Mathematical Geology, Vol. 23, No. 4, 1991

Teacher's Aide

The Variogram Sill and the Sample Variance[1]

Randal J. Barnes[2]

The relationship between the sill of the variogram and the sample variance is explored. The common practice of using the sample variance as an estimate of the variogram sill is questioned, and a conceptual framework for determining the appropriateness of this heuristic is constructed.

KEY WORDS: variogram, sill, sample variance.

THE VARIOGRAM SILL

The variogram is commonly defined to be the expected value of a difference squared (e.g., Isaaks and Srivastava, 1989, p. 222):

$$2\gamma(\mathbf{h}) = E\{[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})]^2\} \tag{1}$$

If the separation vector $\mathbf{h}$ is so large that $Z(\mathbf{x})$ and $Z(\mathbf{x} + \mathbf{h})$ are uncorrelated (i.e., $\mathbf{h} \geq$ range), nothing more than algebra is required to show that the height of the variogram, $\gamma(\mathbf{h})$, equals the population variance. That is, the sill of the underlying variogram equals the variance of the underlying population. (This statement assumes that the random field model has certain nice properties—e.g., a finite population variance, and a variogram with a finite range.)

Starting with Eq. (1), then adding and subtracting the population mean within the inside brackets yields

$$2\gamma(\mathbf{h}) = E\{[Z(\mathbf{x}) - \mu - Z(\mathbf{x} + \mathbf{h}) + \mu]^2\}$$

Next, expanding the square produces

$$= E\{[Z(\mathbf{x}) - \mu]^2 - 2[Z(\mathbf{x}) - \mu][Z(\mathbf{x} + \mathbf{h}) - \mu] + [Z(\mathbf{x} + \mathbf{h}) - \mu]^2\}$$

[1] Received 5 May 1987; accepted 1 September 1987; resubmitted 27 April 1990.

[2] Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive S.E., Minneapolis, Minnesota 55455.

Breaking the expected value into three pieces allows for a significant simplification:

$$= E\{[Z(\mathbf{x}) - \mu]^2\} - 2E\{[Z(\mathbf{x}) - \mu][Z(\mathbf{x} + \mathbf{h}) - \mu]\}$$
$$+ E\{Z(\mathbf{x} + \mathbf{h}) - \mu]^2\}$$

Finally, recognizing that the first term is the population variance, the second term is twice the spatial covariance, and the third term is again the population variance, yields

$$= \sigma^2 - 2 \text{ cov } [Z(\mathbf{h}), Z(\mathbf{x} + \mathbf{h})] + \sigma^2$$

If **h** is larger than the range of the variogram, values separated by a distance **h** are uncorrelated and the spatial covariance "cov [$Z(\mathbf{h}), Z(\mathbf{x} + \mathbf{h})$]" is equal to zero. Thus, the sill of the underlying variogram equals the true population variance:

$$\gamma(\mathbf{h}|\mathbf{h} > \text{range}) = \sigma^2 \tag{2}$$

ESTIMATING THE VARIOGRAM SILL

When dealing with real data, the underlying variogram and the true population variance are not known. As such, the result presented in Eq. (2) is true, but it is not always useful. Unfortunately, confusion abounds throughout the practicing geostatistical community concerning this distinction.

In the past, authors of geostatistics books and papers have suggested the following line of reasoning: the sample variance S^2 is a well-known estimator for the population variance; therefore, it seems reasonable to use the sample variance to estimate the variogram sill. For example, Journel and Huijbregts (1978, p. 231) state: ". . . The experimental dispersion variance

$$S^2 = \frac{1}{N} \sum_{i=1}^{N} (z(x_i) - z)^2$$

can sometimes be used in fitting the sill, . . .".

David (1977, p. 122) is more specific in his statement

> The sill of the variogram is equal to the variance of the samples in the deposit, which can be computed from the samples.

Often, in the technical literature and in practice, a horizontal line is drawn on the experimental variogram plot at the value of the sample variance. The modeled variogram sill is then forced to be equal to this value. Sometimes this approach is a valid and useful modeling heuristic; however, as presented in the next section of this paper, often, it is absolutely improper and will lead to improper variogram models. The final section of this paper discusses when it is safe to use this heuristic and when it must be avoided.

THE SAMPLE VARIANCE

For this discussion, assume that there are "N" available data sample values at locations

$$\{\mathbf{x}_i: i = 1, 2, \ldots, N\}$$

Let the set of available data values be given by

$$\{Z(\mathbf{x}_i): i = 1, 2, \ldots, N\}$$

or for notational brevity,

$$\{Z_i: i = 1, 2, \ldots, N\}$$

The classical arithmetic mean and variance of N values are given in most introductory statistics text (e.g., Benjamin and Cornell, 1970, p. 9–11)

$$M = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

and

$$S^2 = \frac{1}{N} \sum_{i=1}^{N} (Z_i - M)^2 \tag{3}$$

Now, consider the average value of the variogram between all N^2 pairs of available sample data (N data values will generate N^2 pairs when a value can be paired with itself). Notationally, call this value Γ_N. That is,

$$\Gamma_N = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

An equivalence between Eq. (3) and Eq. (4) will be developed in the following paragraphs.

Substituting in the definition of the variogram, Eq. (1), into Eq. (4) yields

$$\Gamma_N = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} E\{(Z_i - Z_j)^2\}$$

Adding and subtracting the sample arithmetic average within the inner parentheses gives

$$= \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} E\{[(Z_i - M) - (Z_j - M)]^2\}$$

Expanding the square within the braces { }, results in

$$= \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} E\{(Z_i - M)^2 - 2(Z_i - M)(Z_j - M) + (Z_j - M)^2\}$$

Since the expected value of a sum equals the sum of the component expected values, this can be rewritten with the expectation operator outside of the summation

$$= E\left[\frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} [(Z_i - M)^2 - 2(Z_i - M)(Z_j - M) + (Z_j - M)^2] |\right]$$

Breaking the sum inside the large brackets into three component sums allows for significant simplification

$$= E\left[\frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_i - M)^2 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_i - M)(Z_j - M) + \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_j - M)^2\right] \tag{5}$$

These three components will now be considered individually and in turn. The first component is

$$\frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_i - M)^2$$

The terms within the parentheses are not functions of j and may thus be factored out of the summation over j, leaving

$$\frac{1}{2N^2} \sum_{i=1}^{N} (Z_i - M)^2 \sum_{j=1}^{N} 1$$

Since the sum of "1" over j varying from 1 to N equals N, the first component of Eq. (5) is reduced to the simple expression

$$\frac{1}{2N} \sum_{i=1}^{N} (Z_i - M)^2$$

The second component is

$$\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_i - M)(Z_j - M)$$

As before, $(Z_i - M)$ is not a function of j and may thus be factored out of the summation over j; this yields

$$\frac{1}{N^2} \sum_{i=1}^{N} (Z_i - M) \sum_{j=1}^{N} (Z_j - M) \tag{6}$$

Interestingly, the right-hand summation in Eq. (6) is equal to zero. This fact is demonstrated as follows.

$$\begin{aligned}\sum_{j=1}^{N} (Z_j - M) &= \sum_{j=1}^{N} Z_j - \sum_{j=1}^{N} M \\ &= \sum_{j=1}^{N} Z_j - M \sum_{j=1}^{N} 1 = \sum_{j=1}^{N} Z_j - MN \\ &= \sum_{j=1}^{N} Z_j - \sum_{j=1}^{N} Z_j = 0\end{aligned}$$

Thus, the contribution of the entire second component of Eq. (5) is zero.

The third component is identical to the first (if the order of summation is exchanged). Therefore, Eq. (5) simplifies to

$$\Gamma_N = E\left[\frac{1}{N} \sum_{i=1}^{N} (Z_i - M)^2\right] \tag{7}$$

Comparing Eq. (7) to the formula for the sample variance, Eq. (3), one can see that the arithmetic average of the variogram between all N^2 pairs of available sample values equals the expected value of the sample variance.

CONCLUSIONS AND CONSEQUENCES

The summary of the preceding lengthy derivation is short. Under fairly general conditions

$$E[S^2] = \Gamma_N = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

In a geostatistical setting, the expected value of the sample variance is a function of the geographic data configuration and the entire variogram, not just the sill of the variogram. This is not a new result (e.g., Journel and Huijbregts, 1978, p. 63–68), but its consequences are often ignored.

There are a number of potential uses for the relationship of Eq. (8). For example, Γ_N and S^2 can be compared to test the validity of a conjectured experimental variogram model: if Γ_N and S^2 differ greatly, the experimental model is suspect. The specific definition of large difference is outside the scope of this paper; and further, the fundamental understanding to be gleaned from Eq. (8) is more important than merely another geostatistical hypothesis-testing result of questionable power.

Equation (8) states that the expected value of the sample variance is equal to the average value of the variogram between all N^2 pairs of sample values. If the N sample values are evenly distributed over an areal extent many times larger than the range of the variogram, then there will be significantly more pairs at long separation distances and fewer pairs at short separation distances. Thus, the average value of the variogram between all pairs, Γ_N, will be the average of many values equal to the sill and a few values less than the sill. In this case, the sample variance is a reasonable first estimate for the variogram sill. This partially justifies Journel and Huijbregts' (1978, p. 231) statement that s^2 "can sometimes be used in fitting the sill."

On the other hand, if most of the N sample values are collected from an area with dimensions equal to or less than the variogram range, Γ_N will be the average of many values less than the sill and a few values equal to the sill. In this case, the sample variance is *not* a reasonable first estimate for the variogram sill. In this case, the sample variance will, on the average, significantly underestimate the variogram sill. This refutes David's (1977, p. 122) statement that "the sill of the variogram is equal to the variance of the samples."

At the other extreme, if the sample values incorporate an otherwise insignificant low-scale trend and the samples are taken over a large area, then despite an apparent graphical emergence of a variogram sill, the sample variance will, on the average, significantly overestimate this apparent sill.

Specifically, when is the sample variance a reasonable first estimate for the variogram sill, as suggested by David and Journel and Huijbregts? A rule-of-thumb is that the data must be somewhat evenly distributed over an area with dimensions greater than three times the range of the variogram. However, the conscientious geostatistician must consider the concept presented in Eq. (8) and its relationship with the specific data at hand (as discussed above).

In general, if a sill is clearly presented by the experimental variogram plot, its value should be used as an estimate of the population variance, and the sample variance should not be used as an estimate of the variogram sill.

REFERENCES

Benjamin, J. R., and Cornell, C. A., 1970, *Probability, Statistics, and Decision for Civil Engineers*: McGraw-Hill, New York, 684 p.

David, M., 1977, *Geostatistical Ore Reserve Estimation*: Elsevier Scientific Publishing, New York, 364 p.

Isaaks, E., Srivastava, R. M., 1989, *An Introduction to Applied Geostatistics*: Oxford University Press, New York, 561 p.

Journel, A. G., and Huijbregts, C. J., 1978, *Mining Geostatistics*: Academic Press, London, 600 p.