Weighted Classical Variogram Estimation for Data With Clustering

Cavan REILLY

Division of Biostatistics
University of Minnesota
Minneapolis, MN 55455-0378
(*cavanr@biostat.umn.edu*)

Andrew GELMAN

Department of Statistics
Columbia University
New York, NY 10027
(*gelman@stat.columbia.edu*)

The classical variogram estimate is convenient but can be unacceptably variable. Improved estimators are possible, especially when the locations of the available data are highly clustered. Using a simple theoretical example, we demonstrate that weighting can dramatically increase the efficiency of classical variogram estimates from clustered data. We give expressions for the weights that lead to minimal variance estimators and indicate some obstacles to the use of these weights. We then introduce a simple iterative weighting scheme intended to approximate optimal weighting. We apply the new weighting to the example that motivated this research—estimating the variogram of home radon levels—and demonstrate its performance in a simulation study.

KEY WORDS: Clustered data; Spatial statistics; Variogram; Weighting.

1. INTRODUCTION

1.1 The Classical Variogram Estimate

The classical variogram estimator is a standard tool in geostatistics (see Journel and Huijbregts 1978; Cressie 1993). It has several purposes including investigating the possibility of spatial correlation without postulating a functional form for the variogram, for exploratory purposes in selecting a parametric form for the variogram that ensures the resulting estimate is conditionally nonnegative definite, and for fitting a parametric variogram model using least squares (or some variant, such as weighted least squares). The standard classical variogram estimate—simple averages within distance bins—is convenient but can be highly variable. In this article we propose using weighted averages in each bin to improve the estimation of the variogram (averaged over these bins). We argue that this weighting is most beneficial when the locations at which there are measurements are highly clustered. With the recent growth in applications of spatial methods to problems in epidemiology and the social sciences (see, e.g., Lawson 2001), there are increasing numbers of datasets that have observations clustered in space. This clustering is a natural consequence of clustering of human populations and is less common in applications of spatial methods in geology. Clustering has established advantages for estimating the variogram in that it allows sampling at small spatial resolution, which is essential for prediction under infill asymptotics (see, e.g., Warrick and Myers 1987). Here we show how accounting for this clustering can lead to an estimate that is less variable than the usual classical variogram estimate. In practice, we would first use tools from the analysis of spatial point patterns (such as Ripley's K function) to investigate the extent of clustering of locations where we had samples. Throughout, we suppose that such an investigation has led to the conclusion that clustering of locations is present.

The effect of variation in the extent of clustering of the measured locations depends on assumptions regarding the stationarity of the field. If the random field is not intrinsically stationary, then clustering of observations will result in a variogram estimate characteristic of the region that is more intensively sampled, at least for some distances (as in Chilès and Delfiner 1999, sec. 2.2.6). We do not consider the issue of nonstationarity in this article.

As we show in this article, clustering can lead to inefficient classical variogram estimates even for data arising from a stationary random field. In this situation, we can potentially obtain a more efficient estimate by using weighted averages instead of the simple binned averages used in the construction of the unweighted classical variogram, because we are averaging correlated quantities. Weighted averages have been suggested in this context before (Omre 1984), but the motivation in the past has been to obtain resistant variogram estimates, not more efficient estimates. (See Cressie 1984 for more discussion of resistant estimators.) The approach of Omre was to use a general method for resistant estimation for spatial processes originally attributable to Switzer (1977).

The use of weighted estimates has been considered in the context of fitting parametric models to the classical variogram estimate (see, e.g., Genton 1998; Lee and Lahiri 2002). That work differs from what is proposed here in that the weights used in that work were defined at the level of the bin, not the level of the paired differences. The aim of those methods is to construct a more efficient estimate of a conditionally nonnegative definite variogram estimate by accounting for correlation in the binned averages. In contrast, here we use weights to estimate the bin averages themselves more efficiently, without necessarily modeling the functional form of the variogram. Further gains possibly could be obtained by combining the two approaches.

Finally, we see the present analysis as a first step toward the more ambitious goal of determining the effect of clustering of observed locations on the estimation of variogram parameters in the context of likelihood-based approaches. Nonetheless, ascertaining the effect of the clustering of observed locations on the classical variogram estimate is of interest in its own right. Although likelihood-based techniques are in general preferable

TECHNOMETRICS, MAY 2007, VOL. 49, NO. 2
DOI 10.1198/004017006000000282

to variogram estimation techniques rooted in the classical variogram estimate, there are three reasons for continued interest in the classical variogram estimate. The first is that likelihood-based techniques in geostatistics are not computationally feasible for large datasets. For example, in the application presented in Section 5, the use of likelihood-based techniques would entail the repeated inversion of a 64,000-dimensional matrix. The second reason for interest in the estimate is model selection for the functional form of the variogram. For example, when fitting a convex form for the variogram when in fact the variogram is concave, identifiability problems may be encountered. Thus a common approach to variogram estimation is to examine the classical variogram estimate and use this as a tool to select a functional form for the variogram, the parameters of which are then estimated using likelihood-based techniques. Finally, investigating the possibility of spatial correlation is a common technique in the context of exploratory data analysis. Graphical tools are indispensable for thoughtful data analysis. In any event, an estimator with lower variability would be preferred over the standard classical estimate, which remains a standard tool in geostatistics. In Section 6 we describe some simulations indicating that the proposed estimate is not only more efficient, but also leads to a higher rate of correct model identification.

1.2 Notation and Background

We assume continuous data y_i, $i = 1, \ldots, n$, observed at point locations $\mathbf{s}_i$ in the plane. The primary tool used by geostatisticians for quantifying the extent of spatial correlation in a random field is the variogram. If $y(\mathbf{s})$ is a stochastic process for $\mathbf{s} \in S$ (here S is some subset of the plane) and we assume that $\gamma(\mathbf{s}_1 - \mathbf{s}_2) = \frac{1}{2}\operatorname{var}(y(\mathbf{s}_1) - y(\mathbf{s}_2))$ is well defined for all $\mathbf{s}_1, \mathbf{s}_2 \in S$, then $\gamma(\mathbf{s})$ is called the semivariogram (or sometimes the variogram). If the process $y(\mathbf{s})$ also has a constant mean, then it is said to be intrinsically stationary. We further assume that the process is isotropic (rotationally invariant), so that the variogram is actually only a function of the distance between any two sites: $\gamma(\mathbf{s}_1 - \mathbf{s}_2) = \gamma(\|\mathbf{s}_1 - \mathbf{s}_2\|)$. In deriving variance estimates, we assume that the data follow a multivariate Gaussian distribution and the process is second-order stationary, which is a standard assumption for applications such as producing prediction intervals in kriging, for which variogram estimates are commonly used (see, e.g., Cressie 1993). Although checking stationarity assumptions can be difficult, a classical variogram estimate that grows markedly more than a quadratic or clearly fails to have a sill indicates that second-order stationarity may not be a useful approximation.

To investigate spatial correlation, a common practice of geostatisticians is to first divide a certain portion of the range of observed distances between sampled points into a number of equal-width bins, then find the average squared difference in each bin (see, e.g., Cressie 1993). Denote the realized value of the process at site $\mathbf{s}_i$ by y_i for $i = 1, \ldots, n$, and use $\mathbf{y}$ for the vector of these realized values. Then the classical estimate of the semivariogram is

$$\hat{\gamma}_{\text{classical}}(d) = \frac{1}{2}\frac{\sum_{(i,j) \text{ such that } \|\mathbf{s}_i - \mathbf{s}_j\| \in (d_k - d_1, d_k + d_1]} (y_i - y_j)^2}{N(d_k)} \tag{1}$$

for k such that $d \in (d_k - d_1, d_k + d_1]$, where $N(d_k)$ is the number of observed distances between sampled points that fall in the bin centered at distance d_k. We label these distances in increasing order as $d_k, k = 1, \ldots, K$. Often the set of points $(d_k, \hat{\gamma}_{\text{classical}}(d_k))$ is taken to be the classical estimate of the semivariogram.

1.3 Weighted Classical Variogram Estimates

We consider weighted estimates of the form

$$\hat{\gamma}(d) = \frac{1}{2}\frac{\sum_{(i,j) \text{ such that } \|\mathbf{s}_i - \mathbf{s}_j\| \in (d_k - d_1, d_k + d_1]} w_{ij}(y_i - y_j)^2}{\sum_{(i,j) \text{ such that } \|\mathbf{s}_i - \mathbf{s}_j\| \in (d_k - d_1, d_k + d_1]} w_{ij}}, \tag{2}$$

for k such that $d \in (d_k - d_1, d_k + d_1]$, where a weight is assigned to each pair of points. We also consider the special case of (2) in which weights are assigned to individual points,

$$\hat{\gamma}(d) = \frac{1}{2}\sum_{(i,j) \text{ such that } \|\mathbf{s}_i - \mathbf{s}_j\| \in (d_k - d_1, d_k + d_1]} w_i(d_k) w_j(d_k)(y_i - y_j)^2 \times \left(\sum_{(i,j) \text{ such that } \|\mathbf{s}_i - \mathbf{s}_j\| \in (d_k - d_1, d_k + d_1]} w_i(d_k) w_j(d_k) \right)^{-1}. \tag{3}$$

Expression (3) has the computational advantage of requiring nK weights rather than of order $\binom{n}{2}$.

In Section 2 we demonstrate, using simple examples, that weighted variogram estimates can be much more efficient than unweighted estimates. In Section 3 we derive the optimal weights for the variogram estimate (2) for any configuration of points; unfortunately, with datasets of the size considered in the application here, these general weights require too much computational effort to be useful in practice and require specification of a parametric form for the variogram (which is undesirable for exploratory data analysis). In Section 4 we present an approximate iterative weighting scheme of the form (3). In Section 5 we illustrate the problem that motivated this research—estimating of the spatial correlation of home radon levels—and in Section 6 we explore the effectiveness of the weighted estimate in a simulation study. We discuss the results and the relation to other approaches in Section 7.

2. CLASSICAL VARIOGRAM ESTIMATES FROM CLUSTERS OF POINTS

When data are sampled at locations that are clustered, the classical semivariogram estimator (1) can be considered a sum of contributions of terms from different clusters. For example, if a cluster of m_1 points is a distance d away from a cluster of m_2 points, then these two clusters contribute $m_1 m_2$ terms to the classical variogram estimate at distance d. Our basic idea is to estimate the variance of the sum of these $m_1 m_2$ terms, as a function of m_1, m_2, and d, and then assign a weight that is inversely proportional to the variance.

In general, we are considering a scenario in which there is clustering in the data at a scale smaller than the scale at which the variogram changes; that is, there are clusters of points

within radius δ for which $\gamma(d) \approx \gamma(d+\delta)$. Thus, if the variogram is estimated within bins of width 2δ, then we are concerned with clustering at a scale smaller than δ. This condition is the background within which we construct our weights. The simulation study in Section 6 checks the performance of the weighted variogram estimate under a range of conditions. Such investigations are instructive because we generally do not require data-dependent bin widths.

2.1 Estimating the Variogram for Small Distances

In classical variogram estimation, $\gamma(0^+)$ is estimated from pairs of points less than some distance δ apart. Consider a cluster of $m \geq 2$ points within a circle of diameter δ. These points together form an estimate of the semivariogram at short distances; hence if $y_1, \ldots, y_m$ are data points in such a cluster, then

$$\begin{aligned}\hat{\gamma}(0^+)_{\text{cluster}} &= \frac{1}{(m-1)m}\sum_{i<j}(y_i - y_j)^2 \\ &= \text{sample variance of } \{y_1, \ldots, y_m\}. \end{aligned} \quad (4)$$

This estimate is approximately unbiased (approximate because of the binning) and has a sampling distribution approximately proportional to χ^2_{m-1}, and thus has expectation $\gamma(0^+)$ and variance $\frac{2}{m-1}\gamma(0^+)^2$.

If we consider our combined estimate of $\gamma(0^+)$ from all of the data to be a weighted average from clusters such as this, then it is optimal to weight each cluster inversely proportional to the variance—that is, the cluster's weight should be proportional to $m-1$. (This weighting is actually optimal only in the absence of correlation between the clusters and thus is only an approximation in reality. In Section 3 we discuss why it is not feasible to determine exactly optimal weights.)

We must now translate the cluster weights into weights on pairs of points as in (2) or on individual points as in (3). Allocating the cluster weight of $m-1$ among the $(m-1)m/2$ pairs in the sum (4) leaves a weight of $2/m$ for each pair. [Recall that the smallest possible value of m for estimating $\gamma(0^+)$ is a single pair, or $m=2$, which then gets a weight of 1.]

For the weights to work out in (3), the product of the individual weights on any two points in a cluster must equal their pair weight of $2/m$, so each individual point gets a weight of $\sqrt{2/m}$, where m is the number of points in the cluster of diameter δ containing this point. Once again, points in clusters of size $m=2$ each get weights of 1.

2.2 Estimating the Variogram for Larger Distances

The classical estimate of the variogram at distance d is derived from all pairs of points approximately d units apart. Clustering in the data results in clustering of the pairs. As before, we would like to assign weights to each pair and then to the individual points, so that each cluster is weighted inversely proportionally to the variance of its contribution to the estimated variogram at distance d.

First, consider a pair of points, y_1 and y_2, separated by a distance d that are sufficiently well isolated from the rest of data to be approximately independent of other values of the process. If we assume that the data follow a multivariate normal distribution and the process is second-order stationary, then the difference $(y_1 - y_2)$ has a normal distribution with mean 0 and variance $2\gamma(d)$; we can then evaluate the variance of this pair's contribution to the classical variogram estimate,

$$\begin{aligned}\text{var}(\hat{\gamma}_{\text{isolated}}(d)) &= \text{var}\left[\frac{1}{2}(y_1 - y_2)^2\right] \\ &= \frac{1}{4}\text{E}[(y_1 - y_2)^4] - \frac{1}{4}[\text{E}(y_1 - y_2)^2]^2 \\ &= \frac{1}{4}\cdot 3(2\gamma(d))^2 - \frac{1}{4}\cdot(2\gamma(d))^2 \\ &= 2\gamma(d)^2. \end{aligned} \quad (5)$$

Next, consider the more general case where the variance of the estimate of $\gamma(d)$ obtained from the $m_1 m_2$ pairs corresponds to a cluster of m_1 points separated from a cluster of m_2 points by approximately distance d. The variance of the estimate based on the clusters also can be approximated using the multivariate normal distribution; see Appendix A for details. (This is an approximation due to the binning that the estimate entails.) The approximation is

$$\begin{aligned}&\text{var}(\hat{\gamma}_{\text{clusters}}(d)) \\ &\quad\approx 2\gamma(d)^2 + 2\left(1 - \frac{3}{4m_1} - \frac{3}{4m_2} + \frac{1}{2m_1 m_2}\right)\gamma(0^+)^2 \\ &\qquad - 4\left(1 - \frac{1}{2m_1} - \frac{1}{2m_2}\right)\gamma(d)\gamma(0^+), \end{aligned} \quad (6)$$

which reduces to (5) in the special case where $m_1 = m_2 = 1$. In the limit of large m_1 and m_2, the variance (6) reduces to

$$\lim_{m_1, m_2 \to \infty} \text{var}(\hat{\gamma}_{\text{clusters}}(d)) \approx 2(\gamma(d) - \gamma(0^+))^2.$$

This perhaps surprising result indicates us that as the cluster sizes increase, their contribution to $\hat{\gamma}(d)$ approaches an asymptote in precision. Thus each of the individual $m_1 m_2$ pairs is contributing, in the limit, an amount of information proportional to $1/(m_1 m_2)$. This implies a drastic downweighting of the contributions of large clusters to the variogram.

For finite m_1 and m_2, we can approximate (6) by

$$\begin{aligned}\text{var}(\hat{\gamma}_{\text{clusters}}(d)) \approx 2&\left(\gamma(d) - \left[1 - \frac{1}{m_1}\right]\gamma(0^+)\right) \\ &\times\left(\gamma(d) - \left[1 - \frac{1}{m_2}\right]\gamma(0^+)\right); \end{aligned} \quad (7)$$

see Section A.3 for details. This approximation is exact for $m_1 = m_2 = 1$ and in general has a relative error of at most 1.

Applying inverse-variance weighting to the clusters implies that each of the $m_1 m_2$ pairs within the sum gets a weight of $\frac{1}{m_1 m_2}$ times a factor inversely proportional to (7). Thus, each pair can be given a weight in (2) of

$$w_{ij} = \frac{1}{\gamma(0^+) + [\gamma(d) - \gamma(0^+)]m_1} \times \frac{1}{\gamma(0^+) + [\gamma(d) - \gamma(0^+)]m_2}.$$

More generally, if n_i is the number of points in the cluster corresponding to point i, then each point can be given a weight in (3) of

$$w_i(d) = \frac{1}{\gamma(0^+) + [\gamma(d) - \gamma(0^+)]n_i}. \quad (8)$$

If we estimate these weights using some semivariogram estimate $\hat{\gamma}(d)$ (that is not necessarily conditionally nonnegative definite), then we use the absolute value of the difference between $\hat{\gamma}(d)$ and $\hat{\gamma}(0^+)$ in the last expression,

$$w_i(d) = \frac{1}{\hat{\gamma}(0^+) + |\hat{\gamma}(d) - \hat{\gamma}(0^+)|n_i}. \quad (9)$$

For isolated points (i.e., $n_i = 1$) the weight (8) reduces to $w_i(d) = \frac{1}{\gamma(d)}$, and for large n_i, the weight assigned to point i is approximately proportional to $1/n_i$.

3. OPTIMAL WEIGHTS FOR THE CLASSICAL VARIOGRAM ESTIMATE

It is straightforward to determine the weights that lead to a minimal variance classical estimate of the variogram under the assumption of multivariate normality and second-order stationarity. Toward this end, we introduce some notation. Let $\mathbf{A}$ denote the $\binom{n}{2} \times n$ matrix that maps $\mathbf{y}$ to the $\binom{n}{2}$-vector of pair differences, and let $\mathbf{B}_d$ be an $\binom{n}{2} \times \binom{n}{2}$ diagonal matrix with the (i, i) element equal to $\tilde{w}_i$ if the ith pair difference is in the bin centered at distance d and 0 otherwise. [Here $\tilde{w}_i$ is the weight for the ith pair difference, i.e., $w_{jk}(d)$ for some j, k.] We suppose that there are p_d differences in bin d. Then the optimally weighted classical semivariogram estimate can be expressed as

$$\hat{\gamma}_{\text{opt}}(d) = \frac{1}{2}\mathbf{y}'\mathbf{A}'\mathbf{B}_d\mathbf{A}\mathbf{y}.$$

If we suppose that $\mathbf{y} \sim \text{N}(\mathbf{0}, \mathbf{\Sigma})$, then, letting $\mathbf{\Phi}_d = \mathbf{A}'\mathbf{B}_d\mathbf{A}$, we find that

$$\text{var}(\hat{\gamma}_{\text{opt}}(d)) = \frac{1}{2}\text{tr}(\mathbf{\Sigma}\mathbf{\Phi}_d\mathbf{\Sigma}\mathbf{\Phi}_d).$$

If we let $\mathbf{w}_{\text{opt}}(d)$ denote the vector of optimal weights for distance d and choose this vector to minimize the variance of the classical variogram estimate, then we find that

$$\mathbf{w}_{\text{opt}}(d) = (\mathbf{1}'\mathbf{\Lambda}_d^{-1}\mathbf{1})^{-1}\mathbf{\Lambda}_d^{-1}\mathbf{1},$$

where $\mathbf{\Lambda}_d = \mathbf{L}_d'(\mathbf{A}\mathbf{\Sigma}\mathbf{A}' \otimes \mathbf{A}\mathbf{\Sigma}\mathbf{A}')\mathbf{L}_d$ and $\mathbf{L}_d$ is a matrix of dimension $\binom{n}{2}^2 \times p_d$ that maps the p_d vector of weights at distance d to the vector of length $\binom{n}{2}^2$ obtained by the vec operation (i.e., stacking of the columns) on the matrix $\mathbf{B}_d$. The derivation uses standard results from linear algebra; we do not provide details here, because we do not use these results in what follows.

Using the optimal weights entails practical difficulties. To start with, these weights depend on the variogram itself (through $\mathbf{\Sigma}$), and hence we would have to solve for the weights iteratively. For this process to work, we must use a valid form for the variogram estimate; otherwise, the covariance matrix can be nonpositive definite, which can lead to the nonexistence of $\mathbf{\Lambda}^{-1}$. But then we are assuming the data are multivariate normal (to derive the optimal weights) and a functional form for the variogram; thus it would be more effective to simply perform maximum likelihood estimation. Moreover, the optimal weights cannot be computed for large datasets (of the sort for which classical variogram estimation is often used), because calculating these weights involves inversion of a matrix that is potentially huge.

4. AN ITERATIVE WEIGHTING SCHEME

4.1 Estimating the Variogram

We use the analytical results from Section 2 to derive an iterative algorithm for improved weighted variogram estimation. As explained in Section 3, exact optimal weighting for all pairs of points is not feasible; instead, we assign a weight $w_i(d_k)$ to each individual point i at each distance d_k and then use the product-weighted estimate (3).

We assign the weights in four steps. First, we compute, for each point i, the number n_i of points within a distance $\hat{\delta}$ of the point, including point i itself (so that $n_i \geq 1$). [More generally, we could use a weighted sum that counts nearby points more than faraway points, e.g., $n_i = \sum_j \exp(-\|\mathbf{s}_i - \mathbf{s}_j\|^2/\hat{\delta}^2)$. As with other nearest-neighbor methods, a more complicated counting scheme of this form can possibly increase efficiency by reducing discontinuities at the edges of clusters.]

Second, we estimate $\gamma(0^+)$, the semivariogram at small distances, using (3), with each point i given the weight $w_i(0^+) = \sqrt{2/n_i}$. These weights are noniterative, and so the estimate $\hat{\gamma}(0^+)$ is done.

Third, we assign initial values for the weights $w_i(d_k)$ to be used for estimating the variogram at distances $d > 0$, most simply by setting $w_i(d_k) = 1$ for all k. We then estimate the semivariogram $\gamma(d)$ for each bin using (3), and then use the resulting estimate $\hat{\gamma}(d)$, along with $\hat{\gamma}(0^+)$, which we have already computed, to construct weights $w_i(d_k)$ from (9).

Fourth, we reestimate the variogram using (3) and the newly calculated weights, and repeat the algorithm until convergence of the estimate occurs. A discussion of the convergence of the algorithm is presented in Section 4.4. It usually converges within several iterations.

4.2 Choice of $\hat{\delta}$

An essential component of the algorithm is specification of the distance $\hat{\delta}$ that defines the local density n_i of points in the neighborhood of each point i. As $\hat{\delta} \to 0$, we get $n_i = 1$ for all i, and as $\hat{\delta} \to \infty$, we get $n_i = n$ for all i, so that in either of these limits, all points have equal weights and (3) reduces to the classical unweighted variogram estimate.

For positive and finite values of $\hat{\delta}$, however, the weighted variogram estimate depends on $\hat{\delta}$. Methods are available to investigate clustering at various distances (e.g., Ripley 1976; Reilly, Schacker, Haase, Wietgrefe, and Krason 2002), and these could be used to select a value of $\hat{\delta}$, but we prefer a model-free approach given the nonparametric setting. Our approach is choose $\hat{\delta}$ to approximately minimize the mean squared error of the weighted estimate; that is, we attempt to minimize

$$\frac{1}{K}\sum_{k=1}^{K} \text{E}\{[\hat{\gamma}_{\hat{\delta}}(d_k) - \gamma(d_k)]^2\}.$$

As we show in Appendix B, minimizing this expression is approximately equivalent to choosing $\hat{\delta}$ to minimize [suppressing the dependence of $\hat{\gamma}_{\hat{\delta}}(d)$ on $\hat{\delta}$]

$$\begin{aligned}&\mathrm{E}[\hat{\gamma}(d_K) - \gamma(d_K)]^2 \\ &\quad - \sum_{k=2}^{K}(k-1)\big\{-2\mathrm{E}[\hat{\gamma}(d_k) - \hat{\gamma}(d_{k-1})][\gamma(d) - \hat{\gamma}(d_{k-1})] \\ &\quad + \mathrm{E}[\hat{\gamma}(d_k) - \hat{\gamma}(d_{k-1})]^2\big\}/K.\end{aligned}$$

This is a weighted sum with larger weights assigned to large k. If K is large, then the first term will be negligible compared with the sum, so we ignore this term here. The rest of the expression depends on the unknown $\gamma(d)$, which we estimate by substituting $\hat{\gamma}(d)$. We then choose $\hat{\delta}$ to minimize $\sum_{k=2}^{K}\frac{k-1}{K}[\hat{\gamma}(d_k) - \hat{\gamma}(d_{k-1})]^2$. Because this objective function gives more weight to large d, it tries to take advantage of the near-0 derivative of the variogram for large d, which we expect for a variogram with a finite effective range (where the effective range is the range for a variogram with a finite range and the distance at which the variogram has reached 95% of its sill for variograms with asymptotic sills). If the unweighted classical estimate suggests that the variogram does not have a finite effective range, then this procedure's intuitive basis is somewhat suspect. A classical variogram estimate suggesting that the process does not have a finite effective range would also cast doubt on whether the process has a constant mean and is second-order stationary, and hence the assumptions that motivate our weights would be suspect as well. For computational purposes, we typically use a grid search to find the minimizing value of $\hat{\delta}$.

4.3 Switching Estimates

At the start of Section 2, we considered the situation in which the extent of clustering is at a scale smaller than the rate of change of the variogram [i.e., $\gamma(0^+) \approx \gamma(\delta)$]. If we treated $\hat{\delta}$ as an estimate of the range of clustering witnessed in the data, then we would expect our weighted estimate to outperform the unweighted classical estimate for distances greater than this distance, but it is not clear whether the weighted estimate would outperform the classical estimate for shorter distances. For this reason, we use the weighted estimate for bin averages only for distances $>\hat{\delta}$, and we continue to use the unweighted estimate for distances $<\hat{\delta}$. If there is clustering at large distances, then the weighted estimate will reduce to the unweighted estimate, because the justification for the choice of weights would no longer be relevant. Alternatively, we could use the estimate of δ to make a decision regarding the bin width, but then the estimate may not be relevant for variogram estimation if this bin width is large relative to the rate of change in the variogram.

4.4 Convergence of the Algorithm

We can show that the dynamical system obtained through the recursive algorithm of Section 4.1 is stable with high probability. Recall from (9) that for each bin, our estimate is a weighted average where the weights depend only on the weighted classical variogram estimate at short distances and d itself. Hence to consider the behavior of the algorithm, it suffices to consider what happens in a single bin—that is, what happens for one of these weighted averages. We must show that the algorithm converges for each bin separately. Let

$$h(x) = \frac{\sum_i w_i(x) z_i}{\sum_i w_i(x)},$$

where z_i is the ith squared difference in a bin and $w_i(x)$ is the weight for this squared difference. Then our iterative scheme takes the form

$$x_{k+1} = h(x_k),$$

where h is a random function of the point process giving rise to the measurements and the real valued process on the plane whose variogram we are trying to estimate.

To demonstrate convergence, we show that there exists a unique solution to the equation

$$x^* = h(x^*)$$

with high probability, and the derivative of the function h is <1 in absolute value at this solution with high probability. Now

$$\begin{aligned}h'(x) &= \frac{\sum_i w_i(x)\sum_i w_i'(x) z_i - \sum_i w_i'(x)\sum_i w_i(x) z_i}{(\sum_i w_i(x))^2} \\ &= \frac{\sum_i w_i'(x)(z_i - h(x))}{\sum_i w_i(x)}.\end{aligned}$$

Taking expectations and denoting the set of functions $w_i(x)$ by w, we have that

$$\begin{aligned}\mathrm{E}h'(x) &= \mathrm{E}\big[\mathrm{E}[h'(x)|w, w']\big] \\ &= \mathrm{E}\left[\frac{\sum w_i'(x)\mathrm{E}[z_i - h(x)|w]}{\sum_i w_i(x)}\right] \\ &\approx 0,\end{aligned}$$

because $\mathrm{E}[h(x)|w] \approx \mathrm{E}[z_i|w]$ (where the approximation is due to the binning). Hence the expected value of the derivative of the function is approximately 0. This implies that if there are sufficient squared differences in each bin, then h will be roughly constant, and the algorithm will converge immediately to the unique fixed point. In practice, h usually deviates slightly from a constant function, and so the algorithm executes several iterations before convergence. It is indeed possible, however, that $|h'(x)| > 1$, and so the algorithm can fail to converge. Because any more precise statement would depend on properties of the spatial processes involved, we simply note that failure of the algorithm to converge can be remedied by pooling bins, which reduces the variation in h'. This strategy has worked every time that we have used the procedure. Finally, we note that our weight functions $w_i(x)$ are actually nondifferentiable functions (because they involve the absolute value), but this poses no problem, because we can always suppose that in practice we are actually using a regularization of the absolute value function that has the necessary differentiability.

5. APPLICATION TO HOME RADON MEASUREMENTS

We now apply the method to the problem that motivated the research: variogram estimation for a dataset of radon measurements in England. This dataset consists of approximately 64,000 observations in homes in a 40-km circular area (see Fig. 1). We work with the log radon level because this is better approximated by a normal distribution. Prolonged exposure to radon has been associated with lung cancer; hence identifying houses with high radon levels is a public health concern. Prediction of home radon levels is an important first step that a homeowner should take in ascertaining health risks in a dwelling (Nazaroff and Nero 1988). If a neighbor's radon level is known, then this information is potentially useful for predicting one's own radon level. This is because radon is frequently associated with certain soil features and is known to have a higher concentration near certain deposits, such as uranium deposits. The presence of such geological features is likely to impact the radon levels in multiple nearby houses. To use the neighbor's radon level, one must have an estimate of the correlation in home radon levels as a function of distance. For our application, we then use this correlation function (obtained from the English dataset) in the United States to predict radon levels in houses based on neighboring houses. This is necessary because we do not have a comparable dataset for the United States but nonetheless would like to make predictions using information available on any neighbor's radon level. Thus, although our application does involve spatial interpolation, it would not be accurately described as kriging, given that the correlation structure is estimated from an entirely different location.

Figure 2 shows the unweighted and weighted variogram estimates (with δ estimated using the method from Sec. 4.2). The estimated value of δ (here $\hat{\delta} = 138$) is such that we should use the unweighted estimate for the first two bins; nonetheless, we display both sets of estimates here for illustration. One undesirable feature of the unweighted estimate is the apparent periodic behavior of the estimated variogram. Such behavior is difficult to interpret in a spatial setting, although this is mathematically possible. The estimated variogram using the weighted method has a more reasonable interpretation and suggests that an exponential model is suitable. This difference is attributable to the lower variability of the weighted estimate. Such periodic behavior is often witnessed due to the substantial correlations that exist between bin averages in variograms.

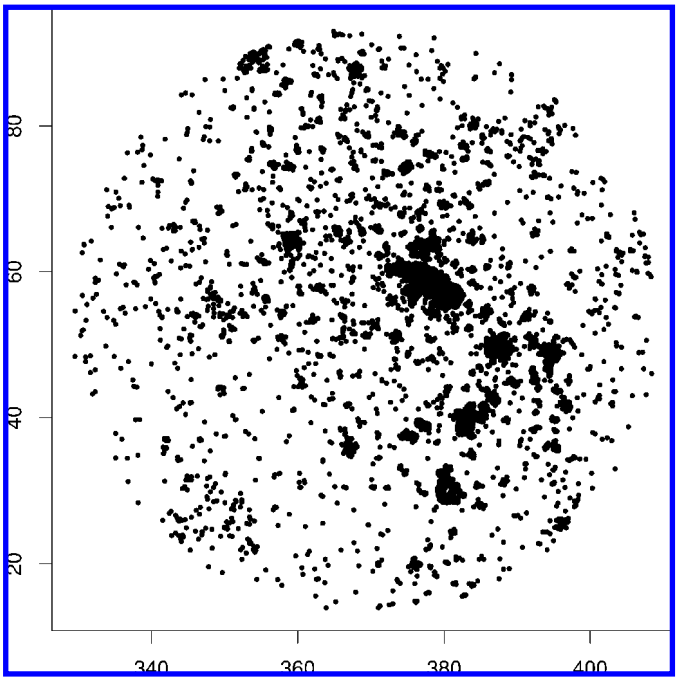

Figure 1. The Sampled Locations for the Radon Data, Which Were Recorded at All Houses in a Small Circular Area in England. The axes are in kilometers from an arbitrary zero. Each point represents the location of a radon measurement.

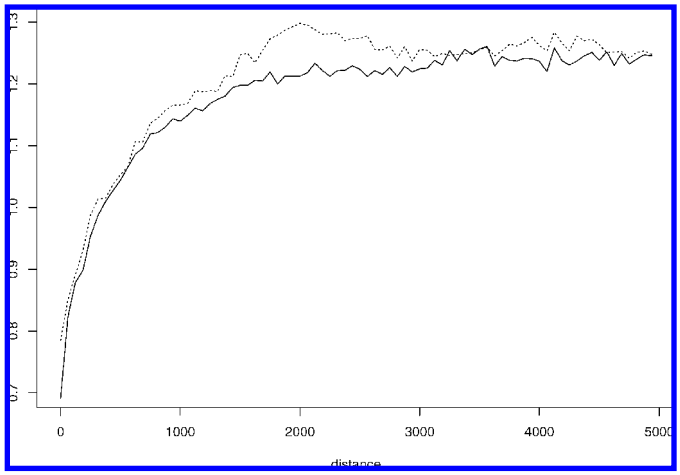

Figure 2. The Weighted (—) and Unweighted (···) Classical Semivariogram Estimates for the Radon Dataset, With δ Estimated From the Data. The weighted estimate has not been smoothed in any way; its cleaner appearance derives from the greater efficiency of the estimate at each distance.

The weighted classical variogram estimate in Figure 2 was used to fit an exponential variogram model using Cressie's (1985) approximation to weighted least squares, which was then used to derive a correlation function. This function in turn can be used by homeowners to update the predictive distributions for their home radon levels given measurements from nearby houses. This updating is performed using a previously-fitted hierarchical linear regression of home radon measurements (Lin, Gelman, Price, and Krantz 1999).

Suppose that we are interested in a particular house's log radon level, given a vector of measurements $\mathbf{x} = (x_1, \ldots, x_m)$ at m nearby houses. We define θ_1 as the log radon level that we are trying to predict, $\boldsymbol{\theta}_2$ as the m-vector of log radon levels for the nearby houses, $\mathbf{X}_1$ as the vector of regression predictors for the target home, and $\mathbf{X}_2$ as the matrix of predictors for the m other homes. We are interested in $p(\theta_1|\mathbf{X}_1, \mathbf{X}_2, \mathbf{x})$. (We implicitly condition on the characteristics of the house for which we desire the prediction and the characteristics of the nearby houses.) We suppose that we observe the radon level in any home subject to independent measurement error with variance σ^2; hence $\mathbf{x} = \boldsymbol{\theta}_2 + \boldsymbol{\epsilon}$ for some mean-0 error $\boldsymbol{\epsilon}$. Here the variance of this noise is sufficiently well understood to be taken as known ($\sigma = .47$), but if this were not the case, then it could be estimated from data. From the model of Lin et al. (1999) we have approximately (given the home characteristics and location) $p(\theta_1|\mathbf{X}_1, \mathbf{X}_2) = p(\theta_1|\mathbf{X}_1) = \mathrm{N}(\theta_1|m_1, s_1^2)$ for some val-

ues m_1 and s_1^2 that depend on the location and the house-level predictors. Similarly, $p(\boldsymbol{\theta}_2|\mathbf{X}_1, \mathbf{X}_2) = \mathrm{N}(\boldsymbol{\theta}_2|\mathbf{m}_2, \mathbf{S}_2)$ for some m-vector $\mathbf{m}_2$ and $m \times m$ matrix $\mathbf{S}_2$, the nondiagonal elements of which we estimate using the standard deviations from the regression model and the correlations based on our estimated variogram; thus $\mathbf{x}|\mathbf{X}_1, \mathbf{X}_2 \sim \mathrm{N}(\mathbf{m}_2, \mathbf{S}_2 + \sigma^2\mathbf{I})$, where I represents an $m \times m$ identity matrix. If we use the notation $\mathrm{cov}(\theta_1, \boldsymbol{\theta}_2|\mathbf{X}_1, \mathbf{X}_2) = \mathbf{S}_{12}$, then $\mathrm{cov}(\theta_1, \mathbf{x}|\mathbf{X}_1, \mathbf{X}_2) = \mathbf{S}_{12}$, and so we find (after an application of the theorem on the regression of components of a multivariate normal vector on one another) that $\theta_1|\mathbf{X}_1, \mathbf{X}_2, \mathbf{x} \sim \mathrm{N}(m_1 + \mathbf{S}_{12}(\mathbf{S}_2 + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \mathbf{m}_2), s_1^2 - \mathbf{S}_{12}\mathbf{S}_2^{-1}\mathbf{S}_{21})$. For the most common setting for this particular application of $m = 1$ (a measurement from a single neighbor), we have implemented these computations on a radon measurement and remediation website (*www.stat.columbia.edu/radon*).

6. SIMULATION STUDY

Because our estimator does not have any guaranteed optimality properties, here we demonstrate the gains in efficiency that are possible using this estimator. Because we are primarily interested in the efficiency of our estimator as a function of the extent of clustering, we vary the extent of clustering (we use two levels here). We also vary the extent of spatial correlation (two levels: long-range and short-range autocorrelation) and the sample size (with mean number of points either 250 or 500). We consider two types of point processes to model the locations where there are measurements: homogeneous Poisson and a Poisson cluster process, both defined on the unit square. Samples from the Poisson cluster process are generated by first simulating the parent locations uniformly over the region. Then offspring locations around each parent are simulated from a bivariate normal distribution with standard deviation .03 and no correlation. We always use 10 parents and vary the sample size by varying the number of offspring. Throughout we assume that the semivariogram has the exponential form

$$\gamma(d) = 1 - \mathrm{e}^{-\phi d},$$

where $\phi = 2$ for the long-range variogram and $\phi = 20$ for the short-range variogram. Finally, we set the marginal variance of the process to 1. The choice of marginal variance and the presence of a nugget effect have no impact on the simulation results. For each simulation condition, we used 1,000 independent samples. We repeated the simulations with different seeds, and repeated some conditions with 10,000 samples. These exercises indicate that the Monte Carlo error does not affect the overall conclusions.

In general the performance of the estimator improves with the extent of clustering and with larger sample sizes. Simulation results for the long- and short-range variograms are shown in Figures 3 and 4. The weighted estimate is more efficient for some bins even when the point process is homogeneous Poisson. On average, not much is gained in efficiency for this case; the average efficiency across bins is about 1.1 for these simulations. In contrast, substantial gains in efficiency are possible when there is considerable clustering (as high as 1.8 for one bin average and on average 1.3–1.5). The simulations demonstrate that the estimate is more efficient for the short-range variogram than for the long-range variogram. The relative efficiency is always 1 in the first bin because if $\hat{\delta} = 0$, then the weighted estimate equals the unweighted estimate, whereas if $\hat{\delta} > 0$, then we use the unweighted estimate in the first bin; hence the estimates coincide in this case as well. Finally, we also computed Cressie's robust estimate (Cressie 1984) in these simulations, but the efficiency of this estimate was invariably lower than the

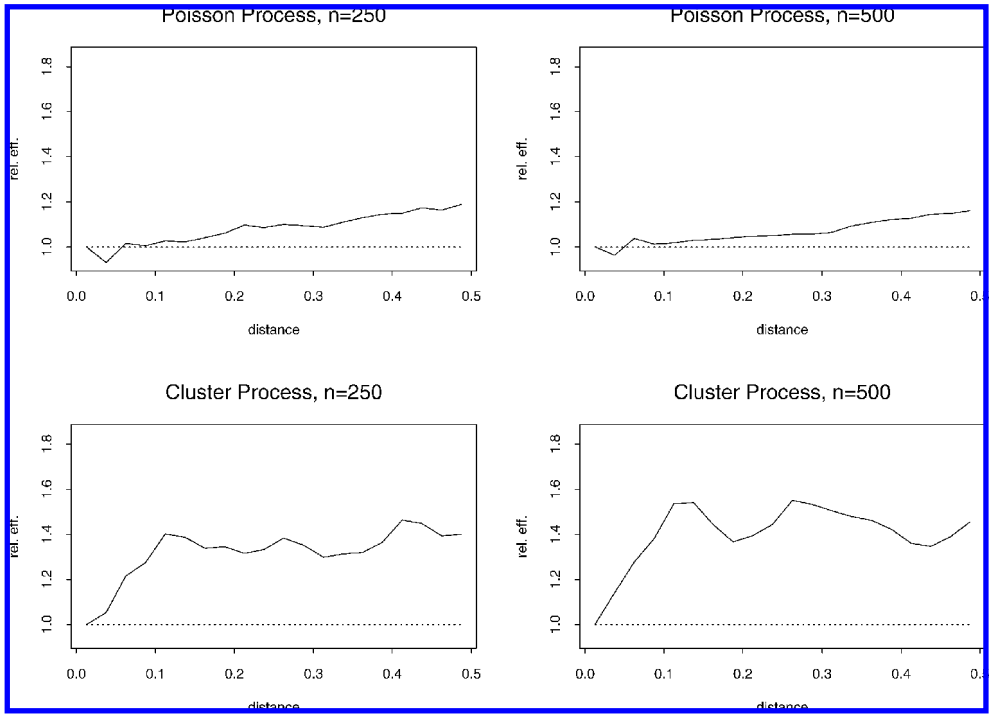

Figure 3. Efficiency of the Weighted Classical Semivariogram Estimate Relative to the Unweighted Average for a Simulation Study Assuming Long-Range Spatial Correlations. We consider two point processes and two sample sizes.

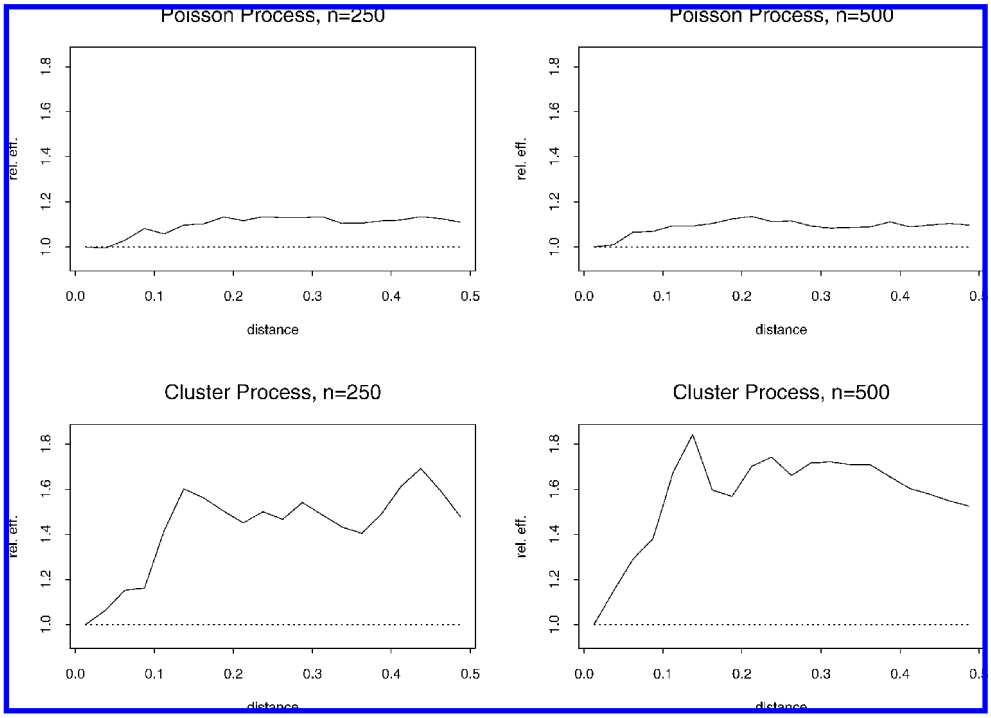

Figure 4. Efficiency of the Weighted Classical Semivariogram Estimate Relative to the Unweighted Average for a Simulation Study Assuming Short-Range Spatial Correlations. We consider two point processes and two sample sizes.

unweighted estimate. This is not surprising given that the goal of that estimator is to construct a robust estimate, not a more efficient estimate, and as such is more variable than the simple unweighted average.

6.1 Improved Model Selection

Although a more efficient estimate of the classical variogram estimator is of interest in its own right, because this estimator is often used as an exploratory tool to investigate the possibility of spatial correlation and as an intermediate step in prediction, as discussed in Section 1, we also note that the improved estimator also leads to higher rates of correct model identification. To investigate this, we extended the previous simulations to fit several parametric forms (exponential, spherical, and power) to the various binned average estimators (i.e., the unweighted estimate, the weighted estimate introduced here, and Cressie's robust weighted estimate) using the approximation to weighted least squares mentioned previously. The parametric form that resulted in the lowest weighted error measure was then deemed the form selected by that particular binned average–type estimate. So for each binned average–type estimator, given a set of simulation conditions, we can determine the proportion of the time that a particular estimator identified the correct parametric form. For datasets with no clustering of observed locations, all three binned average estimators performed comparably, but for the cluster process described earlier, gains of 5–10% over the unweighted estimate were typical for the weighted estimate, whereas gains over Cressie's weighting were typically in the 10–15% range for the weighted estimate.

7. SUMMARY

With spatially nonuniform data, clusters of points become highly influential in the classical variogram estimate (especially for distances that happen to equal the distance between two clusters). When standard practice is followed and equal weights are given to all pairs of points, the resulting variogram estimate can be inefficient and thus highly unstable. These problems are reduced using a weighted variogram estimate that downweights data points in denser areas. We have demonstrated that one tractable form of these weights can lead to improved estimates.

We have made a number of assumptions in the present work, and generalizing these could add to the value of the method. It is not clear how the assumption of intrinsic stationarity in lieu of second-order stationarity would affect the weights, but other assumptions clearly would have a meaningful impact. As a simple example, if the observations had heavier tails than a normal distribution, then the sample variance of y would be greater than in the case of normally distributed data. Hence when we estimated $\gamma(0^+)$, we should downweight large clusters even more strongly than in the Gaussian case. Another extension would be to consider the effect of anisotropy. In that case, we could use ellipsoidal neighborhoods to compute the n_i for each i. This necessitates further investigation.

To summarize our method, the goal is to compute a semivariogram estimate $\hat{\gamma}(d)$ as a function of distance d using the weighted-average formula (3), with local point densities n_i computed using an estimated distance scale $\hat{\delta}$. The steps are as follows:

1. Begin with an initial estimate of the semivariogram, $\hat{\gamma}^{(0)}(d_k)$ in bins d_k using the simple method of unweighted averages within bins, and choose an initial value for the distance scale $\hat{\delta}$.

2. For each point i, compute n_i, the number of points (including point i itself) within a radius $\hat{\delta}$ of the point.
3. For computing $\hat{\gamma}(0^+)$, the semivariogram in the first bin, assign to each point i the weight $w_i(0) = \sqrt{2/n_i}$.
4. In each bin d_k for $k > 1$, compute $\hat{\gamma}^{(j+1)}(d_k)$ given the previous estimate, $\hat{\gamma}^{(j)}(d_k)$, assigning to each point i a weight equal to $w_i(d_k) = (\hat{\gamma}(0^+) + |\hat{\gamma}^{(j)}(d_k) - \hat{\gamma}(0^+)|n_i)^{-1}$.
5. Reestimate the semivariogram using (3) and the newly computed weights.
6. Recompute $\hat{\delta}$ to minimize the cross-validatory measure of mean squared error as described in Section 4.2, a measure motivated by the assumed smoothness of the variogram for large distances.
7. Repeat steps 4–6 until convergence.

This procedure is algorithmic and could be used automatically for classical variogram estimation (e.g., replacing the `vgram` command in the fields library for the statistical software R). An R function that implements the algorithm (calling C routines) is available for free at *www.biostat.umn.edu/~cavanr*.

A related idea is to use weighting to adjust for sampling bias, as is done in the survey literature (see Lohr 1999). In our context, this would be appropriate for estimating the average variogram of a nonstationary process. For example, in the radon problem, one might be interested in a geographical rather than a population-weighted average. This would presumably suggest even further downweighting of the points from densely populated areas, with the amount of downweighting depending on the extent of the estimated nonstationarity of the process. Furthermore, these considerations naturally lead to considerations regarding optimal sampling schemes for geostatistical problems in which the goal is to estimate the variogram; such considerations are currently being pursued. Finally, another avenue for investigation of these matters is provided by the theory of marked point processes, but we leave these considerations to future research.

ACKNOWLEDGMENTS

We thank the National Science Foundation for grants SBR-9708424, SES-99-87748, and Young Investigator Award DMS-97-96129. We thank Phillip Price for discussions and Jon Miles from the U.K. National Radiation Protection Board for the data.

APPENDIX A: DERIVATION OF THE SAMPLING VARIANCES FOR THE CLUSTERED VARIOGRAM ESTIMATE

A.1 Setting up the Problem Using the Multivariate Normal Distribution

We derive the variance of $\text{var}(\hat{\gamma}_{\text{clusters}}(d))$, the estimated semivariogram from the clustered data in the simple example of Section 2, under the assumption that the data come from a multivariate normal distribution with correlations as specified by the variogram γ; in fact, we need only work with $\gamma(0^+)$ and $\gamma(d)$.

We will need the following simple moment calculation: If u and v are jointly normally distributed with means 0, standard deviations σ, and correlation ρ, then $\text{E}(u^2v^2) = (1 + 2\rho^2)\sigma^4$. This can be derived using the iterated expectation,

$$\begin{aligned}\text{E}(u^2v^2) &= \text{E}(\text{E}(u^2v^2|u)) \\ &= \text{E}(u^2\,\text{E}(v^2|u)) \\ &= \text{E}\big(u^2(\rho^2u^2 + (1-\rho^2)\sigma^2)\big) \\ &= 3\rho^2\sigma^4 + (1-\rho^2)\sigma^4 \\ &= (1 + 2\rho^2)\sigma^4.\end{aligned}$$

A.2 Evaluating the Variance by Decomposing the Summation

We are now ready to estimate $\text{var}(\hat{\gamma}_{\text{clusters}}(d))$, the variance of the classical semivariogram estimate from two clusters containing m_1 and m_2 points and separated by a distance d. Let y_{ik} represent the ith value of the process in cluster k for $k = 1, 2$.

Our basic strategy is to evaluate $\text{var}(\hat{\gamma}_{\text{clusters}}) = \text{E}(\hat{\gamma}^2_{\text{clusters}}) - (\text{E}(\hat{\gamma}_{\text{clusters}}))^2$. From the definition of the variogram, we have approximately (due to binning) $\text{E}(y_{i1} - y_{i2})^2 = 2\gamma(d)$, and thus $\text{E}(\hat{\gamma}_{\text{clusters}}) = \frac{1}{2m_1m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} 2\gamma(d) = \gamma(d)$, which makes sense; $\hat{\gamma}_{\text{clusters}}$ is an approximately unbiased estimate (approximate due to the binning).

To determine $\text{E}(\hat{\gamma}^2_{\text{clusters}})$, we decompose the square of sums,

$$\hat{\gamma}^2_{\text{clusters}} = \frac{1}{4m_1^2m_2^2}\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(y_{i1} - y_{j2})^2\right) \times \left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(y_{i1} - y_{j2})^2\right). \quad \text{(A.1)}$$

This summation has $m_1^2m_2^2$ terms, which fall into three categories:

1. m_1m_2 terms of the form $(y_{i1} - y_{j2})^4$: products of two identical factors
2. $m_1m_2(m_1 + m_2 - 2)$ terms of the form $(y_{i1} - y_{j2})^2(y_{i'1} - y_{j2})^2$ or $(y_{i1} - y_{j2})^2(y_{i1} - y_{j'2})^2$ for $i \neq i'$ and $j \neq j'$: one of the points is common to both pairs
3. $m_1m_2(m_1 - 1)(m_2 - 1)$ terms of the form $(y_{i1} - y_{j2})^2(y_{i'1} - y_{j'2})^2$ for $i \neq i'$ and $j \neq j'$: two different pairs of points.

We separately figure out the expectation of each of these three terms, nothing that, by symmetry, all of the terms in each category have the same expectation:

1. The squared term, $\text{E}(y_{i1} - y_{j2})^4$, is simplest. The difference $(y_{i1} - y_{j2})$ has approximately mean 0 and variance $2\gamma(d)$. Thus, assuming normality,
$$\text{E}(y_{i1} - y_{j2})^4 \approx 12\gamma(d)^2. \quad \text{(A.2)}$$
2. To evaluate the term $\text{E}((y_{i1} - y_{j2})^2(y_{i'1} - y_{j2})^2)$, we work with the joint normal distribution of the two factors $(y_{i1} - y_{j2})$ and $(y_{i'1} - y_{j2})$. Each approximately has mean 0 and variance $2\gamma(d)$, and their covariance is approximately
$$\text{E}((y_{i1} - y_{j2})(y_{i'1} - y_{j2})) \approx 2\gamma(d) - \gamma(0^+).$$

We can now use the $\mathrm{E}(u^2v^2)$ result derived at the beginning of this appendix to obtain

$$\mathrm{E}\big((y_{i1}-y_{j2})^2(y_{i'1}-y_{j2})^2\big) \approx 12\gamma(d)^2 + 2\gamma(0^+)^2 - 8\gamma(d)\gamma(0^+). \quad \text{(A.3)}$$

3. We determine $\mathrm{E}((y_{i1}-y_{j2})^2(y_{i'1}-y_{j'2})^2)$ in a similar way. The two factors $(y_{i1}-y_{j2})$ and $(y_{i'1}-y_{j'2})$ have a joint normal distribution, each with approximate mean 0 and variance $2\gamma(d)$, and their covariance is approximately

$$\mathrm{E}((y_{i1}-y_{j2})(y_{i'1}-y_{j'2})) \approx 2(\gamma(d)-\gamma(0^+)).$$

The $\mathrm{E}(u^2v^2)$ result then yields

$$\mathrm{E}\big((y_{i1}-y_{j2})^2(y_{i'1}-y_{j'2})^2\big) \approx 12\gamma(d)^2 + 8\gamma(0^+)^2 - 16\gamma(d)\gamma(0^+). \quad \text{(A.4)}$$

Expressions (A.2)–(A.4) give the approximate variances of the individual terms of (A.1); putting them all together yields

$$\mathrm{E}(\hat{\gamma}^2_{\text{clusters}}) \approx 3\gamma(d)^2 + 2\left(1-\frac{3}{4m_1}-\frac{3}{4m_2}+\frac{1}{2m_1m_2}\right)\gamma(0^+)^2 - 4\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)\gamma(d)\gamma(0^+).$$

We can now determine the approximate variance of the estimator,

$$\begin{aligned}&\mathrm{var}(\hat{\gamma}_{\text{clusters}}(d))\\ &= \mathrm{E}(\hat{\gamma}^2_{\text{clusters}}) - (\mathrm{E}(\hat{\gamma}_{\text{clusters}}))^2\\ &\approx 2\gamma(d)^2 + 2\left(1-\frac{3}{4m_1}-\frac{3}{4m_2}+\frac{1}{2m_1m_2}\right)\gamma(0^+)^2\\ &\quad - 4\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)\gamma(d)\gamma(0^+). \quad \text{(A.5)}\end{aligned}$$

A.3 Approximate Factorization of the Variance

We can approximate expression (A.5) with the simpler form,

$$\mathrm{var}_{\text{approx}}(\hat{\gamma}_{\text{clusters}}(d)) = 2\left[\gamma(d)-\left(1-\frac{1}{m_1}\right)\gamma(0^+)\right] \times\left[\gamma(d)-\left(1-\frac{1}{m_2}\right)\gamma(0^+)\right], \quad \text{(A.6)}$$

which conveniently factors into a term for each cluster. This approximation is exact for $m_1 = m_2 = 1$ and in general has a relative error of at most 1. The relative error of (A.6) as an approximation to (A.5) [i.e., $(\mathrm{var}(\hat{\gamma}_{\text{clusters}}(d)) - \mathrm{var}_{\text{approx}}(\hat{\gamma}_{\text{clusters}}(d)))/\mathrm{var}(\hat{\gamma}_{\text{clusters}}(d))$] can be expressed as

$$\frac{1}{4}\,\frac{\frac{1}{m_1}+\frac{1}{m_2}-\frac{2}{m_1m_2}}{x^2-2x(1-\frac{1}{2m_1}-\frac{1}{2m_2})+1-\frac{3}{4m_1}-\frac{3}{4m_2}+\frac{2}{m_1m_2}},$$

where $x = \gamma(d)/\gamma(0^+)$. Note that

$$\begin{aligned}&x^2-2x\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)+1-\frac{3}{4m_1}-\frac{3}{4m_2}+\frac{2}{m_1m_2}\\ &= \left(x-\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)\right)^2\\ &\quad -\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)^2+1-\frac{3}{4m_1}-\frac{3}{4m_2}+\frac{2}{m_1m_2},\end{aligned}$$

so that if $\gamma(d) \geq \gamma(0^+)$ (as would usually be the case), then $x \geq 1$, and so

$$\left(x-\left(1-\frac{1}{2m_1}-\frac{1}{2m_2}\right)\right)^2 \geq \left(\frac{1}{2m_1}+\frac{1}{2m_2}\right)^2,$$

which implies that the denominator in our relative error expression is bounded from below by $\frac{1}{4m_1}+\frac{1}{4m_2}+\frac{2}{m_1m_2}$. Hence the relative error is bounded from above by the expression

$$\frac{\frac{1}{m_1}+\frac{1}{m_2}-\frac{2}{m_1m_2}}{\frac{1}{m_1}+\frac{1}{m_2}+\frac{2}{m_1m_2}},$$

and this latter expression is bounded from above by 1 (the value to which it converges in the limit $m_1 = 1, m_2 \to \infty$).

APPENDIX B: CALCULATIONS RELATING TO THE CHOICE OF $\hat{\delta}$

Here we provide details on the computations for approximating the average variance of the semivariogram. These computations provide a motivation for the method used to determine the distances over which to summarize clustering in the data. The dependence of $\hat{\gamma}_{\hat{\delta}}(d)$ on $\hat{\delta}$ is suppressed throughout.

For $k > 1$,

$$\begin{aligned}&\mathrm{E}[\hat{\gamma}(d_k)-\gamma(d_k)]^2\\ &= \mathrm{E}\big\{[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})]^2\\ &\qquad -2[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})][\gamma(d_k)-\hat{\gamma}(d_{k-1})]\\ &\qquad +[\gamma(d_k)-\hat{\gamma}(d_{k-1})]^2\big\},\end{aligned}$$

but

$$\begin{aligned}&\mathrm{E}[\gamma(d_k)-\hat{\gamma}(d_{k-1})]^2\\ &= \mathrm{E}\big\{[\gamma(d_k)-\gamma(d_{k-1})]^2\\ &\qquad -2[\gamma(d_k)-\gamma(d_{k-1})][\hat{\gamma}(d_{k-1})-\gamma(d_{k-1})]\\ &\qquad +[\hat{\gamma}(d_{k-1})-\gamma(d_{k-1})]^2\big\}\\ &\approx \mathrm{E}\big\{[\gamma(d_k)-\gamma(d_{k-1})]^2+[\hat{\gamma}(d_{k-1})-\gamma(d_{k-1})]^2\big\},\end{aligned}$$

because $\hat{\gamma}(d_k)$ is approximately unbiased (due to binning) for any $\hat{\delta}$. Hence if we define

$$g(k) = \mathrm{E}[\hat{\gamma}(d_k)-\gamma(d_k)]^2,$$

then we find that

$$\begin{aligned}g(k) &\approx \mathrm{E}[\gamma(d_k)-\gamma(d_{k-1})]^2\\ &\quad -2\mathrm{E}[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})][\gamma(d_k)-\hat{\gamma}(d_{k-1})]\\ &\quad +\mathrm{E}[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})]^2+g(k-1)\end{aligned}$$

for $k > 1$. If we define

$$\begin{aligned}a(k) &= \mathrm{E}[\gamma(d_k)-\gamma(d_{k-1})]^2\\ &\quad -2\mathrm{E}[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})][\gamma(d_k)-\hat{\gamma}(d_{k-1})]\\ &\quad +\mathrm{E}[\hat{\gamma}(d_k)-\hat{\gamma}(d_{k-1})]^2,\end{aligned}$$

then we are choosing $\hat{\delta}$ to minimize $\frac{1}{K}\sum_{k=1}^{K} g(k) \approx \mathrm{E}[\hat{\gamma}(d_1) - \gamma(d_1)]^2 + \frac{1}{K}\sum_{k=2}^{K}(K-k+1)a(k)$, and because one term of a is not a function of $\hat{\delta}$, we define

$$b(k) = -2\mathrm{E}[\hat{\gamma}(d_k) - \hat{\gamma}(d_{k-1})][\gamma(d_k) - \hat{\gamma}(d_{k-1})] + \mathrm{E}[\hat{\gamma}(d_k) - \hat{\gamma}(d_{k-1})]^2$$

and minimize $\mathrm{E}[\hat{\gamma}(d_1) - \gamma(d_1)]^2 + \frac{1}{K}\sum_{k=2}^{K}(K-k+1)b(k)$. We can use the recursion for g to solve for $g(k)$ as a function of $g(K)$: $g(k) \approx g(K) - \sum_{j=k+1}^{K} a(j)$. Hence

$$\frac{1}{K}\sum_{k=1}^{K}\mathrm{E}[\hat{\gamma}(d_k) - \gamma(d_k)]^2 \approx g(K) - \sum_{k=2}^{K}\frac{(k-1)a(k)}{K},$$

and we choose $\hat{\delta}$ to minimize $g(K) - \sum_{k=2}^{K}(k-1)b(k)/K$.

[Received August 2004. Revised May 2006.]

REFERENCES

Chilès, J., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley.

Cressie, N. A. C. (1984), "Towards Resistant Geostatistics," in *Geostatistics for Natural Resource Characterization*, *Part 1*, eds. G. Verly, M. David, A. G. Journel, and A. Marechal, Dordrecht: Reidel.

——— (1985), "Fitting Variogram Models by Weighted Least Squares," *Journal of the International Association for Mathematical Geology*, 17, 563–586.

——— (1993), *Statistics for Spatial Data* (2nd ed.), New York: Wiley.

Ecker, D., and Gelfand, A. (1997), "Bayesian Variogram Modeling for an Isotropic Spatial Process," *Journal of Agricultural, Biological and Environmental Statistics*, 2, 347–369.

Genton, M. G. (1998), "Variogram Fitting by Generalized Least Squares Using an Explicit Formula for the Covariance Structure," *Mathematical Geology*, 30, 323–345.

Journel, A. G., and Huijbregts, C. J. (1978), *Mining Geostatistics*, London: Academic Press.

Lawson, A. (2001), *Statistical Methods in Spatial Epidemiology*, New York: Wiley.

Lee, Y., and Lahiri, S. (2002), "Least Squares Variogram Fitting by Spatial Subsampling," *Journal of the Royal Statistical Society*, Sec. B, 64, 837–854.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. (1999), "Analysis of Local Decisions Using Hierarchical Modeling, Applied to Home Radon Measurement and Remediation" (with discussion), *Statistical Science*, 14, 305–337.

Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury.

Nazaroff, W. W., and Nero, A. V. (eds.) (1988), *Radon and Its Decay Products in Indoor Air*, New York: Wiley.

Omre, H. (1984), "The Variogram and Its Estimation," in *Geostatistics for Natural Resources Characterization, Part 1*, eds. G. Verly, M. David, A. Journel, and A. Marechal, Dordrecht: Reidel, pp. 107–125.

Reilly, C., Schacker, T., Haase, A., Wietgrefe, S., and Krason, D. (2002), "The Clustering of Infected SIV Cells in Lymphatic Tissue," *Journal of the American Statistical Association*, 97, 943–954.

Ripley, B. D. (1976), "The Second-Order Analysis of Stationary Point Processes," *Journal of Applied Probability*, 13, 255–266.

——— (1981), *Spatial Statistics*, New York: Wiley.

Stein, M. L. (1999), *Interpolation of Spatial Data*, New York: Springer-Verlag.

Switzer, P. (1977), "Estimation of Spatial Distributions From Point Sources With Application to Air Pollution Measurement," *Bulletin of the International Statistical Institute*, 47, 123–137.

Warrick, A., and Myers, D. (1987), "Optimization of Sampling Locations for Variogram Calculations," *Water Resources Research*, 23, 496–500.