AN ABSTRACT OF THE DISSERTATION OF

Cynthia Cooper for the degree of Doctor of Philosophy in Statistics presented on February 14, 2006.

Title: Developing a Basis for Characterizing Precision of Estimates Produced from Non-probability Samples on Continuous Domains.

Abstract approved:

____
Don L. Stevens, Jr.

This research addresses sample process variance estimation on continuous domains and for non-probability samples in particular. The motivation for the research is a scenario in which a program has collected non-probability samples for which there is interest in characterizing how much an extrapolation to the domain would vary given similarly arranged collections of observations. This research does not address the risk of bias and a key assumption is that the observations could represent the response on the domain of interest. This excludes any hot-spot monitoring programs. The research is presented as a collection of three manuscripts. The first (to be published in *Environmetrics* (2006)) reviews and compares model- and design-based approaches for sampling and estimation in the context of continuous domains and promotes a model-assisted sample-process variance estimator. The next two manuscripts are written to be companion papers. With the objective of quantifying uncertainty of an estimator based on a non-probability sample, the proposed approach is to first characterize a class of sets of locations that are similarly arranged to the collection of locations in the non-probability sample, and then to predict variability of an estimate over that class of sets using the covariance structure indicated by the non-probability sample (assuming the covariance structure is indicative of the covariance structure on the study region). The first of the companion papers discusses characterizing classes of similarly arranged sets with the specification of a metric density. Goodness-of-fit tests are demonstrated on several types of patterns (dispersed, random and clustered) and on a non-probability collection of locations surveyed by Oregon Department of Fish & Wildlife on the Alsea River basin in Oregon. The second paper addresses predicting the variability of an estimate over sets in a class of sets (using a Monte Carlo process on a simulated response with appropriate covariance structure).

Developing a Basis for Characterizing Precision of Estimates Produced from Non-probability Samples on Continuous Domains.

by
Cynthia Cooper

A DISSERTATION

submitted to

Oregon State University

In partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented February 14, 2006
Commencement June 2006

Doctor of Philosophy dissertation of Cynthia Cooper presented on February 14, 2006.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Cynthia Cooper, Author

ACKNOWLEDGEMENTS

It is impossible to name all the people who have contributed to my education at Oregon State University. I am grateful to Don L. Stevens for allowing me the opportunity to research a stimulating topic.

CONTRIBUTION OF AUTHORS

Dr. Stevens developed the Side-Vertex-Boundary metric on domains of aerial extent and provided an R script to compute this metric.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

DISSERTATION INTRODUCTION

This research addresses sampling and estimation on responses observed on spatial domains. In this context the response is defined on a continuous domain, as opposed to a finite population. Generally the response observed at two locations in close proximity will have covariance, which typically decreases as distance between the locations increases. Two types of approaches have been developed and applied to estimation on continuous-domain responses – model-based and design-based. The model-based approaches treat the response as a realization of a random field (see, for example, Cressie (1993)). These approaches obtain minimum mean-square error estimates, conditional on the observations, without accounting for a selection process, random or otherwise, for what elements on the domain are observed. The design-based approach focuses on randomization of the sampling process and treats the response as fixed.

The two approaches tend to serve different objectives. Design-based methods are often employed to estimate a status on a population total or mean. The unobserved response on the domain is typically extrapolated from the observed elements by expansion estimators. The expansion factors are based on marginal and pairwise inclusion probabilities (or densities, on continuous domains) specified by the sampling design - presumably extrapolating from the observations by the amount of the domain they are representative of. Model-based methods are often employed to predict an average response on the domain or at unobserved locations on the domain. In the latter, a typical precursor step is deciding on an appropriate model (form and parameters) of the mean and covariance structure and estimating the parameters involved. The form of the covariance between two locations depends on a range and rate of decay of correlation, usually as a function of distance and possibly orientation between the two locations.

The different objectives encompass dual aspects of monitoring programs. For near-term policy decisions for harvest, for example, an estimate of status and trend are useful. The estimate is a characterization of the response for a particular window of time. The design-based estimates often employed disregard the stochastic behavior of the response over time, treating the response as fixed. For longer-term policies, resource managers may be interested in understanding the stochastic behavior of the response, both in terms of variability spatially and over time – and any interaction of these components of variability. The model-based approach would typically use a likelihood approach to estimate parameters of candidate models to develop one that usefully describes the mean and covariance structure of the response. In this scenario, there is typically some interest in drawing relationships between the mean response and predictors and/or modeling an intrinsic mechanism of the response that influences covariance and that might usefully represent propagation of the response in space and/or time.

Two examples of design-based monitoring programs are the Environmental Protection Agency's (EPA) Environmental Monitoring and Assessment Program (EMAP) (Peterson *et. al.* (1999)); and the Oregon Monitoring Plan of the Oregon Department of Fish and Wildlife (ODFW) (ODFW (2002)), which is an augmented rotating panel sampling design (developed by EPA) for monitoring Oregon-Coast-Natural coho salmon. Examples of typical model-based sampling are found in mining geological surveys and soils and hydrology surveys (Cressie (1993); Oliver and Webster (1986)).

Each approach addresses a different component of variation in an estimate or prediction. The design-based approach specifically quantifies the variability of an estimate due to which elements are sampled – the sample-process variance. The model-based approach quantifies the variation in a prediction (the mean square prediction error (MSPE)) as induced by the variation in increments in response from one location to another as influenced by the covariance structure of the response as realized from a stochastic process.

Key to sample process variance estimation on a spatial domain is the combination of the covariance structure that generally exists for responses on a continuous domain and the interaction of the covariance structure and the sample structure – the resolution and arrangement of sample locations as well as the dimension and size of the region being monitored. At one scale, sparsely spaced observations may not appear to have covariance. A finer resolution of sampling may manifest the covariance within a narrower range than the sparsely spaced locations show (Oliver *et. al.* (1986)). Also the interaction between sample resolution and covariance in the observations may depend on orientation of pairs of sample locations and the orientation of a covariance structure in the response.

The goal of this research is to provide a foundation for assessing the uncertainty of an estimate that is an extrapolation from a non-probability sample. For the purposes of characterizing a population status, non-probability samples are to be avoided whenever possible, as they have been shown to lack adequate representation of the domain response due to selection bias (see for example, Paulsen *et. al.* (1998) and Peterson *et. al.* (1999)). Nevertheless, due to costs and time constraints, agencies often do have data that has not been collected at locations selected by a formal sampling process involving an explicit random mechanism. Also, it is not uncommon that data collected with a modeling objective in mind will be of interest for potential to estimate a summary of the response. For modeling, the choice of locations may have been driven by the goal to optimize the parameter estimation and may not have been chosen by a random mechanism.

Since the non-probability sample is not the result of a randomization of what elements to observe, any estimate based on the observed responses has no sample process variance. However, since the estimate is an extrapolation from the observed responses to the domain, clearly the extrapolation will not represent the domain's response exactly. The consumers of the estimate might be interested to know how much an extrapolation would vary if the response had been observed at other locations – particularly for sets of locations with spatial arrangement similar to that of the non-probability sample, *i.e.* among sets of locations with the same number of points and with similar degrees of clustering or dispersion or randomness.

An assessment for this latter question is only reasonably entertained if the agency has a reason to believe that the mean and covariance of the sample might adequately represent that of the rest of the domain. Monitoring at hot-spots precludes concluding anything about the rest of the domain based directly on the responses observed in the sample. Supposing that the non-probability sample is a collection of observations at locations that would not differ in response from the rest of the domain due to systematic causes (potentially related to the choice of the locations), the agency might have some not-unreasonable data for ascertaining how much the estimate would vary if the observations had been taken at different but similarly arranged sets of locations.

The general approach taken in this research to characterize the variability of such a non-probability estimate is to characterize classes of similarly arranged patterns and to predict the sample process variance that would be observed over that class of patterns. The classes of patterns are defined by specifying a univariate density on a point pattern metric that then imposes a set measure on the universe of sets of *n* points taken on the domain. The variance prediction relies on the covariance structure observed at the observed locations in the non-probability sample.

The chapters are organized as follows. There are three chapters in the body of the dissertation, before the final Conclusion chapter. Each chapter is developed as a separate manuscript, intended to be self-contained. The first manuscript, "Sampling and variance estimation on continuous domains" (to be published in *Environmetrics* in 2006; Chapter 2 of the dissertation), discusses idiosyncracies of, and model- vs. design-based approaches to sampling and estimation on continuous domains. The manuscript promotes a model-assisted sample process variance estimator, illustrated on a stratified sampling design on a continuous domain with a simulated response with exponential covariance structure.

The second manuscript, "Characterizing classes of similarly arranged point patterns as a reference of variability on non-probability samples" (Chapter 3 of the dissertation), develops the construction of a measure (joint density, if it exists) on locations in patterns of points. The measures on the sets of n locations are implied by probability densities on point pattern metrics, where the metrics are statistics on the distances between locations in the point patterns that characterize the degree of clustering or repulsion (manifested as regularity of point spacing) in patterns of points. Empirical metric densities of various classes of point patterns are tested for effective assessment of goodness-of-fit (GOF) of arbitrary patterns to the classes of patterns. Three point pattern metrics are tested on spatially regular and random patterns on a domain of areal extent; and three point pattern metrics are tested on regular, random and clustered patterns on a linear stream-network domain (the Alsea River basin in the Coast range in Oregon). The steps for assessing GOF and suitability of a class as one that has patterns similarly arranged to a non-probability sample is illustrated on a non-probability sample collected by Oregon Department of Fish and Wildlife (ODFW) on the Alsea basin. The ODFW sample is found to have suitable GOF in a class of clustered patterns, suggesting the choice of this class for assessing variability of an estimate on sets of locations similarly arranged to the ODFW non-probability sample.

The third and final manuscript "Estimator variance over similarly-arranged random or non-random locations on continuous domains" (Chapter 4 in the dissertation) develops an approach to predicting sample process variability. The proposed approach is to simulate a response with a covariance structure as suggested by the covariance in the sample, and then derive a Monte Carlo estimate of the sample process variance of an estimate taken on samples from a prescribed class of patterns. The approach is demonstrated with reasonable results on an areal domain with a simulated response with exponential covariance for stratified and random patterns. The approach is also illustrated on the Alsea stream network for a simulated moving-average response, for stratified and random patterns. However, on the stream network, the assessment of variability has poor relative error.

The last chapter, Chapter 5, is the Conclusion.

REFERENCES

Cressie N 1993 *Statistics for Spatial Data*, Wiley.

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-2002-07).

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.

Paulsen SG, Hughes RM, Larsen DP 1998 "Critical elements in describing and understanding our nation's aquatic resources" *Jo. Of the American Water Resources Association* 34, 995-1005.

Peterson SA., Urquhart NS, Welsh, E. B. 1999 "Sample representativeness: a must for reliable regional estimates of lake condition" *Environmental Science and Technology* 33: 1559 - 1565.

SAMPLING AND VARIANCE ESTIMATION ON CONTINUOUS DOMAINS

Cynthia Cooper

Environmetrics
John Wiley & Sons, Ltd.
1 Oldlands Way
Bognor Regis
West Sussex PO22 9SA
UK
To be printed in 2006

SAMPLING AND VARIANCE ESTIMATION ON CONTINUOUS DOMAINS

Cynthia Cooper
Oregon State University
Dept. of Statistics
44 Kidder Hall
Corvallis OR 97330
cooper@science.oregonstate.edu
541-737-1981
Fax: 541-737-3489

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

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.

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 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 locations 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, Madow and Tepping (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 describes an application extending mean and spatial covariance structure models to include abiotic factors to predict bird species diversity on the Indian subcontinent (Carroll (1998)). 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 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 (Ver Hoef (2002)). He observes that the confidence intervals resulting from block kriging are between 20-40% 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 i^{th} 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) and Royall (1988)) or to achieve a numerically well-conditioned system of equations to provide stable estimation of parameters or coefficients (see for example 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), and Journel and Huijbregts (1977), 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 "*s*" defined on the continuous 2- or 3-D domain. A vector of sampled locations will be denoted in bold ***z(s)***.

On a continuous domain, the probability measure of any sample must be defined for a continuous domain. The inclusion density (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 s_i (or just 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) and Olea (1994)). 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, Thompson and Godambe (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, Madow and Tepping (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 Horvitz-Thompson 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).

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

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 Horvitz-Thompson (Yates and Grundy (1953)). When $\pi_{ij} \le \pi_i \pi_j$ (as in RTS design), the YG estimator (below) has the advantage that it will not produce negative estimates.

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

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 *et. al.* (2003)), as this would put substantial weight on those associated cross-product terms. Stevens *et. al.* (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 2 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 (J.N.K. 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

$$v_{win} = b - c_{avg} = b - \int\limits_{\substack{h \in \left\| s_i - s_j \right\| \\ \forall s_i , s_j \in A}} c(h) f(h) \; dh \tag{3}$$

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 ($v_{win} = b - c_{avg}$). The model-assisted variance estimate of a linear estimator $a'z$ is $\hat{v} = \sum_{i=1}^{n} a_i^2 v_{win,i} \overset{\substack{congruent \\ tesselation}}{=} \sum_{i=1}^{n} a_i^2 v_{win}$, 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 RandomFields package available in R (Schlather (2001)). The strata are defined by a regular 10 x 10 grid of 20 x 20 square strata overlaid on the 200 x 200 field. Each element in the field is $(0.1)^2$ distance-units square, so the areal extent of the field is 20 x 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 8 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 2x2 distance-units squared, one observation per stratum – resulting in an average sampling interval of 2 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 model-assisted (MA), Yates-Grundy (YG) and the Horvitz-Thompson (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 4 and sill of 2 are shown in Figure 2-1, for illustration. Histograms from the other combinations are similar.

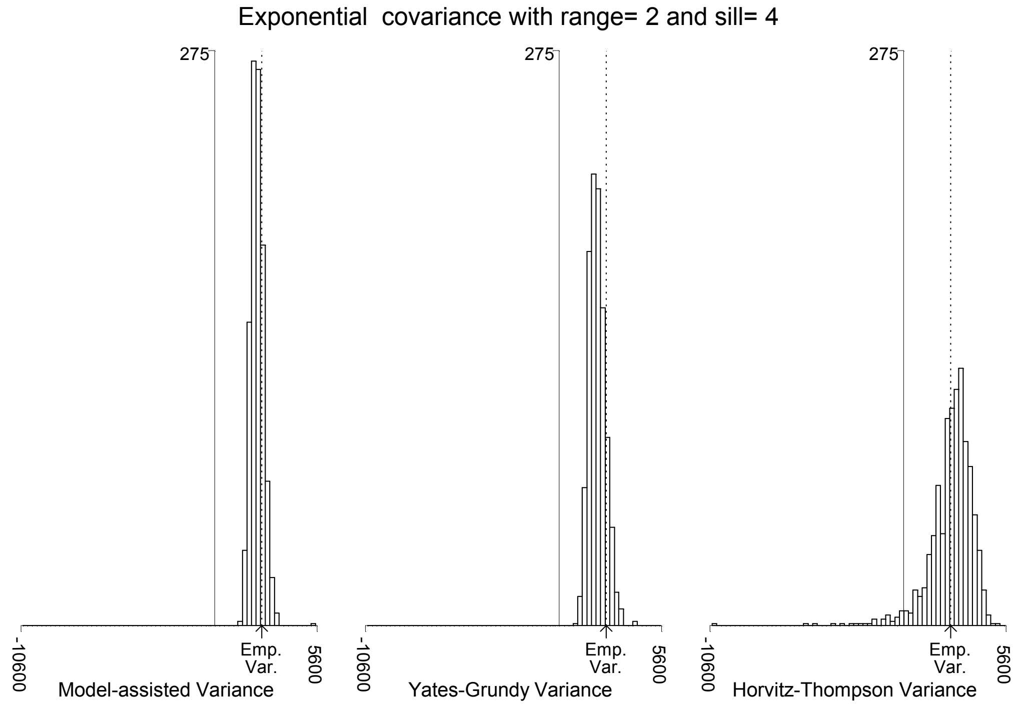

Figure 2-1 Histograms of the variance estimates (1000 trials with randomly located grids). "Emp. Var." is the empirical variance of the HT estimates of total.

The variance estimators are compared to the empirical variance of the HT estimate of total. Table 2-1 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 of 2) and 13.8% (MA for sill of 1 and range of 2). More often than not, the MA variance estimator out-performs 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 2-2.

Table 2-1 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	0.185	*0.070	0.372
V_{YG}	*-0.045	-0.122	-0.117	-0.176	-0.052	*-0.032	-0.228	-0.133
V_{MA}	0.055	*-0.024	*-0.001	*-0.035	0.049	0.084	-0.138	*0.004

* Indicates the smallest absolute median relative error of the range-sill combination.

Table 2-2 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 2-3 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 2-3. The largest absolute median relative error is 12% for a sill of 1 and range of 4. 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 4.

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.

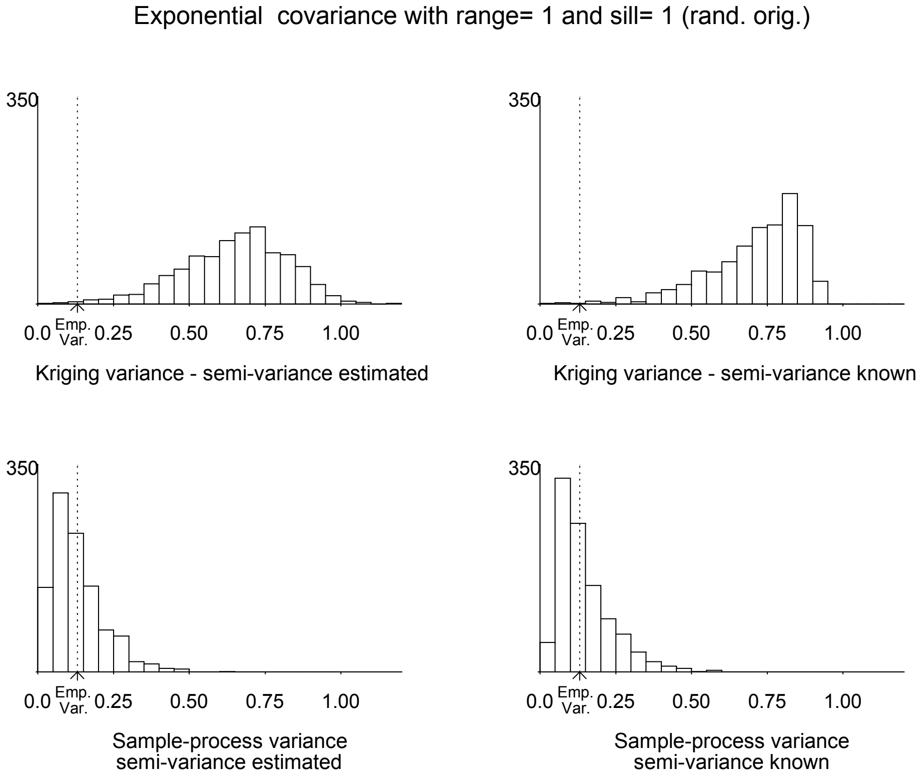

Figure 2-2 Histograms of MSPE (upper panels) and approximate sampling-process variance (lower panels).

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 2-2 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 1 and range of 1, 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 CONCLUSION

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 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) and 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 *et. al.* (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.

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.

Dalenius T, Hájek J, Zubrzycki S 1961 On plane sampling and related geometrical problems *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 125-150 Neyman, J (ed.) University of California Press (Berkeley, CA).

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

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-2002-07).

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).

Rawlings JO, Pantula SG, Dickey DA 1998 *Applied Regression Analysis – A Research Tool* (2nd Ed.) (Springer).

Ripley BD 1981 *Spatial Statistics* (Wiley).

Royall RM 1988 The prediction approach to sampling theory. *Handbook of Statistics* Vol. 6 (Ed. Krishnaiah PR and Rao CR): 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. *Handbook of Statistics Volume 12 Environmental Statistics* 263-306.

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).

Ver Hoef, JM 2002 Sampling and geostatistics for spatial data *Ecoscience* 9(2): 152-161.

Wolter KM 1985 *Introduction to Variance Estimation* (Springer-Verlag).

Yates F, Grundy PM 1953 Selection without replacement from within strata with probability proportional to size *J. of Royal Stat. Soc. Ser. B* 1: 253-261.

Zimmerman DL, Cressie N 1992 Mean Squared Prediction Error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math*. 44(1): 27-43.

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 C[*z(s),z(t)*] = C[||*s-t*||] = 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*; 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 $\mu = |A| w \mu = w \int_A \mu \; ds$. Within-stratum variance is expressed as follows:

$$V[z \mid Z, A] = w\int_A (z - z_A)^2\, ds = w\int_A ((z-\mu) - (z_A - \mu))^2\, ds$$
$$= w\int_A \left\{ (z-\mu)^2 + (z_A - \mu)^2 - 2(z-\mu)(z_A-\mu) \right\}\, ds \tag{1}$$

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

$$\int_A (z_A - \mu)^2\, ds = \int_A \left(w\int_A z(t)\ dt - \mu \right)^2 ds = |A| \left(w\int_A z(t)\ dt - \mu \right)^2$$
$$= \left(w\int_A (z(t) - \mu)\ dt \right) \left(\int_A (z(s) - \mu)\ ds \right) \tag{2}$$

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

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

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

CHARACTERIZING CLASSES OF SIMILARLY ARRANGED POINT PATTERNSAS A REFERENCE OF VARIABILITY ON NON-PROBABILITY SAMPLES

Cynthia Cooper

Don L. Stevens, Jr.

CHARACTERIZING CLASSES OF SIMILARLY ARRANGED POINT PATTERNS AS A REFERENCE OF VARIABILITY ON NON-PROBABILITY SAMPLES

Cynthia Cooper

Don L. Stevens, Jr.

Oregon State University

ABSTRACT

Design-based variance addresses variance of estimates induced by the sampling process of observing a random subset of the response over its domain. The variance of the estimate is the expected squared deviation from its mean over all possible samples taken on the domain. The probability (or "measure" on a continuous domain) need not be uniform over all samples. For non-probability (a.k.a. purposive) samples – such as convenience samples or observations taken at haphazardly selected locations, an estimate based on purposive elements has no sample-process variance. Nevertheless, a stakeholder might reasonably ask, using information inferred from observations about an assumed-stationary response covariance structure, how much would an estimate derived from other sets of observations at similarly arranged patterns of locations (elements) vary? A class of similarly arranged elements or a sample process can be characterized by set measures on the universe of point patterns. This study is part of on-going research to derive the variability of an estimator over similarly arranged collections of observations of a regionalized response on a continuous domain of areal extent or on a linear network domain such as a stream network. In this paper, several point pattern statistics are examined for their utility to provide a set measure on sets of elements taken from the universe of possible point patterns on each domain. Results in this study show a reversal in tendencies in efficiency between metrics incorporating either all or only neighboring point pair distances on areal domains vs. on linear network domains. Three point-pattern statistics are examined for a domain of areal extent – an inner product statistic applied to ordered point-pair distances, the "Side-Vertex-Boundary" (SVB) Dirichlet tile statistic, which measures regularity of point patterns, and a statistic derived from Ripley's K(t) functions. Three statistics are examined for linear networks – a statistic derived from the theoretical exponential distribution of completely random (Poisson process) consecutive-location distances; a stochastic rank statistic based on the cumulative distribution of the consecutive-location distances at a proportion of the sampling resolution; and a 2D version of SVB. The utility of these metrics to perform goodness-of-fit (GOF) assessment of patterns to classes of point patterns is evaluated by Monte Carlo methods for spatially regular (stratified), and random (Complete Spatial Randomness (CSR)) patterns on areal extents and also for these and clustered patterns on a stream network. The inner-product and exponential-fit statistics are fairly good; the K(t)-derived, SVB and stochastic-rank statistics are very good as discriminators.

Keywords: K(t)-functions, Dirichlet tile, spatial point pattern, linear network sampling, Goodness-of-fit

1 INTRODUCTION

1.1 Establishing a Foundation for Quantifying Variability on Non-probability Samples

When one reports a summary statistic of a population, in the field of Statistics, it is understood that when the statistic is based on a sample from the population, that there is variability of the outcome of the statistic based on the sample process – that is, based on which elements in the population were used to compute the statistic. Considerable theory is established to describe sampling distributions. There are two typical models to derive the behavior of a sample statistic – one which supposes an underlying distribution of the response being observed; and one which regards the response as fixed and focuses on the randomization of sample selection. In a typical introductory development of the sampling distribution as described by the model-based paradigm, if observations are drawn from a theoretically infinite domain, supposing the observations are independently identically distributed (iid), a sufficient statistic is derived to estimate a distributional parameter - say, the mean. The behavior of the statistic is examined by taking the expected value and variance of the estimator, which will be derived with respect to a presumed distribution. As an illustration, suppose the sample mean is used to estimate the population mean, where the population is assumed to have random behavior described by a particular distribution. In the following familiar development, x_{i} is an observed value of a random variable X, which it is supposed, has some distribution indexed by a (vector) parameter: $\Phi(\underline{\beta})$ (e.g. the normal distribution). If the observations are iid, the examination of the theoretical performance of the statistic $W(\underline{X})$ is simplified. In these expressions and throughout, the operator E[] represents taking expectation and V[] represents taking the expectation of the second central moment i.e. – the variance.

$$X \overset{iid}{\sim} \Phi(\underline{\beta}) \text{e.g.}\, X \overset{iid}{\sim} N(\mu, \sigma^2)$$

$$W(X) = \sum_{i=1}^{n} w_i x_i = \underline{w}'\underline{x}$$

$$E\left[\sum_{i=1}^{n} w_i x_i\right] = \sum_{i=1}^{n} w_i E[x_i] = \mu \sum_{i=1}^{n} w_i$$

$$V\left[\sum_{i=1}^{n} w_i x_i\right] \overset{iid}{=} \sum_{i=1}^{n} w_i V[x_i] = \sigma^2 \sum_{i=1}^{n} w_i$$

In the second model, the response *X* is treated as fixed (there may be a qualifying index period for which the response might be regarded as fixed). To examine the quality of a sample statistic (which can be interpreted as an extrapolation from a sample to the unobserved response on the fixed domain), a model of the randomization process is useful. This is developed extensively in Survey Methodology. As a simple illustration, consider a response *X* on a fixed domain – which could be a finite population or a bounded surface. (For all practical purposes, the response is always finite). Indicator variables on the indices of the elements in the domain are distributed according to the probabilities (or densities) that the elements are included in a sample.

For example, on a bounded study region sampled uniformly, the density of a point at location *s* is modeled as $f(\underline{s}) = \frac{1}{|A|}$, where |A| denotes the area of the region (so that the probability of including a point from a subset of the domain with area |a| is |a|/|A|). Then the analysis of the sample statistic *W(X)* used to estimate a summary of the domain involves the indicator variables. For example, for a finite population of size N, the estimator involves the indicator variable of the *i*th unit and the behavior is examined with respect to the distribution of the indicator variable: $E\left[\sum_{i=1}^{N} w_i x_i I[i \in S]\right] = \sum_{i=1}^{N} w_i x_i E[I[i \in S]]$. In theory, since the practitioner has control over how units get sampled, the weights in the estimator can be chosen to provide an unbiased estimator. The variance of the estimator is also with respect to the behavior of the indicator variable on the domain, and involves the second-order properties – the pairwise inclusion probabilities (or densities) on the domain. Cordy (1993) extends some finite population design-based estimators to the continuous domain.

Godambe and Thompson (1988) provide an interesting alternative interpretation to describe the behavior of a sample statistic. In this development, supposing the objective is to establish the mean of a population, the variation among individuals from the population mean is regarded as nuisance parameters. They use the random effects model to show how the randomization of the sampling process is a way to adjust for the nuisance parameter. They conclude that the randomization provides a model-free basis for the long-run frequency behavior of a sample statistic. This property is an important foundation for the preference of employing a sample design to produce a probability sample of observations on which population statistics are estimated.

A probability sample is one for which elements in the domain are randomly selected to be in a sample, by explicitly employing a random mechanism. The domain is represented by a frame - a list of units for a finite population or a map for a domain of spatial extent (e.g. a list of segments for a domain of linear extent)). Each element on the frame is assigned a probability, or an inclusion density is defined on the domain, such that a random mechanism is employed to determine inclusion of domain elements in the sample with the probability or inclusion density associated with each element. The way that the inclusion probabilities or inclusion density are defined is formally the sample design. The sample design may assign variable inclusion probabilities, sometimes as a function of auxiliary variables which could be continuous or categorical (such as in cluster or stratified sampling). (There are sampling and estimation procedures for cases in which the target population elements are not itemized - referred to as distance sampling, which is not covered in this study (see Buckland *et. al.* (1993)).

An important underlying premise of using a sample to characterize a population is that the sample is representative of the population. A potential disadvantage of the randomization paradigm is that the sample may only be guaranteed to be representative on average with respect to all possible samples obtained from the prescribed sample design. Royall and Cumberland (1985; 1981) illustrate bias (i.e. error, not expected error) of unrepresentative samples which occur by chance under least restrictive sampling designs on numerous examples involving finite populations (from actual surveys, such as hospital data). In their developments, Royall and Cumberland suggest that a robust strategy of any sample design is to restrict the sampling process in some way in order to achieve a sample cumulative distribution that closely approximates that of the population: $F_s(Y_s) \approx F(Y)$, where Y represents the random response in the finite population and Y_s represents the random response observed in the sample.

Taking a sample without employing a formal sample design – specifically, without a random mechanism is never recommended. There are examples in abundance of misleading extrapolations from convenience samples to a population or domain. A very famous example involves predicting the outcome of the 1936 United States presidential election (Freedman *et. al.* 1991). Haphazard samples, convenience samples and observations taken on selected elements thought to represent the domain are usually afflicted with selection bias. Peterson *et. al.* (1999) and Paulsen *et. al.* (1998) discuss examples in natural resource monitoring of non-probability samples that do not adequately represent the response over the domains of interest. Other sources of bias include survey non-response (which can include landowner denial of access in spatial studies) and measurement error. Selection bias and other sources of bias and adjustments for bias are not addressed in this study. Simcox *et. al.* (2004) demonstrate an analysis of representativeness of non-probability samples for water quality on USGS monitoring stations in a New Hampshire watershed, using techniques similar to post-stratification.

Though not recommended, non-probability samples do happen to good agencies, frequently. Doing a proper sample design and survey can be prohibitively expensive. An agency may have some observations from pilot studies, or from index monitoring stations that might hopefully represent a carefully specified subset of the domain, or from observations from encountered phenomena. While any extrapolation from the observations to a domain would most certainly have to be treated as preliminary without a probability sample design, a stakeholder would naturally be inclined to ask how much the extrapolation from the observations in the non-probability sample might change if a similar sample of elements on the domain had been observed. This is the question of focus in this research.

The crux of the question is that there is no sampling distribution model to characterize the sample statistic from the methods used in practice today. Survey design-based methods require that the inclusion probabilities or densities be specified. For the non-probability sample, technically, the inclusion probabilities of the observed elements are one, meaning that the event that the observed elements were included in the sample is a sure event, providing no information about the long-run frequency behavior of the sample based on those observations.

While model-based methods conditional on observations do not require the inclusion densities, there are the disadvantages of having to defend a prescribed distribution model. Hansen, Madow and Tepping (1983) warn that model misspecification can be difficult to detect. They demonstrate that a model-based extrapolation may be asymptotically inconsistent when the model is misspecified. Furthermore, even when a model is correctly specified, one almost always needs to estimate the parameters. The outcome of the estimate depends on what elements were observed. Smith (1984) cites several authors including Kish on design effects (the effects of how the elements were selected for observation), customarily ignored in conditional inferences, that can result in substantially poor confidence interval coverage or bias.

The other difference between the model-based approach and the design-based approach is the interpretation of the variability estimated. The model-based approach explicitly models variability as induced by the underlying random process of the response. The design-based approach explicitly addresses the variability induced by the randomization of the sampling process. For the question of focus – how much would the agency's extrapolation vary over other sets of observations – the variability of concern is that induced by which elements were observed. As noted, the problem with the non-probability sample is that current available methods do not apply, because the specification of the inclusion probabilities (density) is pathologic.

Not a lot of researchers have addressed non-probability samples. Where non-probability samples have been employed, the application is essentially using the additional information from the non-probability samples as auxiliary data to adjust estimates based on observations from an additional subsequent probability sample. Brus and de Gruijter (2000) employ a non-probability sample to do a post-stratification adjustment. When the response of interest is linearly correlated to an auxiliary variable, post-stratification is a process to adjust a sample statistic of the response for bias (error), in which the difference between the total or mean estimates for the auxiliary and that of the known population total or mean of the auxiliary variable is used to estimate a correction to the sample statistic. In their application, Brus *et. al.* use the kriged value (the best linear unbiased predictor interpolation) conditional on the non-probability sample at the probability sample locations as the auxiliary information. Overton, *et. al.* (1993) also use a non-probability sample (described as "found" data) to do similar post-stratification adjustments on a probability sample. They explore using non-probability data for augmenting sample size and also for inference on a response not available from a probability sample on the same region.

In the present research, the objective is to provide an assessment of the variability of an extrapolation from a non-probability sample to the rest of the domain, perhaps as a preliminary data point, accepting that the extrapolation includes risk of selection bias nevertheless. Because the non-probability sample has pathologic inclusion densities, there is not a basis for the application of the conventional design-based metrics. The assessment of bias and efficiency (variance) on the extrapolation does not have a long-run frequency basis of interpretation as described by Godambe and Thompson (1988). However, the observations and the order of the observations as provided by their arrangement on the domain will impact what kind of bias (in the sense described by Royall and Cumberland) an extrapolation would have when the observations are taken on a regionalized response – a continuous response with a covariance structure such that locations in close proximity are more similar than locations beyond the range of covariance. A question of interest that has useful and natural interpretation is how much the extrapolation would vary on other sets of observations that are arranged similarly to the given non-probability sample. From this natural idea, we develop a basis for characterizing the stability of the extrapolation (conversely, the variability), by first providing a precise, mathematical characterization of a class of patterns of similar arrangement. In a subsequent paper (Dissertation Chapter 4), methods are explored for estimating the variability of an estimate over a class of patterns.

1.2 Developing a method to characterize classes of point patterns

Many spatial point patterns arise as the result of some stochastic point process. Examples abound in ecology – e.g. patterns of locations of trees and nests (Ripley (1981)). Examples are found in astronomy (Ripley (1977)). Studies of disease transmission and extent will involve arrangements of locations (Besag and Diggle (1977)). Quite often, a goal is to characterize a point pattern with the ultimate objective of modeling some underlying stochastic process (such as dispersion or inhibition or competition). The metrics that characterize a stochastic point process can be applied to specify a class of similarly arranged patterns, though the objectives differ. The goal of this study is to develop a way to characterize classes of similarly arranged elements on spatial domains. (The terms elements and points are used interchangeably throughout.)

A class of similarly arranged elements or a sample process can be characterized by metrics that partition point patterns for useful features such as regularity of spacing or clustering of points within the pattern. Ripley (1981 (Ch. 8)) discusses some popular point pattern metrics and provides numerous examples on data such as tree and nest locations. Many point-pattern statistics are based on inter-point distances or on nearest neighbor distances. The Clarke-Evans (CE) statistic, the sum of the nearest-neighbor distances, is one of the earlier devised metrics Ripley (1981). Ripley (1981) also describes two functions that characterize the second-moment properties manifested by inter-point distances and distances from arbitrary points to points in the point pattern. These are P(t) – the cumulative distribution of the distance t to the nearest event from any arbitrary point; and the popular K(t) – the average number of events within a distance t of a point, normalized by the intensity λ (expected rate of events per unit area) of the process, so that λ K(t) is the expected number of events within a distance t. For a 2D spatial Poisson process, $\lambda\, K(t) = \pi\ t^2$. Ripley (1981) cites several studies that suggest Besag's linearized version of K(t) $\left(B(t) = t\sqrt{\pi/\lambda}\right)$ is more sensitive to departures from Poisson behavior than CE or P(t) (Ripley (1981)). The distribution of quadrat counts is also sometimes employed (Ripley (1981)). Quadrat counts record the number of events occurring within a fixed-dimension fixed-area frame (usually circular or square), for random placements of the quadrats (as described in Ripley (1981)). Statistics on the Dirichlet tiles can also be useful summaries. The Dirichlet tile on a point is the enclosed area defining the part of the domain closer to that point than any other point in the point pattern.

A useful point-pattern metric partitions the universe of point patterns in some way meaningful to the application – in this case, by separating similar arrangements from those not similar enough to that of the purposive sample. The metric is a measurable mapping from an n-vector of locations $\underline{s}$ to a finite or non-negative subset of the Real numbers. Refer to Figure 3-1.

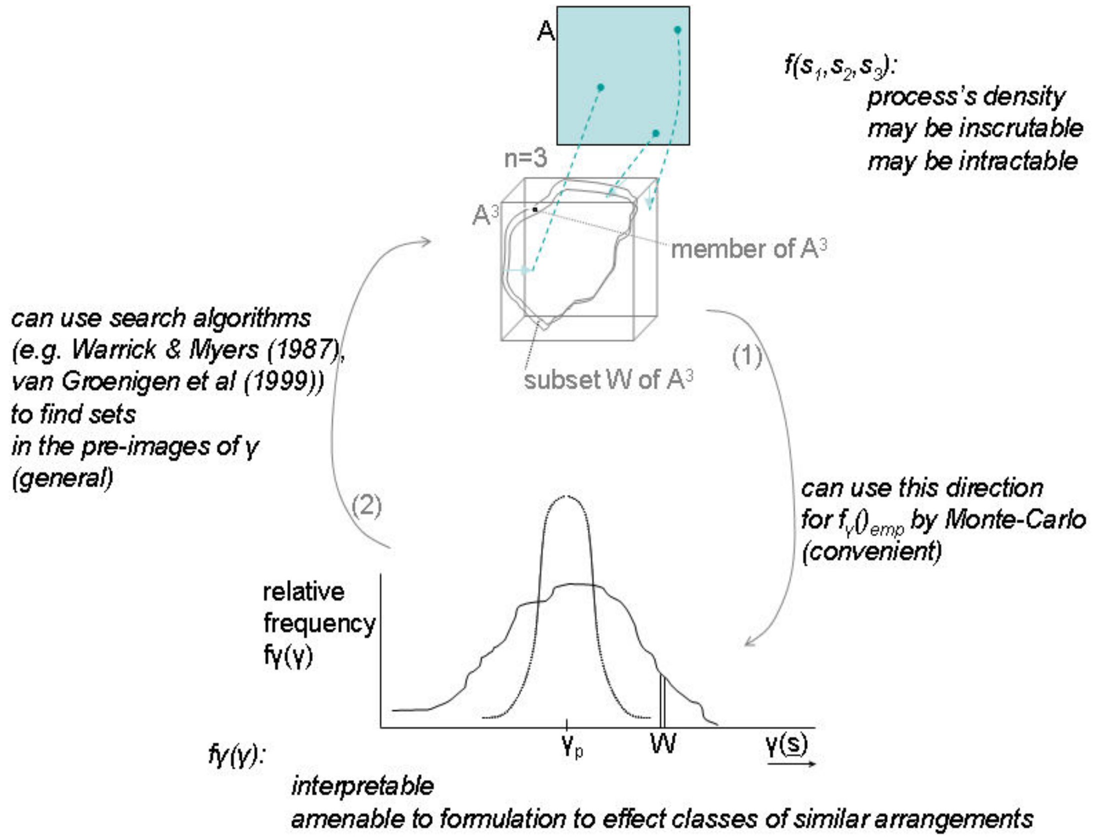

Figure 3-1 Schematic of dual mappings from set space A^3 to metric mapping $\gamma(\underline{s})$ (arrow (1)) and from metric mapping to pre-images in set space A^3 (arrow (2)). The conceptual wedge W in the set space represents the pre-image of the event that the metric falls within the interval labeled W. A formulation of metric density $f(\gamma(\underline{s}))$ defines a class of point patterns in A^3 (arrow (2)), providing a unique (a.e.) reference on which to characterize variability of an extrapolation from a non-probability sample. Specificity of GOF of various metrics is examined by realizing point patterns from representative types of processes (arrow (1)) and examining separation of the empirical metric densities that result.

Consider size-n realizations of some stochastic process, where $n \geq 2$. Let ζ^n denote the n-fold product space of the domain A – where A is some bounded spatial domain. Let $\underline{s}$ denote a member of ζ^n, where the components s_i of $\underline{s}$ are each an ordered pair (x,y) denoting a location on A. Note that this notation of a member of ζ^n treats the components as ordered, although in application, inference on the domain should be invariant to the sample index order of the locations in the sample. The notation simplifies the development that follows, without limiting generality, as long as permutations of sample index are accounted for. That is, in this notation, there are *n!* ways to represent a specific point process outcome *{s₁, s₂, … , sₙ}* (excluding any multiple occurrences at a location). The members of ζ^n will have some probability measure, denoted $f_{sp}(\underline{s})$, depending on the stochastic process from which the point patterns are realized.

Let $\gamma(\underline{s})$ denote a bounded or non-negative metric of the pattern of locations, i.e. a measurable mapping from ζ^n to the Real numbers. Let a default stochastic process be the Complete Spatial Randomness (CSR) process, for which the density $f_{sp}(\underline{s})$ is constant. Let Γ^o denote the range of $\gamma(\underline{s})$ for the point pattern domain ζ^n. For the default process let $f_\gamma^o = f_\gamma(\gamma(\underline{s}))$ denote a measure on the range of $\gamma(\underline{s})$ induced by the measure on $\underline{s}$ for a CSR process ($f_{CSR}(\underline{s})$). This measure is guaranteed by requiring that the metric $\gamma(\underline{s})$ is measurable. The measure $f_\gamma^o = f_\gamma(\gamma(\underline{s}))$ is a probability measure and so is non-negative everywhere on its support (the range Γ^o of $\gamma(\underline{s})$). Also in general and for f_γ^o, $\int\limits_{\gamma\in\Gamma} f_\gamma(\gamma(\underline{s})) = 1$.

Define G to be a subset of Γ^o that is a member of a family of sets that are finite unions of *M* intervals {I_i} covering Γ^o. If the *M* intervals of {I_i} are overlapping, represent G by a finite union of M^* disjoint intervals {J_j}. The set G is a member of the Borel sets generated on Γ^o, so G is measurable with respect to the density f_γ^o on Γ^o. Let the measure of G be denoted as $\Delta(G)$.

The evaluation of the measure $\Delta(G)$ is expressed as follows:

$$\Delta(G) = \int\limits_{\gamma\in G} f_\gamma^o(\gamma(\underline{s}))d\gamma = \sum_{j=1}^{M^*} \int\limits_{\gamma\in J_j} I[\gamma(\underline{s})\in J_j] f_\gamma^o(\gamma(\underline{s}))d\gamma = \sum_{j=1}^{M^*} \int\limits_{\gamma\in J_j} f_\gamma^o(\gamma(\underline{s}))d\gamma$$

Because $\gamma(\underline{s})$ is a measurable functions mapping from ζ^n to the Real numbers, the measure that characterizes the point pattern realizations of a stochastic process $f_{sp}(\underline{s})$ induces a measure $f(\gamma(\underline{s}))$ on the metric range. For the CSR process, $f_{CSR}(\underline{s}) \xrightarrow{\underline{s}\longrightarrow\gamma(\underline{s})} f_\gamma^o(\gamma(\underline{s}))$.

The goal, for providing a basis for describing variability of an extrapolation over similarly arranged locations of observations, is to characterize classes of point patterns similar to the purposive sample. A distribution of a class similar to the purposive sample should include outcomes where the range of $\gamma(\underline{s})$ of the class is similar to the metric of the purposive sample's: $\gamma_p \equiv \gamma(\underline{s}_p)$. That is, the interest is for the description of the metric to define a class on the universe of (size-n) point patterns, so that a subset of the metric range would imply certain arrangements of points (the pre-image of the subset) (as indicated by arrow (2)). $f(\gamma(\underline{s}))\xrightarrow{\gamma(\underline{s})\longrightarrow\{\underline{s}\}} f_{\zeta^n}(\underline{s})$ where $\{\underline{s}\}$ denotes some set of point patterns.

To this end, suppose a measure $f(\gamma(\underline{s}))$ is to be specified on the metric range, such that the support Γ of the new measure is a subset of the range Γ^o and such that $f(\gamma(\underline{s})) \geq 0 \quad \forall \gamma \in \Gamma, \Gamma \subseteq \Gamma^o$ and $\int\limits_{\Gamma} f(\gamma(\underline{s}))d\underline{s} = 1$. Let ζ^* denote the inverse image of Γ under $\gamma(\underline{s})$ - *i.e.*- $\zeta^* = \{\underline{s} : \gamma(\underline{s}) \in \Gamma\}$. Since $\Gamma \subseteq \Gamma^o$, $\zeta^* \subseteq \zeta^n$. The Appendix is intended to show that there is a unique (a.e.) measure on ζ^n induced by a measure $f(\gamma(\underline{s}))$ defined as above. Thus specifying a certain density $f(\gamma(\underline{s}))$ on the metric $\gamma(\underline{s})$ defines a class of point pattern arrangements on the universe of point patterns. Specifying the class of point patterns provides a reference against which the variability of an extrapolation from a non-probability sample can be measured.

The joint density $f(\underline{s})$ may not have a closed-form expression. The Strauss model (Strauss (1975)) characterizes the joint density of point patterns realized from inhibiting or clustering stochastic processes as a probability density on the number of nearest neighbors (y), where two points are neighbors if they are within a fixed distance r. In this model, $f(s_n| s_1, \ldots, s_{n-1})$ is characterized by addition of t_n nearest neighbors by the addition of s_n, or $t_n = y_n - y_{n-1}$. Strauss (1975) shows that the unconditional form of the joint density is $f(\underline{s};v) = f(\underline{s};v=0)\dfrac{e^{vy}}{M_Y(v)}$, where v is expected number of neighbors and $M_Y(v)$ is a normalizing function. Kelly and Ripley (1976) derive a recursive formulation of Strauss' model $f(\underline{s};a,b,c) = ab^n c^{y_n}$ which they then use to propose a birth and death process to realize a Strauss model. Geyer (1999) shows that, conditional on the number of points n, in the limit on a parameter of expected number of neighbors v, the Strauss process is either completely regular or a one-ball cluster.

In these developments, there is typically interest in describing a stochastic mechanism. In the present application, there is not a stochastic mechanism associated with the non-probability sample. The goal is to develop a set measure in order to give explicit definition to classes of similarly arranged patterns. Class definition is accomplished by partitioning the universe of sets of spatially-arranged points with suitable metrics.

In application, going from some subset *G* of the support of $f(\gamma(\underline{s}))$ to the pre-image $\{\underline{s}\}_G$ in ζ^n that maps to *G* (Figure 3-1 arrow (2)) can be achieved by various search algorithms with varying degrees of efficiency. One strategy is doing an Accept/Reject method of trying out patterns at random, accepting or rejecting them with a probability proportional to the density $f(\gamma(\underline{s}))$ evaluated at the metric $\gamma(\underline{s})$ produced by each. Van Groenigen *et. al.* (1999) use spatial simulated annealing (SSA) to search for patterns with particular properties, by iteratively perturbing an initial (size-n) set (one element at a time), allowing a new sample either when an optimizing fitting function improves or with some (decreasing) probability. The fitting function for effecting the class of point patterns might involve the metric density $f(\gamma(\underline{s}))$ and the observed metric $\gamma(\underline{s})$ resulting after each perturbation. Warrick and Myers (1987) have a search algorithm for achieving particular distributions of point pair distances, by which they take sums of squares of discrepancies in the realized and desired distributions and select an outcome with a minimum sum of squares.

A question of interest in this research involves evaluating various metrics for effectiveness in partitioning the space of size-n sets of points into meaningful classes that would be of interest for comparison with a purposive sample's pattern. To this purpose, it is convenient to start with sets of various representative types – e.g. regular, random, clustered and highly clustered, on which to examine empirical distributions of each candidate metric resulting from the representative types (in the direction of arrow (1) in Figure 3-1). The best metrics will have densities $f(\gamma(\underline{s}))$ with very little overlap from one type to the next. For this study, the representative types of patterns are produced by stochastic processes (described in the Methods section) to generate repeated realizations of each type of pattern to get empirical densities $f(\gamma(\underline{s}))$. Thus, Monte-Carlo methods are used to evaluate specificity of goodness-of-fit (GOF) of arbitrary patterns to classes of various types of patterns. A good GOF means the pattern is typical of those in that type and excludes outcomes more extreme than the 20th and 80th quantiles of the empirical densities of the patterns.

In the process of evaluating metrics, the stochastic processes implemented produce sets of points in the point pattern space, from which the locations are mapped into a one-dimensional metric range (arrow (1)). This is an expedient way to study effectiveness of candidate metrics on the representative types of patterns to evaluate performance for assessing GOF. For the objective of defining a class of point patterns on which to reference the variability of an extrapolation, specifying a density on the metric range provides a most general class formulation that is interpretable and manageable, without relying on stipulating a process to realize that class of points. That is, specifying a density and its support on a metric is a tractable way to explicitly define a class of similarly arranged patterns (where the coercion is in the direction of Figure 3-1 arrow (2)).

The domain is fundamental to what point patterns are possible to observe. The usual domain of point patterns is one of areal extent. For this study, the domain is a square area with no holes. The other domain examined in this study is a linear network – the actual example used is a network of stream segments on the Alsea basin, in Oregon. The frame provided is an ARC shapefile provided by Oregon Department of Fish & Wildlife (ODFW). The Alsea network section is illustrated in Figure 3-2 overlaid with the locations of a non-probability survey taken by ODFW.

On the domain with areal extent, a measure on point patterns can be characterized by densities on point-pattern statistics such as the point-pair distances, or on features of the Dirichlet tiles, or on characteristics of Ripley's K(t) functions, or functions related to these metrics. In the case of the linear network domain, the consecutive-point distances are a useful statistic to derive more concise point-pattern statistics (this point is revisited in the Discussion). The empirical densities of candidate statistics considered (described in more detail in the next section) are produced using Monte Carlo methods. The measure induced on the universe of point patterns characterizes point patterns similarly arranged to a (non-random) point pattern for which there are observations. The characterization of the arrangement of points and estimated response covariance structure are to be applied in a subsequent paper (Dissertation Chapter 4) to characterize the variation in an estimate produced from a class of similarly arranged patterns of elements.

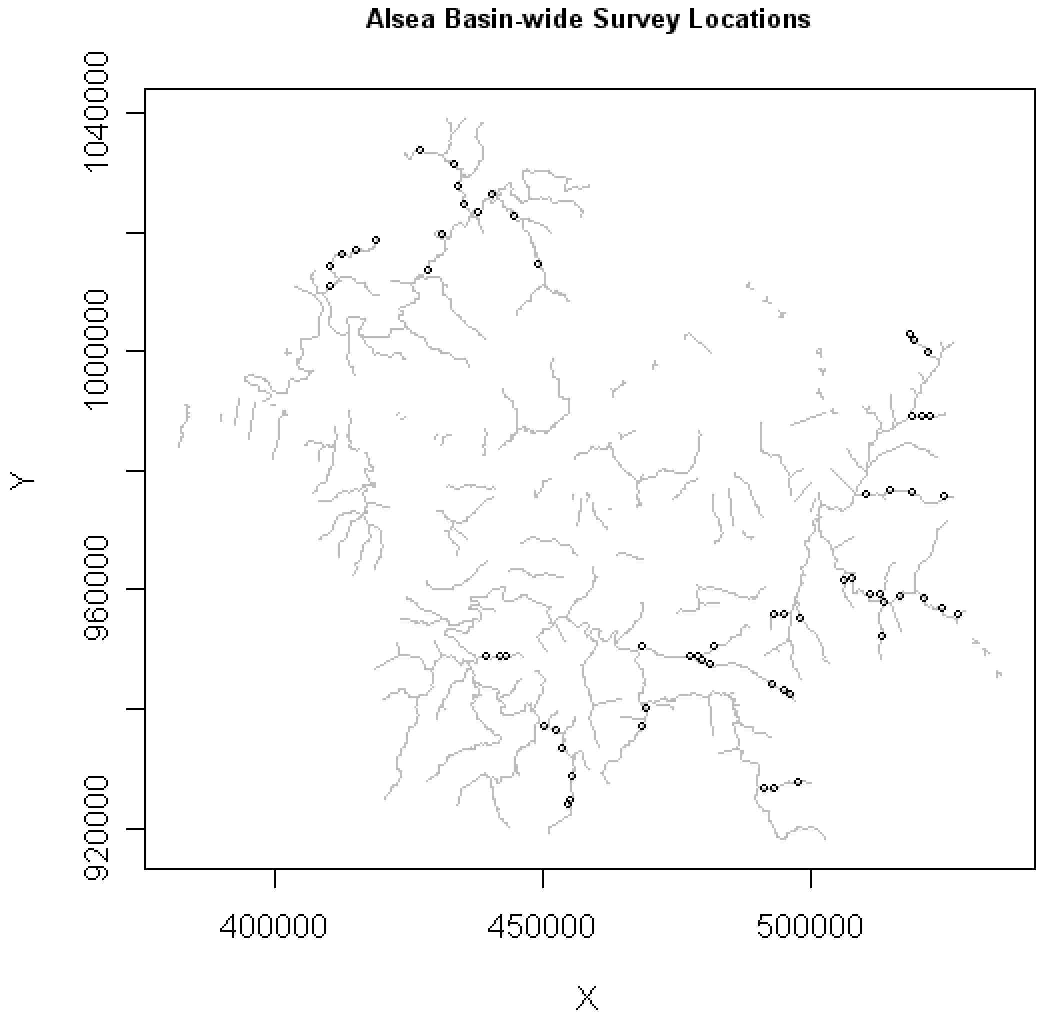

Figure 3-2 A linear-network domain – a section of Alsea basin, Oregon. Points overlaying the network represent a non-probability sample surveyed by ODFW.

2 ASSESSING METRIC SPECIFICITY FOR GOODNESS-OF-FIT TESTS

The goal is to characterize point patterns with arrangements similar to the purposive sample's arrangement. If the arrangements are similar, for isotropic point processes, the distribution of functions and metrics characterizing arrangements should also be similar. For example, the distribution of inter-point distances, the SVB statistic, K(t) functions and derived metrics should be similar. In the following, a sample or set refers to the collection of elements observed.

A density of a point pattern metric imposes a measure on point patterns, as developed in the Introduction. To get a class of point patterns that is useful in the sense of being similarly arranged to the purposive pattern, the density of the metric should be centered on the metric value of the purposive pattern and vanishing for patterns that are considerably different from the purposive pattern. In general terms, similarity means that the patterns match in degree of regularity or clustering or randomness, which can be characterized specifically by the differences in their metric values.

It is possible to construct a density on a metric such that the values of the metric closest to that of the purposive pattern are assigned the greatest density values. Let $\underline{s}_P$ denote the vector of the non-probability sample locations. Suppose $\gamma(\underline{s})$ is some metric (a measurable mapping from ζ^n to the real numbers) that would be useful to partition the universe of samples and $h(\gamma(\underline{s}),\gamma_p)$ some function of the discrepancy between the metric of sample $\underline{s}$ and that of the purposive sample $\gamma_p \equiv \gamma(\underline{s}_p)$. In a general form, a density based on the discrepancy between the metric on the non-probability sample and on some candidate pattern $\underline{s}$ can be expressed by the form $f(\gamma(\underline{s})) = \dfrac{e^{-h(\gamma(\underline{s}),\gamma_p)}}{\int\limits_{s \in g(A)} e^{-h(\gamma(s),\gamma_p)} \partial \underline{s}}$, where the denominator is the normalizing constant that makes the measure on the point pattern $\underline{s}$ integrate to one (so that integrating a function with the constructed measure gives the weighted average of that function). The form is expressed as the exponential of a discrepancy function $h(\gamma(\underline{s}),\gamma_p)$ so that the density is always non-negative.

For example, one way to construct such a measure is to make the density decrease linearly with the increase in absolute deviation from the purposive pattern's statistic. This penalizes deviations proportionately to the size of the deviation. To make the density a proper probability density, divide each absolute deviation by the integral of the absolute deviations of the metric over the range of metrics on the class, so that the probability density function (PDF) integrates to one. In the notation above, $h(\gamma(\underline{s}),\gamma_p)$ would be a function of the natural log of the absolute value of the deviation: $h(\gamma(\underline{s}),\gamma_p) = -\ln(c - |\gamma(\underline{s}) - \gamma_p|)$, where c is the maximum absolute difference for the range of the constructed density. An alternative is to impose a bell-shaped density centered on the purposive pattern's statistic, so that smaller deviations are not penalized as much as greater ones. This can be achieved by making the log density proportional to the negative squared deviation. In the notation above, $h(\gamma(\underline{s}), \gamma_p)$ is proportional to the square of the deviation: $h(\gamma(\underline{s}), \gamma_p) = \frac{1}{c}(\gamma(\underline{s}) - \gamma_p)^2 \quad c > 0$. As for all probability densities, the values would be divided by a normalizing constant so that the density integrates to one.

For this study, goodness-of-fit specificity is analyzed on empirical densities of metrics on classes of patterns realized by various point processes. That is, instead of using a closed form density as described above, the goodness-of-fit is examined by comparing a metric of a pattern to the empirical quantiles of metrics from the representative types of patterns realized by simulated point processes. As will be illustrated on the ODFW non-probability sample, these empirical metric densities can be applied as the working specifications of classes of patterns. That is , similarly arranged classes of patterns can be defined by the empirical metric densities of processes of which metrics close to that of the purposive's happen more frequently and metrics relatively different happen least frequently.

For defining a class of similarly-arranged patterns, any critical value can be set to exclude patterns with statistics in the tails of the density. Tolerance or intolerance to more dissimilar patterns may be field-specific, much the way an acceptable Type I error rate (i.e. what p-value is considered significant) varies between fields, from fairly liberal (for ecological applications) to fairly strict (for medical applications) depending on the objectives of the field. For an assessment of goodness-of-fit (GOF), typically the Type I error rate is relatively large – i.e. the tails are made large so that a candidate is considered consistent with the definition of the class only if the probability of an outcome as or more extreme than the statistic (the discrepancy between a metric and the purposive metric) is relatively large – say, 20% for example.

In this paper, several point pattern metrics are examined for their utility to provide a set measure on sets of elements taken from the universe of possible point patterns on each domain. The point-pattern metrics examined for a continuous domain include an inner product metric applied to the ordered point-pair distances, the Side-Vertex-Boundary (SVB) Dirichlet tile metric, which measures regularity of point patterns, and a metric derived from K(t) functions. On the linear network domain, three metrics are examined: a metric derived from the exponential joint density of consecutive-location distances; the cumulative distribution ("stochastic rank") of distances at a proportion of the sample resolution; and a 2D version of the SVB metric.

Each metric is tested to determine its specificity to point patterns generated by two processes on the areal domain and three processes on the stream network domain – a Complete Spatial Randomness (CSR) process generating uniformly distributed arrangements of points; a simple grid-tessellation stratified (GTS) process (on the continuous domain) or a simple stratified process (on the linear network) generating relatively regularly-spaced collections of points; and a clustering process (on the stream network only) generating clusters of points from a hierarchical distribution of parent and child point element processes. The three processes provide a range of characteristics of point patterns that are considered here as representative examples of classes of similarly arranged sets of elements.

For each of the metrics considered, the empirical density of the metric over one process is used to assess goodness-of-fit of patterns generated by the other processes. That is, the distribution is evaluated by Monte Carlo methods for specificity for assessing GOF to a class of sets meant to be similarly arranged to some realization from the former process. Besag *et. al.* (1977) use Monte Carlo methods to generate confidence bands on various statistics characterizing spatial randomness of point patterns. They use the confidence bands to detect changes in point patterns over time – for example, examining association between occurrences happening close in time vs. close in proximity to test for contagion. They apply the Monte Carlo method for numerous examples, noting that closed-form models of point patterns may be intractable, and that the method is conveniently applied for a study region with any arbitrary ragged outline. They cite Hope (1968), who shows that an MC test for significance would have only a little less power than that of a UMP test.

The six metrics are described in more detail in the subsequent paragraphs.

2.1 Inner-product metric

One approach to measuring similarity is to examine the inner product of ordered distances of an arbitrary set of locations and that of the purposive sample. The idea is similar to looking for registration in sinusoidal signals. For simplicity, assume that an arbitrary set and the purposive set have the same number of points so that the inner product is defined. The normalized inner product of the ordered distances provides a statistic that should be closer to one when two point arrangements match closely (the normalizing constant is the squared norm of the distance vector of the purposive collection). The inner product statistic is produced as follows. Let $\underline{s}$ and $\underline{t}$ denote the dx1 vector of d inter-point distances of the size-n purposive sample and a 2nd size-n sample, respectively. Let $s_{(i)}$ and $t_{(i)}$ denote the ith largest distances of the purposive and 2nd samples. The inner product statistic is $\sum_{i=1}^{n} s_{(i)} t_{(i)} \Big/ \sum_{i=1}^{n} s_{(i)} s_{(i)}$.

For this study, the purposive samples used are in fact simulated from stochastic point processes, although the concepts would be applied where observations had been taken on a non-probability sample.

2.2 Side-Vertex-Boundary (SVB) metric

The "Side-Vertex-Boundary" (SVB) metric measures regularity of the Dirichlet tiles. With perfect regularity, the boundary of a Dirichlet tile would be as close to that of a circle centered on the point as possible. On a unit square study area, the area of each tile centered on each point of the most regularly spaced size-n point pattern would be 1/n. The radius of a circle with this area is $1/\sqrt{n\pi}$. The SVB is the mean squared deviation between distances to points along the boundary of a Dirichlet tile and the radius of a circle of area 1/n (see Figure 3-3). In other words, if $B(D_i)$ denotes the boundary of the Dirichlet tile of the ith point, the SVB is computed by $\frac{1}{n}\sum_{i=1}^{n}\left\{ \int_{B(D_i)} \left(s - \frac{1}{\sqrt{n\pi}} \right)^2 ds \right\}$. This is approximated in implementation as the mean squared deviation between {the distances from the points to the Dirichlet tile vertices, sides, and boundary} and {the nominal distance corresponding to maximal regularity $1/\sqrt{n\pi}$ }. That is, if v_{ij} denotes the distance between the ith point and the jth vertex of the ith tile, and if n_{ij} denotes the perpendicular distance between the ith point and the jth side of the ith tile, the SVB is approximated as $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2r_i}\sum_{j=1}^{r_i}\left(\left(v_{ij} - \frac{1}{\sqrt{n\pi}} \right)^2 + \left(n_{ij} - \frac{1}{\sqrt{n\pi}} \right)^2 \right)$.

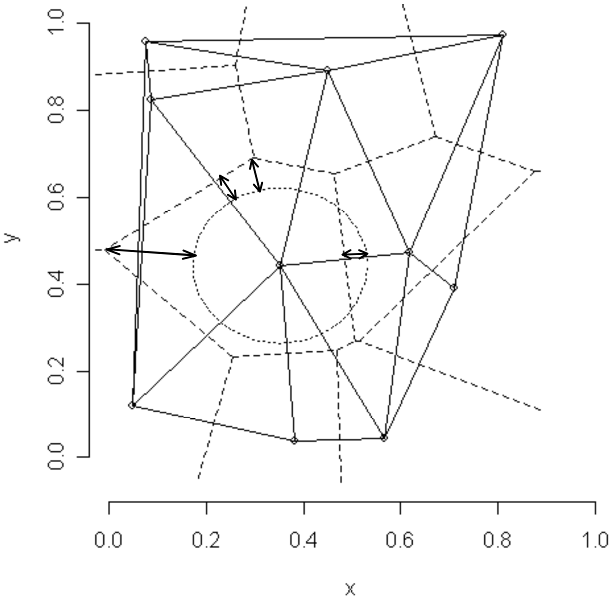

Figure 3-3 The SVB metric measures deviation of the Dirichlet tile boundary from a circumference of a circle of area 1/n for an arrangement of n points on a unit-square area. Dashed lines indicate the boundaries of Dirichlet tile for 10 points. A circle of area 1/n is centered on one of the points. The statistic approximates the average of the squared distance between the tile boundary and the circle boundary as the average distance between the tile vertices and sides and the circle circumference. Examples of the distances are indicated by the segments with arrows for two of the six sides and two of the six vertices.

2.3 K(t) Deviations metric

Ripley's K(t) functions and variants describe the expected number of events (points) within a distance t of some arbitrary location, normalized for the overall intensity of a process. For complete spatial randomness (CSR) as produced by a Poisson process, the expected number of points is directly proportional to the area – i.e. $K(t) \propto \pi t^2$. Clustered processes would have higher K(t) values at distances within the range of the clustering influence ($K(t) > \pi t^2$) and inhibition processes (processes with more regularity) would have lower K(t) values within the range of the inhibiting or repulsive force ($K(t) < \pi t^2$). The functions are essentially variously estimated by method-of-moment estimators (binning inter-point distances and taking the observed average number of points within t_i of auxiliary points), with corrections applied for boundary effects. See Ripley (1981) and Baddeley and Turner (2005). To apply the information in the K(t) functions for partitioning the point pattern universe, it is useful to summarize a key statistic derived from them.

On the continuous domain for the CSR and GTS processes being examined here, it is useful that the GTS process tends to depart from the theoretical value of K(t) almost immediately starting from very small distances, since there are typically few events within a fraction of the average sampling increment of a spatially grid-stratified design. The CSR process on the other hand should not depart much. Figure 3-4 shows typical profiles of the deviation from the theoretical K(t), plotted vs. distance for each process (shown here is one example from many that looked much the same). On examining such plots, it is apparent that the GTS discrepancies have notable magnitude within the range of the average sampling increment. The CSR discrepancies of notable magnitude (random in nature) may not manifest until some distance greater than the average sampling increment. On this basis, a useful metric to separate the CSR and GTS processes is the minimum distance t at which the magnitude of the discrepancy exceeds some threshold value. For this study the threshold is set to two.

The K(t) discrepancy statistic is computed as $min\{t : |K(t)_{obs} - K(t)_{theo}| \geq 2\}$, where *t* is distance; $K(t)_{obs}$ is the fitted K(t) function and $K(t)_{theo}$ is the theoretical K(t) function of a Poisson (CSR) process.

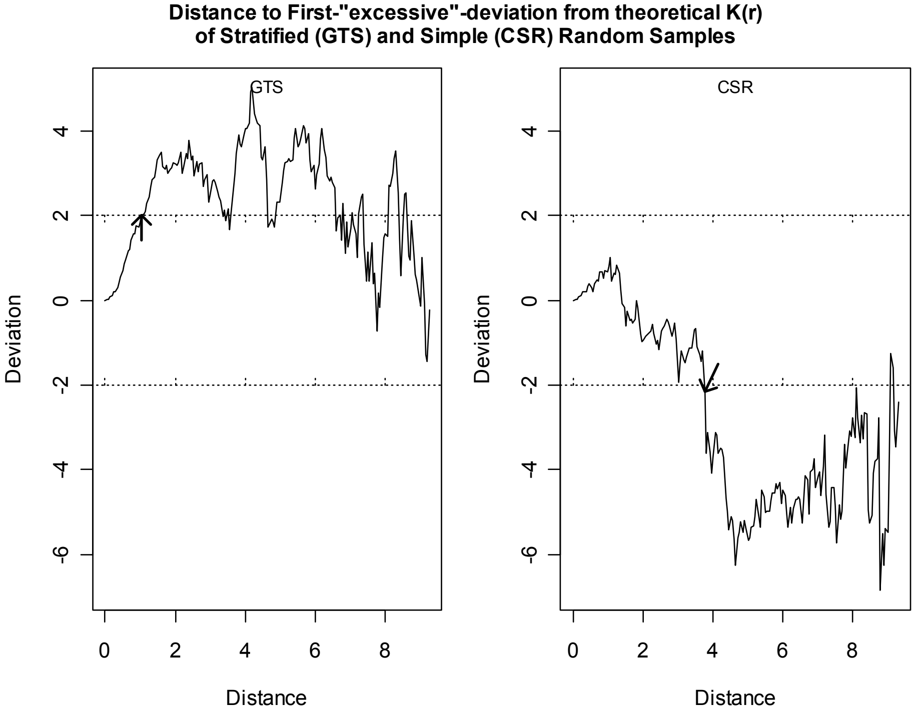

Figure 3-4 Distance of first occurrence of "excessive" deviation from theoretical K(t) function of a GTS and CSR point process.

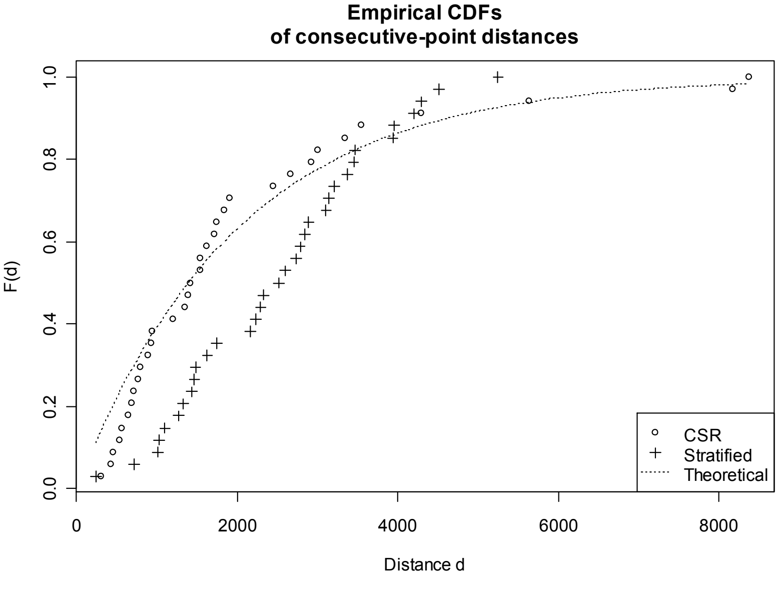

Figure 3-5 Examining deviation from exponential consecutive-point distances along a section of a stream network. The theoretical distribution of the CSR consecutive-point distances is an exponential distribution with rate equal to number of sample points per unit stream length.

2.4 Exponential consecutive-point distance metric

In the case of point processes examined on segments from linear networks, it is more useful to examine distance between two consecutive locations on the network (vs. all inter-point distances). The consecutive point distances include both the up- and down-stream distances from each point to the next points up- and down-stream on the network. The theoretical consecutive-point distance for locations distributed independently uniformly along a linear domain (a Poisson process) is exponentially distributed with mean equal to the sample resolution (the average number of points per unit length). The CSR process should follow an exponential cumulative distribution function (CDF) with only random deviations, while a stratified process will not. This is illustrated in Figure 3-5. That is, the joint density of CSR-point-pattern consecutive distances should be greater than (i.e. more likely to have been produced by a process with exponential inter-occurrence behavior than) that of a more regular pattern. The joint density of the consecutive-point distances is the product of the univariate exponential density of each distance, where the mean is estimated to be the total stream length (as represented by the sampling frame) divided by the sample size: $\prod_{j=1}^{n-1} \frac{1}{\theta} e^{-\frac{d_j}{\theta}}$, where θ denotes the mean (estimated as the sample resolution) and d_j denotes the jth consecutive-point distance.

One metric examined derived from the log likelihood of a pattern is the ratio of average consecutive distance to average intensity, a multiple of the random variable term $\frac{\sum_{j=1}^{n-1} d_j}{\theta}$ in the log likelihood: $\log_e\left(\prod_{j=1}^{n-1}\frac{1}{\theta}e^{-\frac{d_j}{\theta}}\right)=\left(-(n-1)\ \log_e\theta-\frac{\sum_{j=1}^{n-1} d_j}{\theta}\right)$. For the CSR process, this should be centered around 1.

2.5 Stochastic rank metric

Another metric on the point patterns on linear networks comes from two observations about the empirical CDFs of the consecutive-point distances (referring again to Figure 3-5). (1) The empirical CDF of the CSR-process consecutive distances is less than that of the stratified process at the higher range of the support (i.e.- the CSR process is stochastically greater than the stratified at this part of the support). (2) In contrast, at the lower ranges on the support (the lower consecutive-point distances), the CDF of the CSR-process consecutive distances is greater than that of the stratified process. The threshold defining the switch in stochastic rank will depend on the sampling resolution. For the stratified process on the linear stream-network domain, the resolution is characterized by the length of the strata, or in other words, the average consecutive-point distance. A straight-forward metric is the value of the empirical CDF $\hat{F}$ at which a point pattern's distances (d_j) exceed some proportion or multiple of the sample resolution – here, for example, the value is the proportion of the consecutive distances less than or equal to 1/5th the sample resolution (θ), *i.e.* - $\hat{F}(\theta/5)$.

2.6 2D SVB metric

On linear networks, the locations of a collection of points with the highest degree of regularity would be equidistant (where distance is along the stream). On a network of unit total length, the locations in a perfectly regular size-n collection would all be 1/n units apart. The 2D analog of the Dirichlet tile is a line segment (or a bent segment represented by several subsegments) with endpoints at the midpoints between consecutive sample points. The 2D analog of the SVB is the average squared distance between the midpoints (endpoints of a 2D Dirichlet tile) between consecutive locations and the endpoints of nominal tiles of length 1/n centered on each point in the sample.

Let θ denote the sample resolution (stream network total length divided by number of points n). Let t_i denote the stream location of the ith point. Let m_{ij} denote the down- and up-stream midpoints between the ith point and the down- and up-stream sample points consecutive to the ith point. Let $\lambda(\ )$ denote the stream-flow distance between two stream locations. The 2D SVB metric is computed as

$$\sqrt{\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{2}\left(\lambda\left(m_{ij},t_i+\left(I(j=2\equiv upstream)-I(j=1\equiv(downstream))\right)\frac{\theta}{2}\right)\right)^2}\Bigg/\theta,$$

where the operator $\left(I(j=2\equiv upstream)-I(j=1\equiv(downstream))\right)=\pm1$ centers the nominal tile of length θ at the ith location.

3 METHODS

For each combination of process and point-process metric, an empirical distribution of the metric is obtained from simulated realizations of the processes. For each point process, each point pattern metric is observed for 1000 realizations of the point process. On the stream-network, five pairs of point processes are examined to evaluate the effectiveness of each metric to provide a meaningful assessment of GOF – see Table 3-1. For each pair of processes, the proportion of realizations of the designated arbitrary process patterns that would not exceed the 20th and 80th quantiles of the designated purposive process is reported. These quantiles would be the threshold for which the metric of a realization from the process used to produce the hypothetical purposive sample would result in incorrectly rejecting the null that a realization is consistent with the process (Type I error rate). The relatively large Type I error rate is accepted here because a GOF test should provide confidence that a sample is easily consistent with a process – specified by the metric's distribution – which produces the class of similarly arranged samples.

The samples are realized as follows.

3.1 Areal-extent spatial domain

Basic stratified samples were drawn repeatedly from a square area. The strata are defined by a regular 10 x 10 grid of 20 x 20 square strata overlaid on the 200 x 200 area. Each sample contains 100 observations, with one observation per stratum, giving an average sample resolution of 2 distance units. On each trial, the grid is offset by a random amount and wrapped around the end of the area 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. Samples with complete spatial randomness (CSR) are realized for each trial by selecting 100 (x,y) coordinate pairs, each coordinate selected independently and uniformly along the 20-unit x and y dimensions.

3.2 Linear-network spatial domain

A single set of realizations of the five sample processes (stratified, random, clustered, highly-clustered and non-probability-translated), taken on a section of the Alsea basin in Oregon are illustrated in Figure 3-6. Stratified samples were effected by drawing uniform numbers repeatedly and independently from a length equal to the specified sample resolution (27,344 meters (the average intensity of the ODFW non-probability sample)). The number of points depends on the sample resolution. In the results reported, for 1,667,989 meters of stream length in the section of the Alsea basin, there were 61 points in every sample. To determine the location of each point, the stream network is mapped to one line segment, stringing together the various branches, and the strata are then super-imposed on this auxiliary mapping of the stream network. This mapping may introduce an artifact wherever a stratum overlaps a position on the mapping where two disjoint stream reaches are strung together. No attempt has been made to evaluate the influence of this artifact on the results. The location on the stream network then involves looking up the appropriate segment corresponding to the position along the mapping, and determining the position along the segment, from the segment's up-stream node, corresponding to the portion of the random offset within the stratum overlapping the segment within which a stratum location falls on the mapping.

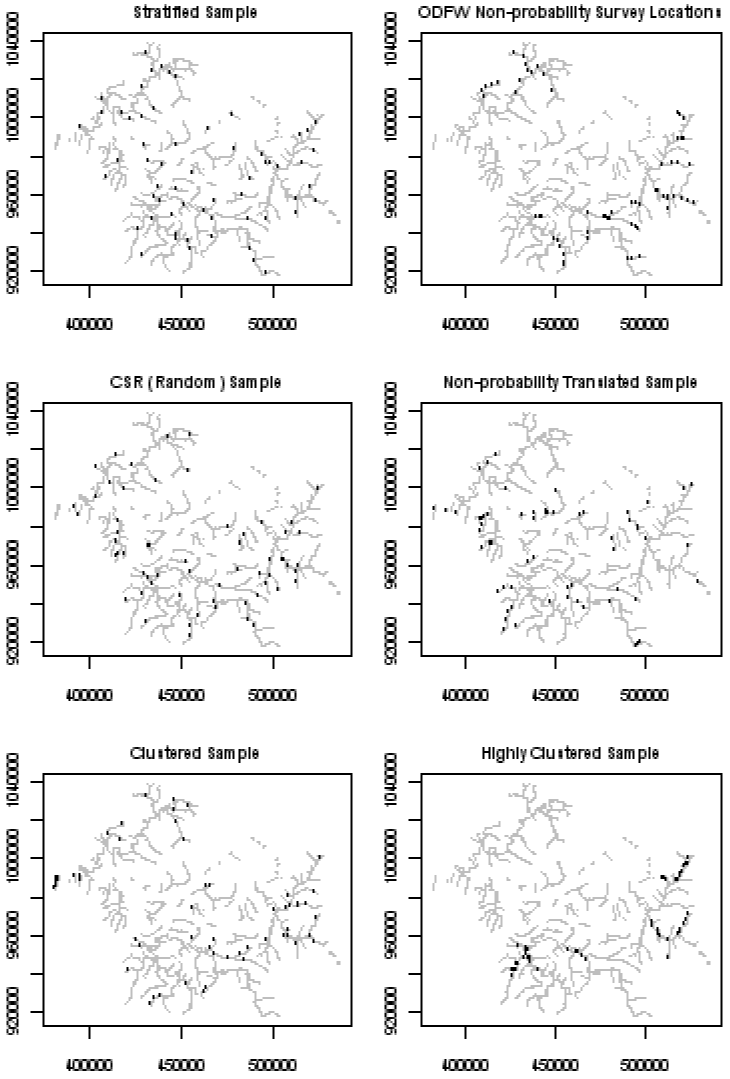

Figure 3-6 Examples of patterns of points realized on a section of the Alsea basin in Oregon. The clustered and highly clustered patterns are generated with 20% and 15% parent points and average child point dispersions of 1% and 1/3 % of total stream length, respectively. See text for details.

A similar logic is used to effect CSR samples on the linear network. In this case, sample size is again determined by dividing the total stream length by the sample resolution and rounding the number. This number of independent uniform random numbers are drawn from an interval with length equal to the total stream network length. As above, there is a mapping of the branches in the stream network to one line segment. The locations are determined by looking up the segment corresponding to the distance along the mapping determined by each random number, and then determining the offset within the segment from its upstream node as the remaining distance from the segment node to the random location within that stream segment's interval on the mapping.

The clustered process is produced by a hierarchical model. A specified proportion of the size-n locations are assigned to be parent locations. The locations are determined as for the CSR process. For each parent location a Poisson distribution is used to choose a random number of child locations to be dispersed near the parent location. In order to keep the sample size constant, more children are added, randomly choosing the number of additional children by the Poisson distribution, until the total of the parent and child locations makes a size-n sample. The last parent to get additional child points when the number of total locations would exceed *n* is assigned a reduced number of child points to limit the sample to *n* locations. The child locations are dispersed by random offsets that follow an exponential distribution with a mean that controls the degree to which the child locations cluster around the parent. The offsets are alternately added to or subtracted from the parent location to disperse the child points in both directions (up- and down-stream) from the parent. The clustered and highly clustered patterns are generated with 20% and 15% parent points and average child point dispersions of 1% and 1/3 % of total stream length, respectively.

The non-probability-translated samples are obtained by translating the linear mapping of the ODFW non-probability locations a random amount and mapping back to the stream segments, as for the other processes.

The various statistics that characterize the two processes on the areal and linear-network domains are computed as follows.

3.3 Areal extent domain metrics

The inner-product metric is the inner-product of the point pattern's ordered distances with that of the purposive point pattern (for this study, generated with one of the point processes), normalized by the square norm (the square length) of the ordered distances of the purposive pattern (the inner product of that vector with itself).

The SVB metric is obtained from an R routine provided by Don L. Stevens, using the R-package "deldir" to generate the Dirichlet tiles for the point pattern (with (x,y) point coordinates scaled to a unit-square area). The routine extracts the distances to the tile vertices and boundaries, and computes the mean squared deviation from the nominal distance of a most regular pattern ($1/\sqrt{n\pi}$ on a unit square).

The K(t) deviations metric is obtained by first getting the estimated K(t) functions using the R-package "spatstat" (Baddeley et. al. (2005)), and then observing the minimum distance at which the average of the estimated K(t) functions deviates from the theoretical K(t) function by 2 or more units.

3.4 Linear-network domain metrics

The log-exponential-fit metric is the ratio of the average of the consecutive-point distances to the average sample resolution. This is the critical part of the log-likelihood, computed as $m\log(\gamma)-\gamma*\sum_{i=1}^{m}d_i$, where *m* is the number of consecutive-point distances, d_i is the ith consecutive-point distance, and γ is the rate of the exponential density – estimated to be the sample size divided by the total stream length of the frame.

The stochastic rank metric is produced using the empirical CDFs of the consecutive-point distances (each distance's rank divided by the total number of distances). The metric is the value of the empirical CDF at which the consecutive-point distance exceeds 1/5th the average sampling resolution.

The 2D SVB is computed as the ratio of the square-root of average deviations relative to the sample resolution. The numerator inside the square root is the average of the squared differences of the along-stream locations of midpoints between sample locations and of the locations corresponding to the nominal tile boundaries – the locations half the sampling resolution distance up- and down-stream from the sample point locations.

4 RESULTS

4.1 Metrics on point patterns on a continuous domain of areal extent

In the best cases, the range of the metric of patterns from one process contains barely any support for the density of the same metric resulting from the other process. Figure 3-7 shows the histograms of the inner-product metric for a situation in which the purposive sample is a spread-out collection of locations (simulated by a GTS point process) for 10000 realizations each of the GTS and CSR processes. The metric is the normalized inner product of the sorted distances with that of the GTS purposive sample. A metric equal to one indicates an exact match of the sorted distances with those of the GTS purposive sample.

Tested for goodness-of-fit where a pattern is considered not atypical when its metric is within the range between the 20th and 80th quantiles of the metric of the class of patterns to be matched, CSR realizations would have an acceptable GOF for a regular class of patterns (where the points are spread out – as in the GTS process) about 14% of the time when the inner product metric is used. In fact, some of the time a CSR process will produce a pattern that appears spread out just by chance. On the other hand, GTS realizations never fail the GOF test by the inner product metric when the class of patterns is completely random (CSR). This is revisited in the Discussion.

The SVB metric for CSR realizations is less than the observed maximum of the GTS process only 49 out of 10000 times (about 0.5 % of the time). The distributions of the SVB metric for each process are summarized in the histograms in Figure 3-8. Neither process's realizations pass a GOF test based on the SVB empirical density from the other process's realizations. The SVB metric is an excellent discriminator between patterns from the GTS process vs. those from the CSR.

Figure 3-9 shows the histograms of the K(t)-derived "discrepancy-distance" metric – the minimum distance t at which the magnitude of discrepancy between the theoretical and observed values of normalized expected number of events exceeded a threshold (set to two). The CSR discrepancy-distance metric was less than the maximum GTS only 0.77% of the time. This discrepancy distance is an excellent GOF metric to eliminate CSR realizations from a class of GTS-produced patterns (or vice versa to reject GTS realizations if the purposive sample has a CSR-like arrangement).

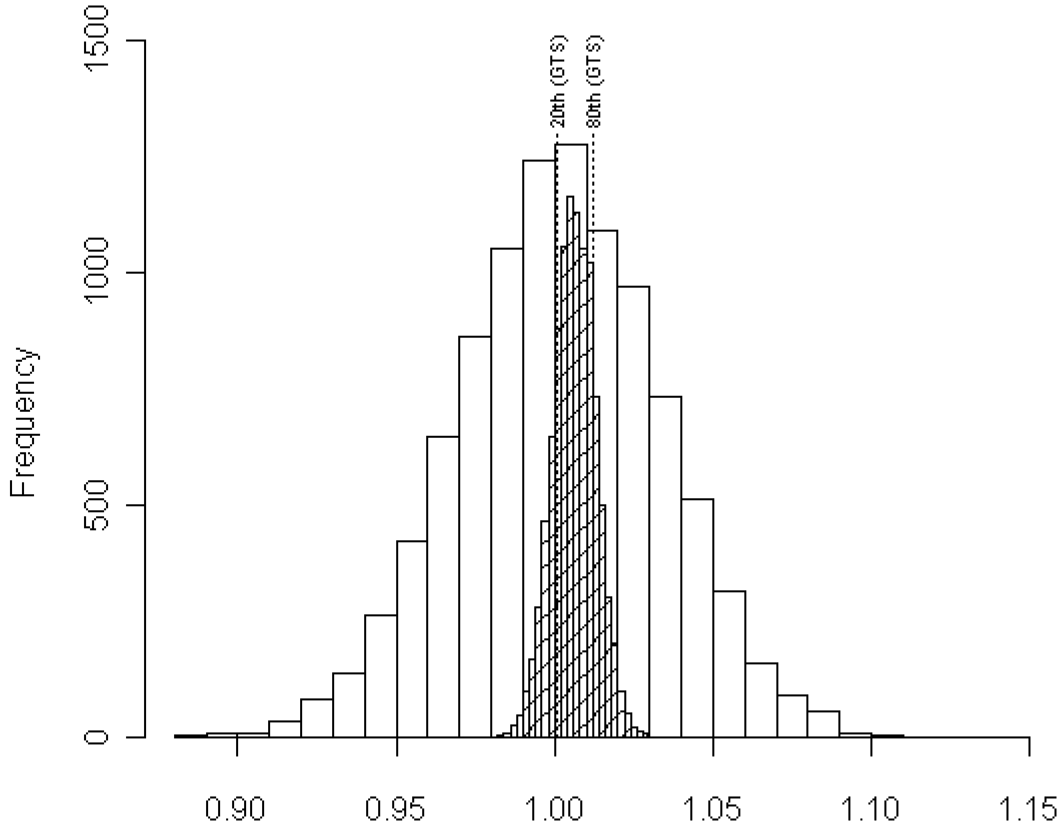

Figure 3-7 Histograms of the inner product score from 10000 GTS and CSR point patterns; the narrower bin intervals on the GTS histogram and wider bin intervals on the CSR histogram represent approximately 1/20 of each of their respective ranges (both areas are 10000 units).

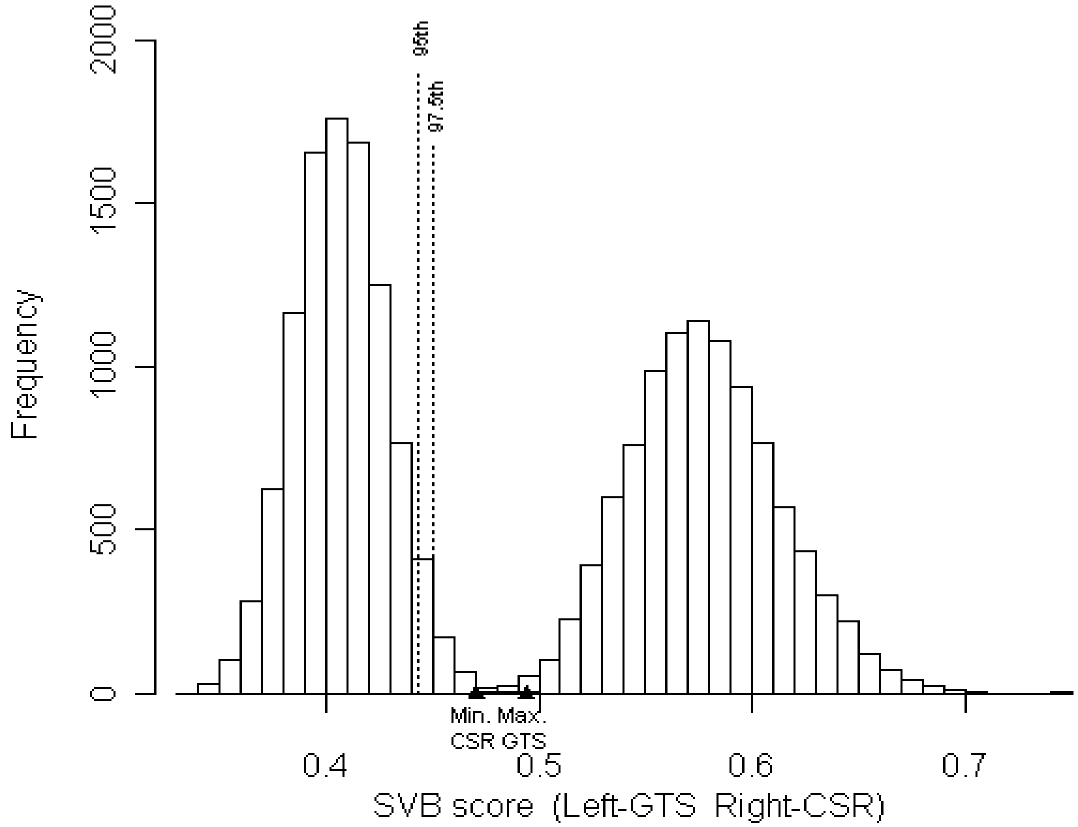

Figure 3-8 Side-Vertex-Boundary (SVB) histograms show very little overlap, making this an excellent GOF metric to discriminate more regular (GTS) patterns from more random (CSR) patterns.

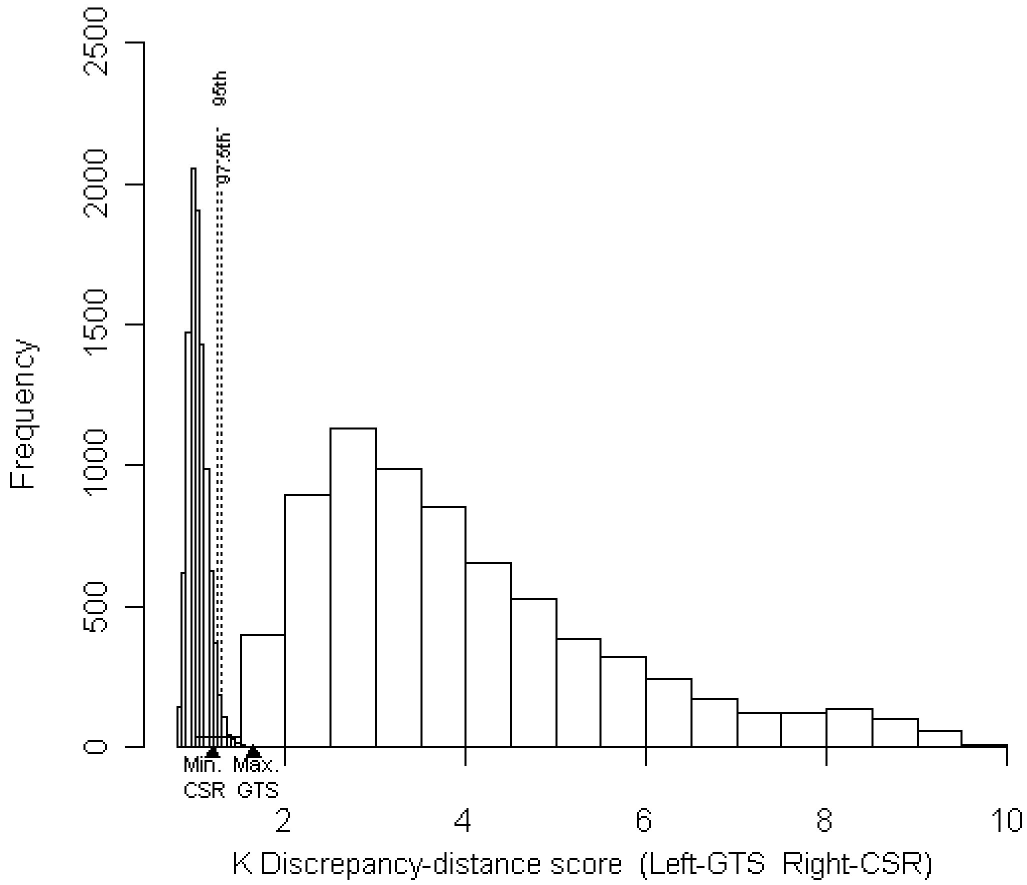

Figure 3-9 Histograms of K discrepancy distance score (the minimum distance at which the observed K(t) deviates by 2 or more units from the theoretical K(t)) show little overlap between outcomes for spatially regular (GTS) and more random (CSR) point patterns.

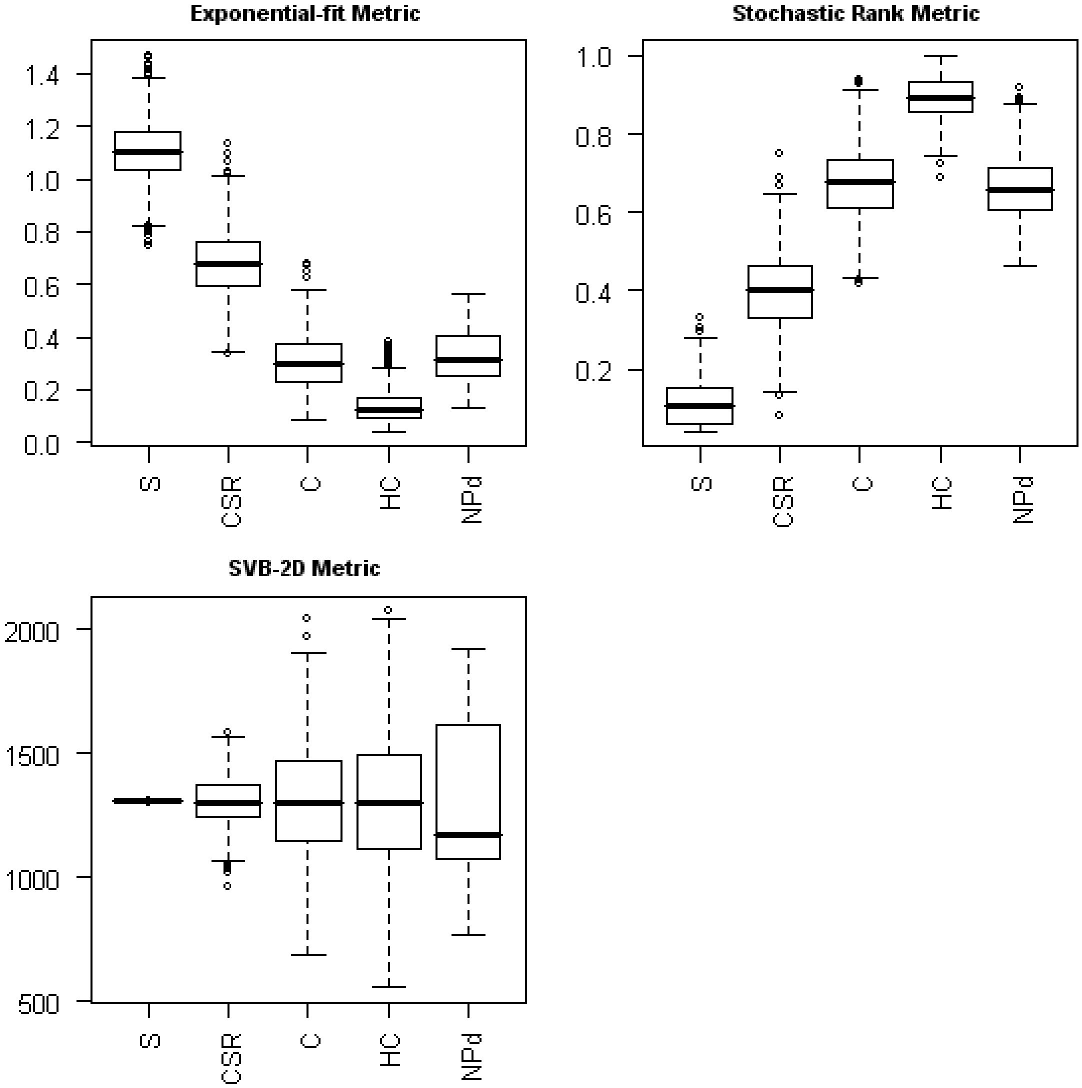

Figure 3-10 Boxplots of the linear stream network metrics examined. S= Stratified; CSR= Random; C= Clustered; HC= Highly clustered; NPd= derived from ODFW non-probability sample.

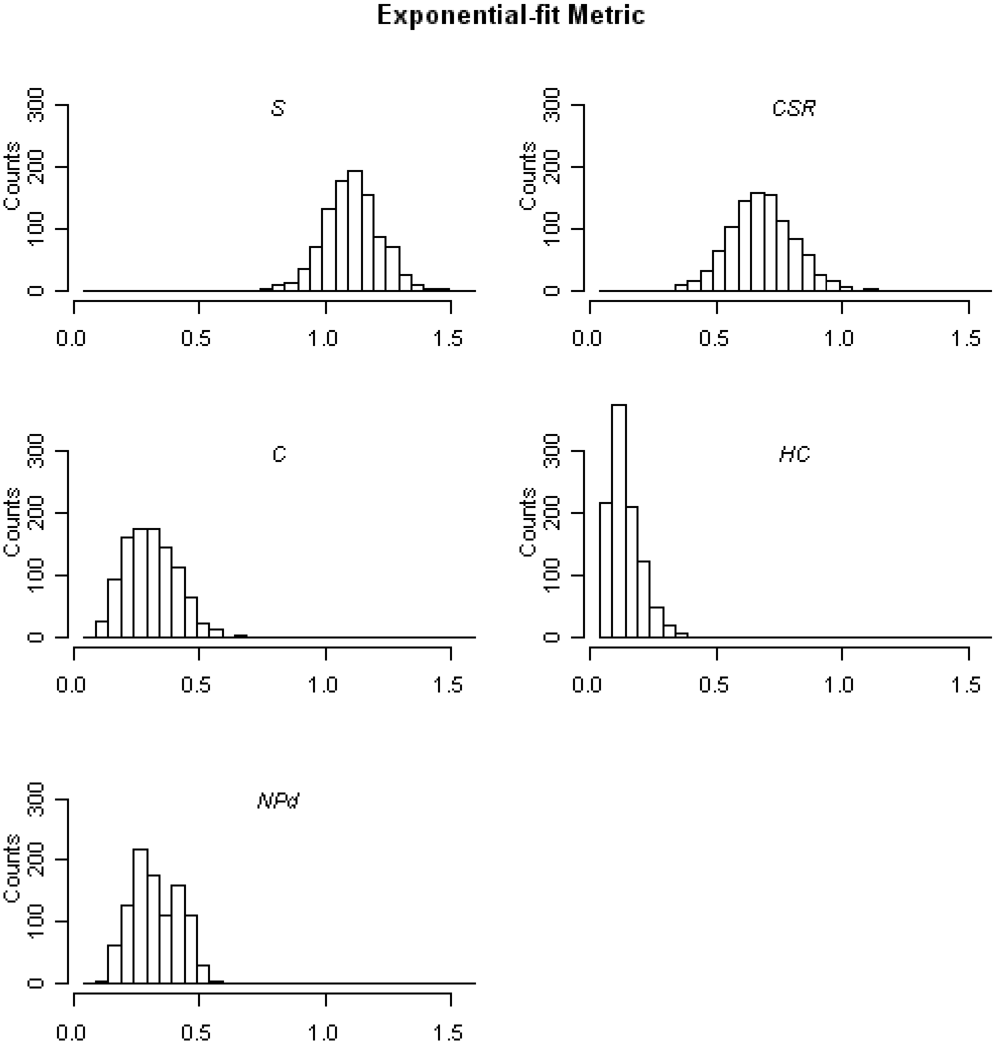

Figure 3-11 Histograms of the exponential-fit metric for point processes realized on the linear network domain. S= Stratified; CSR= Random; C= Clustered; HC= Highly clustered; NPd= derived from ODFW non-probability sample.

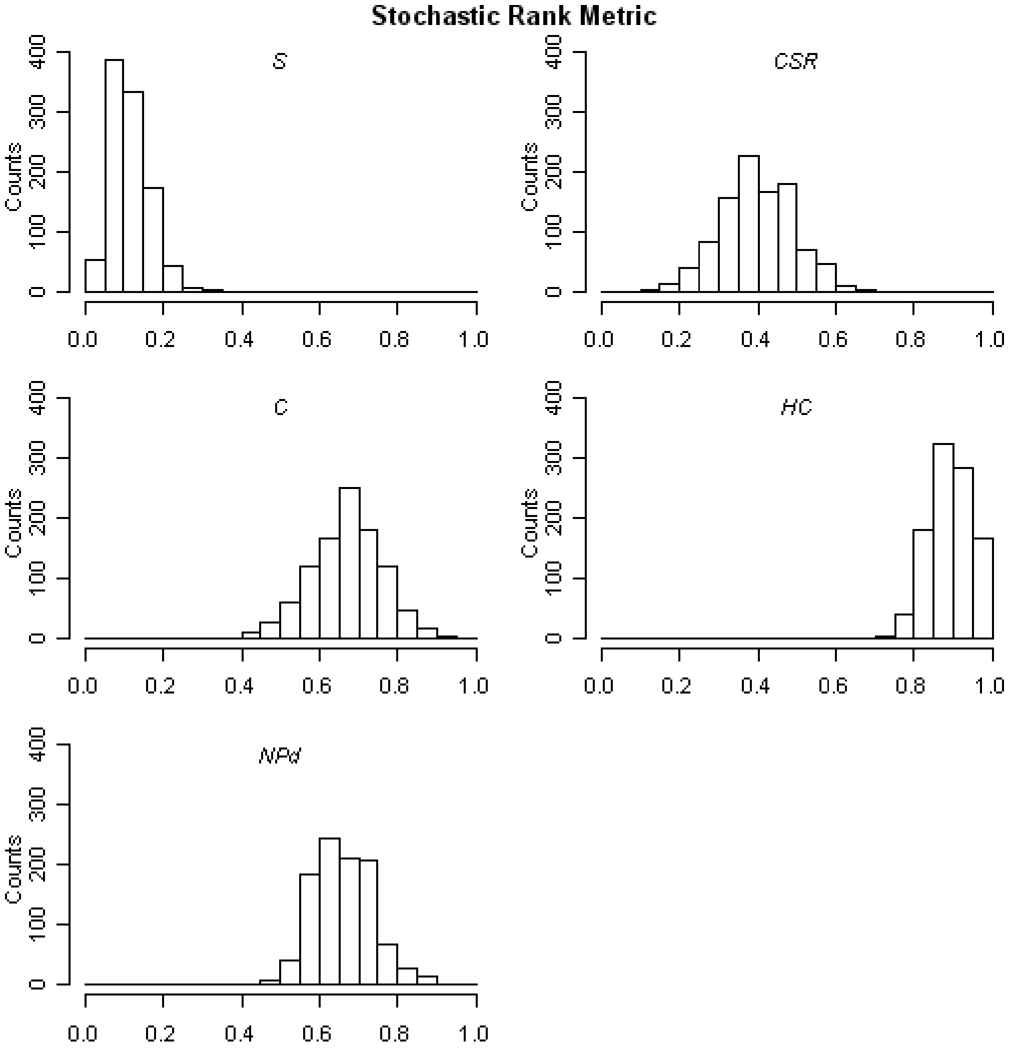

Figure 3-12 Histograms of the stochastic rank metric for point processes realized on the linear network domain. S= Stratified; CSR= Random; C= Clustered; HC= Highly clustered; NPd= derived from ODFW non-probability sample.

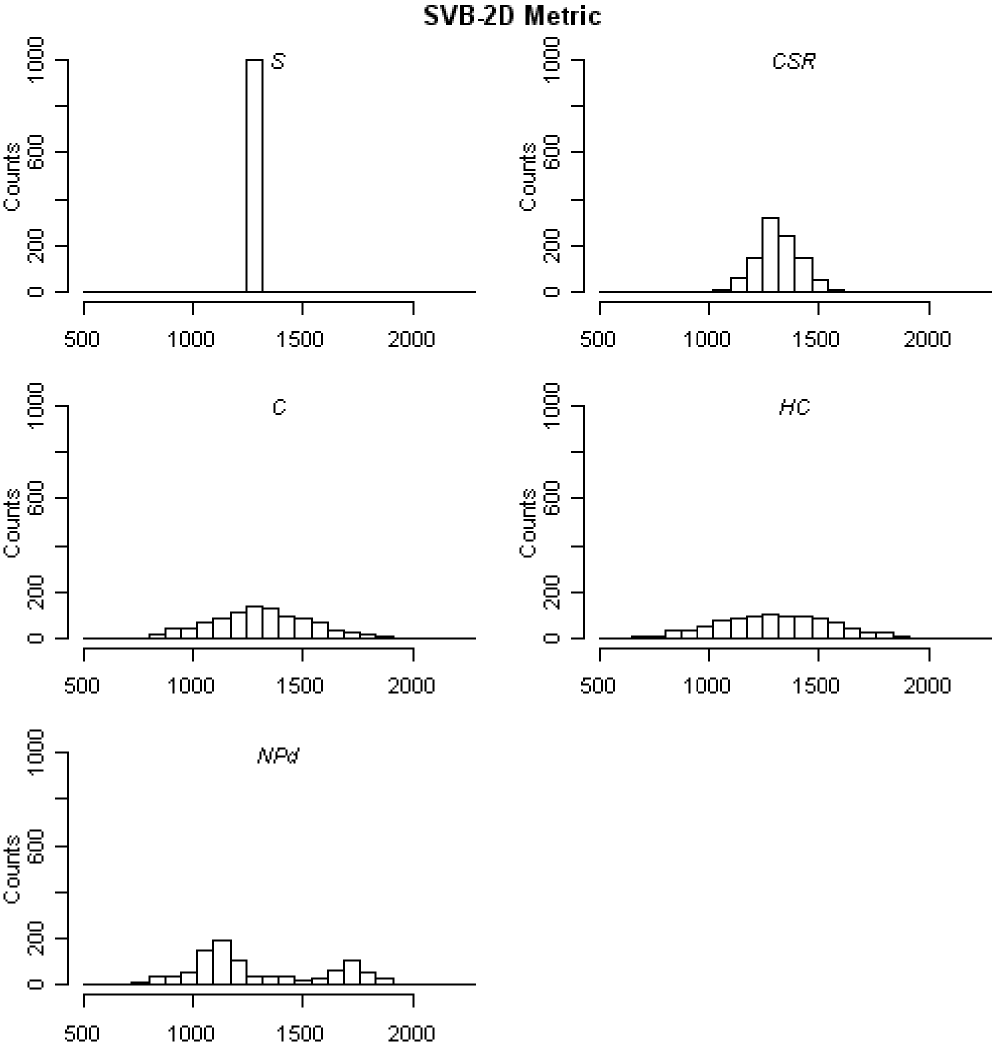

Figure 3-13 Histograms of the 2D Side-Vertex-Boundary SVB metric for point processes realized on the linear network domain. S= Stratified; CSR= Random; C= Clustered; HC= Highly clustered; NPd= derived from ODFW non-probability sample.

4.2 Metrics on point patterns on a linear network

Figure 3-10 shows boxplots of metrics computed on 1000 realizations for each the five point processes produced on the Alsea stream network. Figures 3-11 to 3-13 show the histograms for the three metrics for each of the point process patterns. Table 3-1 reports the proportion of time that a pattern from one process (the arbitrary pattern) would be considered consistent with the class of patterns designated as similar to a purposive pattern. A pattern is judged not atypical if its metric does not exceed either the 20th or 80th quantiles of the purposive class' metrics.

Table 3-1 Assessment of metrics for effectiveness in evaluating goodness-of-fit (GOF) of point patterns realized from various processes. S= Stratified; CSR= Random; C= Clustered; HC= Highly clustered; NPd= derived from ODFW non-probability sample.

Metric	Arbitrary	Purposive	GOF	Conclusion
Log-exponential	S	CSR	0.003	A pattern with regularity is not typical of a class of more random patterns.
	CSR	S	0.004	A typical random pattern is not consistent with a class of patterns with regularity.
	C	CSR	0.007	A clustered pattern is not typical of a class of more random patterns.
	C	NPd	0.568	A clustered pattern is similar to patterns derived by translating the ODFW non-probability sample.
	HC	NPd	0.072	A highly clustered pattern is not consistent with a class of patterns derived by translating the ODFW non-probability sample.
Stoch. Rank	S	CSR	0.002	(same as Log-exponential)
	CSR	S	0.004	"
	C	CSR	0.013	"
	C	NPd	0.524	"
	HC	NPd	0.002	"
SVB 2D	S	CSR	1.000	Random patterns include patterns with regularity by chance (but only about 1-2% of the time – see next entry)
	CSR	S	0.015	A typical random pattern is not consistent with a class of patterns with regularity.
	C	CSR	0.289	A clustered pattern is not typical of a class of more random patterns.
	C	NPd	0.804	A clustered pattern is similar to patterns derived by translating the ODFW non-probability sample.
	HC	NPd	0.726	A highly clustered pattern is similar to patterns derived by translating the ODFW non-probability sample.

5 DISCUSSION

Assessing GOF of point patterns to classes of point patterns allows us to consider the amount of information or precision an extrapolation from a particular arrangement of observations might afford us. Table 3-2 shows the empirical sample-process variances of an estimate of the total response, produced on the realizations from each class of patterns, for a moving-average response simulated along the Alsea network. The estimate is the scaled sum of the observations, where the scale is the inverse of the sampling intensity. Not surprisingly, the more clustered patterns have higher variance than the more spatially balanced patterns. Estimating the variance for a class of patterns is developed in a companion paper (Dissertation Chapter 4).

Table 3-2 Empirical sample process variances

Process	Empirical sample process variance
Stratified (S) (more regular patterns)	1,113,010
Random (CSR) (typically neither clustered nor regular)	1,221,466
Clustered (C)	2,376,003
Highly Clustered (HC)	4,012,697
ODFW Non-probability-derived (NPd)	1,673,838

The table illustrates that describing the variability of an extrapolation for similarly arranged patterns depends on how the class of similarly arranged patterns is defined. For example, any pattern could have come from a CSR class, although the more highly clustered and more spatially balanced patterns in that class are less typical of that class. This is indicated by the metrics of these realizations relative to the range in the CSR class, or in the case of the 2D SVB for the stratified process, as indicated by the narrow range of the metric for these realizations relative to the range overall (this is addressed again below). The variability of clustered or highly clustered patterns is more extreme than the CSR class, on a regionalized response such as the moving-average simulated response. This happens because if a cluster of locations happens to fall in a part of the domain with an extreme high or extreme low response, that range of the response is over-represented. The resulting estimate over- or under-estimates the overall response and both extremes are not infrequent outcomes in the class of clustered patterns.

The ODFW non-probability sample (with locations arranged as shown in Figure 3-2) is clearly clustered, but not as clustered as the class of patterns produced with tight dispersion around the parent points (the HC class). The metrics computed on the actual ODFW set of locations observed on Alsea are: log exponential metric 0.27; stochastic rank 0.64; and 2D SVB 1090.6. The first two metrics are most consistent with the clustered pattern and (not surprisingly) the class derived from translating the mapping of the non-probability sample. In assessing the range of values that an extrapolation could have over similarly arranged patterns of points, the most useful assessment would consider the variability over a class of patterns with a clustered characteristic.

The purposive metrics of the ODFW Alsea non-probability sample can be used directly to define classes of point patterns, using constructed densities as described at the beginning of Section 2. The exponential form of a constructed density is applied here for the squared- and absolute- relative-deviation. The relative deviation is the difference between a metric of an arbitrary pattern and that of the Alsea sample, divided by the Alsea sample metric. Metrics with a relative error exceeding 50% are assigned zero density. The two constructed densities are shown in Figure 3-14, plotted against the relative deviation. The corresponding ranges of metrics in the support of the constructed densities (the ranges assigned non-zero measure) would be from 0.136 to 0.407 for the exponential-fit metric; from 0.311 to 0.933 for the stochastic rank metric; and from 545.3 to 1635.9 for the 2D SVB metric.

The probability of a metric being as or more extreme – as determined by each of these two densities – is determined for three arbitrary point patterns – one realization each from the CSR, clustered and highly clustered processes (the same processes as for GOF specificity tests above). The measures (p-values) produced from the constructed metric densities are summarized in the Table 3-3.

There is little difference in the probability outcomes between the densities based on the squared- or absolute-relative deviations. As assessed by the constructed densities, the clustered or highly clustered process realizations would be considered consistent with the class of point patterns for two of the three metric densities. The highly clustered is nearly within the 20% cut-off for its worst case (the stochastic rank measure).

Table 3-3 Assessment of goodness-of-fit (GOF) of three arbitrary point patterns using two constructed metric densities derived from ODFW non-probability sample metrics. Values are the probabilities of a metric as or more extreme than that observed on each arbitrary pattern, as measured by the constructed densities. The densities are centered on the metric values of the ODFW non-probability sample (values in the last row).

	Metric Value			p-values		
				Squared (Absolute) Rel. Dev.		
	Log-exp.	Stoch. Rank	SVB 2D	Log-exp.	Stoch. Rank	SVB 2D
CSR	0.80	0.20	1148.5	0.0 (0.0)	0.0 (0.0)	0.44 (0.43)
Clustered	0.53	0.54	1157.0	0.0 (0.0)	0.36 (0.34)	0.43 (0.42)
Highly Clustered	0.24	0.81	1118.5	0.37 (0.35)	0.18 (0.17)	0.47 (0.47)
ODFW	0.27	0.62	1090.6			

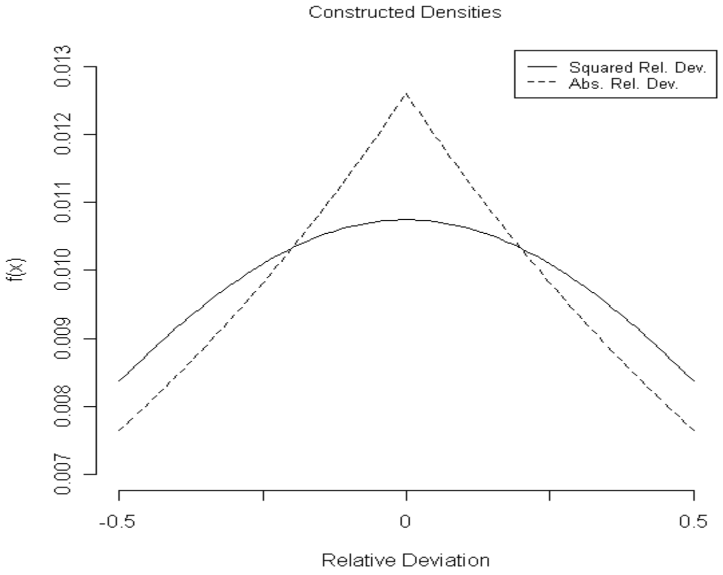

Figure 3-14 Two constructed densities centered on the metric values of the ODFW non-probability sample metrics. The relative deviation corresponds to metric values from 0.136 to 0.407 for the exponential-fit metric; from 0.311 to 0.933 for the stochastic rank metric; and from 545.3 to 1635.9 for the 2D SVB metric

The following subsections discuss some details of the specific performance of some of the metrics.

5.1 Metrics on point patterns on a continuous domain of areal extent

The inner-product density seems to perform acceptably for the GTS null (comparing a pattern with a GTS-simulated purposive arrangement). If the purposive arrangement is simulated from a CSR process, only the SVB and K(t)-derived metrics are useful to eliminate a GTS pattern from a class of CSR patterns. The SVB and K(t)-derived metrics both show excellent power for discriminating the two classes of patterns.

In the case that the metric is based on the inner product statistic, for the case that the purposive arrangement comes from a GTS process, the purposive vector of sorted distances is relatively more uniform than a vector of distances from a CSR process. The inner product for comparing a CSR arrangement to the GTS in this case does the desirable thing – it's either too small or too large, compared to the inner product of two GTS distance vectors, since a CSR pattern is likely to have substantially more closer points than the more spatially regular GTS process. However, for the inner product of a GTS vector on the purposive pattern produced by a CSR process, the inner product operation behaves like averaging of the distances, with little difference in weighting compared to the averaging effect of a CSR vector on the purposive CSR vector. Thus, when all the metrics are compared, the most extreme ones will virtually always be from a pair of CSR vectors (as evident in Figure 3-7).

Alternatively, for the processes simulated in this study, the SVB and K(t)-derived metrics bifurcate patterns from the GTS and CSR process very effectively. The supports of the empirical densities are nearly non-overlapping (Figure 3-8 and Figure 3-9). A GOF test of a realization of one process for consistency with patterns from the other process fails in all cases, using either the SVB or K(t) metric. Either metric is a useful discriminator to separate patterns with regularity from those that are random (Poisson-process) spatially distributed events.

5.2 Linear network point patterns– consecutive-point distances vs. all inter-point distances

On a domain restricted to a linear network, metrics used to characterize point patterns should be relevant to the universe of samples of points taken on the linear network. Restricting the focus to up- and down- stream distances constrains the metrics the way the possible point patterns are constrained to be on the domain.

It would seem natural to examine all inter-point up-/down-stream distances, since on the continuous domain there is sometimes more power to be gained with a function of all point-pair distances. Such a statistic would contain more information about an arrangement of points than merely the nearest-neighbor distances. Ripley (1981 (Ch. 8)) cites several studies that report that Besag's linearized–K(t) statistic is more powerful than the Clarke-Evans nearest-neighbor statistic and other statistics based on the expected cumulative distribution of nearest-neighbor distances, for rejecting a CSR process in favor of clustered or more regular (inhibiting) processes. Diggle (1979) compares several point pattern statistics including the Clarke-Evans, a quadrat-count statistic and maximum-departure from untransformed and linearized versions of K(t). The quadrat-count statistic is an excellent metric for clustering, but overall, Besag's linearized-K(t) maximum discrepancy provides the best power for detecting departures from complete spatial randomness (Diggle (1979)).

Intuitively, the inter-point distances of a point pattern provide essential information about its arrangement. For example, a cluster of points and a set of points in a line might have the same nearest-neighbor distances, but would have different inter-point distance distributions, with the cluster of points having only smaller distances and the line of points have distances ranging from the smallest of the nearest-neighbor distances to the length between the first and last point in a line. As a simple example, consider the two arrangements of three points each in Figure 3-15.

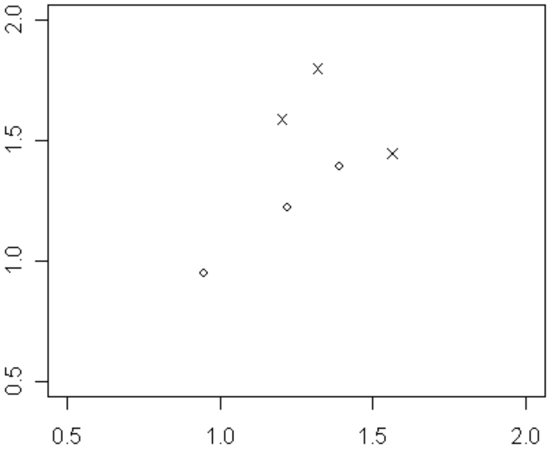

Figure 3-15 Two point patterns with the same nearest-neighbor distances but different distributions of inter-point distances.

The nearest neighbor distances in each pattern are the same, but the inter-point distances of the x's are (0.4293318, 0.2415141, 0.3863482), while those of the circles are

(0.3863482, 0.6278622, 0.2415141). In contrast, distance as measured by up-/down-stream (stream-flow) distance will not indicate information about angles between points, and on the linear network, orientation between two locations would be interpreted only as either up-stream or down-stream.

Distributions of inner-products on sorted inter-point stream distances examined for CSR and stratified stream network samples showed the (empirical) support for the CSR containing that of the stratified sample population (data not shown). In contrast to continuous domains of areal extent, for the linear network domains, metrics derived on all pairwise stream distances do not appear to partition the sample universe usefully for discriminating randomness from regularity among point patterns. Consecutive-point distances, similar to but not exactly the same as the nearest-neighbor distances of areal-extent point processes, seem to contain the more relevant information. It would seem that the information contained in the consecutive-location distances is diluted when all the pairwise distances are incorporated into a metric. It is an interesting reversal of the performance of the derived metrics between including all pairwise distances in producing metrics on areal domains or restricting metrics to local consecutive-point distances on linear network domains.

5.3 Point patterns on a linear network – 2D SVB

The 2D SVB histogram of the CSR process completely contains that of the stratified process. It is not impossible for a Poisson process to produce a pattern that manifests regularity, just by chance. Depending on the interested parties, this may or may not be regarded as a limitation for defining a class of similarly arranged points. If a stakeholder wants to be conservative about the potential variability of an estimate over the class of patterns, the specification of a class to be "more random than regular" might be important to avoid diluting the effect of the more extreme estimates that a non-regular sample could have on a domain with a regionalized (covarying) response. In this case, the 2D SVB is less desirable to define a class of CSR-like patterns because it allows more regularly spaced patterns (these have adequate frequency of occurring on the support of the CSR 2D SVB). If a conservative quantification of variability is important, the stochastic rank metric would be the preferred choice. If the goal is to define the class of patterns with regularity in the spacing of points, either the stochastic rank metric or the 2D SVB metric will serve well to reject CSR patterns from the more regular patterns.

In this study, the utility of several point process metrics is examined to assess GOF empirically evaluated for discriminating regular (GTS or stratified) vs. random (CSR) spacing of points in samples taken on domains of areal extent and for stratified, CSR and clustered patterns on a linear network. The GOF process is illustrated for a non-probability sample collected by ODFW. On the areal-extent domain, the metrics incorporate all inter-point distances. On the linear network, the metrics are based on consecutive-point distances. For the areal-extent domain, the SVB and K(t)-deviation metrics perform best. The inner-product metric is demonstrated to be inferior for detecting that a CSR process is "not similar to" a GTS process, and this metric is of no use to detect that a GTS process is "not similar to" a CSR pattern. For the linear-domain network, illustrated on a section of the Alsea River basin in Oregon, the stochastic rank metric and log-exponential-fit metric perform well. The non-probability sample is most similar to a class of clustered samples, which would be a more useful reference class for considering the variability of an estimate. The 2D SVB is an excellent indicator of non-regularity, but cannot be used to reject a stratified pattern as "not similar to" a CSR pattern.

ACKNOWLEDGEMENTS

Thanks to Tony Olsen (EPA) and Erin Gilbert (ODFW) for help with Alsea River basin and other stream section shapefiles; and to ODFW for stream basin data.

REFERENCES

Baddeley A, Turner R 2005 "spatstat: An R package for analyzing spatial point patterns" *Journal Of Statistical Software* 12 (6) 1-42.

Besag J, Diggle PJ 1977 "Simple Monte Carlo tests for spatial pattern" *Appl. Statist*. 26 (3) 327-333.

Buckland ST, Anderson DR, Burnham KP and Laake JL 1993 *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall (London).

Chung, KL 2001 *A Course in Probability Theory* Academic Press (San Diego).

Cordy, CB 1993 "An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe" *Statistics & Probability Letters* 18, 353-362.

Diggle PJ 1979 "On parameter estimation and Goodness-of-fit testing for spatial point patterns" *Biometrics* 35, 87-101.

Geyer C 1999 "Likelihood inference for spatial point processes" in *Stochastic geometry: likelihood and computation*, edited by Barndorff-Nielsen O, Kendall WS, Lieshout MNM; Chapman & Hall (Boca Raton, FL).

Hansen MH, Madow WG, Tepping BJ 1983 An Evaluation of model-dependent and probability-sampling inferences in sample surveys. *JASA* 78: 776-793.

Hope ACA 1968 "A simplified Monte Carlo significance test procedure" *Journal of the Royal Statistical Society Ser. B* 30, 582-598.

Kelly FP, Ripley BD 1976 "A note on Strauss's model for clustering" *Biometrika* 63(2), 357-360.

Overton JMcC, Young TC, Overton WS 1993 "Using 'found' data to augment a probability sample: procedure and case study" *Environmental Monitoring and Assessment* 26 65-83.

Paulsen SG, Hughes RM, Larsen DP 1998 "Critical elements in describing and understanding our nation's aquatic resources" *Jo. Of the American Water Resources Association* 34, 995-1005.

Peterson SA, Urquhart NS, Welsh EB 1999 "Sample representativeness: a must for reliable regional estimates of lake condition" *Environmental Science and Technology* 33: 1559 - 1565.

Ripley BD (1977) "Modelling spatial patterns" *Journal of the Royal Statistical Society B* **39** 172-212.

Ripley BD 1981 *Spatial Statistics* (Wiley).

Simcox AC, Whittemore RC 2004 "Environmental index for assessing spatial watershed sampling networks" *Journal of Environmental Engineering* 130 (6) 622-630.

Strauss DJ 1975 "A model for clustering" *Biometrika* 62(2), 467-475.

van Groenigen JW, Siderius W, Stein A 1999 "Constrained optimization of soil sampling for minimization of the kriging variance", *Geoderma* v87, 239-259.

Warrick AW, Myers DE 1987 "Optimization of sampling locations for variogram calculations", *Water Resources Research* 23, 496-500.

APPENDIX

A METRIC MEASURE INDUCES A SET MEASURE ON THE POINT PATTERN SPACE

Let ζ^n denote the n-fold product space of the domain A – where A is some bounded spatial domain. Let $\underline{s}$ denote a member of ζ^n, where the components s_i of $\underline{s}$ are each an ordered pair (x,y). Let $\gamma(\underline{s})$ denote a measurable mapping from ζ^n to the Real numbers. Let the range of $\gamma(\underline{s})$ be denoted by Γ^o. Let G denote a member of the Borel sets generated on Γ^o., with probability measure of G denoted $\Delta(G)$. Suppose a measure $f(\gamma(\underline{s}))$ is to be specified on the metric range, such that the support Γ of the measure is a subset of the range Γ^o and such that $f(\gamma(\underline{s}))\geq 0 \quad \forall \gamma \in \Gamma, \Gamma \subseteq \Gamma^o$ and $\int_{\Gamma} f(\gamma(\underline{s}))d\underline{s} = 1$. Let ζ^* denote the inverse image of Γ under $\gamma(\underline{s})$ - *i.e.*- $\zeta^* = \{\underline{s} : \gamma(\underline{s}) \in \Gamma\}$. Since $\Gamma \subseteq \Gamma^o$, $\zeta^* \subseteq \zeta^n$. The following paragraphs are intended to show that there is a unique (a.e.) measure on ζ^n induced by a measure $f(\gamma(\underline{s}))$ as defined.

For the discussion, let G now denote a member of the Borel sets on Γ. Define a class of sets of pre-images of G $\left\{H : \gamma(\underline{s}) \in G \quad \forall \underline{s} \in H\right\}$. Since the sets G are a member of Borel sets on a subset of Γ^o, every such G has a pre-image H in ζ^*. Consider a field Ψ generated by the class of pre-images (Ψ includes the null set and is closed to complements and finite unions of sets in the class of pre-images). Later in the development, the Carathéodory Theorem will be applied to extend a measure on the field Ψ to a unique a.e. measure on a σ-field generated by Ψ.

Assign a function to a set H in Ψ as $\mu(H) = \Delta[G]$. This function has the properties that $\mu(\phi) = 0$ and $\mu(H_\Gamma) = 1$, where H_Γ denotes the pre-image of the entire support Γ of $f(\gamma(\underline{s}))$. The function $\mu(\ )$ is now to be applied to members of Ψ that include finite disjoint unions. Let H_U be shorthand notation for a disjoint union: $H_U = \bigcup_{i=1}^{n} H_i : H_i \cap H_j = \phi \quad \forall i,j : i,j \in 1...n, i \neq j$. The function on the disjoint union is shown to be well-defined by first completing a few steps: first the existence of an inverse of

H_U is verified; then the individual inverses of the disjoint pair of sets H_i, H_j are shown to be disjoint.

(Claim) The pre-image of H_U (denoted H_U^{-1}) exists. Consider any arbitrary element $\underline{h}$ $\underline{h} \in H_U$. Since H_U is the union of *n* disjoint sets, $\underline{h}$ is an element of one and only one of the sets H_i involved in the union. By definition of our field, each H_i is a pre-image of some Borel set G on Γ. This is true for any $\underline{h} \in H_U$, so H_U^{-1} exists.

Next, if two pre-images H_1 and H_2 are disjoint, then so are the corresponding inverses G_1 and G_2. To verify this, suppose there is an element *g* that is an element of both G_1 and G_2. There is at least one element $\underline{s}$ in ζ^n that maps to *g*, since it is an element of sets G_1 and G_2 from the Borel field on Γ. Then $\underline{s}$ is an element of H_1 and also H_2, as these are the pre-images of G_1 and G_2. Then H_1 and H_2 are not disjoint. Therefore, H_1 and H_2 are disjoint only if G_1 and G_2 are. This is true for all pairs of disjoint sets from the field Ψ.

Using (a) the definition $\mu(\;)$ on Ψ and (b) the definition of a set H and (c) finite additivity of the measure $\Delta(G)$, the measure defined for the finite disjoint union H_U then, is $\mu(H_U) \overset{(a)}{=} \Delta(H_U^{-1}) = \mu\left(\dot{\bigcup_{i \in 1..n}} H_i\right) \overset{(b)}{=} \Delta\left(\dot{\bigcup_{i \in 1..n}} G_i\right) \overset{(c)}{=} \sum_{i=1}^{n} \Delta(G_i) \overset{(a)}{=} \sum_{i=1}^{n} \mu(H_i)$. This shows the finite additivity of the function $\mu(\;)$ on Ψ.

The finite additivity extends to countable additivity by induction. This is true because the function $\mu(\;)$ is defined for the finite disjoint union of the set H_U and any set H_k such that H_k and H_U are disjoint. Since the function $\mu(\;)$ is a non-negative function on sets in Ψ, with countable additivity, it is a measure. Since it integrates to one, it is a probability measure. For a field with $\mu(\phi) = 0$ and $\mu(H_\Gamma) = 1$ and for which the measure $\mu(\;)$ has countable additivity, by the Carathéodary Theorem (see Chung (2001)), the measure on the field extends uniquely to a measure on a σ-field generated on the field (or on a σ-field generated on domain that generates the field).

Now let the specified measure $f(\gamma(\underline{s}))$ be defined on all of the range Γ^o by setting it to zero anywhere in Γ^o outside of Γ. Then for the set H_c of all elements $\underline{u}$ in ζ^n that map to $\Gamma^o \cap \Gamma^c$, $\mu(H_c) = \Delta(\Gamma^o \cap \Gamma^c) = \int\limits_{\gamma \in \Gamma^o \cap \Gamma^c} f_\gamma(\gamma) d\gamma = 0$.

Finally, it is shown that the spaces ζ^n and $\sigma(\Psi)$ are the same. The metric $\gamma(\underline{s})$ was defined to map from ζ^n to Γ^o. Since the class of sets H are pre-images of the Borel sets on Γ^o (the range of the mapping from ζ^n to the Real numbers), the field Ψ is contained in ζ^n. Since any element in any set H or union of sets H_i is a an element of a pre-image of a value in the range of $\gamma(\underline{s})$ defined on ζ^n, $\sigma(\psi) \subset \zeta^n$. For any element $\{\underline{s} : \underline{s} \in \zeta^n\}$, the mapping $\gamma(\underline{s}) \in \Gamma^o \subset \mathrm{B}(\Gamma^o)$, where $\mathrm{B}(\Gamma^o)$ denotes the Borel field generated on Γ^o. For such an element $\underline{s}$ this implies there exists a set G in $\mathrm{B}(\Gamma^o)$ such that $\gamma(\underline{s})$ is an element of G. This in turn implies that there is a pre-image H of G for which $\underline{s}$ is an element. This implies that $\zeta^n \subset \sigma(\psi)$.

ESTIMATOR VARIANCE OVER SIMILARLY-ARRANGED RANDOM OR NON-RANDOM LOCATIONS ON CONTINUOUS DOMAINS

Cynthia Cooper

ESTIMATOR VARIANCE OVER SIMILARLY-ARRANGED RANDOM OR NON-RANDOM LOCATIONS ON CONTINUOUS DOMAINS

Cynthia Cooper

Oregon State University

ABSTRACT

Design-based variance addresses variance of estimates induced by the sampling process of observing a random subset of the response over its domain. The variance of the estimate is the expected squared deviation from its mean over all possible samples taken on the domain. The probability (or "measure" on a continuous domain) need not be uniform over all samples. For non-probability (purposive) samples – such as convenience samples or observations taken at haphazardly selected locations, technically, purposive elements do not contribute anything to the sample-process variance (as they would be in every sample, with probability one). Nevertheless, a stakeholder might reasonably ask, using information inferred from observations about an assumed-stationary response covariance structure, how much would an estimate derived from other "similarly arranged" patterns of elements vary? A class of similarly arranged elements or a sample process can be characterized by set measures on the universe of point patterns. This study is part of on-going research to derive the variability of an estimator over similarly arranged collections of observations of a regionalized response on a continuous domain of areal extent or on a linear network domain such as a stream network. In this paper, a process to quantify the variability of estimators over a class of similarly arranged patterns is proposed and demonstrated. The process is designed to be general and avoids imposing any stochastic point-pattern process to accomplish the objective of describing variability over similarly arranged sets of locations.

Keywords: Non-probability sample, Semi-parametric variogram, Monte Carlo

1 INTRODUCTION

When a population is characterized by estimates based on observations from a sample of the population, the estimate will depend on what elements of the population were in the sample. This is variability due to sample process. Variability in the sample also depends on variability of the response in the population or on a domain. If the population response were uniform, the estimate would not change from one sample to the next (assuming the estimator function executed on the data stays the same). On the other hand, for responses that are patchy, meaning there are some contiguous areas of higher values and some of lower values, the arrangement of locations at which observations are collected can have a lot of impact on the estimate and therefore on the variance of the estimate. Spreading the points out improves the chances of observing a range of values representative of the range and distribution of the response overall. Taking observations at clusters of points might result in higher values or lower values being over-represented in the sample, causing the outcome of the estimator to have a greater occurrence of over- and under-estimating, and having greater variance over the implemented sample space.

Royall and Cumberland (1985; 1981) refer to the degree to which the observations are representative of the distribution of the response on the whole population or domain as sample balance. They discuss bias that may exist in any particular sample due to sample imbalance. The bias among samples that Royall and Cumberland refer to is the component of variance due to sampling process.

Model-based sampling, or restricted sampling, controls for bias from imbalance (or, directly related, for sample process variance). In particular, if there is covariance in the response among population elements or over a continuous domain, some configurations of sample units or elements will provide more information than others about the domain's overall response. The observed variability in a sample might be misleading for assessing the variability of the population overall if a sample is taken that includes covarying units. For example, there may be a natural hierarchy in the population or block-effects on a domain such that clusters of elements have more similar response than elements observed throughout the domain. Additionally some samples may be less efficient than others, because elements with co-varying response provide partly redundant information about the response overall.

On a continuous domain with covariance diminishing as distance increases, as sampling resolution (average number of points per unit length or area, a.k.a. intensity) increases, the magnitude of covariance within sample response increases. The resulting effective covariance within the sample impacts the estimator variance over the sample process. This interaction of sample arrangement and covariance structure of response is exploited to design optimal samples, and to predict the sample process variance of various sample designs. This latter application can be particularly useful where there is not an adequate configuration of observations to get direct estimates of variance within clusters or strata. For example, design-based methods employ sums of squares to estimate variance for responses with exchangeable covariance (where the covariance within a cluster is constant); a single observation within a stratum or cluster does not permit direct estimation of the within-stratum (-cluster) variance. (See Cordy and Thompson (1995); Cooper (2006))

The conventional design-based methodology quantifies variability of an estimate due to the sampling process. That is, the response on the domain is treated as fixed (interpreted as a snapshot in time for many applications), and the random variable in the equation that computes the estimate is the indicator variable for whether the element is or is not included in the sample. The variance of the estimate is the expected squared deviation from its mean over all possible samples taken on the domain. The probability (or measure, on a continuous domain) of any sample outcome need not be uniform over all samples. In design-based methodology, the form of the variance is most often derived in closed form (or approximated) as some function of the square and cross-product terms of the response, the terms of which are often estimated by scaling up by weights inversely proportional to marginal and pairwise inclusion probabilities (or inclusion densities on continuous domains).

For a non-probability (a.k.a. purposive) sample, such as convenience samples or observations taken at haphazardly selected locations, technically, an estimate based on the observations from a purposive sample has no sample-process variance. In the scenario of this study, some pattern of locations (represented as points) is presumed to be the result of a study in which some observations have been collected on some spatial domain, where the locations visited were *not* selected by a sample design and the sample is not a probability sample.

A probability sample is one for which elements in the domain are randomly selected to be in a sample, by explicitly employing a random mechanism. The domain is represented by a frame - a list of units for a finite population or a map for a domain of spatial extent (e.g. a list of segments for a domain of linear extent)). Each element on the frame is assigned a probability, or an inclusion density is defined on the domain, such that a random mechanism is employed to determine inclusion of domain elements in the sample with the probability or inclusion density associated with each element. The way that the inclusion probabilities or inclusion density are defined is formally the sample design. The sample design may assign variable inclusion probabilities, sometimes as a function of auxiliary variables which could be continuous or categorical (such as in cluster or stratified sampling). (There are sampling and estimation procedures for cases in which the target population elements are not itemized - referred to as distance sampling, which is not covered in this study (see Buckland *et. al.* (1993)).

Peterson *et. al.* (1999) and Paulsen *et. al.* (1998) discuss selection bias in examples of non-probability samples that do not adequately represent the response over the domains of interest. Simcox *et. al.* (2004) demonstrate an analysis of representativeness of non-probability samples for water quality on USGS monitoring stations in a New Hampshire watershed, using techniques similar to post-stratification. Though not recommended, non-probability samples do happen to good agencies, frequently. Doing a proper sample design and survey can be prohibitively expensive. An agency may have some observations from pilot studies, or from monitoring stations that might hopefully represent a carefully specified subset of the domain, or from observations from encountered phenomena. Data may have been collected in a particular configuration that optimizes fitting a model, and then subsequently there is interest in using the non-probability data for extrapolating a characteristic of the response across the study area.

While any extrapolation from the observations to a domain would most certainly have to be treated as preliminary without a probability sample design, a stakeholder would naturally be inclined to ask how much the extrapolation from the observations in the non-probability sample might change if a similar sample of elements on the domain had been observed. This is the question of focus in this research. The problem with the non-probability sample is that current available design-based methods do not apply, because the specification of the inclusion probabilities (density) is pathologic.

In the present research, the objective is to provide an assessment of the variability of an extrapolation from a non-probability sample to the rest of the domain, perhaps as a preliminary data point, accepting that the extrapolation includes risk of selection bias nevertheless. Because the non-probability sample has pathologic inclusion densities, there is not a basis for the application of the conventional design-based metrics. The assessment of bias and efficiency (variance) on the extrapolation does not have a long-run frequency basis of interpretation as described by Godambe and Thompson (1988). However, the observations and the order of the observations as provided by their arrangement on the domain will impact the degree of potential bias (in the sense described by Royall and Cumberland) an extrapolation would have when the observations are taken on a regionalized response – a continuous response with a covariance structure such that locations in close proximity are more similar than locations beyond the range of covariance. A question of interest that has useful and natural interpretation is how much the extrapolation would vary on other sets of observations that are arranged similarly to the given non-probability sample. From this natural idea, a basis is proposed for characterizing the stability of the extrapolation (conversely, the variability). In a previous paper, a mathematical characterization of a class of patterns of similar arrangement is developed ("Characterizing classes of similarly arranged point patterns as a reference of variability on non-probability samples"; Dissertation Chapter

3). In this paper, methods are explored for estimating the variability of an estimate over a class of patterns.

2 BACKGROUND

A sample can be characterized by how representative it is of the domain response. Response covariance within the sample impacts how well the distribution of the response within the sample matches the distribution of the response throughout the domain (i.e. described as "balance" by Royall & Cumberland (1985; 1981)). Covariance within the sample depends on the range of covariance structure of response and sample resolution (relative to range) as well as the size/dimension of study area (relative to range). Oliver & Webster (1986) explore how pure nugget (sometimes referred to as variance due to measurement error) on a coarse scale may manifest covariance at a finer scale. A pattern with complete spatial randomness (CSR) would have relatively more covariance in the observed responses than a more regular (i.e.-spatially balanced) pattern (see, for example, Cochran (1946)).

A formal model of the characteristics of observations from a collection of locations or elements on a domain has three parts: a component representing the domain ("A"), a component representing all finite sets of points on the domain, denoted ζ^n, and a set measure on members of ζ^n (which could be probabilities or densities (when they exist) for sets generated by a stochastic process, though this is not meant to suggest that a collection of locations be treated like a random sample if it is not). As an example, the universe of size-n sets for complete spatial randomness (CSR) is the n-product (Cartesian product) space of domain A. The model indicates that a covariance structure in the response on domain A will be relevant to the characteristics of a sample, as well as to the variance over estimators that are functions of the responses at the samples' observed elements. This is true regardless of properties of a sample process (e.g. inclusion being independent from one point to the next). For the context of this study, for a class of point patterns characterized by a measure on the set universe ζ^n, the covariance structure of the response on domain A impacts the variability of the estimator function applied to sets within that class of point patterns.

In this development the collection of observations is assumed to be taken from a stationary regionalized response structure on a continuous domain – that is, the mean and covariance structure of the response is assumed to be independent of location. For this study, the covariance structure is assumed to be isotropic (not dependent on the orientation of any two points). An important qualification of any application of the process described here is that a sample must be defensibly representative of the domain. A sample which includes observations collected at "hot spots" is not representative of the response overall. Such a sample would violate the assumption of response stationarity: at "hot spots", the mean of the response is not independent of location. (Refer to Simcox *et. al.* (2004)) A sample that avoids hot spots might be defensibly representative. Sources of bias due to frame error or non-response (in the context of continuous domain studies, due to accessibility problems that may or may not be related to the response) are not addressed here.

3 APPROACH

The process to quantify variability should be general to any collection of points without requiring, suggesting or imposing a stochastic generating process. For example, the collection of points may resemble an outcome of a random process with complete spatial randomness, or it may be more regular or more clustered, resembling the outcome of a random process with inhibition or clustering. Since the collection of points in the scenario of interest here is not a random sample, conclusions drawn from the observations should be reviewed with careful consideration to any (inadvertent) bias due to the selection of locations without a random mechanism. Avoiding application of a model of any stochastic point process serves to keep the process general to any collection of points at the same time that it does not encourage one to forget the non-random origin of the data.

The first part of the process is constructing measures on point sets to characterize classes of similarly arranged points, since variability of any estimate based on the non-probability sample is with respect to how the estimate could vary over samples from that class. Methods of constructing measures are based on metrics that partition the point-set universe ζ^n in some useful manner that distinguishes point sets based on characteristics of the point patterns such as regularity and clustering. This is addressed in the first paper (Dissertation Chapter 3).

Once a measure is defined on the set universe ζ^n, the process exploits the covariance structure of the domain's response to predict variability over sets of points observed on that domain. The proposed procedure is to model the response's covariance structure given the collection of observations, and then to use the Monte Carlo (MC) method to estimate variability over the specified class of point patterns. The MC step is done by simulating a response with the fitted covariance structure and observing the estimate obtained from many sets of observations taken from the class of sets arranged similarly to the non-probability set of observations. Webster and Oliver (1992) use a similar approach to demonstrate the variability of fitted variogram models at each lag, in which they simulate many realizations of the response to obtain Monte Carlo confidence bands (in this case the variability is over realizations of the response's generating random process, not due to sample process variability).

Besag and Diggle (1977) demonstrate the Monte Carlo (MC) approach in several examples of point pattern analysis. They cite Hope (1968), who finds an MC test for significance would have only a little less power than that of a UMP test. The joint densities of the point patterns and of observations taken on these points would be complicated or possibly intractable. Monte Carlo methods in the analysis of point samples on continuous domains provide a way to examine the probability law of estimators and statistics on the point patterns, possibly on domains with irregular boundaries and/or including holes. Besides all the practicality of the MC method, the application is general to any collection of observations and without consideration of any stochastic point process that the non-probability collection of points might resemble.

The process of simulating a response using a modeled covariance structure is analogous to the concept of reproducing equivalent sets of data from a sufficient statistic. An interpretation of a sufficient statistic is that it provides all the information to partition the data space into "equivalence classes" that would result in equivalent inferences. The original data is not necessary given sufficient statistics. The sufficient statistic could be used to generate data equivalent to the original data. For the context of this study, the data on the entire domain is not observed (or there would be no need for estimating the desired summary characteristic). A response simulated from the modeled covariance structure is "equivalent" data in terms of the behavior of the covariance structure. The newly produced response is sufficient to analyze the variability of the estimator on the specified class of point patterns. Strictly speaking, whether the modeled covariance structure adequately captures the behavior of the underlying response depends on how much the sample observations represent that behavior of the response on the domain. Assuming this is a viable assertion, the modeled covariance structure is used to produce a data set equivalent to the one for which there is not complete information.

4 METHODS

The procedure is applied in two scenarios. In the first, the response is a simulated regionalized response with an exponential covariance structure on a continuous domain of areal extent. In the second, the procedure is applied to a simulated response on a stream network, for which the covariance structure is modeled using a semi-parametric covariance structure derived as the covariance of a moving-average process (see Barry and Ver Hoef (1996)). The response on the stream network mimics summer parr per KM as modeled by personnel at Oregon Department of Fish and Wildlife (ODFW). For the exponential-covariance response on the areal-extent domain, the parameters of the exponential structure are estimated using REML. For the stream-network response, the moving average coefficients are estimated using a non-linear least squares fit.

The covariance structure is used to simulate a realization of a response on the domain represented by the frame. On the simulated realization, the MC process is used to examine variability of an estimator over many sets drawn from the class of sets defined as similarly arranged to the purposive sample. The GaussRF() function in the RandomFields package of R (Schlather (2001)) was used to simulate the response on the domain of areal extent. To simulate a response on a stream network, sequences of innovations (i.e. independent Gaussian random variables with suitable variance to mimic the observed response's variance) were produced and a moving-average filter with the estimated coefficients was applied to each of the network's segments. In the case of segments downstream of multiple confluences, the contribution of each of the upstream segments was averaged, with equal weighting for simplicity.

Specific details on the methods to test the proposed process are described for each domain in the next two sections.

4.1 Variability on a continuous domain of areal extent

To test the process, a "true" response of exponential structure on a 200 x 200 unit-square study region is simulated. The exponential structure is simulated for a zero-mean process with range of 2 and sill of 4. This response is sampled with two configurations of 100 points – one with complete spatial randomness (CSR) and one with a more regular interval between points – a grid tessellation stratified (GTS) point pattern. The point patterns are used as the non-probability samples even though they are produced with stochastic processes. For each configuration, an exponential covariance structure is fit using REML and this modeled structure is used to simulate another realization of the domain response, as though the true response were not known. For 1000 simulations, similarly-arranged patterns of size-100 samples are generated by translating and rotating the original pattern, wrapping the pattern around the edges of the study region from bottom to top and right to left when the rotated, translated pattern goes outside the boundary. For the original and each newly generated pattern, a Horvitz-Thompson estimate (Horvitz and Thompson (1952); or see S. Thompson (1992)) of total is produced assuming constant weights proportional to the sample intensity (which averages 1 point per 4 unit-square area). For the study, this is done with response observed both on the original response and also on the equivalent response. The empirical variance from the generated response and the true response are compared. This process is done for each of the two generated purposive patterns.

4.2 Variability on a simulated stream network

The approach to estimate variability of an estimate over a class of similarly arranged locations for a stream network is the same as the approach for that on an areal extent. The method in application requires some modification, but the overall steps are the same – model the covariance, simulate a response on the domain and use the Monte-Carlo method to estimate the variability on a class of sets of locations arranged similarly to the (non-probability) collection of locations.

There are two main specific differences from the areal-extent example – (1) the method chosen to model the covariance and (2) a modification to average over a number of simulated realizations in order to guard against inadequate representation of the original response's covariance structure. The method to model the variogram with a moving-average process variogram is described in Barry and Ver Hoef (1996). The modification to average over a number of simulated response realizations was determined to be necessary from preliminary results (discussed in subsequent sections).

The moving-average variogram is a piece-wise linear variogram that would model the average squared differences of a moving-average process response as a function of stream-flow distance between any point pair in the domain. The number k of piece-wise linear nodes are specified such that there are some average number (e.g. 30) of observed squared-differences in each stream-flow distance bin between 0 and range c. The range c in this context refers to the lower endpoint of the bin interval containing the largest distances. The bins are defined to be even intervals of *c/k* up to distance *c*, with one additional bin between *c* and the maximum inter-point distance. Coefficients of the moving average producing the observed variogram are estimated using non-linear least squares. A more extensive description of fitting a moving average variogram is described by Barry *et. al.* (1996).

The stream network response is simulated as a moving average process that will be realized with a covariance structure similar to that of the response observed. The simulated response is produced by a moving average process, where the fitted coefficients are applied to a sequence of independent, identically distributed (iid), zero-mean, finite-variance innovations. Although the response on the stream network is continuous, it is modeled here as the result of a moving average process on discretely spaced innovations. Any artifact introduced by the quantization is ignored. The Appendix describes how to determine the variance of the innovations and how a suitable resolution of the simulated response can be achieved by interpolating additional coefficients in between the fitted MA coefficients.

The performance of the suggested process to quantify variability of the estimator is evaluated by examining relative error. To analyze the performance, the procedure is applied to a stream network simulated response that is treated as the true response. The steps are summarized as follows:

1. Produce a true simulated moving-average process response.
2. Fit a moving-average variogram to the true response.
3. Repeat the analysis (above) for MC (=100) trials:
    - a. For each of J (=20) trials
        - i. Simulate a moving average response (using coefficients estimated in (2)) to mimic the covariance of the true response.
        - ii. For each of I (=10) trials, generate a collection of locations from the class of arrangements similar to the (non-probability) arrangement of locations.
        - iii. Save the variance of the I realized estimates on both the true response and on the mimic response.
    - b. Save the observed average (over J trials) variance (over I trials) for both estimates from the true and mimic responses.
    - c. Calculate the relative error from each MC trial as

      {difference between average observed variances}/{true response average observed variance}

Values for I (number of replicate samples per mimic realization (10)) and J (number of replicate mimic realizations (20)) were chosen after comparing panels of point plots of the realized estimates from each sample within each realization and on the true response. Qualitatively, the ranges of estimates seemed to show more consistency within the samples than between realizations. Differences in characteristics of the realizations are described in more detail in the Discussion.

The performance of the variance estimator is examined on one segment of stream in the Alsea basin in Oregon. Figure 4-1 and Figure 4-2 show examples of samples on the segment. The purposive sample is produced here as the result of a process with complete spatial randomness (CSR)), by mapping the stream network to one line segment of appropriate length, choosing the intervals between locations along the linear map by a Poisson process, and then transforming the locations on the mapped line segment back to the stream network. The class of similarly arranged sets of locations is effected with a process similar to that used for the areal-extent example, except without applying rotation. That is, a random offset is added to the locations in the mapped line segment and the new locations are transformed back to the stream network by the appropriate mapping.

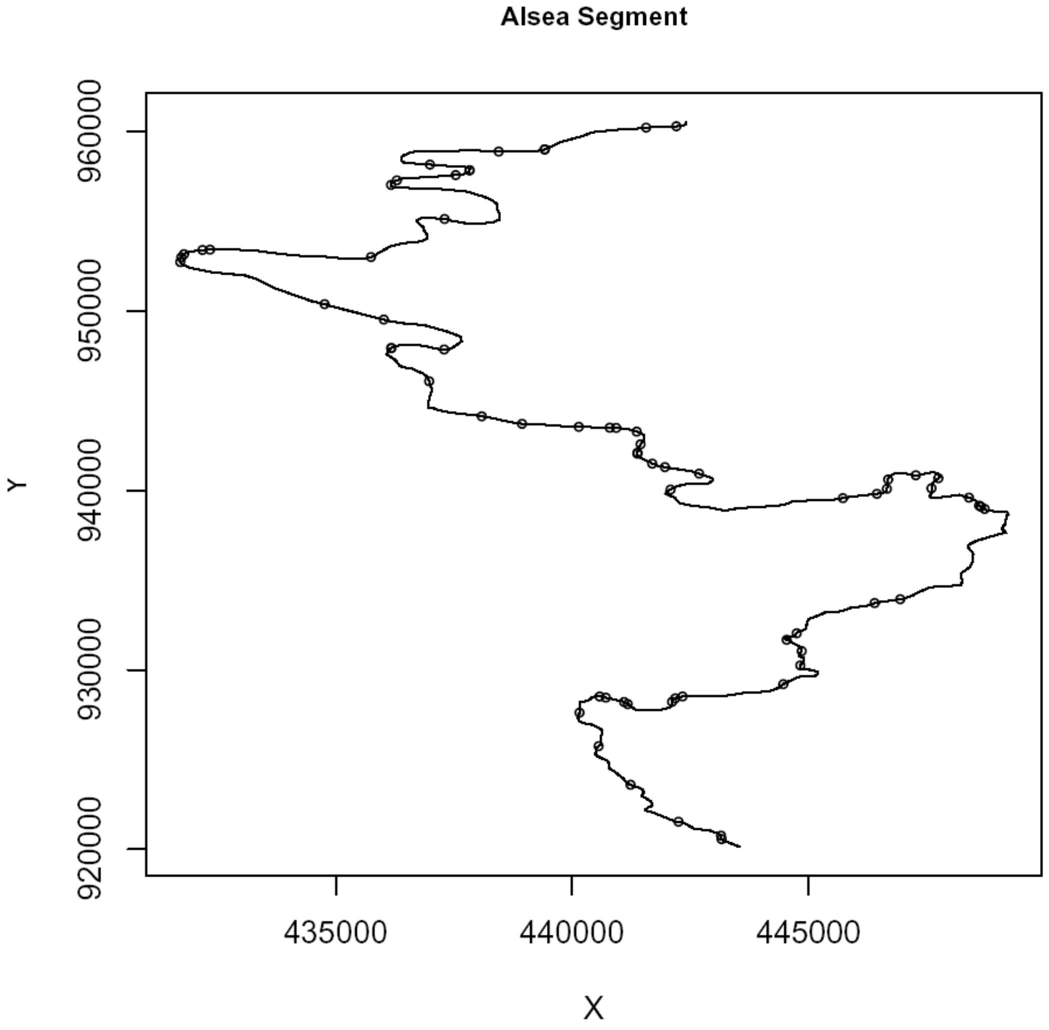

Figure 4-1 A CSR sample taken on a segment of stream in the Alsea basin of Oregon. This is treated as the non-probability collection of locations and translated random amounts to achieve a class of similarly arranged locations of points.

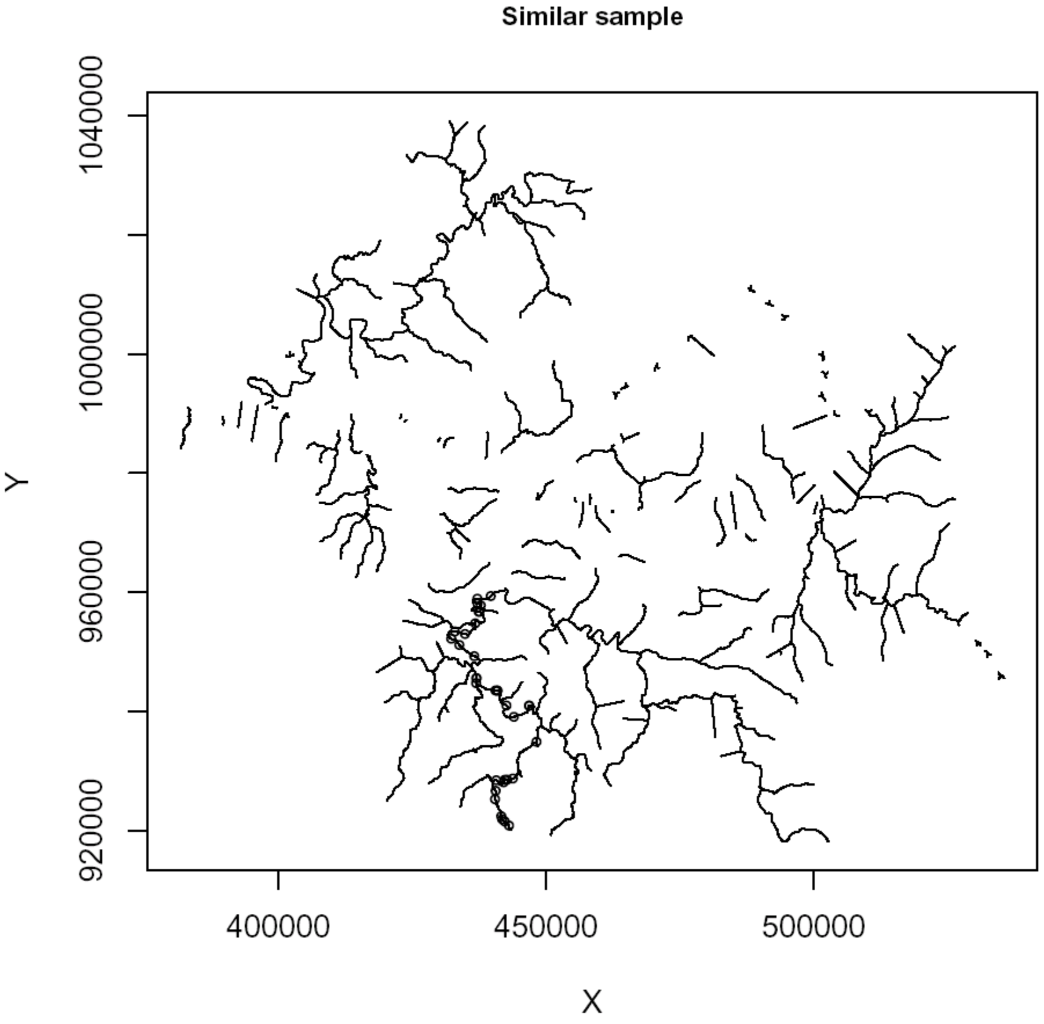

Figure 4-2 A sample taken on the studied stream segment in the Alsea basin stream network in Oregon. This sample is from a class of similarly arranged sets of locations generated from a "non-probability" sample (produced by a Poisson process for this study).

5 RESULTS

The following subsections contain a brief summary of the results of the process applied on the areal and stream-network domains. Discussion of results of the process as applied to each domain follows in Section 6.

5.1 Variability on a continuous domain of areal extent

For the regularly spaced (GTS) collection of points, the observed empirical variance on the true surface was 4244.7 vs. that on the "equivalent generated" response, which was 4639.8. For the more random (CSR) collection of points, the observed empirical variance on the original true response was 10240.3, vs. that on the equivalent generated surface, which was 10059.9.

5.2 Variability on a simulated stream network

Table 4-1 below summarizes the results of observed relative error for 100 trials each of three ranges *c* specified to be either the thirtieth, fortieth or fiftieth percentiles (*cq* = .3, .4, .5) of the purposive inter-point distances in the sample along the segment of Alsea (as in Figure 4-2, for example). The range *c* in this context refers to the lower endpoint of the bin interval containing the largest distances.

Table 4-1 Stream network relative error statistics for "range" set to the 30th, 40th and 50th percentiles of observed inter-point stream-flow distances (where "range" is the lower endpoint of the upper-most stream-flow distance bin):

cq	Relative Error Summary Statistics					
	Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.3	-0.16	0.05	0.14	0.17	0.31	0.62
0.4	-0.27	-0.10	0.01	0.02	0.09	0.44
0.5	-0.39	-0.24	-0.14	-0.11	-0.01	0.31

6 DISCUSSION

6.1 Variability on a continuous domain of areal extent

The empirical variance of the equivalent generated response is about 10% greater than that of the original data for the regularly spaced (GTS) pattern of observations. The equivalent generated empirical variance of the CSR pattern was about 2% less than that of the original data. These results are one outcome of a Monte Carlo study on 1000 trials for each point pattern. They suggest that the approach would provide a stakeholder with a useful approximation of how much an estimate would vary on a domain had the data been collected at different but similarly arranged locations.

6.2 Variability on a simulated stream network

There are deficiencies in the performance of this methodology as applied to stream networks, regarding bias and efficiency. Bias is discussed first, followed by efficiency.

To achieve unbiasedness, it is critical that the variogram be fit with an appropriately selected range *c* that results in a moving-average fitted variogram that most closely matches the empirical variogram generated on the same intervals. The empirical variogram is produced as the observed averages within the same bins defined by the nodes of the moving average variogram, with the observed average plotted at the average distance within the bin. In the example, the observed relative error indicates positive bias for mimic responses generated as the result of a variogram fit with the range set to the 30th percentile of the distances (with the highest bin including the top 70th (100-30) percentile distances). It indicates negative bias for the mimic responses generated from a variogram fit with the range at the 50th percentile (the highest bin defined to include the top 50th (100-50) percentile of the distances). When the range *c* is specified to be the 40th percentile, although there is substantial variability in the outcomes of the relative error, on average the relative error indicates that the variability as observed on the mimic response is approximately unbiased for estimating the variability of the estimator as observed for the same samples on the true response. The moving average variogram fitted for *c* set to the 40th percentile of the distances achieved the closest match to the empirical variogram in terms of being most parallel to and having least vertical deviation from the empirical variogram.

The same process repeated on another ODFW modeled response showed the direction of bias being the reverse of how it occurred in the above example – that is, as the percentile defining *c* increased, the bias went from being negative to being positive (data not shown). As with the example above, the amount of bias as indicated by the relative error is least severe for the fitted moving average variogram that most closely matches the empirical variogram fit on the bins specified by *c* and *k*.

The efficiency of the method to indicate variability on the true response by the empirical variability observed on a mimic response is probably not adequate for any but the most preliminary studies, even when the moving-average fitted variogram aligns closely with the empirical variogram. To improve efficiency, it is critical to average the empirical Monte Carlo variance over multiple realizations of the mimic response. A review of various realizations of the moving average process on the segment shows that, between realizations, there is enough variability due to naturally occurring longer and shorter periods of persistence that it is risky to take the empirical variability over similarly arranged locations on any one realization as being "representative enough" of the true response's characteristic. As an example, consider the two realizations shown in Figure 4-3 from the same 5th -order moving average process, with coefficients fit by non-linear estimation of a moving average variogram to one derived response on the Alsea basin survey data. The ranges in response on the two realizations are approximately the same (about 7500 units or so), but there is longer persistence in the gradually increasing trend in the realization to the left, and more incidents of cycling to extremes in the one on the right. Thus samples taken on the left could indicate less variability over all compared with those taken on the right. Averaging over numerous realizations helps to mitigate the effect of this phenomenon.

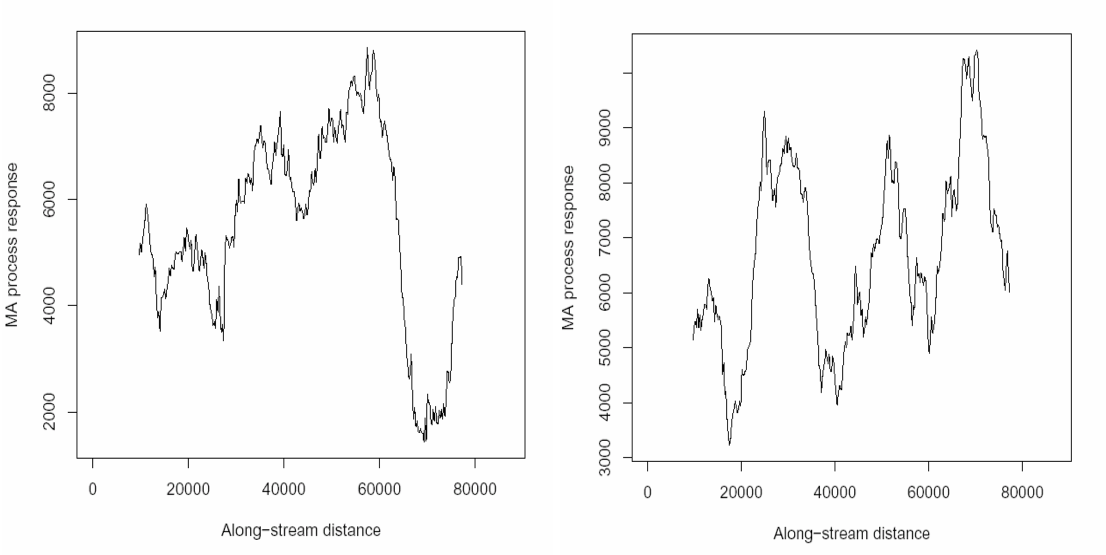

Figure 4-3 These two realizations of a particular 5th-order moving average process illustrate the variability in persistence of outcomes of a moving average process.

7 CONCLUSION

In this study, a method to characterize the variability of an estimate over a class of sets of locations similarly arranged to a non-probability sample is proposed and illustrated. The method addresses an interest to characterize the information in non-probability samples that might defensibly be representative of a response on an underlying domain. The method avoids any suggestion that a non-probability sample is the result of a stochastic point process and is general to any configuration of domain or arrangement of points. The paper gives a basis of describing the variability of estimates taken on sets of locations similarly arranged to a non-probability sample by noting the relationship between the spatial order of elements in the samples and the covariance of the sample responses as impacted by that ordering.

The covariance structure in the response on the domain influences variability over an estimator on samples taken from that domain. The proposed method to quantify variability over a class of sets of locations determined to be similar to a non-probability sample uses a modeled covariance structure to produce an "equivalent" response on which to obtain a Monte Carlo estimate of variability on the class of sets of locations. This is similar to an approach used by Webster and Oliver (1992) to simulate realizations of a stochastic process (of a continuous response with covariance) to generate confidence intervals on fitted variograms. The method proposed here is illustrated for examples on an areal domain with a simulated response with exponential covariance, and on a stream-network domain with a simulated moving-average response.

The results suggest that the proposed method is viable for the response on the areal domain, but only marginally useful for the response on the stream network. On the areal domain, the variability on the equivalent and original response over 1000 trials differed by 2% for spatially random arrangements and by 10% for more regular arrangements of locations. On the stream network, using the coefficients from the fitted moving-average variogram to produce an equivalent moving average response on the stream network had the limiting feature that the moving average response for the segment studied had substantially varying degrees of persistence among the realizations of the moving average process, which resulted in differing indications of variability over the class of similarly arranged locations. Averaging over numerous realizations helped to mitigate but did not overcome the effect of the varying nature of the moving average response. Additionally, without a carefully chosen combination of range and number of nodes to define the piece-wise moving average variogram, the stream-network results showed substantial risk of bias.

ACKNOWLEDGEMENTS

Thanks to Tony Olsen (EPA) and Erin Gilbert (ODFW) for help with ODFW Basin Survey stream section shapefiles; and to ODFW for stream basin data.

REFERENCES

Barry RP, Ver Hoef JM 1996 "Blackbox kriging: spatial prediction without specifying variogram models" *JABES* 1(3) 297-322.

Besag J, Diggle PJ 1977 "Simple Monte Carlo tests for spatial pattern" *Appl. Statist.* **26** (3) 327-333.

Buckland ST, Anderson DR, Burnham KP and Laake JL 1993 *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall (London).

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.

Cooper C 2006 "Sampling and estimation on continuous domains" *Environmetrics (in press).*

Cordy CB, Thompson CM 1995 "An application of the deterministic variogram to design-based variance estimation" *Mathematical Geology* 27(2) 173-205.

Hope ACA 1968 "A simplified Monte Carlo significance test procedure" *Journal of the Royal Statistical Society Ser. B* 30, 582-598.

Horvitz DG, Thompson DJ 1952 "A generalization of sampling without replacement from a finite universe" *JASA* 47, 663-685.

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.

Paulsen SG, Hughes RM, Larsen DP 1998 "Critical elements in describing and understanding our nation's aquatic resources" *Jo. Of the American Water Resources Association* 34, 995-1005.

Peterson SA, Urquhart NS, Welsh EB 1999 "Sample representativeness: a must for reliable regional estimates of lake condition" *Environmental Science and Technology* 33: 1559 - 1565.

Ripley BD 1981 *Spatial Statistics* Wiley (New York).

Royall, R.M., Cumberland, W.G. 1985 "Conditional coverage properties of finite population confidence intervals" *JASA* 80, 355-359.

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.

Simcox AC, Whittemore RC 2004 "Environmental index for assessing spatial watershed sampling networks" *Journal of Environmental Engineering* 130 (6) 622-630.

Thompson SK 1992 *Sampling* Wiley.

Webster R, Oliver MA 1992 "Sample adequately to estimate variograms of soil properties", *Journal of Soil Science* v43, 177-192.

APPENDIX – SIMULATING A MOVING AVERAGE STREAM NETWORK RESPONSE

The goal of the analysis is to simulate a response with a covariance structure that closely mimics that of the response being sampled. For the stream networks, the covariance is modeled by fitting a moving-average covariance structure, using nonlinear estimation to fit the moving average coefficients that achieve a fitted variogram matching the empirical variogram. The simulated response is produced by a moving average process, where the fitted coefficients are applied to a sequence of independent, identically distributed (iid), zero-mean, finite-variance innovations. Although the response on the stream network is continuous, it is modeled here as the result of a moving average process on discretely spaced innovations. Any artifact introduced by the quantization is ignored. The first subsection describes how to determine the variance of the innovations. The second discusses how a suitable resolution of the simulated response can be achieved by interpolating additional coefficients in between the fitted MA coefficients.

A.1 - Determining variance of the innovations

The variance of the innovations depends on the order *k* and range *c* of the moving average (MA) process (or in application, on the order and range we choose to apply to fit the semi-variogram). The relationship is determined as follows. Let z_i denote iid innovations with zero mean and variance σ_z^2. Let E[] and V[] denote the expectation and variance operators. Let the vector ***a*** be the vector of moving average coefficients and let *m* denote the discrete lag for which we are evaluating the variogram. The responses this lag distance apart are each the result of the moving average on the innovations, so that one response is a sum of innovations a lag of *m* away from the innovations of the other response (i.e. $\sum_{i=1}^{r} a_i z_i$ and $\sum_{i=1}^{r} a_i z_{i-m}$). The expectation of the squared increment between the two responses with lag *m* is equated to the fitted variogram at that lag distance to arrive at the relationship between the variance of the innovations and the modeled order *k* and range *c*.

$$E\left[\left(\sum_{i=1}^{r} a_i z_i - \sum_{i=1}^{r} a_i z_{i-m}\right)^2\right] = E\left[\left(\sum_{i=1}^{r} a_i z_i\right)^2 + \left(\sum_{i=1}^{r} a_i z_{i-m}\right)^2 - 2\left(\sum_{i=1}^{r} a_i z_i\right)\left(\sum_{i^*=1}^{r} a_{i^*} z_{i^*-m}\right)\right]$$

$$\overset{iid}{=} 2E\left[\left(\sum_{i=1}^{r} a_i z_i\right)^2\right] - 2\sum_{i\neq\left(i^*-m\right)} E\left[a_i a_{i^*} z_i z_{i^*-m}\right] - 2\sum_{i=\left(i^*-m\right)} E\left[a_i a_{i^*} z_i z_{i^*-m}\right]$$

$$\overset{iid}{=} 2E\left[\left(\sum_{i=1}^{r} a_i z_i\right)^2\right] - 2\sum_{i=\left(i^*-m\right)} E\left[a_i a_{i^*} z_i z_{i^*-m}\right]$$

Using the identities $E\left[z_i\right]=0$ and $E\left[\sum_{i=1}^{r} a_i z_i\right]=0$ and $V\left[X\right]=E\left[X^2\right]-\left(E\left[X\right]\right)^2$, the expression reduces to the following:

$$2E\left[\left(\sum_{i=1}^{r} a_i z_i\right)^2\right] - 2\sum_{i,\ i^*\in\{1:r\}\ |\ i=i^*-m} a_i a_{i^*} E\left[z_i^2\right]$$

$$= 2V\left[\sum_{i=1}^{r} a_i z_i\right] - 2\sum_{i,\ i^*\in\{1:r\}\ |\ i=i^*-m} a_i a_{i^*} V\left[z_i\right]$$

$$\overset{Z\overset{iid}{\sim}\left[0,\sigma_z^2\right]}{=} 2\sigma_z^2\left(\sum_{i=1}^{r} a_i^2 - \sum_{i,\ i^*\in\{1:r\}\ |\ i=i^*-m} a_i a_{i^*}\right)$$

Setting this equal to the fitted variogram value for lag *m*, referring to Barry and Ver Hoef (1996), the variogram at lag *m* is equal to

$$2\sigma_z^2\left(\sum_{i=1}^{r} a_i^2 - \sum_{\substack{i,i^*\in[1..r] \\ |\ i=\left(i^*-m\right)}} a_i a_{i^*}\right) = 2\hat{\gamma}\left(h_m\right) = 2\frac{c}{k}\left(\sum_{i=1}^{r} a_i^2 - \sum_{\substack{i,i^*\in[1..r] \\ |\ i=\left(i^*-m\right)}} a_i a_{i^*}\right)$$ which implies that

the variance of the innovations should be set to the lag interval – i.e. the range *c* divided by the number of nodes *k* in the fitted variogram.

A.2 – Manipulating the resolution of the simulated response

In the original study, Oregon Department of Fish and Wildlife collected "basin survey" data on certain segments of the Alsea basin (Figure 4-4). The range *c* of the process for which the variogram was fit to the observed variogram data of the original response was set to include some proportion (either 30%, 40% or 50%) of the observed inter-point stream-flow distances. (Locations on disconnected segments are modeled as having infinite inter-point distances). The nodes are specified such that there are an average number (30) of observed squared-differences in each bin between 0 and range *c*. The bins are defined to be even intervals of *c/k* up to distance *c*, with one additional bin between *c* and the maximum of the inter-point distances observed.

The 30th, 40th and 50th percentiles of the original ODFW survey on Alsea were about 5400, 7600 and 10900 meters respectively. (Total stream length included in the Alsea basin frame is 1,666,234 meters.) This may be beyond a scale of local covariance, though the scale could be reasonable for response associated with land-use (though an effect at this scale could more conveniently be modeled as a main effect). To understand how a covariance structure might be exploited to examine variability over a class of similarly arranged sets of locations, simulations were done on one long segment with a finer sample resolution to examine a more local-scale covariance.

The studied stream segment covers a length of close to 86,000 meters (the segment in Figure 4-2 with the overlaid points depicting one sample). The sample resolution was set to have an average point spacing at about one-fourth the distance equal to the interval distance between nodes of the variogram fit to the original basin survey data. If the observed data is binned to have approximately 30 observations per bin, for the original data set, this allows for 2- or 3-node variograms to be fit to the observed data, where the first and last nodes are at distance 0 and maximum distance observed, and the second-to-last node is set at the one of 30th/40th / or 50th percentile of the distances. This produces a lag interval on the original data of about 5400, 7600 and 5400 meters respectively (where the lag interval is *c/k* and *k* depends on the specified average number of observations per bin). By design, the average point spacing in the purposive sample is approximately equal to this distance, so that for a CSR arrangement there are an adequate number of smaller distances to provide information about the covariance structure at close distances, where the gradient of the variogram is steepest.

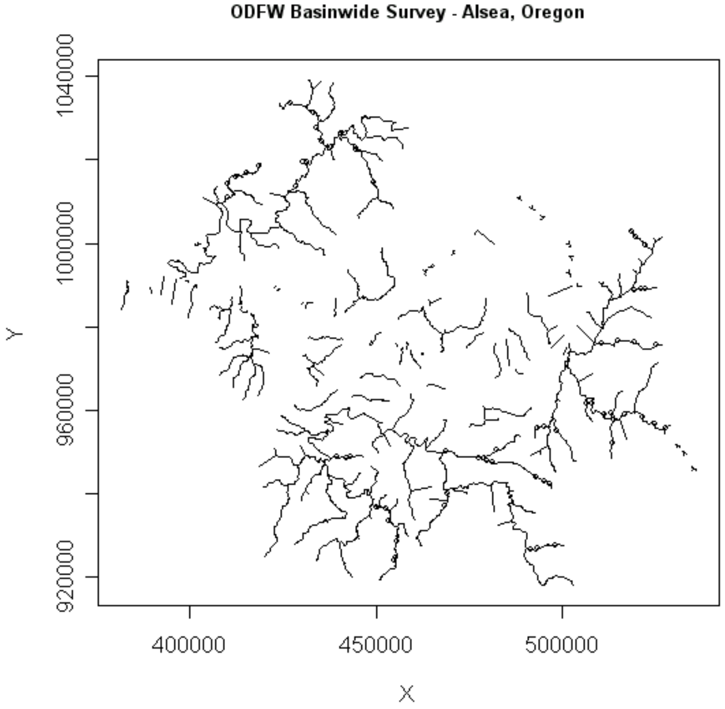

Figure 4-4 Open circles indicate the locations of the Oregon Department of Fish & Wildlife Alsea basin-wide survey on the stream network in the Alsea basin in Oregon.

On the local scale (on the segment studied), the sample resolution was about four times finer than the average sample resolution in the original data provided on the Alsea basin. To make the sample process on the simulated response behave as reasonably representative of an actual response, the response is simulated at intervals about 5% of the inter-node distance of the fit variogram – giving a simulated response resolution of about 270 to 380 or so meters. Since the variogram on the original data had 2-3 nodes, the moving average process is modeled as a 2nd or 3rd order process with lag interval about 20 times the distance of the desired interval of the simulated response. To achieve a simulated response at a finer resolution, a higher-order process is derived by interpolating between the original coefficients and applying the higher order process to innovations spaced at the desired interval.

The immediate consequence of this procedure is that the variance of the innovations must be adjusted to behave as though the lag scale *c/k* is appropriate for the innovations at this higher-resolution spacing. For simplicity, the variance is set to the original lag scale divided by 20 – the number of points to be simulated on each inter-node distance interval.

Although this introduces some error not quantified here, if the number of nodes of the fitted variogram had been chosen such that the lag scale would be the same distance apart as the desired interval, this would be the theoretical variance used for the innovations. A comparison of realized samples from each of the true and mimic simulated responses suggests that the adjusted variance of the innovations is reasonable. That is, the approximate ranges and low, medium and high frequency fluctuations in the samples of each of the responses are similar for many trials (an example is shown in Figure 4-5).

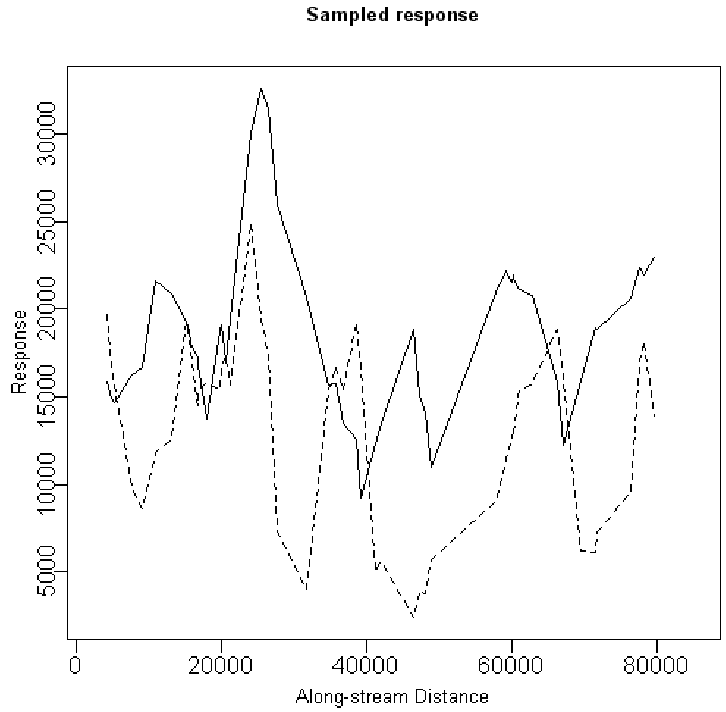

Figure 4-5 A typical pair of sample profiles taken on the true and mimic moving-average response on the studied segment of the Alsea stream network. The general agreement of ranges and degree of fluctuations validates the variogram fitting and response simulation methods.

DISSERTATION CONCLUSION

This research provides basis and justification for describing the uncertainty in an extrapolation from a non-probability sample collected on a continuous domain, where the uncertainty is characterized as the sample process variability that would occur if the extrapolation from observed elements to the domain were made with observations collected at different but similarly arranged locations in the domain. Very little has been done to address the application of non-probability sample data to estimating population characteristics. Where there has been application, the data were used typically by employing post-stratification or similar analyses to adjust or augment a probability sample.

The basis for describing the uncertainty of an estimate based on a non-probability sample exploits the covariance structure in the regionalized response typical of a continuous domain, and the interaction between the sample intensity and arrangement and the range of the covariance of the response. Unlike finite population unit identifiers, the locations of the elements observed on a continuous domain – the element-identifying information, are not ancillary to estimating the sample process variance. This is explored in the first manuscript (Chapter 2).

In this manuscript, the model- and design-based approaches to sampling and estimation are compared. The models of stochasticity, emphases of applications, estimation methodologies (method-of-moments vs. likelihood approaches) and the component of variation addressed by each approach are discussed and compared. Since design-based survey methodologies have traditionally been developed and applied on finite populations, the paper describes the idiosyncrasies of sampling and estimation on continuous domains, including a change from finite inclusion probabilities to continuous inclusion densities. The non-exchangeable covariance structure typical of continuous domains is introduced, emphasizing that information provided by locations being observed is relevant to characterizing variance, which is in contrast to the irrelevance of unit IDs of finite populations (except where the unit ID indicates group membership). The Horvitz-Thompson and Yates-Grundy design-based variance estimators are compared with a model-assisted variance estimation approach that exploits the response's covariance structure, for tessellation-stratified samples taken on a continuous domain with a simulated exponential covariance structure. The model-assisted variance estimator is demonstrated to be more efficient than the two purely design-based variance estimators.

The interaction between the sample process and the underlying response is made explicit by a model of the probability space of the domain, the universe of finite sets generated on that domain and a set measure on the universe of finite sets. The fact that the set locations are taken from a domain for which the observations within a range of covariance are correlated is essential to establishing that the arrangements of the locations and the underlying covariance of the response on the domain are interacting. This suggests that for a certain class of patterns of locations, there would be a predictable covariance structure and therefore a predictable sample process variance *with respect to* that class of locations.

If an agency has a non-probability sample from which a preliminary extrapolation is made to the domain, a reasonable concept of uncertainty would be the amount of variability such an extrapolation would have from observations collected over similarly arranged sets of locations. A critical premise of entertaining this characterization of the uncertainty is that the response's covariance structure is stationary – the mean and the covariance must not depend on the location in the domain. If a practitioner has data collected at locations that are known to be contaminated or distinguished from the baseline response in some other way, this immediately violates this assumption, for the expected response at a hotspot depends on the location. Supposing that, at least for a preliminary study, the response has been observed at locations not specified by a probability sample but neither specifically different from the rest of the region being explored, the responses and the covariance observed between them might be representative of the region. Then the arrangement of the locations on the region and the covariance structure on the response on the region would provide useful information to describe what kind of variability the interested parties might see for extrapolations on similarly arranged locations of observations.

From this setup, an approach is proposed that the uncertainty be characterized by explicitly defining the class of similarly arranged locations and then predicting variability with respect to that class. In Chapter 3, a process of characterizing classes of sets of locations is developed from point pattern metrics. Point pattern metrics are selected based on their potential to differentiate patterns of points that are clustered, dispersed or random. The relationship between a probability density on a point pattern metric and a set measure on sets of locations on the domain is established. Examining the model of the domain and the finite sets of elements on the domain, an argument is presented that by restricting the frequency of measurable sets of outcomes of a metric, this imposes a (unique) measure on the sets of locations and so precisely characterizes a class of patterns of locations.

Chapter 3 examines the utility of various point pattern metrics to assess goodness-of-fit (GOF) of patterns to classes of patterns with regularity, clustering or complete spatial randomness. Three metrics are compared on a domain of areal extent on regular and random classes of patterns – an inner-product metric, a Side-Vertex-Boundary (SVB) Dirichlet-tile metric and a metric based on Ripley's K(t) functions (Ripley (1977)). The SVB and K(t)-derived metrics are demonstrated to have very good utility for assessing GOF to exclude regular patterns from random patterns and vice versa.

The manuscript in Chapter 3 introduces an interesting reversal in tendencies in efficiency between metrics incorporating either all or only neighboring point pair distances on areal domains vs. on linear network domains. The chapter discusses how on areal domains, those metrics that incorporate all inter-point distances are usually more powerful than those based on only the nearest-neighbor distances. On a stream network, the reverse is true – metrics incorporating consecutive distances are more useful than metrics incorporating all inter-point stream-flow distances.

Three other metrics are compared on a stream network domain for regular, random and clustered patterns of locations, based on consecutive stream-flow distances: a metric derived from an exponential distribution of consecutive point distances, a metric based on stochastic rank of consecutive point distances and a 2D version of the SVB metric. The first two of these are demonstrated to have very good specificity for assessing GOF. The 2D SVB is demonstrated to be excellent for excluding non-regular patterns from a class of regular (stratified) patterns. The process of examining the GOF is illustrated on a non-probability sample collected on the Alsea River basin by Oregon Department of Fish and Wildlife (ODFW). The non-probability sample is shown to be consistent with a class of sets of clustered locations.

Having specified a class of sets that are similarly arranged to that of the non-probability sample, a consumer of the data analysis will have a specific reference for which to describe the variability in the extrapolation from the non-probability sample observations and similarly arranged samples. Chapter 4 establishes a basis for describing variability of estimates taken on sets of locations similarly arranged to a non-probability sample by noting the relationship between the spatial order of elements in the samples and the covariance of the sample responses as impacted by that ordering. A Monte Carlo (MC) approach is proposed and examined for efficacy to quantify the sample process variability of an estimate for a particular class of sets of locations and on a response with a particular covariance structure.

The process in application is to model the response covariance structure conditional on the observed data, with which a response is simulated on the entire domain with the appropriate covariance structure. On the domain, sample process variance over the class of point patterns is estimated by MC sampling from the specified class of sets of locations.

The proposed approach is evaluated by examining the relative error between estimates of sample process variance and the empirical sample process variance on a simulated response treated as the true response. The process is tested on a square domain with a simulated response with exponential covariance structure. The empirical variance is observed for 1000 samples realized from each of a grid tessellation stratified (GTS) and Complete Spatial Randomness (CSR) sample process on both the true and mimicking simulated responses. The process is illustrated to work reasonably well on the domain of areal extent, with relative error being 10% and 2% for the two processes, respectively.

On the stream network domain, a moving average process was simulated for the section of the Alsea River basin. For this response and domain, it was necessary to extend the above process to average over multiple moving-average realizations produced to mimic the initial moving average response covariance structure. The relative errors observed were not adequate to recommend this approach for a stream network domain.

Ideally no population characteristic estimates would be based on non-probability samples, given the multitudes of reports that show selection bias introduces a serious risk that the data in the non-probability samples is not representative of the response on the rest of the domain. Estimates are likely to be biased. Due to time and cost constraints, agencies will hope to glean as much useful information as can be mined from data that has been collected by whatever means. The data may have been intended for an objective other than characterizing the response over the rest of the domain, and subsequently there might be interest in attempting to characterize the response if only for preliminary information. The preliminary information may be applied to improve the design of subsequent data collection.

This research develops the opportunity – for studies on continuous domains with regionalized responses – to exploit the interaction between the sample design (resolution and order) and the covariance structure of the response on the domain that influence the sample process variance. The interaction is formalized by stating the probability model for the domain, the universe of finite sets generated on the domain and the set measure on the generated sets. By developing a precise characterization of classes of point patterns, a reference for characterizing uncertainty of an extrapolation from a non-probability sample is established. Characterizing the variability of an extrapolation requires the specification of the class of similarly arranged sets over which the variability in the extrapolation would occur.

REFERENCES

Ripley BD (1977) "Modelling spatial patterns" *Journal of the Royal Statistical Society* B 39 172-212.

BIBLIOGRAPHY

Baddeley A, Turner R 2005 "spatstat: An R package for analyzing spatial point patterns" *Journal Of Statistical Software* 12 (6) 1-42.

Barry RP, Ver Hoef JM 1996 "Blackbox kriging: spatial prediction without specifying variogram models" *JABES* 1(3) 297-322.

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.

Besag J, Diggle PJ 1977 "Simple Monte Carlo tests for spatial pattern" *Appl. Statist.* **26** (3) 327-333.

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.

Buckland ST, Anderson DR, Burnham KP and Laake JL 1993 *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall (London).

Carroll, SS 1998 Modelling abiotic indicators when obtaining spatial predictions of species richness. *Environmental and Ecological Statistics* 5(3): 257-276.

Chung, KL 2001 *A Course in Probability Theory* Academic Press (San Diego).

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.

Cooper C 2006 "Sampling and estimation on continuous domains" *Environmetrics (in press).*

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.

Dalenius T, Hájek J, Zubrzycki S 1961 On plane sampling and related geometrical problems *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 125-150 Neyman J (ed.) University of California Press (Berkeley, CA).

Diamond P, Armstrong M 1984 Robustness of variograms and conditioning of kriging matrices. *Mathematical Geology* 16: 809-822.

Diggle PJ 1979 "On parameter estimation and Goodness-of-fit testing for spatial point patterns" *Biometrics* 35, 87-101.

Geyer C 1999 "Likelihood inference for spatial point processes" in *Stochastic geometry: likelihood and computation*, edited by Barndorff-Nielsen O, Kendall WS, Lieshout MNM; Chapman & Hall (Boca Raton, FL).

Hansen MH, Madow WG, Tepping BJ 1983 An Evaluation of model-dependent and probability-sampling inferences in sample surveys. *JASA* 78: 776-793.

Hope ACA 1968 "A simplified Monte Carlo significance test procedure" *Journal of the Royal Statistical Society Ser. B* 30, 582-598.

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

Kelly FP, Ripley BD 1976 "A note on Strauss's model for clustering" *Biometrika* 63(2), 357-360.

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-2002-07).

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.

Overton JMcC, Young TC, Overton WS 1993 "Using 'found' data to augment a probability sample: procedure and case study" *Environmental Monitoring and Assessment* 26 65-83.

Paulsen SG, Hughes RM, Larsen DP 1998 "Critical elements in describing and understanding our nation's aquatic resources" *Jo. Of the American Water Resources Association* 34, 995-1005.

Peterson SA., Urquhart NS, Welsh, E. B. 1999 "Sample representativeness: a must for reliable regional estimates of lake condition" *Environmental Science and Technology* 33: 1559 - 1565.

Rao JNK 2003 *Small Area Estimation* (Wiley).

Rawlings JO, Pantula SG, Dickey DA 1998 *Applied Regression Analysis – A Research Tool* (2nd Ed.) (Springer).

Ripley BD 1981 *Spatial Statistics* (Wiley).

Royall RM, Cumberland WG 1981 An empirical study of the ratio estimator and estimators of its variance. *JASA* 76(373): 66-80.

Royall RM, Cumberland WG 1985 "Conditional coverage properties of finite population confidence intervals" *JASA* 80, 355-359.

Royall RM 1988 The prediction approach to sampling theory. *Handbook of Statistics* Vol. 6 (Ed. Krishnaiah PR and Rao CR): 399-413.

Schlather, M 2001 "Simulation and analysis of random fields" *R News* 1/2, 18-20.

Simcox AC, Whittemore RC 2004 "Environmental index for assessing spatial watershed sampling networks" *Journal of Environmental Engineering* 130 (6) 622-630.

Stehman SV, Overton WS 1994 Environmental sampling and monitoring. *Handbook of Statistics Volume 12 Environmental Statistics* 263-306.

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.

Strauss DJ 1975 "A model for clustering" *Biometrika* 62(2), 467-475.

Thompson SK 1992 *Sampling* (Wiley).

van Groenigen JW, Siderius W, Stein A 1999 "Constrained optimization of soil sampling for minimization of the kriging variance", *Geoderma* v87, 239-259.

Ver Hoef, JM 2002 Sampling and geostatistics for spatial data *Ecoscience* 9(2): 152-161.

Warrick AW, Myers DE 1987 "Optimization of sampling locations for variogram calculations", *Water Resources Research* 23, 496-500.

Webster R, Oliver MA 1992 "Sample adequately to estimate variograms of soil properties", *Journal of Soil Science* v43, 177-192.

Wolter KM 1985 *Introduction to Variance Estimation* (Springer-Verlag).

Yates F, Grundy PM 1953 Selection without replacement from within strata with probability proportional to size *J. of Royal Stat. Soc. Ser. B* 1: 253-261.

Zimmerman DL, Cressie N 1992 Mean Squared Prediction Error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math*. 44(1): 27-43.