



STANFORD UNIVERSITY

STANFORD, CALIFORNIA 94305-2225

DEPARTMENT OF APPLIED EARTH SCIENCES
School of Earth Sciences

(415) 723-0847
Telex: 3722871 STANUNIV
Fax: 415-725-0979
October 15, 1992

Prof. Robert Ehrlich
Editor Math Geology
Dept. of Geological Sciences
University of South Carolina
Columbia, SC 29208

Dear Bob:

Upon your request, and a bit reluctantly, I went through the various notes of J. W. Merks. All these notes are similar in content sharing the same figures and examples. It seems to me that Mr. Merks' anger arises from a misreading of geostatistical theory, or a reading too encumbered by classical "Fischerian" statistics. The following attempt at clarification addresses the three recurring points made in Mr. Merks' notes.

1. Data and degrees of freedom
2. Unbiased estimation of spatial variance
3. The smoothing effect of kriging

1 - Data and degrees of freedom

The very reason for geostatistics or spatial statistics in general is the acceptance (a decision rather) that spatially distributed data should be considered a priori as dependent one to another, unless proven otherwise. It is that spatial dependence which allows differentiated local interpolation and mapping in general. Were the data independent one from another then only global statistics can be retrieved. In presence of dependence the classical notion of degrees of freedom vanishes: n spatially dependent data do not provide n degrees of freedom.

It is not correct to state categorically that kriging, or for that matter any other interpolation algorithm, does not add any information to the system. It does through the implicit or explicit model of correlation. Indeed, change the variogram model yet keep the same data, the kriging estimates change. This correlation/covariance/variogram model can be borrowed from another field or outcrop, it is then genuine new information. Or, it can be inferred from the same data used for kriging. In the latter case, new information is introduced through aspects of the sample bivariate (two-point) distribution. The important question is, of course, how representative of the unsampled area is that bivariate information, i.e., how appropriate is the prior "decision" of stationarity.

2 - Unbiased estimation of the spatial variance

The reason for the denominator $(n - 1)$ in the classical expression of the variance estimator

$$\text{Var}^* \{Z\} = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$$

is indeed correction of the bias introduced by the prior estimation of the population mean \bar{z} by the same n data z_i .

However, that correction is valid *if and only* if the n data Z_i can be considered as independent one from another. In a spatial context with spatial dependence, the case is a bit more complex requiring some more detailed notations:

Let $Z(\mathbf{x})$ be the random variable defined at location \mathbf{x} within the finite¹ field A . Let $z(\mathbf{x}_i)$, $i = 1, \dots, n$ be the n data values at the n sample locations $\mathbf{x}_i \in A$. The global statistics over A are:

$$m_A = \frac{1}{|A|} \int_A z(\mathbf{x}) d\mathbf{x} \quad (1)$$

$$\sigma_A^2 = \frac{1}{|A|} \int_A [z(\mathbf{x}) - m_A]^2 d\mathbf{x} \quad (2)$$

After interpretation of the variable $\{z(\mathbf{x}), \mathbf{x} \in A\}$, by the random function $\{Z(\mathbf{x}), \mathbf{x} \in A\}$, the previous integrals become stochastic:

$$M_A = \frac{1}{|A|} \int_A Z(\mathbf{x}) d\mathbf{x} \quad (3)$$

$$S_A^2 = \frac{1}{|A|} \int_A [Z(\mathbf{x}) - M_A]^2 d\mathbf{x} \quad (4)$$

Under a stationary model for the random function $\{Z(\mathbf{x}), \mathbf{x} \in A\}$, consider the two "naive" equal-weighted estimators:

$$\hat{M}_A = \frac{1}{n} \sum_{i=1}^n Z(\mathbf{x}_i) \quad (5)$$

$$\hat{S}_A^2 = \frac{1}{n} \sum_{i=1}^n [Z(\mathbf{x}_i) - \hat{M}_A]^2 \quad (6)$$

No matter the statistical dependence between the n random variables $Z(\mathbf{x}_i)$, the estimator (5) is unbiased since:

$$\begin{aligned} E\{\hat{M}_A\} &= \frac{1}{n} \sum_{i=1}^n E\{Z(\mathbf{x}_i)\} = \frac{1}{n} \sum_i m = m \\ &= E\{M_A\} = \frac{1}{|A|} \int_A E\{Z(\mathbf{x})\} d\mathbf{x} = m \end{aligned}$$

with $m = E\{Z(\mathbf{x})\}$ being the stationary model mean. Note that unbiasedness refers to identification of the model mean m , not necessarily to identification of the field average m_A as defined by integral (1).

As for the variance estimator (4), the development goes as:

$$\begin{aligned} E\{\hat{S}_A^2\} &= \frac{1}{n} \sum_i E\{Z^2(\mathbf{x}_i) - 2\hat{M}_A Z(\mathbf{x}_i) + \hat{M}_A^2\} \\ &= \frac{1}{n} \sum_i E\{Z^2(\mathbf{x}_i)\} - \frac{1}{n^2} \sum_i \sum_j E\{Z(\mathbf{x}_i) Z(\mathbf{x}_j)\}, \end{aligned}$$

¹To avoid any problem of integral definition, one could as well discretize the field A into a very large, yet finite, number of cells of support equal to that of the n data.

while:

$$E \{ S_A^2 \} = \frac{1}{|A|} \int_A E \{ Z^2(\mathbf{x}) \} dx - m^2$$

Under the model stationarity:

$$E \{ Z^2(\mathbf{x}_i) \} = E \{ Z^2(\mathbf{x}) \} = l^2, \text{ for any } \mathbf{x}_i, \mathbf{x} \in A.$$

Note that the stationary model variance is:

$$\text{Var} \{ Z(\mathbf{x}) \} = \sigma^2 = l^2 - m^2, \text{ for all } \mathbf{x} \in A$$

The bias attached to the variance estimator \hat{S}_A^2 is :

$$\begin{aligned} \sigma^2 - E \{ \hat{S}_A^2 \} &= (l^2 - m^2) - \frac{1}{n} \sum_i E \left\{ [Z(\mathbf{x}_i) - \hat{M}_A]^2 \right\} \\ &= (l^2 - m^2) - \frac{1}{n} \sum_i Z^2(\mathbf{x}_i) + \frac{1}{n^2} \sum_i \sum_j E \{ Z(\mathbf{x}_i) Z(\mathbf{x}_j) \}, \end{aligned}$$

and finally:

$$\begin{aligned} \sigma^2 - E \{ \hat{S}_A^2 \} &= \frac{1}{n^2} \sum_i \sum_j E \{ Z(\mathbf{x}_i) Z(\mathbf{x}_j) \} - m^2 \\ &= \frac{1}{n^2} \sum_i \sum_j C(\mathbf{x}_i - \mathbf{x}_j) \end{aligned} \quad (7)$$

with $C(\mathbf{h}) = \text{Cov} \{ Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h}) \} = E \{ Z(\mathbf{x}) Z(\mathbf{x} + \mathbf{h}) \} - m^2$ being the stationary covariance of the random function model.

Expression (7) is that given page 195 in Journel & Huijbregts (1978).

Remarks:

- Again, the previous bias refers to indentification of the model variance σ^2 , not the field variance as defined by integral (4). However, under appropriate ergodic hypotheses (about the random function model) one can identify the integral (4) to the model variance σ^2 when A is large with regard to the covariance range.
- If the n random variables "data", $Z(\mathbf{x}_i), i = 1, \dots, n$, can be considered as independent two by two, then

$$E \{ Z(\mathbf{x}_i) Z(\mathbf{x}_j) \} = \begin{cases} E \{ Z(\mathbf{x}_i) \} \cdot E \{ Z(\mathbf{x}_j) \} = m^2, & \text{for } i \neq j \\ E \{ Z^2(\mathbf{x}_i) \} = l^2 = \sigma^2 + m^2, & \text{for } i = j \end{cases}$$

The bias (7) reduces to the classical expression:

$$\sigma^2 - E \{ \hat{S}_A^2 \} = \frac{\sigma^2}{n},$$

in which case an unbiased estimator of the model variance σ^2 is the expression (oft quoted by J. W. Merks):

$$\hat{S}_A^{*2} = \frac{n}{n-1} \hat{S}_A^2 = \frac{1}{n-1} \sum_{i=1}^n [Z(\mathbf{x}_i) - \hat{M}_A]^2 \quad (8)$$

In presence of spatial dependence, there is no advantage (nor does it make any practical difference if n is large) in using estimator (8) instead of (6).

Most geostatisticians make (implicitly) the argument that because the field A of observations is large with regard to the covariance range, the bias term (7) is negligible, indeed:

$$\frac{1}{n^2} \sum_i \sum_j C(\mathbf{x}_i - \mathbf{x} - j) \rightarrow 0, \text{ as } |A| \rightarrow \infty,$$

provided the n data locations are evenly (no clusters) distributed over the field A .

- A proposal was made to test for spatial correlation at lag h , i.e. to test whether $C(h)$ is zero or not, using a F -test to compare the experimental covariance $\hat{C}(h)$ calculated from the "ordered" data set and that $\hat{S}^2(h)$ calculated from the "randomized" data set. Give the understanding of "ordered" as considering the data values at their actual locations, and "randomized" as randomizing the locations of the data values, then the proposal is a mere (trivial) check for correlation, not a formal test. Indeed:
 - the F -test requires normality or quasi-normality of the populations whose variances are being compared. A prior normal score transform of the data values could achieve this, but then one would be testing the correlation of the normal score transforms. This is a minor point though.
 - the F -test requires independence of the data values used to calculate the variances in both populations. One cannot invoke independence as a constitutive hypothesis of a test used to check for spatial dependence.
 - as to compare the experimental covariance $\hat{C}(h)$ to a value $\hat{S}(h)$ obtained by randomizing, i.e., ignoring the actual data locations, one is better off comparing $\hat{C}(h)$ to the sample variance (6) or (8) which is none other than an estimate of $C(0) = \text{Var}\{Z(\mathbf{x})\}$. Such a comparison would amount to plot the various estimates $\hat{C}(h)$ versus h and check that this plot differs from $\hat{C}(0)$; indeed this is exactly what all geostatisticians do.

3 - The smoothing effect of kriging

In Merks' notes the term "kriged variance" is used alternatively for two different meanings:

1. as a "kriging variance", i.e. the "estimation" variance $E\{[Z(\mathbf{x}_0) - Z^*(\mathbf{x}_0)]^2\}$ resulting from a linear kriging process" (Olea, 1992, p. 41)
2. as the spatial variance of the estimates resulting from the kriging process.

These two different meaning are fundamentally different: in the first case it is a model-derived error variance, in the second case it is a measure of variability in space of the kriging estimates.

That the kriging variance has been oversold as a measure of accuracy of the kriging estimate is an unfortunate fact. Most geostatisticians now recognize that the kriging variance, being data values-independent, does not generally inform about the accuracy of the corresponding kriging estimate (Olea, *ibid.*).

As for the spatial variance of kriging estimates, except in the case of a regular grid, it is a meaning-

less statistic essentially because it is non-stationary depending on the specific data configuration prevailing at each location being estimated. Let \mathbf{x} be the location being estimated, the estimate is:

$$z^*(\mathbf{x}) = \sum_{\alpha=1}^{n(\mathbf{x})} \lambda_{\alpha}(\mathbf{x}) z(\mathbf{x}_{\alpha})$$

where $n(\mathbf{x})$ is the number of data $z(\mathbf{x}_{\alpha})$ retained in the neighborhood of \mathbf{x} , and $\lambda_{\alpha}(\mathbf{x})$ the set of weights given, e.g. but necessarily, by kriging. The corresponding estimator is written:

$$Z^*(\mathbf{x}) = \sum_{\alpha=1}^{n(\mathbf{x})} \lambda_{\alpha}(\mathbf{x}) Z(\mathbf{x}_{\alpha}) \quad (9)$$

In the vast majority of applications, the weights $\lambda_{\alpha}(\mathbf{x})$ are positive and are made to add up to one: $\sum \lambda_{\alpha}(\mathbf{x}) = 1$, defining a convex estimator $Z^*(\mathbf{x})$. The variance of this estimator is:

$$\text{Var} \{Z^*(\mathbf{x})\} = \sum_{\alpha=1}^{n(\mathbf{x})} \sum_{\beta=1}^{n(\mathbf{x})} \lambda_{\alpha}(\mathbf{x}) \lambda_{\beta}(\mathbf{x}) C(\mathbf{x}_{\alpha} - \mathbf{x}_{\beta}) \quad (10)$$

assuming unbiasedness, $C(\mathbf{h})$ being the stationary model covariance.

Remarks:

- In general, since the estimator $Z^*(\mathbf{x})$ is convex, there is a smoothing effect, i.e.

$$\text{Var} \{Z^*(\mathbf{x})\} < \text{Var} \{Z(\mathbf{x})\} = C(0) = \sigma^2 \quad (11)$$

Moreover that smoothing effect is non-stationary, i.e. varies from one location \mathbf{x} being estimated to another.

Stochastic simulation (Olea, 1991, p. 71.) has been introduced, precisely, to correct for such smoothing effect:

$$\text{Var} \{Z_s(\mathbf{x})\} = \text{Var} \{Z(\mathbf{x})\}, \text{ for all } \mathbf{x},$$

where $z_s(\mathbf{x})$ is the simulated value at location \mathbf{x} .

- Because of the non-stationarity of the variance (11) it is difficult to interpret the spatial variance ν^2 of kriging estimates called "kriged variance" by Merks:

$$\nu^2 = \frac{1}{K \text{ or } (K-1)} \sum_{k=1}^K [z^*(\mathbf{x}_k) - \bar{z}^*]^2 \quad (12)$$

with: $\bar{z}^* = \frac{1}{K} \sum_{k=1}^K z^*(\mathbf{x}_k)$, and the \mathbf{x}_k being the estimated locations.

Because of the smoothing effect of kriging, it is an error (as pointed out by J. W. Merks) to take ν^2 as an estimate of the spatial variance σ^2 , or σ_A^2 as defined in (2). It is also an error to mix within a same variance calculation of type (12), kriging estimates $z^*(\mathbf{x}_k)$ and actual data values $z(\mathbf{x}_{\alpha})$.

Once again, the correct approach if reproduction of the actual spatial variability matters is stochastic simulation *not* kriging. If this is what J. W. Merks alludes to, then he is 100% correct. In modern geostatistics, the smoothing effect of kriging and/or the smoothing effect due to volume averaging is not approached anymore through variance correction of the histogram of estimates, but through fine scale conditional simulations.

This answer took more than I expected originally, but I hope its length will not detract too much from the intended clarification goal. I'll leave it to you to decide whether this letter should be sent to J. W. Merks; however, I strongly feel that Math Geology has had more than its share of detracting invectives.

Yours sincerely,

A. G. Journel

A. G. Journel
Professor

References:

Journel, A. G. and Huijbregts, Ch., 1978, *Mining Geostatistics*, Academic Press, 600p.

Olea, R. (editor), 1991, *Geostatistical Glossary and Multilingual Dictionary*, Oxford Press, 177p.