac{\Delta(s_i) I_R(s_i)}{\pi(s_i)} - \frac{\Delta(s_j) I_R(s_j)}{\pi(s_j)} \right]^2 \right\}. \qquad (4)$$

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

4. The Geometry of Variance

In Section 2, we made the claim that an RTS design has the property of ensuring an even spatial distribution of sample points, in the sense that the achieved number of sample points in any arbitrary subregion of the universe will be nearly proportional to the size of the subregion. Of course, for an SRS sample, the *expected* number of sample points in a subregion is exactly proportional to the size of the subregion. The same is true for an RTS design, but in addition, the variance of the number of sample points will be small relative to the SRS variance. The following theorem, easily proved from the definitions, provides us with the means to investigate the variance in sample number.

Theorem 2: Let $S = \{s_1, s_2, ..., s_n\}$ be a sample from a continuous universe U with joint density function $f(s_1,s_2,...,s_n)$, marginal density functions $f_1(s)$, $f_2(s)$, ..., $f_n(s)$, and inclusion functions $\pi(s)$ and $\pi(s,t)$, and let $R \subset U$. Let $\tilde{n}_R = \sum_{i=1}^{n} I_R(s_i)$ be the number of samples that fall in R. Then $\bar{n}_R = E[\tilde{n}_R] = \int_R \pi(s)\, ds$ and

$$V(\tilde{n}_R) = \int_R\int_R \pi(s,t) ds dt - \bar{n}_R(\bar{n}_R - 1) \ .$$

We can contrast the behavior of an RTS and SRS designs by calculating $V(\tilde{n}_R)$ for a variety of regions. For an SRS, of course, $V(\tilde{n}_R)$ depends only on the area of the region relative to the area of the universe. For an RTS design, the variance in achieved sample size does depend on the geometry of the subregion, mostly through the perimeter to area ratio. We illustrate this with a variety of subregions of the unit square, ranging from simple geometric shapes to regions with complex irregular boundaries. The irregular boundary case included convex regions, ones with concavities, ones with long, narrow extensions, disconnected regions, and ones with holes. Specifically, the following seven regions were used: (1) Regular - a circular region with area 0.5, (2) Frag-2 - a region composed of two disconnected equal-area circles with a total area 0.5, (3) Frag-4 - a region composed of four disconnected equal-area circles with a total area 0.5, (4) Frag-8 - a region composed of eight disconnected equal-area circles with a total area 0.5, (5) Holes - a circular region with area 0.5 from which seven equal-area circles (holes) with a total area of 0.1 were removed, (6) Irregular - an irregularly-shaped polygon, and (7) L&N - an irregularly-shaped polygon with a long and narrow form. The seven regions are shown in Figure 2.

Table 1. Sample size variance for the regions given in Figure 2.

Region	$V_{SRS}(\hat{n}_R)$	$V_{RTS}(\hat{n}_R)$	P/A
Regular	25.00	4.02	5.02
Frag-2	25.00	5.57	7.08
Frag-4	25.00	7.93	10.02
Frag-8	25.00	10.78	14.18
Holes	23.99	8.79	13.68
Irregular	24.46	5.15	7.93
L & N	22.90	11.30	20.37

We took U to be the unit square, and used a grid size in the RTS design corresponding to a sample density of 100 points in the unit square. The comparisons to the SRS are made on the basis of a total sample size of 100, so that $\tilde{n}_R = 100 \times \frac{\textbf{Area of } R}{\textbf{Area of } U}$ for both the SRS and RTS. Table 1 gives $V(\tilde{n}_R)$ for the SRS and RTS, along with the perimeter to area ratio, for each of the seven regions. We note that even for the extreme cases of L & N and Frag-8, the sample size variance for the RTS is less than half that for the SRS.

We can gain some further insight into the impact of region boundaries by rewriting **(2)** as

$$V_{YG}(\hat{\mu}_z) = \int_R \left\{ \frac{1}{2} \int_R \left[\pi(s)\pi(t) - \pi(s,\ t)\right] \left[\frac{\Delta(s)}{\pi(s)} - \frac{\Delta(t)}{\pi(t)}\right]^2 ds \right. \qquad (5)$$

$$\left. + \left[\frac{\Delta(s)}{\pi(s)}\right]^2 \int_{U-R} \left[\pi(s)\pi(t) - \pi(s,t)\right] ds \right\} dt \ / \ | \ \hat{R} \ |^2$$

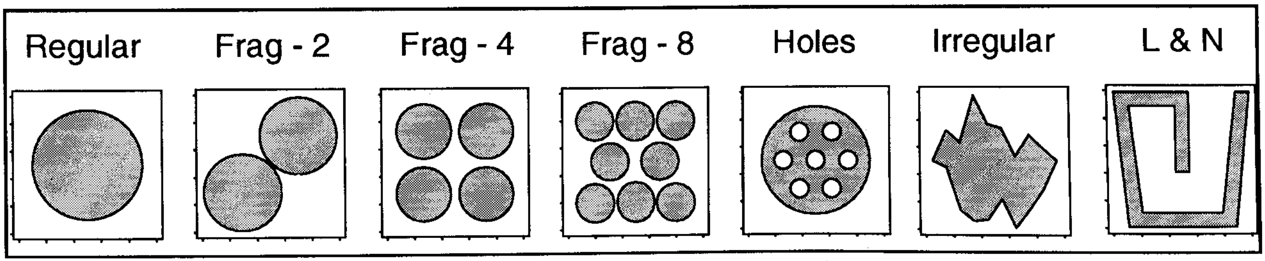

Figure 2. The seven regions used in the regional geometry and variance simulation studies.

For the RTS design, the joint inclusion function is

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

denotes the tessellation polygon centered on s. It follows that

$$\pi(s)\pi(t) - \pi(s,\ t) = \pi(s)\pi(t)\left(\frac{|\ C(s)\cap C(t)\ |}{|\ C\ |}\right) \geq 0 \text{,}$$

and hence that the inner integral is also always non-negative. Thus, we can express the YG variance as

$$V_{YG}(\hat{\mu}_z) = \int_R v(s)\,ds \quad \text{where}$$

$$v(s) = \left\{\frac{1}{2}\int_R \left[\pi(s)\pi(t) - \pi(s,\ t)\right]\left[\frac{\Delta(s)}{\pi(s)} - \frac{\Delta(t)}{\pi(t)}\right]^2 dt \right. \tag{6}$$

$$\left. + \left[\frac{\Delta(s)}{\pi(s)}\right]^2 \int_{U-R}\left[\pi(s)\pi(t) - \pi(s,t)\right]dt \right\} / \ |\ \hat{R}\ |^2 .$$

The function $v(s)$ can be interpreted as the contribution to the variance that arises at the point $s \in R$, so we can use $v(s)$ to examine the impact of regional geometry on variance. We have done this for the "regular" region for a very simple surface. We chose a tilted plane for the surface, because we did not want to confound the effect of the regional boundary with local surface variation. Figure 3 is a perspective plot of $v(s)$. Note that the interior of the region makes essentially no contribution to the variance; instead, all the variance arises from points near the boundary of R.

A similar sort of boundary effect is observed when the boundary is not a regional boundary, but a boundary where the sampling density changes. We illustrate with an annular region R_1 surrounding a circular region R_2 where the sample density is 7 times the sample density in R_1. In this case, of course, the $v(s)$ function is a bit more complicated, but can again be interpreted as representing the contribution to the variance that arises from the point s. Figure 4 is a perspective plot of $v(s)$, note the ridge that follows the boundary between R_1 and R_2.

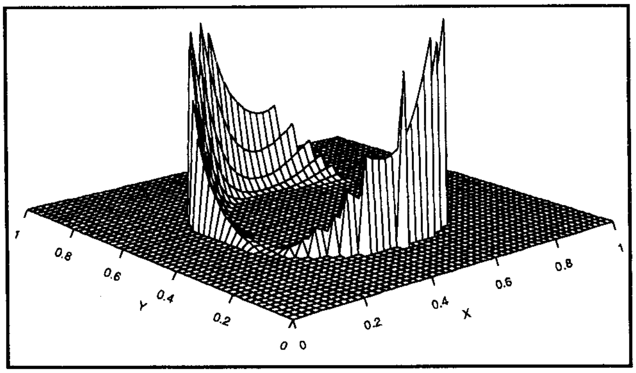

Figure 3. Perspective of $v(s)$ for an RTS sample from a circular region on tilted plane.

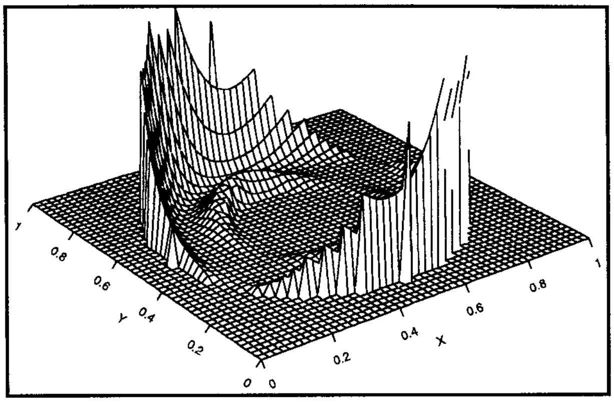

Figure 4. Perspective of $v(s)$ for an MD-RTS sample from an annular region enclosing a circular region on a tilted plane.

To summarize this section, the RTS design does achieve an even spatial distribution, so that it should be a good choice for a design where identification of important subpopulations is likely to occur post-design. However, there is a price to be paid: *every* population becomes a "post-design-specified" population, that is, we control the total number of points in the universe, which is unlikely to coincide with any population of intrinsic interest. Thus, even populations identified prior to design have random sample sizes. Furthermore, the region near a subpopulation boundary makes a substantial and disproportionate contribution to variance. Heuristically, the inflation of variance near boundary occurs because of local variation in sample density: average density over repeated samples is still constant, but the boundary induces variation on the scale of the tessellation polygon. Furthermore, the boundary effect manifests itself at the boundaries of designed density changes.

5. Variance Estimators for Domains and Sub-Populations

The design approach does result in well-dispersed sample points, and does permit great flexibility in varying sample point spatial density. Stevens (1997) showed that the joint inclusion functions are non-zero almost everywhere, so that, in theory, the HT and YG variance estimators are available and unbiased. However, in practice, the case is not so straightforward. The inclusion probability formulae are somewhat complicated. The joint inclusion probability obtained by treating the sample as an independent random sample (IRS) is computationally convenient, but tends to overestimate the variance. The HT variance estimator is unbiased, and is easily applied to subpopulations, but has an unfortunate tendency to yield negative estimates. The YG estimator is unbiased and is guaranteed to be positive for the RTS design, but is not easily applied to arbitrary subpopulations because it involves sums over the entire sample, and so cannot be computed from knowledge of only the sample points in *R*. However, we can rewrite the estimator as

$$\hat{V}_{YG}(\hat{\mu}_z) = \Bigg\{ \sum_{s_i \in R} \sum_{\substack{s_j \in R \\ j>i}} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i, s_j)}{\pi(s_i, s_j)} \right] \left[\frac{\Delta(s_i)}{\pi(s_i)} - \frac{\Delta(s_j)}{\pi(s_j)} \right]^2 \qquad (7)$$

$$+ \sum_{s_i \in R} \left[\frac{\Delta(s_i)}{\pi(s_i)} \right]^2 \sum_{s_j \notin R} \left[\frac{\pi(s_i)\pi(s_j) - \pi(s_i, s_j)}{\pi(s_i, s_j)} \right] \Bigg\} / |\hat{R}|^2$$

Thus, we do not need the response outside of *R*, but only the location and inclusion probability functions for sample points outside *R*.

To derive other variance estimators, rewrite (2) as

$$V_{YG}(\hat{\mu}_z) = \Bigg\{ \frac{1}{2} \int_R \int_R \left[\pi(s)\pi(t) - \pi(s, t)\right] \left[\frac{\Delta(s)}{\pi(s)} - \frac{\Delta(t)}{\pi(t)} \right]^2 dt\,ds \qquad (8)$$

$$+ \int_R \left[\frac{\Delta(s)}{\pi(s)} \right]^2 \int_{U-R} \left[\pi(s)\pi(t) - \pi(s,t)\right] ds\,dt \Bigg\} / |\hat{R}|^2$$

The first term in (8) depends entirely on points in *R*, and will be denoted $V_{YG|R}$, while the second, denoted $V_{YG|\tilde{R}}$ depends on points in $\tilde{R} = U - R$. From this perspective, (7) can be viewed as the sum of two estimators, say $\hat{V}_{YG|R}$ of $V_{YG|R}$ and $\hat{V}_{YG|\tilde{R}}$ of $V_{YG|\tilde{R}}$. In the following, we develop an alternative estimator for $V_{YG|\tilde{R}}$.

First, note that

$$V_{YG|\tilde{R}} = \int_R \left[\frac{\Delta(s)}{\pi(s)} \right]^2 g(s)ds \,/\, |\hat{R}|^2 \ , \qquad \text{where}$$

$g(s) = \int_{U-R} \left[\pi(s)\pi(t) - \pi(s, t)\right] dt$ depends only on the design and *R*. Presumably, π(•) and π(•, •) are known (or knowable), so that if the boundary of *R* is also known, then *g(s)* can be computed for any point $s \in R$. We can then estimate $V_{YG|\tilde{R}}$ by applying the HT theorem to yield $\hat{V}_{YG|\tilde{R}} = \sum_{j \in R} \frac{\Delta(s_j)^2 g(s_j)}{\pi(s_j)^3} \,/\, |\hat{R}|^2$. The practical difficulty in applying this observation is the computation of *g(•)* for each sample point $s_j \in R$. In those cases where sample information is used to determine membership in *R*, exact computation of *g(s)* is not even feasible. However, *g(s)* has a simple, intuitive interpretation that leads to a feasible approximation. Using the definitions of inclusion and joint inclusion probability, we have that

$$\begin{aligned} g(s) &= \int_{U-R} \left[\pi(s)\pi(t) - \pi(s, t)\right] dt \\ &= \int_U \left[\pi(s)\pi(t) - \pi(s,t)\right] dt - \int_R \left[\pi(s)\pi(t) - \pi(s,t)\right] dt \\ &= n\pi(s) - (n-1)\pi(s) - \bar{n}_R \pi(s) + \int_R \pi(s, t)dt \qquad (9) \\ &= \int_R \pi(s, t)dt - \pi(s)(\bar{n}_R - 1) \ . \end{aligned}$$

Set $\pi(t \mid s) = \pi(s, t)/\pi(s)$, so that $\pi(t \mid s)$ is the conditional inclusion density at *t* given a sample point at *s*. Then

$$\int_R \pi(s, t)dt = \pi(s) \int_R \pi(t \mid s)dt = \pi(s)[\bar{n}_R(s) - 1] \quad ,$$

where $\bar{n}_R(s)$ is the expected number of sample points in *R* given a sample point at *s*. Substituting this result into (9), we have that

$$g(s) = \pi(s)[\bar{n}_R(s) - \bar{n}_R] = \pi(s)\Delta\bar{n}_R(s). \qquad (10)$$

As for *g(s)*, $\Delta\bar{n}_R(s)$ is computable given π(•) and π(•,•) and the boundary of *R*, but all of the previous objections regarding *g(s)* apply. However, in the case of an RTS design, we can get a crude but easily computable approximation to $\Delta\bar{n}_R(s)$ with relative ease. The approximation is based on two observations: first, $\Delta\bar{n}_R(s) = 0$ for *s* in the "interior" of *R*, that is, if a grid cell centered on $s \in R$ does not intersect a grid cell centered anywhere on the boundary of *R*, and second, because of the tight control over "local" spatial density of the RTS design, each point in the interior of *R* is guaranteed to have neighbors that are not "too far" away, where the precise meaning of "too far" depends on the geometry of the underlying grid. The surrogate for $\Delta\bar{n}_R(s)$ that we propose is based on counting neighboring sample points in *R* for each sample point in *R*, and comparing that number to the expected number of neighbors. The approximation takes the form

$$\Delta\bar{n}_R(s) = \max\left(0,\ 1 - \frac{\textbf{achieved number of neighbors}}{\textbf{expected number of neighbors}} \right) .$$

The hope, of course, was that points with fewer than expected neighbors were in that condition because their neighbors fell outside of *R*.

The definition of "neighbor" that we use is based on the observation that, for the RTS design, $C(s) \cap C(t) = \emptyset$ implies $\pi(s, t) = \pi(s)\pi(t)$, where *C*(*x*) is a grid-cell shaped polygon centered on *x*, so that a neighbor of *s* is any point *t* such that $\pi(s, t) \neq \pi(s)\pi(t)$. Geometrically, this is any point *t* in a polygon centered on *s*, similar to *C* but with four times the area. We will refer to this estimator as YG-Neighbor (YG-N).

6. Simulation Study Design

We investigated the behavior of these approximations and the resultant variance estimators by simulating random surfaces on irregularly-shaped regions. The random surfaces were generated using two distinct procedures. The first procedure produced a surface by interpolating between points that were uniformly distributed over the unit square, with values independent N(0,1) variates. By varying the number of points in the unit square, we can control the spatial variability of the resulting surface, with more random points giving a surface with more rapid changes. Two of the surfaces utilized in the simulations were created using the first procedure. One surface, which will be referenced as Surface I, was produced using 20 points in the unit square. The other surface, which will be referenced as Surface II, was produced using 100 points in the unit square. The second procedure was designed to produce a surface with essential no spatial structure. A regular square grid composed of a large (greater than 10,000) number of grid points was employed to cover the unit square. A N(0,1) variate was generated independently for each grid point, and the variate value was assigned to all points in the tessellation polygon (square cell) associated with the grid point. Thus, the surface was composed of a large number of flat plates. The third surface utilized in the simulations, which will be referenced as Surface III, was created using the second procedure. The three surfaces are shown in Figure 5.

The rationale for using a design that controls the spatial dispersion of the sample points is to exploit the structure of the surface in an attempt to increase precision. Thus, we expect the RTS design to have less variance than an IRS design, and anticipate that the IRS variance approximation to the RTS variance will overstate the variance for Surfaces I and II. Surface III has essentially no structure to exploit, so that an RTS design should have roughly the same behavior as an IRS design, making the IRS variance approximation appropriate.

The cumulative distribution function (cdf) was used to evaluate performance of the variance estimators. For each combination of surface and region, a set of points that covered the range of possible sample values was determined. The true cdf for the surface was calculated for this set of points. A total of 1,000 replicates were obtained for each combination of surface and region. For each replicate a sample was selected from the region, and estimates of the true cdf and variance estimators were calculated for each of the points. For each variance estimator composite estimates of variance were obtained by calculating the sample mean of the 1,000 variance estimates at each point in the cdf. Estimates of the true variance of the cdf were obtained by calculating the sample variance of the 1,000 cdf estimates at each of the points. Two measures were used to evaluate performance of the variance estimators: (1) relative bias and (2) 90% confidence interval coverage. Bias was estimated as the difference between the mean of the variance estimates and the estimated true variance of the cdf. Relative bias was obtained by dividing the bias estimate by the estimated true variance of the cdf. A relative bias of 1 indicates a variance estimate inflated by a factor of 2. Coverage was taken as the proportion of the confidence intervals that included the true cdf value, where the confidence intervals were obtained using Normal distribution theory. Finally, the sampling plan employed in the simulations used the same sampling density for the entire region. The expected size of the sample in the region was 30.

7. Simulation Study Results and Discussion

Performance of each of the variance estimators will be discussed in this section. Due to the relatively small

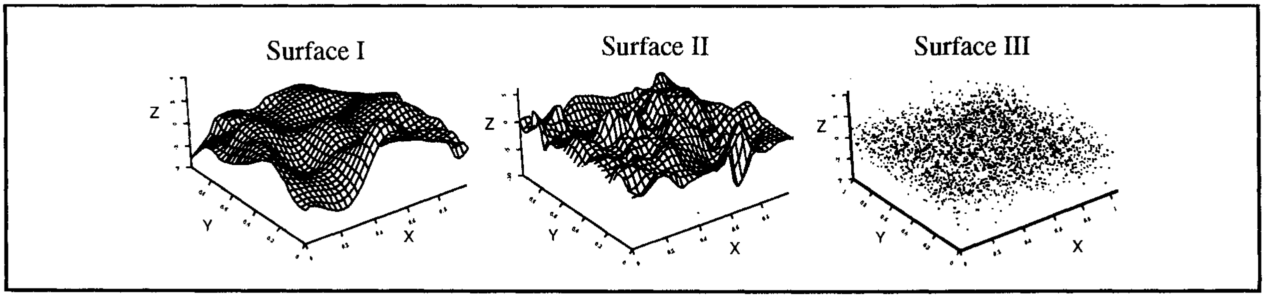

Figure 5. The three surfaces used in variance simulation.

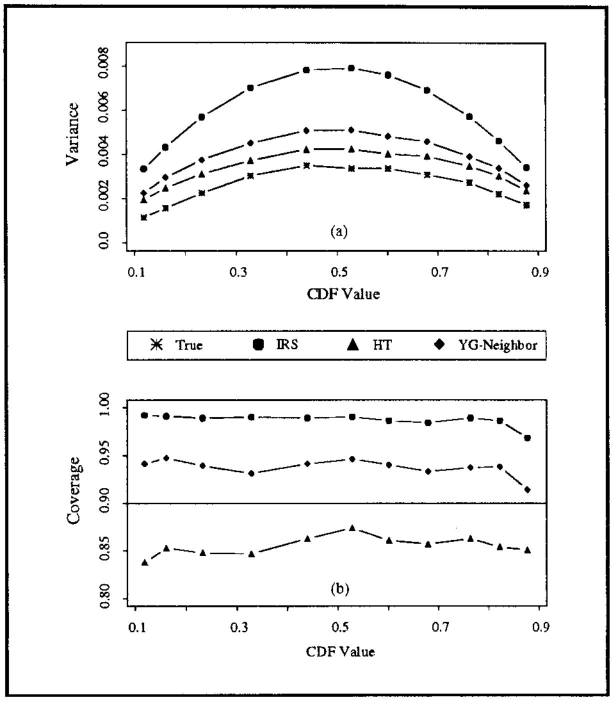

Figure 6. Variance and coverage of variance estimators for Surface I.

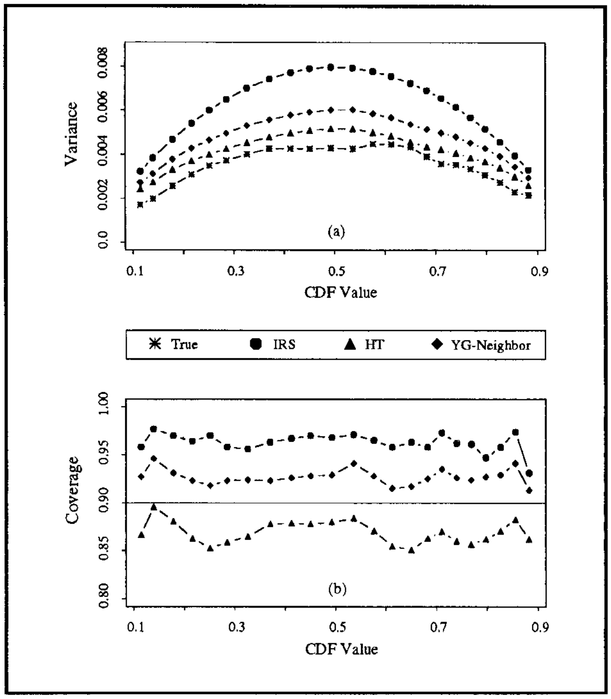

Figure 7. Variance and coverage of variance estimators for Surface II

sample sizes used in the simulations, all of the estimators produced poor performance in the tails of the cdf. Thus, performance of the estimators will be evaluated only for values of the cdf between 0.1 and 0.9, inclusive.

Typical simulation results are illustrated in Figures 6 through 8 for the three surfaces, respectively. The figures show results for region Regular. In each figure part (a) is a plot of the estimated true variance of the cdf and the variance estimators evaluated at each point in the cdf, and part (b) is a plot of coverage of the 90% confidence intervals for the variance estimators evaluated at each point in the cdf.

Results were similar for Surface I and Surface II (Figures 6 and 7) in comparison to Surface III (Figure 8). For Surfaces I and II the IRS estimator produced extensive overestimation of the true variance for most points, whereas the HT and YG-N estimators produced slight to moderate overestimation of the true variance. For Surface III the IRS and HT estimators produced slight overestimation of the true variance, and the YG-N estimator produced slight to moderate overestimation of the variance for most points in the cdf. Confidence interval coverage for the surfaces mirrored the variance results. For Surfaces I and II coverage was much greater than the nominal value for the IRS estimator, slightly greater than the nominal value for the YG-N estimator, and less than the nominal value for the HT estimator. For Surface III coverage for all three of the estimators was close to the nominal value. Boxplots of the variance estimates for the HT and YG-N estimators for the 1,000 samples from Surface I and region Regular are presented in Figure 9. As expected, the distribution for the HT estimator was skewed to the left and included negative variance estimates. The distribution for the YG-N estimator was skewed to the right, but the variance estimates were always positive. The skewed distribution for the HT estimator provided insight into the reason that coverage for the HT estimator was less than the nominal value even when the estimator was positively biased (see Figures 6 and 7).

Further evaluation of the estimators will utilize means of the relative bias and 90% confidence interval coverage estimates for cdf values between 0.1 and 0.9, inclusive. Means of the relative bias estimates are provided in Table 2 and means of the 90% confidence interval coverage estimates are provided in Table 3.

As expected, the IRS variance estimator performed poorly for Surfaces I and II and performed very well for Surface III. Mean relative bias of the IRS variance estimates indicated extensive overestimation for Surfaces I and II and slight overestimation for Surface III (Table 2). Mean coverage for the IRS estimator was very conservative for Surfaces I and II, i.e., coverage values often were close to 1.0 (Table 3). For Surface III mean coverage was marginally greater than the nominal value.

The HT variance estimator was expected to be unbiased but very unstable. Negative estimates,

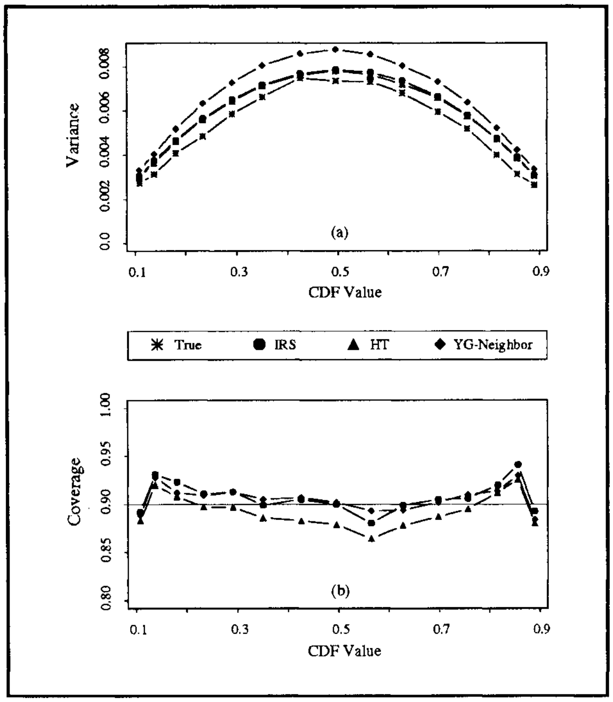

Figure 8. Variance and coverage of variance estimators for Surface III.

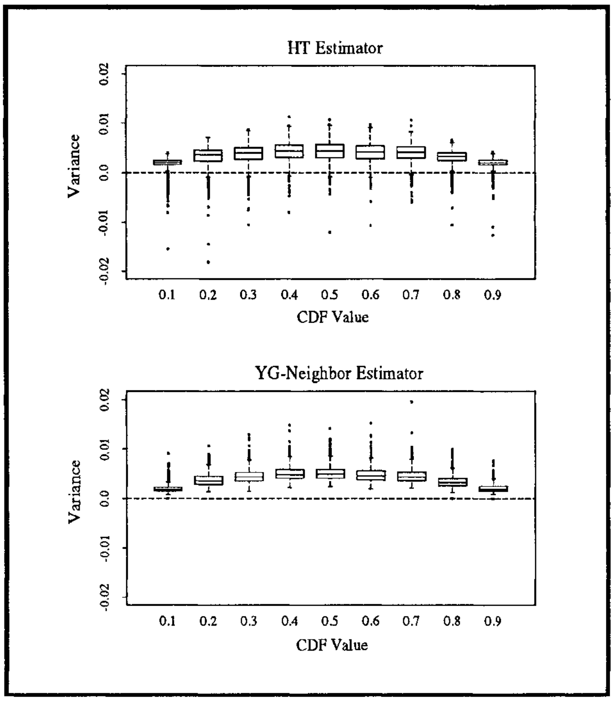

Figure 9. Boxplots of HT and YG-N estimators for Surface I and Regular region.

sometimes very negative, are common. These estimates are a foreseeable consequence of the joint inclusion function for the RTS design, which lets points get close together with vanishingly small probability. Points that are close together then have large leverage on the HT variance estimator and can easily cause it to be negative. Mean relative bias results for the HT estimator indicated that the estimator was positively biased for all three surfaces (Table 2). Positive bias for the HT estimator was not anticipated. Recall that the ratio estimator for variance is only approximately unbiased. For the surfaces, regions, and sample size employed in the simulations, the approximation seems to be less than optimal for the HT estimator. Additional simulations indicated that the amount of bias for the HT estimator decreased as the sample size increased. Confidenceinterval coverage for the HT estimator consistently was less than the nominal value for Surfaces I and II and very close to the nominal value for Surface III (Table 3).

The YG-N estimator was stable, never negative, and much closer to the true variance than the IRS estimator. Mean relative bias values indicated that the YG-N estimator moderately overestimated the true variance for all three surfaces (Table 2). Confidence interval coverage for the YG-N estimator was moderately greater than the nominal value for Surfaces I and II and slightly greater than the nominal value for Surface III (Table 3).

Regarding the three surfaces, mean relative bias decreased for all of the estimators as the smoothness of

Table 2. Mean of the relative bias of the IRS, HT and YG-N estimators evaluated at all values of the cdf between 0.1 and 0.9 for the two surfaces and the seven regions.

Region	Surface I			Surface II			Surface III		
	IRS	HT	YG-N	IRS	HT	YG-N	IRS	HT	YG-N
Regular	1.345	0.349	0.580	0.762	0.193	0.384	0.127	0.115	0.243
Frag-2	1.335	0.378	0.588	0.845	0.262	0.471	0.174	0.168	0.291
Frag-4	1.088	0.344	0.547	0.834	0.249	0.454	0.169	0.155	0.304
Frag-8	1.074	0.314	0.521	0.783	0.274	0.446	0.116	0.098	0.229
Holes	1.204	0.332	0.586	0.827	0.229	0.475	0.125	0.105	0.254
Irregular	1.511	0.385	0.643	0.676	0.193	0.370	0.203	0.203	0.337
L&N	1.374	0.371	0.736	0.591	0.170	0.408	0.140	0.114	0.301

Table 3. Mean of the ninety percent confidence interval coverage of the IRS, HT and YG-N estimators evaluated at values of the cdf between 0.1 and 0.9 for the three surfaces and the seven regions.

Region	Surface I			Surface II			Surface III		
	IRS	HT	YG-N	IRS	HT	YG-N	IRS	HT	YG-N
Regular	0.987	0.855	0.937	0.963	0.869	0.927	0.908	0.869	0.927
Frag-2	0.983	0.862	0.937	0.968	0.870	0.933	0.914	0.902	0.913
Frag-4	0.975	0.867	0.936	0.967	0.876	0.931	0.913	0.897	0.912
Frag-8	0.973	0.859	0.934	0.964	0.882	0.933	0.906	0.890	0.903
Holes	0.980	0.860	0.935	0.967	0.867	0.936	0.906	0.893	0.907
Irregular	0.985	0.851	0.935	0.960	0.870	0.924	0.917	0.909	0.919
L&N	0.984	0.862	0.948	0.952	0.863	0.931	0.912	0.897	0.919

the surface decreased (Table 2). Similarly, for all three estimators, mean confidence interval coverage became closer to the nominal value as the smoothness of the surface decreased, which means that mean confidence interval coverage decreased for the IRS and YG-N estimators and increased for the HT estimator (Table 3). For each combination of surface and estimator, results were consistent among the seven regions, i.e., for a given surface the type of region had little impact on performance of the estimator. This consistency occurred for both mean relative bias and mean confidence interval coverage. Increasing the fragmentation of the region produced no consistent pattern regarding bias or coverage. Similarly, there was no consistent pattern of difference in bias or coverage between the regular region and the irregular regions or the region with holes.

Among the three estimators examined in the simulations, the YG-N estimator produced the best overall performance. Although the YG-N estimator was positively biased due to the skewed distribution of the estimates, coverage for the estimator was close to the nominal value for all cases. The IRS estimator produced extensive overestimation of the true variance and conservative confidence interval coverage for Surfaces I and II. Although the HT estimator was positively biased for Surfaces I and II, confidence interval coverage was less than the nominal value. For Surface III the IRS and HT estimators produced excellent performance, but the YG-N estimator also performed very well for that surface.

References

Bellhouse, D.R. (1977). 'Some optimal designs for sampling in two dimensions'. *Biometrika* 64, 605–611.

Cordy, C. (1993). 'An extension of the Horvitz-Thompson theorem to point sampling from acontinuous universe'. *Probability and Statistics Letters* **18**, 353–362.

Dalenius, T., J. Hájek, and S. Zubrzycki. (1961). 'On plane sampling and related geometrical problems'. *Proceedings of the 4th Berkeley Symposium on Probability and Mathematical Statistics* **1**, 125–150.

Hájek, J. (1971). 'Comment on a paper by D. Basu. In: Godambe, V. P., and Sprott, D. A. (eds.) *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston, p. 236.

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

Olea, R.A. (1984). 'Sampling design optimization for spatial functions'. *Mathematical Geology* **16**, 369–392.

Overton, W.S. and S.V. Stehman. (1993). 'Properties of designs for sampling continuous spatial resources from a triangular grid'. *Communications in Statistics Part A: Theory and Methods*, **22**, 2641–2660.

Sen, A.R. (1953). 'On the estimate of the variance in sampling with varying probabilities'. *Journal of the Indian Society of Agricultural Statistics* **7**, 119–127.

Stehman, S.V. and W.S. Overton. (1994). 'Environmental sampling and monitoring'. In *Handbook of Statistics* **12**, eds. G.P Patil and C.R. Rao, 263–305. Amsterdam, The Netherlands: Elsevier Science.

Stevens, D. L., Jr. (1994). 'Implementation of a national environmental monitoring program'. *Journal of Environmental Management* **42**, 1–29.

Stevens, D. L., Jr. (1997). 'Variable density grid-based sampling designs for continuous spatial populations'. *Environmetrics* **8**, 167-197.

Thompson, S.K. (1992). *Sampling*. New York: John Wiley & Sons.

Yates, F. and P.M. Grundy. (1953). 'Selection without replacement from within strata with probability proportional to size'. *Journal of the Royal Statistical Society* **B15**, 253–261.