Estimating The Theoretical Semivariogram From Finite Numbers of Measurements

Li Zheng
Kansas Geological Survey, University of Kansas, Lawrence, KS

Stephen E. Silliman
Department of Civil Engineering and Geological Sciences, University of Notre Dame, Notre Dame, IN

Abstract

We investigate from a theoretical basis the impacts of the number, location, and correlation among measurement points on the quality of an estimate of the semivariogram. The unbiased nature of the semivariogram estimator, $\hat{\gamma}(\boldsymbol{r})$, is first established for a general random process, $Z(\boldsymbol{x})$. The variance of $\hat{\gamma}_Z(\boldsymbol{r})$ is then derived as a function of the sampling parameters (the number of measurements and their locations). In applying this function to the case of estimating the semivariograms of the transmissivity and the hydraulic head field, it is shown that the estimation error depends on the number of the data pairs, the correlation among the data pairs [which in turn are determined by the form of the underlying semivariogram, $\gamma(\boldsymbol{r})$], the relative locations of the data pairs, and the separation distance at which the semivariogram is to be estimated. Thus, design of an optimal sampling program for semivariogram estimation should include consideration of each of these factors. Further, the function derived for the variance of $\hat{\gamma}_Z(\boldsymbol{r})$ is useful in determining the reliability of a semivariogram developed from a previously established sampling design.

Introduction

Given a trend-removed random field, $Z(\boldsymbol{x})$, the theoretical semivariogram, $\gamma_Z(\boldsymbol{r})$, is defined as

$$\gamma(\boldsymbol{r}) = \frac{1}{2}var[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{r})] = \frac{1}{2}\langle[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{r})]^2\rangle \quad (1)$$

where $\boldsymbol{x}$ is location, $\boldsymbol{r}$ is the vector lag distance, var stands for the variance and the bracket designates the ensemble averaging. Although the determination of the theoretical semivariogram requires ensemble averaging, we are often limited in practice to one realization and a finite number of measurements of $Z(\boldsymbol{x})$. Thus, the ensemble average is commonly estimated from the spatial average by assuming ergodicity. For example, given n measurements $\{Z(\boldsymbol{x}_i),\ i=1,...,\ n\}$ from a single realization, the most commonly used semivariogram estimator, $\hat{\gamma}(\boldsymbol{r})$, for a specified vectorial lag distance, $\boldsymbol{r}$,is (*Matheron*, 1965),

$$\hat{\gamma}(\boldsymbol{r}) = \frac{1}{2N(\boldsymbol{r})}\sum_{i=1}^{N(\boldsymbol{r})}[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})]^2 \quad (2)$$

where $N(\boldsymbol{r})$ is the number of the data pairs, $Z(\boldsymbol{x}_i)$ and $Z(\boldsymbol{x}_i+\boldsymbol{r})$, available from the n measurements.

The quality of the estimate of the semivariogram for a given $\boldsymbol{r}$ thus depends on the bias and precision of $\hat{\gamma}(\boldsymbol{r})$, as an estimator for $\gamma(\boldsymbol{r})$. Many have used $N(\boldsymbol{r})$, the number of the data pairs, $Z(\boldsymbol{x}_i)$ and $Z(\boldsymbol{x}_i+r)$, as an index for measuring the reliability of $\hat{\gamma}(\boldsymbol{r})$. For example, a common “rule” applied to estimation of the semivariogram is that at least 30 pairs of

measurements, $Z(\boldsymbol{x}_i)$ and $Z(\boldsymbol{x}_i + \boldsymbol{r})$, are required for each lag distance $\boldsymbol{r}$ in order to ensure a reliable semi-variogram estimate (e.g., *Journel and Huijbregt*, 1978). Review of the literature indicates that this rule is derived from a simplification of earlier work by *Matheron* (1965). *Webster* and *Oliver* (1992) have studied this rule through numerical sampling of a random field generated with a known semivariogram. Using a uniform, square sampling grid and comparing sample semivariograms against the known underlying semivariogram, these authors argued that at least 200-300 measurements are needed to estimate a semivariogram reliably.

This emphasis on $N(\boldsymbol{r})$ has also resulted in efforts to devise algorithms to maximize $N(\boldsymbol{r})$ by adjusting the placement of a fixed number of measurements. *Russo* (1984) and *Warrick* and *Myers* (1987), for example, present algorithms which optimize the location of sampling points based on a series of constraints, including constraints on the number of sample points in each lag distance. A recent paper by *Conwell et al.* (1997) extended this earlier work by taking into account the role of measurement instruments in the sampling design. *Morris* (1991) argued that accurate estimation of the semivariogram depended not only on $N(\boldsymbol{r})$ but also on the correlation among the measurements. He proposed an alternative index, the maximum equivalent uncorrelated pairs, as a measure of the estimation accuracy. However, this index applies only to the special case of a concave semivariogram.

Other methods for determining the accuracy of semivariogram estimates involve Monte Carlo simulation (*Russo* and *Jury*, 1987; *Corsten* and *Stein*, 1994; etc.) and the subsampling method (*Chung*, 1984; *Shafer* and *Varljen*, 1990). The Monte Carlo approach obtains the confidence interval via repeated sampling from multiple realizations of a random field. The subsampling method involves subdividing a measurement set into sub-samples, and then estimating the semivariogram from each subgroup of samples, thus allowing characterization of

the variation of parameter estimates between sub-samples.

In the present paper, the variance of the semivariogram estimate, $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$, is determined theoretically. This theoretical result is used to obtain a functional relationship between $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$ and the sampling parameters (e.g., the number of the measurements and their locations). This general relationship then enables us to examine the interactions between the magnitude of $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$ and its various controlling mechanisms. The interactions among the number of measurements, sampling locations, the underlying semivariogram, and measurement error are illustrated through applying this relationship to the estimation of the semivariograms of the transmissivity and the hydraulic head field. Finally, the implications of these results are discussed and extended to possible applications for sampling design and interpretation of the reliability of semivariogram estimates.

Theoretical Background

The bias of the semivariogram estimator, $\hat{\gamma}(\boldsymbol{r})$, can be examined for a random field by determining the ensemble mean value of both sides of (2). This leads to

$$\begin{aligned} \langle \hat{\gamma}(\boldsymbol{r}) \rangle &= \langle \frac{1}{2N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2 \rangle \\ &= \frac{1}{N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})} \frac{1}{2} \langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2 \rangle \end{aligned} \tag{3}$$

Applying (1) to (3),

$$\langle \hat{\gamma}(\boldsymbol{r}) \rangle = \frac{N(\boldsymbol{r})\gamma(\boldsymbol{r})}{N(\boldsymbol{r})} = \gamma(\boldsymbol{r}) \tag{4}$$

Thus $\hat{\gamma}(\boldsymbol{r})$, as given in (2), is an unbiased estimator for $\gamma(\boldsymbol{r})$.

Since $\hat{\gamma}(\boldsymbol{r})$ is an unbiased estimator, $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$ can then be written as,

$$\sigma^2_{\hat{\gamma}(\boldsymbol{r})} = \langle [\hat{\gamma}(\boldsymbol{r}) - \gamma(\boldsymbol{r})]^2 \rangle \tag{5}$$

Substituting (2) into (5) provides,

$$\begin{aligned} \sigma^2_{\hat{\gamma}(\boldsymbol{r})} &= \left\langle \left\{ \left[\frac{1}{2N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2 \right] - \gamma(\boldsymbol{r}) \right\}^2 \right\rangle \\ &= \frac{1}{N^2(\boldsymbol{r})} \left\langle \left\{ \sum_{i=1}^{N(\boldsymbol{r})} \left[\frac{[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2}{2} - \gamma(\boldsymbol{r}) \right] \right\}^2 \right\rangle \end{aligned} \tag{6}$$

To simplify the notation, we define a new random variable, $S(\boldsymbol{x}, \boldsymbol{r})$, as,

$$S(\boldsymbol{x}, \boldsymbol{r}) \equiv \frac{[Z(\boldsymbol{x}) - Z(\boldsymbol{x} + \boldsymbol{r})]^2}{2} \tag{7}$$

and obtain

$$\langle S(\boldsymbol{x}, \boldsymbol{r}) \rangle = \gamma(\boldsymbol{r}) \tag{8}$$

Substituting (7) and (8) into (6) results in

$$\begin{aligned} \sigma^2_{\hat{\gamma}(\boldsymbol{r})} &= \frac{1}{N^2(\boldsymbol{r})} \left\langle \left\{ \sum_{i=1}^{N(\boldsymbol{r})} [S(\boldsymbol{x}_i, \boldsymbol{r}) - \langle S(\boldsymbol{x}_i, \boldsymbol{r}) \rangle] \right\}^2 \right\rangle \\ &= \frac{1}{N^2(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})} \sigma^2_{S(\boldsymbol{x}_i, r)} + \frac{2}{N^2(\boldsymbol{r})} \sum_{j>i}^{N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})-1} cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] \end{aligned} \tag{9}$$

where $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ is the covariance between $S(\boldsymbol{x}_i, \boldsymbol{r})$ and $S(\boldsymbol{x}_j, \boldsymbol{r})$, and $\sigma^2_{S(\boldsymbol{x}_i, r)}$ is the variance of $S(\boldsymbol{x}_i, \boldsymbol{r})$.

Equation (9) is essentially identical to equations provided in classical textbooks on statistics for the variance of the sample mean (*e.g.*, equation (4.1.12), Page 378 in *Benjamin* and *Cornell*, 1970) with the exception that here the variable is $S(\boldsymbol{x}_i, \boldsymbol{r})$ rather than $Z(\boldsymbol{x}_i)$. The key to

understanding the importance of (9) is to determine the dependence of $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ and $\sigma^2_{S(\boldsymbol{x}_i, r)}$ on sampling design.

As given in appendix, for a gaussian random variable $Z(\boldsymbol{x})$, we can derive

$$cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] = \frac{1}{2}[\gamma(\boldsymbol{r} + \boldsymbol{R}_{ij}) + \gamma(\boldsymbol{r} - \boldsymbol{R}_{ij}) - 2\gamma(\boldsymbol{R}_{ij})]^2 \tag{10}$$

where $\boldsymbol{R}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ is the separation distance between $S(\boldsymbol{x}_i, \boldsymbol{r})$ and $S(\boldsymbol{x}_j, \boldsymbol{r})$. Since $\gamma(\boldsymbol{r})$ is not a function of **x,** $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ as given in (10) is also independent of **x**, but depends on the relative separation between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. It is noted further that $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ is anisotropic even if $\gamma(\boldsymbol{r})$ is isotropic. When $\boldsymbol{R}_{ij} = 0$ in (10), *S* has variance of

$$\sigma^2_{S(\boldsymbol{x}, r)} = \sigma^2_{S(r)} = 2\gamma^2(\boldsymbol{r}) \tag{11}$$

Substituting (11) into (9), the expression for $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$ may now be written as

$$\sigma^2_{\hat{\gamma}(\boldsymbol{r})} = \frac{2\gamma^2(\boldsymbol{r})}{N(\boldsymbol{r})} + \frac{2}{N^2(\boldsymbol{r})} \sum_{j>i}^{N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})-1} cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] \tag{12}$$

where the expression for $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ is given in (10). A similar expression for $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$ is used by *Cressie* (1984) in the context of performing weighted least-square fitting.

To facilitate the comparison of various $\sigma^2_{\hat{\gamma}(\boldsymbol{r})}$'s at different lag distances, we divide $\sigma_{\hat{\gamma}(\boldsymbol{r})}$ by $\gamma(\boldsymbol{r})$, to obtain the coefficient of variation, $\rho_{\hat{\gamma}(\boldsymbol{r})}$,

$$\rho_{\hat{\gamma}(\boldsymbol{r})} = \left[\frac{2}{N(\boldsymbol{r})} + \frac{4}{N^2(\boldsymbol{r})} \sum_{j>i}^{N(\boldsymbol{r})} \sum_{i=1}^{N(\boldsymbol{r})-1} coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)] \right]^{\frac{1}{2}} \tag{13}$$

where

$$coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)] = \frac{cov[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]}{2\gamma^2(\boldsymbol{r})} = \frac{cov[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]}{\sigma^2_{S(\boldsymbol{x}_i, r)}} \qquad (14)$$

Hence, $\rho_{\hat{\gamma}(\boldsymbol{r})}$ consists of two parts. The first part is completely determined by $N(\boldsymbol{r})$, the number of data pairs. The second part is a function of both the number of pairs and the correlation among the $S(\boldsymbol{x}_i, \boldsymbol{r})$'s.

In two special cases, the expression for $\rho_{\hat{\gamma}(\boldsymbol{r})}$ can be simplified. First, when the $S(\boldsymbol{x}_i, r)$'s are independent, all $coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]$'s are equal to zero, and

$$\rho_{\hat{\gamma}(\boldsymbol{r})} = \sqrt{\frac{2}{N(\boldsymbol{r})}} \qquad (15)$$

In this case, the estimation accuracy depends only on $N(\boldsymbol{r})$. This relationship is identical to the result from classical statistics, where all the samples are taken independently. The second special case is when $S(\boldsymbol{x}_i, r)$'s are completely correlated with each other; all $coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]$'s are then equal to one. Thus $\rho_{\hat{\gamma}(\boldsymbol{r})}$ reduces to a constant as follows,

$$\rho_{\hat{\gamma}(\boldsymbol{r})} = \left\{ \frac{2}{N(\boldsymbol{r})} + \frac{4}{N^2(\boldsymbol{r})} \left\lfloor \frac{N(\boldsymbol{r})^2 - N(\boldsymbol{r})}{2} \right\rfloor \right\}^{\frac{1}{2}} = \sqrt{2} \qquad (16)$$

That is the estimation accuracy becomes independent of sampling. An example of such extreme cases can be found in a random field whose semivariogram grows quadratically with the increasing separation distance.

Other than in these special cases, the estimation precision of $\hat{\gamma}(\boldsymbol{r})$ (defined here as the magnitude of $\rho_{\hat{\gamma}(\boldsymbol{r})}$) is determined not only by $N(\boldsymbol{r})$ but also by the summation of the

$coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]$, the correlation coefficients among the squared increments, $[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2$. This correlation coefficient, according to (10), is determined by the separation distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ ($i, j = 1, \ldots, N(\boldsymbol{r})$), the lag distance, $\boldsymbol{r}$, at which the semivariogram is to be estimated, and the underlying semivariogram, $\gamma(\boldsymbol{r})$. Thus, (13) gives an explicit expression describing the interactions among the variance of the estimate of the semivariogram, the number of data, and the sampling design. It is a general relationship that holds in any random field, $Z(\boldsymbol{x})$, provided that the increments of $Z(\boldsymbol{x})$ are normally distributed.

Illustration Using Hydraulic Head and Transmissivity Fields

In order to illustrate the impact of (13) on problems of interest to hydrogeologist, a comparison is made between estimating the semivariogram of the transmissivity (T) and estimating the semivariogram of the head residual (h, the hydraulic head minus the mean trend in the head). These particular parameters were chosen as they are two of the most frequently studied spatial processes in groundwater hydrology.

For the following illustration, it is assumed that T is log-normally distributed. A new, normally distributed random variable, Y, is therefore defined as $Y=\ln(T)$. It is further assumed that Y is second-order stationary, exists within an infinite spatial flow domain, and is characterized by an isotropic, exponential covariance function,

$$\gamma_Y(\boldsymbol{r}) = \sigma_Y^2[1 - \exp(-r')] \tag{17}$$

where σ_Y^2 is the variance of Y, and $r' = \frac{|\boldsymbol{r}|}{\lambda}$ is the magnitude of the separation vector $\boldsymbol{r}$

normalized by the integral scale, λ, of *Y.*

Dagan (1985) has presented the first and second moments of the distribution of the head residual under mean uniform flow within the type of transmissivity field described above. Of particular interest here is *Dagan*'s expression for the semivariogram of the head residual

$$\gamma_h(\boldsymbol{r}) = \frac{J^2\lambda^2\sigma_y^{\,2}}{2}\cdot\left[(2(\cos\psi)^2-1)\left(\frac{\exp(-r')\cdot(r'^2+3r'+3)-3}{r'^2}\right)+\right. \tag{18}$$

$$\left.\left((\cos\psi)^2-\frac{1}{2}\right)-Ei(-r')+\ln(r')+\exp(-r')+(e-1)\right]$$

where J is the magnitude of the mean regional head gradient $\boldsymbol{J}$, ψ is the direction of $\boldsymbol{r}$ relative to the orientation of $\boldsymbol{J}$, Ei is the exponential integral function, and e is Euler's constant (=0.5227).

Figure 1 shows plots of $\gamma_Y(\boldsymbol{r})$ (normalized by $\sigma_y^{\,2}$) and $\gamma_h(\boldsymbol{r})$ (normalized by $\frac{J^2\lambda^2\sigma_Y^{\,2}}{2}$) versus the normalized separation distance r'. The anisotropic $\gamma_h(\boldsymbol{r})$ is plotted for two different directions, one being parallel with $\boldsymbol{J}$, i.e., $\psi = 0$, and the other perpendicular to $\boldsymbol{J}$, i.e., $\psi = \frac{\pi}{2}$. As shown in the figure, $\gamma_Y(\boldsymbol{r})$ asymptotically approaches a sill as the separation distance increases, whereas $\gamma_h(\boldsymbol{r})$ grows logarithmically with the increasing separation distance. $\gamma_Y(\boldsymbol{r})$ also has a finite integral scale, λ, while the hydraulic head field is correlated over a much longer distance and no finite integral scale can be defined. Hence, these two variables have fundamentally different structure in their semivariograms.

This difference in structure has dramatic impact on the potential to estimate these semivariograms using data collected from a single realization. In order to simplify a comparison

of the estimation of the semivariogram for the Y and h fields, the semivariograms are evaluated only for the cases in which the orientation of the lag $\boldsymbol{r}$ is parallel to the direction of the mean gradient, $\boldsymbol{J}$. (It is straightforward to extend the analysis and its results to other directions if necessary.) Under this condition, and for fixed $\boldsymbol{r}$, the expressions for the theoretical semivariograms {(17) and (18), respectively} can be substituted into (14) and (10) to obtain $coe[S(\boldsymbol{x}_i, r), S(\boldsymbol{x}_j, r)]$ as a function of $\boldsymbol{R}_{ij}$. Figures 2a and 2b illustrate the correlation functions for $S_Y(\boldsymbol{x}, r)$ and $S_h(\boldsymbol{x}, \boldsymbol{r})$. As $\boldsymbol{R}_{ij}$ is a two-dimensional vector, these are three dimensional plots with the horizontal axes defining the directional components of $\boldsymbol{R}_{ij}$, and the vertical axis providing the magnitude of the correlation coefficient. In both figures, $\boldsymbol{r}$ has a magnitude of 10 units and is oriented parallel to $\boldsymbol{J}$. Both $\boldsymbol{R}_{ij}$ and $\boldsymbol{r}$ are normalized by λ. Figures 2a and 2b show that the correlation structures for both $S_Y(\boldsymbol{x}, \boldsymbol{r})$ and $S_h(\boldsymbol{x}, \boldsymbol{r})$ are anisotropic. Further, a local maximum in correlation exists at $\boldsymbol{R}_{ij} = \boldsymbol{r}$. Finally, the figures show that $S_h(\boldsymbol{x}, \boldsymbol{r})$ is correlated over longer distances than is $S_Y(\boldsymbol{x}, \boldsymbol{r})$.

The contribution of the correlation between sample pairs can be illustrated only for a given sampling pattern. Hence, a uniform square grid sampling network is here utilized to illustrate the cumulative contribution of $coe[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ to the magnitude of $\rho_{\hat{\gamma}(\boldsymbol{r})}$. This sampling network is set to be aligned with the direction of $\boldsymbol{J}$, contains n sample points in a square array (where n is the number of points at which T and h are measured) and has minimum spacing, d, between sample points (see Figure 3 for an example in which n=36 and d=1.0). For the discussion below, the minimum separation distance between the $S(\boldsymbol{x}_i, \boldsymbol{r})$'s, m, was set to be equal to d. Thus, specific to this sampling scheme, it is straight forward to show that a general

relationship between the number of $S(\boldsymbol{x}_i, \boldsymbol{r})$'s (i.e., the number of measurement pairs for a particular $\boldsymbol{r}$) and the number of sample points, n, exists as

$$N(\boldsymbol{r}) = n - \frac{r \cdot \sqrt{n}}{d} \tag{19}$$

Using the sampling scheme as defined above, it is possible to calculate $\rho_{\gamma(\boldsymbol{r})}$ according to (13) for a specific $\boldsymbol{r}$. By varying n, it is possible to adjust $N(\boldsymbol{r})$ and evaluate $\rho_{\gamma(\boldsymbol{r})}$ as a function of $N(\boldsymbol{r})$. Figure 4 shows $\rho_{\hat{\gamma}_Y(\boldsymbol{r})}$ and $\rho_{\hat{\gamma}_h(\boldsymbol{r})}$ versus $N(\boldsymbol{r})$ when r=10 along the direction of $\boldsymbol{J}$. This plot also includes the curve for $\rho_{\hat{\gamma}(\boldsymbol{r})} = \sqrt{\frac{2}{N(\boldsymbol{r})}}$, the result obtained by assuming zero correlation among the $S(\boldsymbol{x}_i, \boldsymbol{r})$'s (see 15). From this plot, it is apparent that the correlation among $S(\boldsymbol{x}_i, \boldsymbol{r})$'s strongly influences the rate of reduction of $\rho_{\hat{\gamma}(\boldsymbol{r})}$ with increasing $N(\boldsymbol{r})$. In order to achieve, for example, a value for $\rho_{\hat{\gamma}(\boldsymbol{r})}$ of 0.8, $N(\boldsymbol{r})$ need only be around 5for both uncorrelated data and the transmissivity field. For the head residual, this number increases to approximately 100. Hence, nearly 20 times the data pairs are required to achieve the same coefficient of variation for semivariogram of the head residuals as would be required for an uncorrelated variable or the random fields with short correlation range.

One interesting dependence which was further investigated was the relationship between the coefficients of variation, $\rho_{\hat{\gamma}_Y(\boldsymbol{r})}$ and $\rho_{\hat{\gamma}_h(\boldsymbol{r})}$, and $\boldsymbol{r}$. Based on the same sampling scheme with $N(\boldsymbol{r}) = 64$, $\rho_{\hat{\gamma}_Y(\boldsymbol{r})}$ and $\rho_{\hat{\gamma}_h(\boldsymbol{r})}$ vary with $\boldsymbol{r}$ as shown in Figure 5. Once again, $\rho_{\hat{\gamma}_h(\boldsymbol{r})}$ is consistently larger than $\rho_{\hat{\gamma}_Y(\boldsymbol{r})}$ for all $\boldsymbol{r}$, a result of the head pairs being correlated over longer distances than transmissivity pairs. Further, increasing $\boldsymbol{r}$ appears to have a greater adverse impact on the head residuals than on the transmissivity (i.e., $\rho_{\hat{\gamma}_h(\boldsymbol{r})}$ appears to grow with $\boldsymbol{r}$ whereas $\rho_{\hat{\gamma}_Y(\boldsymbol{r})}$ appears to

be relatively insensitive to $\boldsymbol{r}$). These results imply that not only will the separation distance among sample points need to be modified for estimating the semivariogram of Y versus the semivariogram of h, the basic design of the measurement locations must be modified as well (with approximately uniform distribution of data pairs among lag classes for Y and increasing number of data pairs with increasing lag distance required for h).

Discussion and Conclusions

The theoretical analysis of the sample semivariogram shows that the semivariogram estimator as given in (2) is unbiased, but that the coefficient of variation of the estimator, $\rho_{\hat{\gamma}(\boldsymbol{r})}$, depends not only on the number of the data pairs, $N(\boldsymbol{r})$, but also on the correlation among the data pairs. This correlation is, in turn, related to the form of the underlying semivariogram, $\gamma(\boldsymbol{r})$, the relative locations of the data pairs, and the lag distance, $\boldsymbol{r}$, at which the semivariogram is to be estimated. When the increment, $S(\boldsymbol{x}, \boldsymbol{r})$, is Gaussian, knowledge of $\gamma(\boldsymbol{r})$ is sufficient to define the correlation structure among the squared increments according to equation (10).

Equation (10) leads to at least three significant observations. First, the reliability of a semivariogram estimate derived from measured data is dependent not only on the number of data points collected, but also on the parameter being measured (through the semivariogram of that parameter). Second, random variables exhibiting correlation over large distances are very likely to have squared increments which are highly correlated. Thus the sample semivariogram estimate for a random variable which is correlated at large distances will tend to be unreliable and caution should be used in interpreting sample semivariograms exhibiting a long range correlation structure (e.g., a power law semivariogram). Should a field data set imply such a long range

correlation structure, equation (10), or a modification thereof for non-gaussian random variables, should be used to determine an estimate of the coefficient of variation of the sample semivariogram at each lag distance, thus providing a measure of confidence on the structure observed in the sample data. Third, the optimal distribution of data pairs over the different lag distances at which the sample semivariogram is to be estimated is also a function of the underlying semivariogram. As was shown, a reliable estimate of the semivariogram at various lags for the transmissivity (subject to the constraints outlined above) could be accomplished with relatively uniform numbers of data pairs in each lag class. In contrast, estimating the semivariogram of the head residual, with equal coefficient of variation in each lag class, would require increasing numbers of data pairs as the magnitude of the lag distance is increased.

Appendix: Detailed Derivation for Equation (10)

Given the definition of $S(\boldsymbol{x}_i, \boldsymbol{r})$ as in (7) and its ensemble mean (8), an expression for $cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})]$ can be written as

$$\begin{aligned} & cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] \\ & = \frac{1}{4}\langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i + \boldsymbol{r})]^2 \bullet [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j + \boldsymbol{r})]^2 \rangle - \gamma(\boldsymbol{r})^2 \end{aligned} \tag{20}$$

Using the joint moment generating function, *Papoulis* (example 7-6 in page 158,1984) shows that, if two random variables, X_1 and X_2, are jointly normal with zero mean, the following relationship holds:

$$\langle X_1^2 X_2^2 \rangle = \langle X_1^2 \rangle \langle X_2^2 \rangle + 2\langle X_1 X_2 \rangle^2 \tag{21}$$

Assuming that the increment, $Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{r})$, is jointly normally distributed (which holds, at least, for the case in which $Z(\boldsymbol{x})$ is jointly normal), the application of (21) to (20) leads to,

$$
\begin{aligned}
& cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] \qquad (22)\\
&= \frac{1}{4}\langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})]^2 \bullet [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j+\boldsymbol{r})]^2 \rangle - \gamma(\boldsymbol{r})^2 \\
&= \frac{1}{4}\langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})]^2 \rangle \bullet \langle [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j+\boldsymbol{r})]^2 \rangle \\
&+ \frac{1}{2}\langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})] \bullet [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j+\boldsymbol{r})] \rangle^2 - \gamma(\boldsymbol{r})^2 \\
&= \frac{1}{2}\langle [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})] \bullet [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j+\boldsymbol{r})] \rangle^2
\end{aligned}
$$

Further expanding $[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_i+\boldsymbol{r})] \bullet [Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_j+\boldsymbol{r})]$ into four terms results in

$$
\begin{aligned}
& cov[S(\boldsymbol{x}_i, \boldsymbol{r}), S(\boldsymbol{x}_j, \boldsymbol{r})] \qquad (23)\\
&= \frac{1}{2}\Big\langle \frac{[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j+\boldsymbol{r})]^2}{2} + \frac{[Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_i+\boldsymbol{r})]^2}{2} - \frac{[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)]^2}{2} - \frac{[Z(\boldsymbol{x}_i+\boldsymbol{r}) - Z(\boldsymbol{x}_j+\boldsymbol{r})]^2}{2} \Big\rangle^2 \\
&= \frac{1}{2}[\gamma(\boldsymbol{r}+\boldsymbol{x}_j-\boldsymbol{x}_i) + \gamma(\boldsymbol{r}-\boldsymbol{x}_j+\boldsymbol{x}_i) - 2\gamma(\boldsymbol{x}_j-\boldsymbol{x}_i)]^2 \\
&= \frac{1}{2}[\gamma(\boldsymbol{r}+\boldsymbol{R}_{ij}) + \gamma(\boldsymbol{r}-\boldsymbol{R}_{ij}) - 2\gamma(\boldsymbol{R}_{ij})]^2
\end{aligned}
$$

where $\boldsymbol{R}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$ is the separation distance between $S(\boldsymbol{x}_i, \boldsymbol{r})$ and $S(\boldsymbol{x}_j, \boldsymbol{r})$.

Acknowledgments.

This research is partially supported by U. S. Department of Energy's Subsurface Science Program. The authors also like to thank T. Yegulalp and an anonymous reviewer for their thorough and thoughtful review.

References

Benjamin, J. R., and C. A. Cornell, *Probability, Statistics, and Decision for Civil Engineers.*, McGraw-Hill Book Company, New York, 1970.

Chung, C.F., Use of the Jackknife method to estimate autocorrelation functions (or variograms), in *Geostatistics for Natural Resources Characterization, vol. I,* edited by G. Verly, M. David, A. G. Journel, and A. Marechal, pp. 55-69, D. Reidel Publishing Company, Dordrecht, 1984.

Conwell, P. M., S. E. Silliman, and L. Zheng, Design of a piezometer network for estimation of the variogram of the hydraulic gradient: The role of the instrument, *Water Resour. Res.*, 33(11), 2489-2492, 1997.

Corsten, L.C.A., and A. Stein, Nested sampling for estimating spatial semivariograms compared to other designs, *Applied Stochastic Models and Data Analysis*, 10, 103-122, 1994.

Cressie, N., Statistics for Spatial Data, Rev. ed., Wiley, New York, 1993.

Dagan, G., Stochastic modeling of groundwater flow by conditional and unconditional probabilities: the inverse problem, *Water Resour. Res.*, 21(1), 65-72, 1985.

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

Matheron, G., *Les Variables Regionalisees et leur Estimation*, Masson, Paris, 1965.

Morris, M. D., On counting the number of data pairs for semivariogram estimation, *Math. Geol.*, 23(7), 929-943, 1991.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 2nd edition, McGraw Hill, New York, 1984.

Russo, D., Design of an optimal sampling network for estimating the variogram, *Soil Sci. Soc. Am. J.*, 48, 708-716, 1984.

Russo, D., and W. A. Jury, A theoretical study of the estimation of the correlation scale in spatially variable fields, 1, Stationary fields, *Water Resour. Res.*, 23(7), 1257-1268, 1987.

Shafer, J. M., and M. D. Varljen, Approximation of confidence limits on sample semivariograms from single realizations of spatially correlated random fields, *Water Resour. Res.*, 26(8), 1787-1802, 1990.

Warrick, A.W., and D.E. Myers, Optimization of sampling locations for variogram calculations, *Water Resour. Res.*, 23(3), 496-500, 1987.

Webster, R., and M. A. Oliver, Sample adequately to estimate variograms for soil properties, *J. Soil Sci.*, 43, 177-192, 1992.

List of Figures