ENVIRONMETRICS
Environmetrics 2006; **17**: 539–553
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.763

Sampling and variance estimation on continuous domains

Cynthia Cooper*,†

Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, Oregon 97330, USA

SUMMARY

This paper explores fundamental concepts of design- and model-based approaches to sampling and estimation for a response defined on a continuous domain. The paper discusses the concepts in design-based methods as applied in a continuous domain, the meaning of model-based sampling, and the interpretation of the design-based variance of a model-based estimate. A model-assisted variance estimator is examined for circumstances for which a direct design-based estimator may be inadequate or not available. The alternative model-assisted variance estimator is demonstrated in simulations on a realization of a response generated by a process with exponential covariance structure. The empirical results demonstrate that the model-assisted variance estimator is less biased and more efficient than Horvitz–Thompson and Yates–Grundy variance estimators applied to a continuous-domain response. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: continuous-domain sampling; design-based variance estimation; sample-process variation; kriging; inclusion densities

1. INTRODUCTION

A basic job in resource management is to quantify 'how much' there is of a response (resource) that varies over a continuous domain. Applications of resource management include wildlife, fisheries, and forestry management. Resources may be monitored to assess condition, such as soil contamination. Exploitation of geologic resources requires assessment of average response at unobserved sites. In some applications, model-based methods have historically been employed nearly exclusively of design-based approaches. Design-based approaches address the goal of quantifying 'how much', with the advantage that there is no need to defend a choice of distribution or covariance model.

Design-based and model-based methodologies of sampling and estimation have historically been developed in separate fields of expertise. There are differences in the bases of inference between the two. The contexts under which the two approaches were developed differ. The objectives of the two approaches differ in emphasis. A fundamental part of any estimation job is quantifying the uncertainty (or, conversely, the precision) of the estimate. The interpretation of the uncertainty or variability

*Correspondence to: C. Cooper, Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, OR 97330, USA.
†E-mail: cooper@science.oregonstate.edu

Contract/grant sponsor: U.S. Environmental Protection Agency; contract/grant number: CR82-9096-01.
Contract/grant sponsor: National Research Program.

Received 29 October 2004
Accepted 14 July 2005

depends on the approach (design- or model-based). The variability of an estimator is the result of the estimator being a function of random variables, thus, an estimator is itself a random variable with a distribution. In design-based estimation, the random variables in the estimators are the indicator functions of whether an element in the domain is or is not included in the sample. Uncertainty is based on variability of the estimator due to the sampling process. In model-based predictions, the response on the domain is regarded as random, and uncertainty of the estimator involves some intrinsic covariance structure that characterizes the behavior of the response. The application of the two approaches is compared in the following two examples—one on monitoring Coho salmon, the other on assessing bird species diversity.

The Oregon Department of Fish and Wildlife (ODFW) is following a design-based sampling protocol to monitor Oregon-Coast-Natural Coho salmon (*Oncorhynchus kisutch*) population status and trend. The sampling domain is a network of continuous stream segments. One response observed is the number of spawners in a mile. The response is treated as non-random for determining estimates. The design strategy incorporates an augmented rotating panel design, developed by EPA, such that some sites are visited repeatedly at different intervals over time to monitor trends (ODFW, 2002). A panel is the set of sites visited in the same years. There are 40 panels of varying frequencies of visits.

Within each panel, the sites are spatially balanced to help make the sample of stream segment locations representative of the stream-network domain. The density of sampled stream location guards against small-sample risk of unusual, non-representative samples. Variability of estimates is also controlled by reducing the chances of including pairs of elements with closely correlated responses, accomplished in the ODFW sample by spatially balanced sampling within panels.

Resource managers use the estimated totals and trends for setting harvest policies, and advising policy makers on land-use management. The absence of model specification benefits applications like this one, where policies must withstand stakeholders' possible challenges (Hansen *et al*., 1983).

For applications where a resource is to be assessed in an area with little or no direct observations, modeling a resource's covariance structure is usefully applied. There are many applications in geosciences. The model-based process of kriging predicts a response from a weighted average of observed responses, giving greater influence to those expected to have stronger correlation with the response to be predicted. Carroll (1998) describes an application extending mean and spatial covariance structure models to include abiotic factors to predict bird species diversity on the Indian subcontinent. He demonstrates the improved predictive capability of the universal kriging model with the extended covariance structure. The motivation for the study comes from resource assessment needs where ecological and environmental status is costly to assess and/or is required in areas difficult to access.

Ver Hoef (2002) compares the application of design-based estimators and a modification of block kriging where he treats the domain as a finite population of grid cells, for estimating population totals. He observes that the confidence intervals resulting from block kriging are between 20% and 40% narrower than those produced by design-based estimates applied to stratified samples on a spatial domain. This suggests a gain in efficiency from exploiting covariance structure of the response's underlying random process, although interpretation of the confidence interval depends on the approach. The uncertainty of the model-based approach addresses random variability of the response given its covariance structure, whereas the uncertainty of the design-based approach is derived from the sample process (the estimator varies because the elements from sample to sample vary).

The benefit of model-based concepts has not been fully employed to quantify sample-process variation of estimates, though there is sometimes good reason to do so. Cordy and Thompson (1995) employ the 'deterministic' covariance in a design-based variance estimator, treating the response as a

fixed surface. This paper promotes a model-assisted variance estimator for quantifying the variation due to the sampling process that is of interest in design-based sampling and estimation. The alternative estimator models sample process variance on a continuous domain, taking into account the covariance of the response.

The paragraphs below address, in order, design-based methodology, model-based methodology, idiosyncrasies of sampling and estimation on continuous domains, variance characterized and estimated by design-based methods, and variance characterized and estimated in model-based methods. Following this background material, an alternative model-assisted variance estimator is described for grid-based stratified sampling designs. The empirical behavior of the alternative model-assisted variance estimator is demonstrated in design- and model-based contexts on simulated random fields. The interpretation of sampling process variation for circumstances involving model-based approaches is discussed.

2. COMPARING THE APPROACHES

2.1. Design-based methodology

Design-based methodology was developed in survey methodology, where the applications are nearly entirely on finite populations. Typically the objective is to estimate the total or average of a population (or subpopulation) response.

Obtaining an unbiased estimate is desirable. If an adequate frame exists that effectively enumerates the elements of a population, a random sample implemented by sampling from the frame ensures that the expectation of nominally unbiased design-based estimates—with respect to the sampling process—is the population total or average. Throughout this paper, the term 'sample' refers to the collection of elements or units observed. Sources of bias include frame error, non-random sampling, and 'non-response' or unobserved elements that were meant to be included in the random sample. Non-response, frame error, other sources of bias, and how to adjust for these are not addressed in this study.

In the design-based paradigm, the elements' responses are treated as fixed and are assumed to be observed without error. In design-based inference, the variability of an estimator is induced by the variability in the elements that get sampled from a population or continuous domain. Since the practitioner has control over the sampling process (at least in terms of design, if not in implementation), the properties of the estimators are known exactly. That is, estimates are derived without being obliged to assume a distribution or covariance structure on the population responses.

Estimators are based on scaling the responses of sampled elements to extrapolate from the sample to the entire population. The Horvitz–Thompson (HT) estimator (Horvitz and Thompson, 1952) is a linear combination of elements, weighted by the inverse of their inclusion probabilities. The inclusion probability of an element for finite populations is the sum of the probabilities of all samples that include that element. On a continuous domain, the weight is the inverse of the inclusion density (ID), where the ID is the integral over the measures of samples that include the ith element (see Cordy, 1993).

If every population element has non-zero inclusion probability, the HT estimator is unbiased. The HT estimator provides a design-based estimator that accommodates unequal inclusion probabilities, for applications where some subpopulations are to be sampled more intensely than others.

Because the variability is defined in terms of the sampling-process variance, the variance estimators are based on the variance and covariance of the selection of a pair of elements into a sample. In

practice, only one sample is taken, yet the variability of the estimator is characterized in terms of the variation from sample to sample (referred to here as sampling-process variance). On a finite-population domain, the pair-wise inclusion probability is defined as the sum of probabilities, over the sample universe, that a sample contains both elements in a pair. The HT variance estimator weights the sum of squared- and cross-product responses by the inverse marginal and pair-wise inclusion probabilities. Assuming non-zero pair-wise IDs almost everywhere (a.e.), it is unbiased.

For most interesting applications, there is some hierarchical structure or ordering to the population, and units' responses within a level often are correlated (though not all characteristics observed on each unit need be correlated). The correlation is important to quantifying the variability the practitioner would observe over repeated samples. This is visited again in the section on design-based variance estimation. Knowing something about how the responses between elements are correlated can be useful to design optimal sampling strategies. Statisticians have employed models of correlation structures on populations to compare efficiency of different sampling strategies. Cochran (1946) modeled a finite population ordered in one dimension to show optimality of systematic sampling. His results were extended by others, among them Bellhouse (1977), for finite populations ordered in two dimensions.

Sampling may also be restricted to effect a representative sample in order to reduce bias (Royall and Cumberland, 1981, Royall, 1988) or to achieve a numerically well-conditioned system of equations to provide stable estimation of parameters or coefficients (see e.g. Rawlings *et. al*., 1998). In these contexts, some underlying models are being considered prior to sampling in order to anticipate what sample characteristics will be most useful to the parameter estimation process. The restricted sampling changes the distribution of the samples, which would impact inclusion probabilities derived from their probabilities (or the inclusion densities derived from their measures on continuous domains). A practitioner would want to consider if the restricted subset of samples is leaving out some part of the population that could cause bias in estimators.

2.2. *Model-based methodology*

Model-based inference applies a model of the response as an outcome (a.k.a. realization) of a random process. The random process is characterized by a distribution of the random component that has some covariance structure. For example, the covariance between two points may decay exponentially with distance, with rate of decay characterized by the range parameter. Typically there is a systematic component to the response, which modulates the mean in the distribution of the response and which may also be characterized by a parameterized model. Assuming a particular model, the preliminary objective of model-based work is to estimate model parameters including those of a covariance function that describes the stochastic behavior of the response. Typically the ultimate objective is to predict unobserved elements based on the model and conditional on the responses of the observed elements.

If the stochastic behavior is well characterized by some distribution or covariance structure, model-based estimation can be more efficient than design-based estimation, because knowledge of the structure adds information to what can be expected of unobserved elements.

In model-based methodologies, forecasts or predictions are the expected value of the response. The expectation can often be modeled with a linear model. Conditional on the observed data, assuming the covariance structure is known, the predictions based on the conditional expectations are best linear unbiased predictors (BLUPs), which minimize mean square prediction error (MSPE) (i.e., average squared difference between the observed and predicted values). Zimmerman and Cressie (1992)

discuss the effect of estimating the covariance parameters on the empirical (estimated) BLUP and MSPE.

Kriging produces a best linear unbiased predictor of the response at a location conditioned on the response observed at sample locations. Its application supposes that the response $z(s)$ is a regionalized variable (continuous on the scale of interest). Kriging models a tendency of regression of the response toward the mean (Laslett, 1997). For the current scope, assume the continuous-domain response is the result of an isotropic stationary random process (see Cressie, 1993). An incrementally stationary process is one for which the expected squared-difference in response depends only on distance between the locations, not on the absolute location. A stationary process is a special case, for which the variance and mean of response does not depend on location. A process is described as isotropic if the covariance (or mean squared-difference) does not depend on orientation of the two elements.

A prerequisite to kriging is the specification, estimation, and validation of a semi-variogram or covariogram. The semi-variogram describes the average squared difference of two elements' responses as a function of distance. Kriging coefficients are derived from the system of equations that solve for coefficients which minimize MSPE, subject to the constraint that they sum to one (ensuring uniform unbiasedness). The solution involves the covariance matrix. Refer to Cressie (1993); Thompson (1992); Journel and Huijbregts (1978), among others, for theory and implementation.

In some cases, the distribution of a response may be modeled to depend on auxiliary data, such as for model-assisted estimators. For the discussion here, model-based estimation is with reference to modeling of intrinsic covariance structure and not involving auxiliary data.

3. CONTINUOUS DOMAIN SAMPLING

The first obvious difference between sampling on a continuous domain versus sampling a finite population is that the elements chosen to be in the sample are identified by location instead of unit identification. The notation $z(s)$ will denote the response at location indicated by the 2- or 3-D vector '$\boldsymbol{s}$' defined on the continuous 2- or 3-D domain. A vector of sampled locations will be denoted in bold $\boldsymbol{z}(\boldsymbol{s})$.

On a continuous domain, the probability measure of any sample must be defined for a continuous domain. ID of the ith element is the integral over the measures of samples that include the ith element, and the pair-wise ID of the ith and jth elements is the integral over the measures of samples that include both elements (see Cordy, 1993). The measures are with respect to a measure of a sample on the spatial domain with elements (locations) denoted by vectors $\boldsymbol{s}_i$ (or just $\boldsymbol{s}$). For the scope here, the sampling is non-informative, i.e., the response $z(s)$ does not influence selection of locations included in the sample. Cordy (1993) extends the HT and Yates–Grundy (YG) estimators to the continuous domain.

The response may be continuously varying and described as regionalized. The covariance between two elements is often characterized by the proximity of the two elements. For simulations described in a later section, the random process that characterizes the response is assumed to be incrementally stationary and isotropic.

Since the response on the continuous domain may follow a trend or have spatial covariance, it is often prudent to obtain a spatially balanced design, to maximize efficiency of a sample (minimizing redundancy of observations). This is achieved with either systematic or spatially stratified samples (see Stehman and Overton, 1994; Olea 1984). A feature of these designs is that variance estimation can be problematic. In some cases, there may not be direct estimators of the variance. Stevens (1997) explains

that congruent tessellation stratified designs with constant origin and one observation per stratum have no direct variance estimator. The expectation of the HT estimator does not exist in this case, because samples for which pair-wise IDs equal to zero have non-zero measure (are possible) and the HT variance estimator involves division by pair-wise IDs. Stevens (1997) describes a procedure developed by Dalenius *et al*. (1961) for deriving the pair-wise IDs when the tessellation origin is randomly located. The method is to determine the proportion of a congruent stratum that would not contain a stratum center such that two points would be contained by the same stratum, because by design (of one observation per stratum) those grids so located could not include both points. For the randomly located tessellation, the pair-wise IDs are all non-zero and so the HT and YG variance estimators can be shown to be unbiased.

Unlike finite populations, the response between two elements is rarely exchangeable on the continuous domain. Exchangeable means that any permutation of observations is a sufficient statistic. The joint distribution of an ordered response, such as in a spatial context, depends on the spatial arrangement. In particular, the variance of a linear combination of ordered responses is a function of pair-wise covariance typically depending on proximities. In finite populations, to the extent that the responses are used directly to estimate the sums of squares at a particular level in a nested hierarchy, the responses are implicitly being treated as exchangeable to estimate variability (Bellhouse, *et al*. 1977). As long as the units are exchangeable, the covariance within a particular level is constant and the sums of squares from each level in the structure are a sufficient statistic for variance.

For continuous domains, the sampling process and the response's covariance structure can have an interacting effect on variability of the estimator. If the range of covariance is very small relative to the resolution of points sampled, the sums of squares may be adequate to approximate variance within a particular stratum. The locations of a pair of sampled sites establish a relationship between the sites' responses as either (effectively) independent or correlated. The joint distribution of the sample's responses is generally not adequately handled by treating responses as fixed and exchangeable, as in finite populations. Oliver and Webster (1986) describe a study in which they explore whether what appears to be pure nugget effect (variability due to measurement) at the original sampling resolution would then manifest spatial auto-correlation at a smaller scale. Variance estimation on the continuous domain should account for possible covariance in the responses.

It should be clarified that where sample elements are characterized as having independence due to the sampling process (Hansen *et al*., 1983); Brus and de Gruijter, 1993; 1997), that independence is specific to the selection into the sample of one unit with respect to another, and it does not imply that the responses observed are independent (uncorrelated). Inference on the response surface would involve the distribution of the response surface. The mean and covariance structure, or sufficient statistics of mean, variance and covariance, are called for to reliably quantify estimator variance.

4. CONVENTIONAL DESIGN-BASED VARIANCE ESTIMATORS

These estimators quantify variance due to the sampling process. The HT variance estimator (involving the square and cross-product terms weighted by marginal and pair-wise IDs) is shown here for reference (where π_i and π_{ij} are the marginal and pair-wise inclusion densities; and w'z represents the linear combination of the observations weighted by the inverse marginal IDs).

$$\hat{V}_{\mathrm{HT}}[w'z|S] = V[\hat{\tau}_{\mathrm{HT}}] \overset{\text{Cordy (1993)}}{=} \sum_i \frac{z_i^2}{\pi_i^2} + \sum_i \sum_{j \neq i} \frac{1}{\pi_{ij}} \frac{z_i}{\pi_i} \frac{z_j}{\pi_j} \left(\pi_{ij} - \pi_i \pi_j\right) \tag{1}$$

The HT variance estimator is unbiased with respect to the distribution of samples in the sample universe—provided $\pi_{ij} > 0$ a.e.

Occasionally, the HT estimates turn out to be negative (Yates and Grundy, 1953; Stevens and Olsen, 2003). This is more likely to happen when there are pairs of points for which the pair-wise ID is very small, which can occasionally happen, for example, for random-origin tessellation-stratified (RTS) samples with one observation per stratum. For fixed-sample-size samples, the Yates–Grundy (YG) form of the theoretical variance of a linear estimator is mathematically equivalent to that of the HT (Yates and Grundy, 1953). When $\pi_{ij} \leq \pi_i \pi_j$ (as in RTS design), the YG estimator (below) has the advantage that it will not produce negative estimates.

$$\hat{V}_{\mathrm{YG}}\left[w'z|S\right] = V[\hat{\tau}_{\mathrm{YG}}] \overset{\substack{Cordy \\ (1993)}}{=} \sum_i \sum_{j<i} \frac{1}{\pi_{ij}} \left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right)^2 (\pi_i \pi_j - \pi_{ij}) \tag{2}$$

In the continuous domain and when a RTS design is employed, the YG estimator can still sometimes be destabilized by the occasional sample for which one or more pairs of points happen to have points separated by very small distances (Stevens and Olsen, 2003), as this would put substantial weight on those associated cross-product terms. Stevens and Olsen, (2003) advise that the hazard of instability is even greater for unequal probability sampling.

For stratified sampling with multiple elements per stratum, a design-based variance can be estimated by combining within- and between-strata mean square errors. These direct estimates combine measures of low and high frequency variation—the within-stratum variance measuring the local variation. This estimation of within- and between-variance assumes an exchangeable covariance structure. Systematic samples or constant-origin stratified samples with only one element per stratum do not have a direct estimator of variance. The conventional alternative variance estimators—contrast estimators—typically define quasi-strata containing two or more elements per stratum. There is a plethora of varieties of contrast estimators, altering directions and sizes of the quasi-strata (see Wolter, 1985). The contrast estimators are sometimes interpreted as removing trend (or 1st order correlation for finite-population domains).

5. MODEL-BASED VARIANCE ESTIMATION—MSPE

MSPE measures the variance of the random variable plus squared bias of the estimated mean. Here, the variance is induced by the stochastic behavior of the response, as opposed to sample-process variance. Often forecasting on time domains or prediction in spatial domains involves a covariance structure that is not exchangeable, but depends on lag or distance.

For an incrementally stationary process, the semi-variogram can be characterized by a non-increasing function of distance between the two locations. The best linear unbiased predictor for an unobserved location s_o, conditional on the observed data, is the conditional expectation of the response at that location. If the average square of the increment in response is a decreasing function of distance, the average increment must be decreasing also. The expected value of one location conditional on another will approach the observed value at that other location as distance diminishes. This implies that the BLUP will have a diminishing range of values for locations s_o closer to the sampled locations.

In particular, for the sample resolution and ranges examined in this study, the MSPE is approximated by a linear relationship with the distance from s_o to the nearest observed location, when nearest distances are within the range of the process.

Given incremental stationarity, the increment in response diminishes as distance diminishes, and thus range of the prediction diminishes, i.e., the variability in the prediction due to sample process has some methodical behavior. Samples from the sample universe can be loosely regarded as equivalence classes of samples defined by value and proximity of a sample's closest point to s_o, the value and proximity having important influence on the resulting prediction.

Kriging coefficients can vary substantially from sample to sample (Diamond and Armstrong, 1984), but depending on the range, the kriging prediction may not vary much from sample to sample. Since the prediction is a weighted average of the observations in the sample, the smoothing operation reduces variability.

For resource managers and policy makers, the sampling process variance is of interest as a measure of precision as provided by the sampling and estimation process. In a model-based approach for forecasts and predictions, this measure is often considered irrelevant. The amount of natural variability about the predicted average is estimated by the MSPE, where the natural variability is analogous to the estimated variability about a cell mean in a linear model. An estimate of sampling-process variance would indicate something about the precision with which the average value can be predicted, as afforded by the sampling process. Good or poor precision might be foreseeable, depending on how far the location to be predicted is from the observed locations, relative to the underlying range of covariance. A comparison of the sampling-process variance and the MSPE for various ranges and two sill values is demonstrated in later sections.

6. PROPOSED MODEL-ASSISTED VARIANCE ESTIMATOR

As alluded to above, the HT variance estimator sometimes comes out negative. The YG alternative can occasionally be unstable for spatially balanced samples which happen to have a pair of points very close together. Performance of contrast estimators may depend on choice of orientation and size of quasi-strata and the covariance structure. In what follows, an alternative method to sample-process variance estimation is explored. Spatially balanced sampling designs do not always have a configuration of observations that permit direct estimates of variance, and model-assisted approaches might be useful and justified, as they are in small-area estimation (Rao, 2003).

The proposed approach of a model-assisted (MA) variance estimator is based on some observations about sample designs and estimators. A constrained sample cannot vary as much as a simple random sample. On a continuous domain, estimates from two systematic samples in near proximity, relative to the underlying range of the covariance, may differ very little, depending on the smoothness of the process. Within stratified designs, the variability of observations from each stratum will be limited by the variance within each stratum. Given a reasonable model of the covariance structure for which reasonable estimates of parameters are obtained, the variance of the linear combination of observations (e.g., HT estimators and BLUPs) can be modeled as the sum of squared-coefficients times the average within-stratum variance.

The within-stratum variance is readily modeled as developed in Appendix I, following similar computations for error variance in Ripley (1981, Ch. 3). A general expression of within-stratum variance is

$$v_{\text{win}} = b - c_{\text{avg}} = b - \int_{\substack{h \in \|s_i - sj\| \\ \forall s_i, s_j \in A}} c(h) f(h) \mathrm{d}h \tag{3}$$

where b denotes the sill of the semi-variogram (or the variance of the random process); and where the covariance structure is denoted as $c(h)$, a function of distance h that results in a valid (positive-definite) covariance matrix; and $f(h)$ denotes the density of the distances within stratum area A.

As an example, if the assumed covariance structure is exponential, the average covariance (c_{avg}) is approximated by numerical integration, by averaging $b\exp(-h/r)$ (where r denotes covariance range) over all point-pair distances on a dense grid overlaying the area of the stratum. The average within-stratum variance is the modeled variance of the process reduced by the average covariance of the stratum $(v_{\text{win}} = b - c_{\text{avg}})$. The model-assisted variance estimate of a linear estimator $a'z$ is $\hat{v} = \sum_{i=1}^{n} a_i^2 v_{\text{win},i} \overset{\text{congruent tesselation}}{=} \sum_{i=1}^{n} a_i^2 v_{\text{win}}$, where a is a vector of coefficients, and v_{win} denotes the average within-stratum variance if all the strata are the same dimensions.

Covariance between strata is not relevant to the sample-process variability for a fixed study area completely covered by the stratification grid. Ordinarily, poorly balanced samples from a domain with positive correlation means that, while there is less variability within each sample of positively correlated elements, there is more variability from sample to sample. For the stratified grid overlaying the fixed study region, all the strata are subsampled (at one location) in any sample from the sample universe. If there is positive correlation between strata, that positive correlation will not vary from one sample to the next, on that fixed study region, and does not affect the variability of an estimate from one sample to the next.

Other than within-stratum variance, the only other variability relevant to sample process is variability induced by the definition of the strata, due to randomly locating the grid. Given a stationary process, the average within-stratum variance will not vary due to location, so that a randomized grid origin has no effect on this parameter. Conditional on the grid location, the variance of any element in the sample is the within-stratum variance, as described above. Any effect due to wrapping the boundary strata around the ends of the region is ignored, and in the simulations there is little difference between fixed or randomly located grid stratification, on within-stratum variance or on the empirical variance of the linear estimates.

In the case of a BLUP produced by kriging, the variance estimate is approximated by treating the kriging coefficients as though they are constant, though they are not. The alternative estimator is demonstrated in simulations of both design-based and model-based contexts applied to continuous domains, as described in the following.

7. METHODS

Basic stratified samples were drawn repeatedly from a random field—a single realization of a random process. The random field was generated with an exponential covariance structure using the Random Fields package available in R (Schlather, 2001). The strata are defined by a regular 10×10 grid of 20×20 square strata overlaid on the 200×200 field. Each element in the field is $(0.1)^2$ distance-units square, so a real extent of the field is 20×20 distance-units squared. Each sample contains 100 observations, with one observation per stratum. For comparison, simulations were repeated for both randomized and constant grid origins. In the case of randomized origins (randomized on each trial), the grid is wrapped around the end of the field to continue on the other side, from left to right and bottom to top, so that the strata on the edges straddle the top and bottom or left and right boundary of

the field. The boundary effect in these strata is ignored in the estimation process in this study and the amount of error thus introduced is not quantified here.

For the design-based context, the HT estimate of total response on the domain is computed for each of 1000 trials of stratified sampling on a fixed realization of a random field. For each trial, a semi-variogram is fitted, assuming exponential covariance structure with no nugget, using REML. The model-assisted variance estimate and HT and YG design-based variance estimates are computed for each trial. These are compared to the empirical variance of the HT estimates of the total.

The entire process was repeated for eight combinations of sill and range values (ranges of 0.5, 1, 2 or 4; sill values of 1 or 4). In all cases, the stratum size is 2×2 distance-units squared, one observation per stratum, resulting in an average sampling interval of two distance-units. At present there is no nugget. In previous implementations, there was only a modest effect of model misspecification if the actual covariance was spherical but an exponential form was assumed.

For the model-based scenario, ordinary kriging is used to predict a location (constant over the sampling trials of a particular field). Kriging is implemented as described in Cressie (1993). Sampling and kriging were repeated for 1000 trials per realization. The computed sampling-process variance (estimated by the model-assisted estimator) and the kriging variance (model-based MSPE) were saved for each trial. Means and histograms are compared with the empirical variance of the prediction.

8. RESULTS

8.1. Design-based context results

Histograms of MA, YG, and the HT variance estimates of the 1000 trials for each range-sill combination were examined. In all combinations for stratified samples with a randomly located tessellation grid, the HT variance estimator has a notable negative tail in its distribution. There is a greater prevalence of negative estimates for those samples for which one or more pairs of points are in close proximity. Usually there is not evidence of bias in the HT variance estimator. The MA and YG histograms were generally similar in range, shape, and location for all combinations. The ranges of the MA and YG histograms are typically smaller than those of the HT estimator, even when the negative estimates are ignored. The ranges of the positive parts of the three estimators' histograms are not extraordinarily different. The YG histograms are slightly right-skewed. Those of the MA are for the most part symmetric. Occasionally, but less often than HT or YG, the MA can also get high estimated values of estimator variance, indicating that occasionally the range and sill parameters are poorly estimated. If the range is estimated to be much smaller than the true range, there will be a larger estimate of within-stratum variance, which inflates the estimated variance. The histograms for range of two and sill of four are shown in Figure 1, for illustration. Histograms from the other combinations are similar.

The variance estimators are compared to the empirical variance of the HT estimate of total. Table 1 summarizes the empirical median relative errors of the conventional HT (V_{HT}), the MA (V_{MA}), and the YG (V_{YG}) variance estimates, for the stratified samples taken with a randomly located grid. The empirical median relative error is the difference between the median estimated variance and the empirical variance, divided by the empirical variance.

The median relative errors of the HT variance estimator were all positive. Those of the YG estimator were all negative. Those of the MA estimator were centered around zero. The worst absolute median relative errors were 37.2% (HT for sill of 1 and range of 4); 22.8% (YG for sill of 1 and range

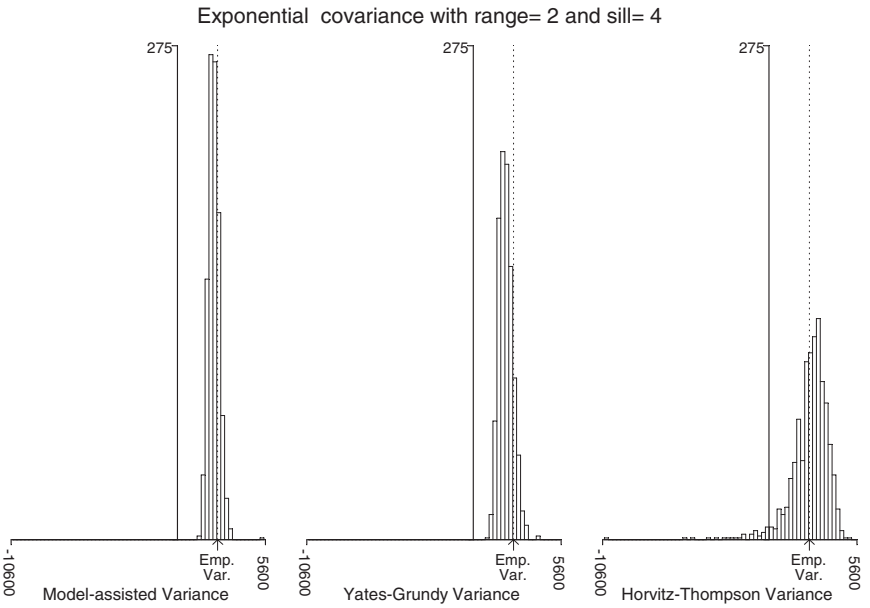

Figure 1. Histograms of the variance estimates (1000 trials with randomly located grids). 'Emp. Var.' is the empirical variance of the Horvitz–Thompson (HT) estimates of total

Table 1. Empirical median relative errors (1000 trials with randomly located grids)

Sill	4				1			
Range	0.5	1	2	4	0.5	1	2	4
V_{HT}	0.068	0.063	0.189	0.161	0.044[a]	0.185	0.070[a]	0.372
V_{YG}	−0.045[a]	−0.122	−0.117	−0.176	−0.052	−0.032[a]	−0.228	−0.133
V_{MA}	0.055	−0.024[a]	−0.001[a]	−0.035[a]	0.049	0.084	−0.138	0.004[a]

[a]Indicates the smallest absolute median relative error of the range-sill combination.

of 2), and 13.8% (MA for sill of 1 and range of 2). More often than not, the MA variance estimator outperforms the YG estimator. In two of the eight combinations, the HT variance estimator has the smallest median relative error.

A comparison of the efficiency of the MA and YG variance estimators relative to the HT variance estimator is given by the ratio of the empirical standard deviations of MA (or YG) variance estimates to that of the HT variance estimates. These are summarized in Table 2.

Table 2. Ratios of empirical standard deviations of variance estimators (1000 trials with randomly located grids)

Sill	4				1			
Range	0.5	1	2	4	0.5	1	2	4
V_{MA}/V_{HT}	0.56	0.43	0.27	0.24	0.66	0.36	0.20	0.14
V_{YG}/V_{HT}	0.77	0.62	0.35	0.28	0.84	0.43	0.27	0.17

Table 3. Empirical median relative errors (1000 trials with fixed grid locations)

Sill	4				1			
Range	0.5	1	2	4	0.5	1	2	4
V_{MA}	−0.050	−0.071	−0.073	0.018	−0.001	−0.019	0.081	0.120

In every case, there is reduction in variability in both the MA and YG variance estimates over the HT estimates. The reduction is greatest for the higher ranges (2 and 4), for which the reduction is on the order of 75%. Reduction at the lowest ranges is on the order of 50%. Variability of the MA variance estimator is on the order of an additional 25% smaller than that of the YG estimator.

In the case that the tessellation grids were not randomly located, the HT variance estimator and YG variance estimator are not available. The empirical median relative error of the MA variance estimator is summarized in Table 3. The largest absolute median relative error is 12% for a sill of one and range of four. Much of the error would be attributed to poor range/sill estimates, as the estimated range is usually smaller than true range for the range of four.

8.2. Model-based context results

The empirical variance indicates sample-process variability of the estimated average response for the kriged location, for the 1000 trial samples. While the MA variance estimator is only an approximation that does not account for variability of the kriging coefficients, the histograms of the estimates produced from 1000 trials for each range-sill combination, with or without randomized origin, do not show any systematic patterns of bias. The approximated estimates would be reasonably useful to suggest a rough idea of the sample-process variability of the estimated expected response.

For the lower ranges, the predicted response tends to be consistent from sample to sample. As the range increases, the amount of variability over the samples starts to have more sizeable magnitude relative to the MSPEs. Figure 2 contains histograms of the MSPE computed using the estimated and known sill and range (upper panels) and of the approximate sampling process variance as estimated by the MA for estimated and known parameters (lower panels) for a sill of one and range of one, for stratified samples with randomized origin. The observed variability in the predicted value is indicated by the reference line labeled 'Emp. Var.'. Histograms for other combinations of range and sill are not notably different.

9. CONCLUSIONS

The basis of inference for design- and model-based approaches to sampling and estimation were compared, and precautions suggested for their applications. The idiosyncrasies of application in the spatial domain were described. The interpretation of the variability described by design- and model-based paradigms was discussed, addressing what source of variation each method quantifies.

There are many applications of employing models to restrict the sampling process, to select samples that will optimize parameter or coefficient estimation, to optimize efficiency and to reduce bias. These applications are found in studies of both design- and model-based objectives. Design-based variance estimation development has focused on variability of inclusion of elements in a sample, with some discussion emphasizing independence due to sampling process. The paradigm seems to

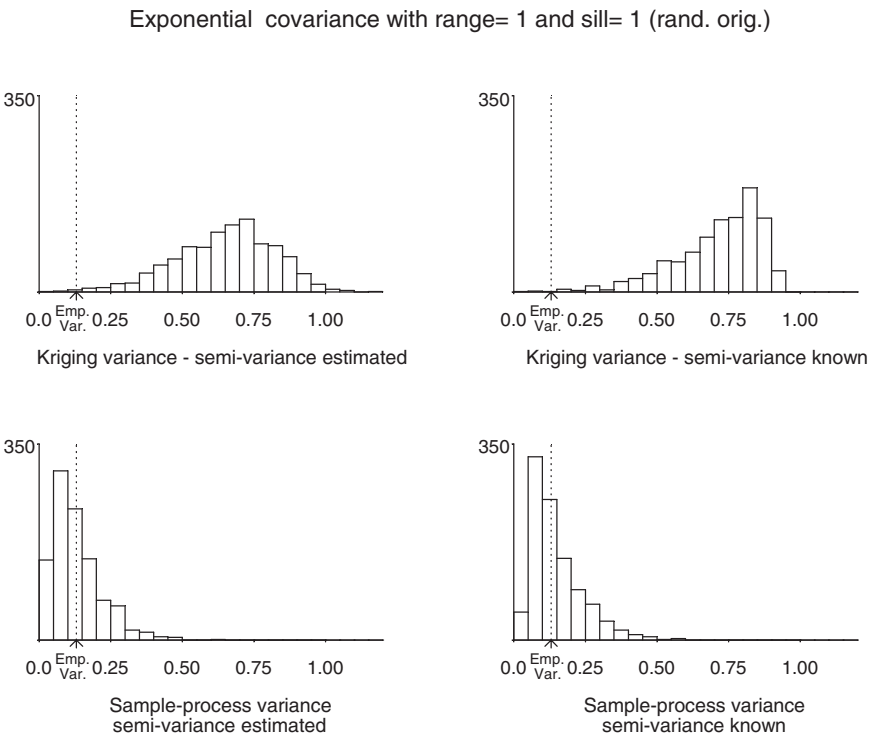

Figure 2. Histograms of mean square prediction error (MSPE) (upper panels) and approximate sampling-process variance (lower panels)

have left out the potential for employing response-covariance models to estimate sample process variance, though this variance is influenced by that covariance structure if the sampling resolution is comparable to the range of covariance.

Besides being efficient, the model-based paradigm that the response is correlated is useful and important for samples taken from the spatial domain. If the sampling resolution is dense relative to the range of the covariance, the exchangeable model of response is less defensible for application to estimating variance. The correlation is not ignored when sampling strategies are compared for optimality (as studied by Cochran, 1946; Bellhouse, 1977). The covariance of the response is fortuitous for providing a potential way to efficiently estimate variances when direct estimators are lacking due to the sampling design structure. The simulations show that explicitly modeling the covariance of the response, to model the restricted variability of observations within strata and of the linear estimators, can provide an efficient and effective approach to estimating sampling process variability. This is consistent with the results in Cordy and Thompson, (1995).

ACKNOWLEDGEMENTS

The author would like to thank Don L. Stevens, Jr. for advising and for support. The research described has been funded by the U.S. Environmental Protection Agency through the STAR Cooperative

Agreement CR82-9096-01. National Research Program on Design-Based/Model-Assisted Survey Methodology for Aquatic Resources at Oregon State University. It has not been subjected to the Agency's review and no official endorsement should be inferred.

APPENDIX I—WITHIN-STRATUM VARIANCE

Computations in this section are similar to error variance modeling described in Ripley (1981). $z(s)$ (abbreviated 'z') is a realization of a stationary, isotropic random process. The covariance of the response $z(s)$ and $z(t)$ is assumed to be a function of the distance 'h' between s and t, denoted $\mathrm{C}[z(s), z(t)] = \mathrm{C}[\|s - t\|] = \mathrm{C}(h)$. Denote the random process mean and variance $E[Z(s)] = \mu$ and $V[Z(s)] = \sigma^2$. To indicate an expectation or variance within a stratum of area A, denote the expectation conditional on the realization Z as $E[z|Z; \mathrm{A}]$; similarly for the variance. The variance of the response within an area A is $E[(z - z_A)^2]$, where z_A is the mean of the response in area A (denoted $|A| = w^{-1}$). Note that $\mu = |A| w\mu = w \int_A \mu \mathrm{d}s$. Within-stratum variance is expressed as follows:

$$\begin{aligned} V[z|Z, A] &= w \int_A (z - z_A)^2 \mathrm{d}s = w \int_A ((z - \mu) - (z_A - \mu))^2 \mathrm{d}s \\ &= w \int_A \left\{ (z - \mu)^2 + (z_A - \mu)^2 - 2(z - \mu)(z_A - \mu) \right\} \mathrm{d}s \end{aligned} \tag{1}$$

The expression in Equation (1) simplifies by combining the 2nd and 3rd terms in the integrand.

$$\begin{aligned} \int_A (z_A - \mu)^2 \mathrm{d}s &= \int_A \left(w \int_A z(t) \mathrm{d}t - \mu \right)^2 \mathrm{d}s = |A| \left(w \int_A z(t) \mathrm{d}t - \mu \right)^2 \\ &= \left(w \int_A (z(t) - \mu) \mathrm{d}t \right) \left(\int_A (z(s) - \mu) \mathrm{d}s \right) \end{aligned} \tag{2}$$

The third term in Equation (1) is expressed as in Equation (2) by noting that $(z_A - \mu) = w \int_A (z(t) - \mu) \mathrm{d}t$ is constant with respect to $\mathrm{d}s$, and can be taken outside the integral. Bringing the original w into the integral, the expression becomes:

$$V[z|Z, A] = w \int_A (z - \mu)^2 \mathrm{d}s - w^2 \int_A (z(t) - \mu) \mathrm{d}t \int_A (z(s) - \mu) \mathrm{d}s \tag{3}$$

In other words, the within-stratum variance is the process variance reduced by the average covariance within the stratum.

REFERENCES

Bellhouse DR. 1977. Some optimal designs for sampling in two dimensions. *Biometrika* **64**(3): 605–611.

Bellhouse DR, Thompson ME, Godambe VP. 1977. Two-stage sampling with exchangeable prior distributions. *Biometrika* **64**(1): 97–103.

Brus DJ, de Gruijter JJ. 1993. Design-based versus model-based estimates of spatial means: theory and application in environmental soil sciences. *Environmetrics* **4**(2): 123–152.

Brus DJ, de Gruijter JJ. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. *Geoderma* **80**: 1–44.
Carroll SS. 1998. Modelling abiotic indicators when obtaining spatial predictions of species richness. *Environmental and Ecological Statistics* **5**(3): 257–276.
Cochran WG. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics* **17**(2): 164–177.
Cordy CB. 1993. An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters* **18**: 353–362.
Cordy CB, Thompson CM. 1995. An application of the deterministic variogram to design-based variance estimation. *Mathematical Geology* **27**(2): 173–205.
Cressie N. 1993. *Statistics for Spatial Data*. Wiley: New York.
Dalenius T, Hájek J, Zubrzycki S. 1961. On plane sampling and related geometrical problems. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Neyman, J (ed.) University of California Press: Berkeley, CA; **1**: 125–150.
Diamond P, Armstrong M. 1984. Robustness of variograms and conditioning of kriging matrices. *Mathematical Geology* **16**: 809–822.
Hansen MH, Madow WG, Tepping BJ. 1983. An Evaluation of model-dependent and probability-sampling inferences in sample surveys. *JASA* **78**: 776–793.
Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *JASA* **47**: 663–685.
Journel AG, Huijbregts CJ. 1978. *Mining Geostatistics*. Academic Press: New York.
Laslett GM. 1997. Discussion of the paper by D.J. Brus and J.J. de Gruijter. *Geoderma* **80**: 45–49.
ODFW 2002. *The Oregon Plan for Salmon and Watersheds 1997—Sampling Design and Statistical Analysis Methods for the Integrated Biological and Physical Monitoring of Oregon Streams* (OPSW-ODFW-2002–2007).
Olea RA. 1984. Sampling design optimization for spatial functions. *Mathematical Geology* **16**(4): 369–392.
Oliver MA, Webster R. 1986. Combining nested and linear sampling for determining scale and form of spatial variation of regionalized variables *Geographical Analysis* **18**: 227–242.
Rao JNK. 2003. *Small Area Estimation*. Wiley: Hoboken, NJ.
Rawlings JO, Pantula SG, Dickey DA. 1998. *Applied Regression Analysis—A Research Tool* (2nd edn). Springer: New York.
Ripley BD. 1981. *Spatial Statistics*. Wiley: New York.
Royall RM. 1988. The prediction approach to sampling theory. In *Handbook of Statistics*, Vol. 6, Krishnaiah PR, Rao CR (eds); 399–413.
Royall RM, Cumberland WG. 1981. An empirical study of the ratio estimator and estimators of its variance. *JASA* **76**(373): 66–80.
Schlather M. 2001. Simulation and analysis of random fields. *R News* 1/2: 18–20.
Stehman SV, Overton WS. 1994. Environmental sampling and monitoring. In *Handbook of Statistics. Environmental Statistics* Vol. 12, pp. 263–306, Patil GP, Rao CR (eds). North Holland: New York
Stevens DL. 1997. Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics* **8**: 167–195.
Stevens DL, Jr., Olsen AR. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* **14**: 593–610.
Thompson SK. 1992. *Sampling*. Wiley: New York.
Ver Hoef JM. 2002. Sampling and geostatistics for spatial data. *Ecoscience* **9**(2): 152–161.
Wolter KM. 1985. *Introduction to Variance Estimation*. Springer-Verlag: New York.
Yates F. Grundy PM. 1953. Selection without replacement from within strata with probability proportional to size *Journal of Royal Statistics Society Serial B* **1**: 253–261.
Zimmerman DL, Cressie N. 1992. Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annual Institute of Statistics and Mathematics* **44**(1): 27–43.

