

Spatial properties of design-based versus model-based approaches to environmental sampling

Don L. Stevens, Jr.

Department of Statistics, 44 Kidder Hall
Oregon State University, Corvallis, OR 97333 USA
Tel.: +001 541 737 3587; Fax +001 541 737 3489
stevens@stat.oregonstate.edu

Abstract

It is widely recognized that an efficient sample of a spatially distributed resource will have some degree of regularity. For example, locating sample points at the nodes of a regular grid is an optimal model-based design for some semivariograms and domain shapes. Locating points becomes more complicated if the domain has an irregular shape or if the design incorporates existing sample points. In this talk, I review some model-based techniques, such as simulated spatial annealing, for incorporating prior knowledge in locating new sample points. These techniques are contrasted with design-based techniques, such as generalized random tessellation stratification, that can also incorporate prior knowledge and existing sample points.

Keywords: spatial simulated annealing, optimal spatial design

1 Introduction

A prevalent need for environmental management agencies is to assess the status of some environmental resource, e.g., the magnitude and extent of contaminated water bodies, the quality of benthic community, or the health of a forest. In most cases, the resource is too extensive to permit a census, so the population characteristics must be inferred from a sample. Two conceptual frameworks are available to support inference from sample properties to population characteristics: design-based and model-based. Design-based inference is rooted in survey methodology. It requires a probability sample, and inference is founded on the resulting sampling distribution. Model-based inference requires that the relationship of sample properties to population properties be specified by a model, and the model itself contains the prescription for the inference to the population. Both frameworks encompass methodologies that prescribe sample selection techniques and inference. Both collections of selection methodologies have techniques to incorporate prior information and knowledge.

Over the last several decades, the sharp distinction between model- and design-based conceptual frameworks has been somewhat blurred by the advent of model-assisted design-based techniques. These have been driven by the recognition that utilizing the insight captured by a model leads to more efficient inference (Särndal, *et al.* 1992). Model-assisted techniques are available both at the sample selection stage and at the analysis stage. In many ways, these developments have led to convergence of results, even though the results are obtained from vastly different theoretical bases. I suggest that a model-based approach to sample selection and a design-based approach for a common objective will result in similar-appearing samples, given that both begin with the same prior information and knowledge.

A common model for environmental variables views the data as a realization of a spatial random field. The model might consist of a deterministic mean function, with the spatial correlation modeled by a semivariogram.

1.1 Study context

The kind of sampling we are envisioning take place in an environmental context, for example, samples of water to check contaminant levels, samples of wetlands to assess condition, samples of benthic communities to assess diversity. Environmental populations have spatial structure: proximate locations tend to be influenced by same set of factors, and they tend to share similar substrates. That structure may be patchy rather than smoothly changing because of influences such as localized management practices, localized contamination, localized development, or natural discontinuities.

Furthermore, many populations of interest have existing samples in place. In the USA, monitoring of surface waters has been mandated by federal law since the early 1970's, and many states have their own long-standing monitoring programs. Most frequently, the existing samples are convenience samples, with little or no consideration given to spatial coverage. In some instances, there is a lengthy historical record of observations attached to the existing sample sites. Preservation of historical continuity is important for assessment of long-term change or trend

1.2 Study strategy

This study will compare a model-based technique for selecting sample points in space to a design-based method. We assume we have only a modest knowledge of the underlying population structure. In particular, we assume only that the population has spatial pattern. In the model-based context, the assumption takes the form of assuming that the population is a realization of a spatial random field (Cressie, 1993) or a regionalized variable (Matheron, 1963). In the design-based context, the assumption is that space coordinates provide ancillary information on which to base the design.

A sensible comparison of model- and design-based sampling designs is possible only with a common objective. By their very nature, the two tend to have somewhat different objective. For example, a spatial statistics model may well require some data to estimate either the form of a semi-variogram, or its parameters, or both. If this were the case, the design objective would include variogram estimation as well as, say, prediction of the mean value for a realization of the random field. Variogram estimation is foreign to a design-based approach, which would be concerned solely with the estimation of the mean value of the fixed population corresponding to the particular realization. So for now, I will restrict attention to the case of a known semivariogram.

From a design-based standpoint, the variance we are concerned about is the variance of the estimate of the mean value over repeaters replications of the design, i.e., over repeated sample selections using the same probability design. One general technique that is used to improve the efficiency of sampling designs is balanced sampling. Briefly, a balanced sample is one with the property that the number of samples in any interval of the range of the response is proportional to the extent of the population in that range. Letting n_I be the number of samples in the interval $I = (z_1, z_2)$, we want $n_I \approx n(F(z_2) - F(z_1))$, where $F(\cdot)$ is the distribution function of the response z . In a perfectly balanced sample, the n_I would be constants, but if we knew enough to ensure that, we wouldn't need to be sampling. Often we can achieve approximate

balance over the target response by balancing the sample over a known ancillary variable correlated with the target response. If the population response has spatial pattern, then space itself can serve as an ancillary variable. A spatially balanced design, that is, one that has a high degree of regularity, will be an efficient design.

There is a great variety of sampling techniques aimed at achieving spatial balance. A strict systematic design is one; spatial stratification is another. The one used here is a generalization of a Random-Tessellation Stratified (RTS) design (see, for example, Overton and Stehman (1993), Olea (1984), Dalenius et al. (1961)). A random tessellation is a randomly placed partition of a domain into disjoint polygons, e.g., the cells associated with a randomly-located regular grid. An RTS design selects one or more points from each polygon of a random tessellation, e.g., from each grid cell. A Generalized Random Tessellation Stratified (GRTS) design (Stevens & Olsen, 2004, 2000, 1999) uses recursive partitioning of grid cells to develop a spatial address. The address sequence is randomized in a way that preserves some spatial proximity, and a systematic sample is selected using the sorted random addresses. A GRTS design preserves the spatial balance inherent in an RTS design, but provides substantially more flexibility, for example, the ability to easily accommodate variable probability requirements.

2 Optimal Spatial Design

There are some general results (Matérn, 1960, Bellhouse, 1977) that suggest that the optimality of a design for predicting the mean value is strongly related to the regularity of the design points. In this study, we will use several criteria of spatial regularity to define optimal design. In an unconstrained case, it has been shown that the optimal design locations are at the nodes of a triangular grid (Yfantis, et al., 1987). However, the presence of a boundary, especially an irregular boundary, influences the optimal site locations. Similarly, the presence of existing sample points will influence the location of additional points. One method for locating sample sites is the use of Spatial Simulated Annealing (SSA) (Van Groenigen and Stein, 1998; Van Groenigen, 2000). Sample points are selected to optimize some criterion that reflects the study objective, for example kriging variance, or a measure of regularity of the resulting spatial point process. SSA begins with a set of random locations, and cycles through the points, perturbing each one in turn. At each step, the optimality criterion is calculated. If the new configuration resulting from the perturbation is better than the prior optimum, it is retained as the new optimum configuration. If it's worse, it is retained with a probability that decreases with the number of cycles. The concept behind retaining the suboptimal configuration is to bump the iteration away from a local optimum. Letting the probability of (temporarily) accepting a suboptimal configuration decrease helps to ensure eventual convergence to the global optimum.

2.1 Optimal spatial design criteria.

A number of recent papers have used simulated annealing to locate sampling points (Di Zio, et al., 2004; Lark, 2002; Van Groenigen and Stein, 1998). In this study, we examine several optimality criteria that optimize spatial regularity. Van Groenigen and Stein (1998) minimized the mean shortest distance (MSD) from an arbitrary point in the domain to a sample point. Di Zio et al. (2004) used the fractal dimension in conjunction with Fractal Simulated Spatial annealing. We also introduce a new measure of spatial regularity based on the regularity of the polygons resulting from a Dirichlet tessellation .

2.1.1 Minimum mean shortest distance

Van Groenigen and Stein (1998) defined a criteria called the mean shortest distance, and then minimized it to obtain an optimal point pattern. For S a set of sample points, x a point in target domain D , let $d(x,S)$ be the distance from x to the nearest point in S . Then mean shortest distance is

$$MSD = \int_D d(x,S) dx / |D| \quad (1)$$

Note that for $C(s)$ the Dirichlet polygon of s

$$MSD = \sum_{s_i \in S} \int_{C(s_i)} d(x,s_i) dx / |D| \quad (2)$$

2.1.2 Fractal dimension

Di Zio, Fontanella & Ippoliti (2004) used a measure of fractal dimension related to Ripley's K -function (Ripley 1976, 1977) as a criterion of regularity. The K -function, $K(r)$, is defined as the average number of sites within radius r of a site divided by the intensity of the process. Under Complete Spatial Randomness $K(r) = \pi r^2$, while under regularity, $K(r) < \pi r^2$ and under clustering, $K(r) > \pi r^2$. Di Zio, Fontanella & Ippoliti noted that D , the slope of the best fitting line produced when $\log(K(r))$ is regressed against $\log(r)$, should approach 2 as sites become more randomly dispersed. As sites get more regular, the slope should increase, so $2-D$ is a measure of regularity.

Boundaries present estimation problems for the K -function, because points near boundaries don't have enough neighbors, and so lead to bias. Several corrections available, such as using weights based on distance from boundary, using a "guard area" or buffer around boundary. For square or rectangular domains, one can also use toroidal geometry. With this geometry, distances are calculated on a torus formed by connecting the top to the bottom, and then bending the resulting cylinder around to connect the left edge to the right edge. See Stevens (1997) for more discussion. The examples use toroidal geometry.

2.1.3 Proposed criterion: Mean squared distance to sides, vertices, and boundaries

Imagine a large number of coins of the same size on a flat surface. If one were to push coins together as tightly as possible, the resulting pattern is a triangular grid with an associated hexagonal tessellation (Figure 1)

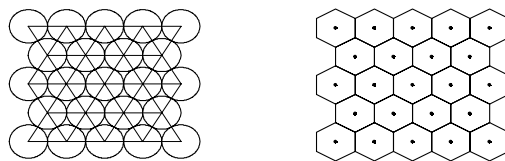


Figure 1 Triangular grid and associated associated hexagonal tessellation.

The hexagonal tessellation for the triangular grid is an example of a Dirichlet tessellation. For an arbitrary point pattern, the Dirichlet tessellation partitions the domain into polygons around

each point such that polygon of a point bounds the area closer to that point than to any other. We might suspect that the Dirichlet tessellation of a compact, regular point pattern should have equal-area polygons, and the polygons should be approximately hexagonal. That suggests that one measure of regularity might be the variance of the polygon area. The area variance is not sensitive to shape, however. A criterion that is sensitive to both variation in area and shape is the variation of the distance from a point to the boundary of its Dirichlet polygon. We define the SVB measure as

$$SVB = \sum_{s_i \in S} \left(\int_{B(D(s_i))} (d(b, s_i) - \bar{d})^2 db \right) / SVB_{NOM} \quad (3)$$

where,

S is a point sample,

$D(s_i)$ is the Dirichlet polygon for point s_i ,

$B(D(s_i))$ is the boundary of $D(s_i)$,

$D(b, s_i)$ is the distance from s_i to the point b ,

\bar{d} is the radius of a circle with area equal to domain area divided by number of samples,

SVB_{NOM} is the mean square deviation for a hexagon with area equal to domain area divided by number of samples.

In practice, the SVB is approximated by the MSD distance from a sample point to Sides, Vertices, and Boundaries relative to a nominal value, where a Side divides adjacent polygons, and a Boundary is a segment of the domain boundary. SVB could be computed locally, because a measure is attached to every sample point. The overall measure is normalized by a single value, but one could use a different value for each point that reflected the target point density, say some measure of local variability, if one had such knowledge. Figure 3 shows a Dirichlet tessellation of a 25 point sample, along with lines that indicate the distance from the point to the polygon vertices, and mid-points of the sides and boundaries.

3 Optimal Designs

SSA was used to derive optimal point patterns for the various criteria. Typical examples of the resulting point patterns for a 50 point sample in the unit square are shown in Figure 2. All of the optimality criteria resulted in points patterns that were obviously much more regular than complete spatial randomness. The fractal dimension criterion without any edge correction shows some obvious edge effects. The edge effects were not evident using toroidal geometry; however, that approach works only for square or rectangular domains. MMSD seemed very slow to compute, but that may be because our algorithm was not as efficient as it might be. The SVB appears to be the most regular, and with a little imagination, the triangular grid axes seem to be emerging. The point patterns produced by the SVB criterion are comparable to those obtained using MMSD, but were much quicker to compute.

4 Simulation Study

To evaluate the benefits of optimal design versus a spatially balanced probability design, we carried out a simulation study. We made the comparison under a model-based perspective: the design points were fixed, and the response surface was varied. In order to create a surface that had spatial pattern but was not necessarily a realization of a stationary, isotropic Figure 2. Optimal point patterns for SVB, Fractal dimension using toroidal distance, MMSD, and a GRTS probability sample.

random field, we created a patchy surface by mixing three smooth surfaces: a plane, a normal density, and a surface with several bumps, plus random noise. Patches were random tessellations of the unit square, generated as Dirichlet polygons of 10 random points. Each patch was then assigned to one of the three surfaces, and a random noise was added. Within a patch, then, the surface was relatively smooth, but severe dislocations could occur at patch boundaries.

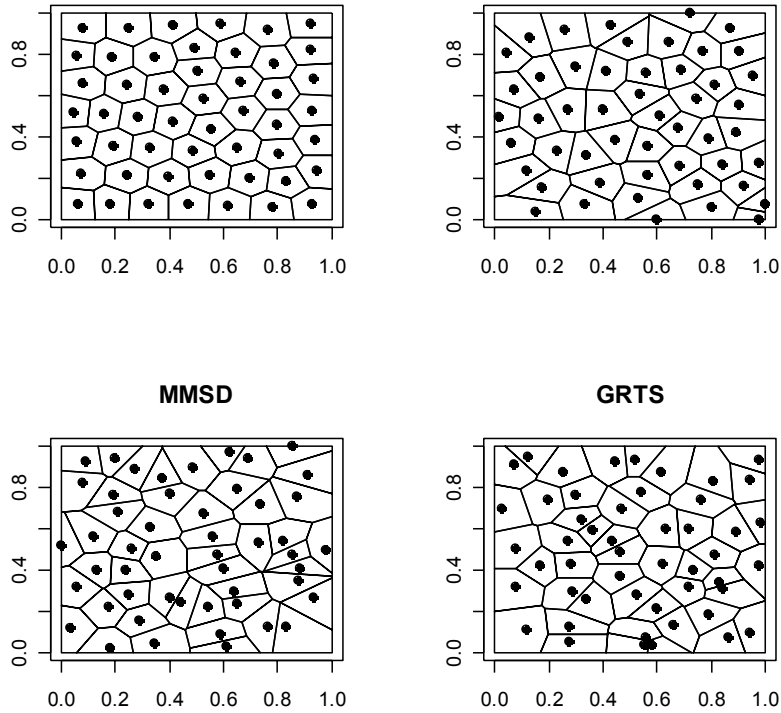


Figure 2 Optimal point patterns for SVB, Fractal dimension using toroidal distance, MMSD, and a GRTS probability sample.

One thousand replicate surfaces were generated, and each replicate was sampled with Uniform Random, Fractal, SVB, and GRTS design points. The results are given in Table 1.

Table 1 Simulation results comparing SRS, Fractal, SVB, and GRTS designs.

	SRS	GRTS	SVB	Fractal
Mean	1.079	1.081	1.091	1.083
Variance	0.0057	0.0039	0.0030	0.0035

Clearly, the designs optimized for regularity performed somewhat better than the probability designs, although the performance of the GRTS design was nearly as good. Not surprisingly,

the designs intended for spatially patterned surfaces did substantially better than a completely random design.

References

- Bellhouse, D.R. 1977, Some optimal designs for sampling in two dimensions, *Biometrika* 64, 605–611.
- Cressie, N., 1993, *Statistics for Spatial Data*. Wiley, New York.
- Dalenius, T., Hájek, J., and Zubrzycki, S., 1961, On plane sampling and related geometrical problems, *Proceedings of the 4th Berkeley Symposium on Probability and Mathematical Statistics* 1, 125–150.
- Di Zio, S., I. Fontenella, and L. Ippoliti., 2004, Optimal spatial sampling schemes for environmental surveys. *Environmental and Ecological Statistics* 11: 397-414.
- Lark, R.M., 2002, Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma* 105:49-80.
- Matérn, B. 1960, *Spatial Variation*, Stockholm, Sweden: Meddelanden från Statens Skogsforskningsinstitut.
- Matheron, G. 1963. Principles of Geostatistics. *Economic Geology* 58:1246-1266.
- Olea, R.A. 1984. Sampling design optimization for spatial functions. *Mathematical Geology* 16: 369–392.
- Overton, W.S. and S.V. Stehman. 1993. Properties of designs for sampling continuous spatial resources from a triangular grid. *Communications in Statistics Part A: Theory and Methods*, 22: 2641–2660.
- Ripley, B.D., 1976, The second-order analysis of stationary point processes, *Journal of Applied Probability* 13:255-266.
- Ripley, B.D., 1977 Modeling spatial patterns, *Journal of the Royal Statistical Society B* 39:983-994.
- Särndal, C., B. Swensen, and J. Wretman .1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York. 694p.
- Stevens, Jr., D.L., 1997, Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*. 8:167-195.
- Stevens, Jr., D. L. and A. R. Olsen, 1999, Spatially Restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics* 4:415-428.
- Stevens, Jr., D.L., and A. R. Olsen, 2000, Spatially-restricted Random Sampling Designs for Design-based and Model-based Estimation. In *Accuracy 2000: Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Delft University Press, The Netherlands. pp. 609-616.
- Stevens, Jr., D.L., and A. R. Olsen, 2004, Spatially-balanced sampling of natural resources. *Journal of the American Statistical Association* 99:262-277
- Van Groenigen, J.W., 2000, The influence of variogram parameters on optimal sampling scheme for mapping by kriging. *Geoderma* 97: 223-236.
- Van Groenigen, J.W., and A. Stein, 1998, Spatial simulated annealing for constrained optimization of soil sampling schemes. *Journal of Environmental Quality* 27:1078-1086.