Mathematical Geology, Vol. 29, No. 2, 1997

Spatial Declustering Weights[1]

G. Bourgault[2]

Because of autocorrelation and spatial clustering, all data within a given dataset have not the same statistical weight for estimation of global statistics such mean, variance, or quantiles of the population distribution. A measure of redundancy (or nonredundancy) of any given regionalized random variable $Z(\mathbf{u}_\alpha)$ *within any given set (of size* N*) of random variables is proposed. It is defined as the ratio of the determinant of the* $N \times N$ *correlation matrix to the determinant of the* $(N-1) \times (N-1)$ *correlation matrix excluding random variable* $Z(\mathbf{u}_\alpha)$*. This ratio measures the increase in redundancy when adding the random variable* $Z(\mathbf{u}_\alpha)$ *to the* $(N-1)$ *remainder. It can be used as declustering weight for any outcome (datum)* $z(\mathbf{u}_\alpha)$*. When the redundancy matrix is a kriging covariance matrix, the proposed ratio is the crossvalidation simple kriging variance. The covariance of the uniform scores of the clustered data is proposed as a redundancy measure robust with respect to data clustering.*

KEY WORDS: determinant, redundancy matrix, declustering weights, uniform scores.

INTRODUCTION

Because of autocorrelation and spatial clustering which create redundancy, all data from any given dataset have not the same statistical weight for the estimation of global statistics such mean, variance, or quantiles of the population distribution. Spatial clustering may occur when the distance between two data locations is smaller than the larger correlation range characterizing the sample dataset spatial continuity. In such a situation, the data values at these two locations are partially redundant, the degree of redundancy being function of the spatial continuity. Within the framework of the spatial random functions (Matheron, 1971), any data value is considered as an outcome of a regionalized random variable (one at each data location) and the spatial continuity is synonymous of correlation between the regionalized random variables. This framework is adopted through the paper. Moreover, it is assumed that stationarity of order 2, univariate stationarity and ergodicity are fulfilled.

A redundancy (or nonredundancy) measure should be attached to any given regionalized random variable $Z(\mathbf{u}_\alpha)$, and therefore to its outcome $z(\mathbf{u}_\alpha)$ (datum

[1]Received 28 March 1996; accepted 1 May 1996.

[2]1761 Chemin Royal, St-Pierre I.O., Québec G0A 4E0, e-mail: gilles@rc2.com

0882-8121/97/0200-0277$12.50/1

at location $\mathbf{u}_\alpha$). This is particularly important in the mineral industry where areas of high values tend to be oversampled for evident economic reasons. The data need to be declustered (weighted) statistically before being used to evaluate any global statistics representative of the whole area (not limited to areas of high values).

An important feature of any declustering procedure is that it should account for the specific spatial continuity of the dataset. In the absence of spatial continuity (white noise) there is no redundancy in the data regardless of the sampling geometry. Without spatial continuity, the data need not be declustered even if the sampling geometry shows clusters. This ideal situation of spatial independence, the cornerstone of classical statistics, is rarely if ever encountered in the earth sciences. Traditional declustering techniques such as polygons of influence, cell declustering (Journel, 1983; Isaaks and Srivastava, 1989, chapt. 10; Deutsch, 1989) and the entropy-based method (Schofield, 1992) do not account for a specific spatial continuity measure. Deutsch (1989) and Isaaks and Srivastava (1989, p. 510) also proposed a global kriging of the field average value, each datum kriging weight then being used as declustering weight. This approach has computational limitations and can not be applied to large datasets. In addition, a somewhat arbitrary kriging grid must be defined calling for a delineation of the limits of the field under study. Such a definition of limits is required by the polygonal approach. In order to overcome the computational limitations of global kriging, Crozel and David (1983) have proposed approximating the global kriging weights by summing the weights from local kriging systems. Then, these approximate global kriging weights are used as declustering weights (see Deutsch, 1989). Note that kriging weights can be negative so they cannot be used as declustering weights because they may yield a nonmonotonic cumulative density function for the declustered data. An algorithm, similar to kriging, but which provides only positive weights and does not call for prior grid definition nor for global neighborhood is proposed.

The degree of nonredundancy attached to datum α, $dnr(\mathbf{u}_\alpha)$, is defined as the ratio of the determinant of the $N \times N$ data location-to-data location redundancy (e.g., covariance or correlation) matrix to the determinant of the $(N - 1) \times (N - 1)$ redundancy matrix excluding data location $\mathbf{u}_\alpha$. The determinant reduces to a single number the redundancy information stored in the redundancy matrix. Being a ratio of determinants, $dnr(\mathbf{u}_\alpha)$ measures the relative change in redundancy matrix when adding the random variable $Z(\mathbf{u}_\alpha)$ to the subset of the $N - 1$ random variables (all other data locations). Thus, it can be used as declustering weight for $z(\mathbf{u}_\alpha)$, outcome of $Z(\mathbf{u}_\alpha)$. The redundancy matrix can be built from any positive-definite data location-to-data location similarity measure. If the data location-to-data location redundancy matrix is taken to be the covariance matrix, then it is shown that the proposed measure, $dnr(\mathbf{u}_\alpha)$, is equivalent to the crossvalidation kriging variance at location $\mathbf{u}_\alpha$. Being a kriging

variance, this nonredundancy measure depends on both the sampling geometry and the spatial continuity measure retained. This declustering alternative is illustrated using the GSLIB dataset **cluster.dat** and is compared to the cell declustering technique proposed by Journel (1983) and Deutsch (1989).

DETERMINANTS: REVIEW

Only symmetric and (semi)positive-definite matrices are considered hereafter. The determinant of $\mathbf{C}_N$, with dimensions $N \times N$, is denoted $|\mathbf{C}_N|$. $\mathbf{C}_N$ models the redundancy between the N variables describing a given stochastic system (or dataset). In this paper, the N variables should be understood as regionalized random variables, one at each data location. The following are relevant properties of determinants of such matrices, which are given in any linear algebra textbook (e.g., Strang, 1988):

(1) $|\mathbf{C}_N| \geq 0$, for (semi)positive definiteness.

(2) $|\mathbf{C}_N| = 0$, if a line (column) of $\mathbf{C}_N$ can be deduced from a linear combination of the others lines (columns). Thus, one of the N variables is in perfect linear relationship with some of the others (complete redundancy).

(3) $|\mathbf{C}_N| \leq \sigma^{2N}$, where σ^2 is the maximum diagonal value of $\mathbf{C}_N$. If $\mathbf{C}_N$ is modeled from a stationary covariance function, σ^2 is the stationary variance, the same for each variable. Equality is obtained for the situation of a diagonal matrix with σ^2 as entry, in which instance there is no redundancy. For a standardized redundancy matrix, $\sigma^2 = 1$, hence $|\mathbf{C}_N| \leq 1$.

(4)
$$\begin{vmatrix} \sigma^2 & \mathbf{0}^t \\ \mathbf{c} & \mathbf{C}_N \end{vmatrix} = \begin{vmatrix} \mathbf{C}_N & \mathbf{c}^t \\ \mathbf{0} & \sigma^2 \end{vmatrix} = \sigma^2 |\mathbf{C}_N|.$$

where $\mathbf{0}$ is a vector of N zeros and $\mathbf{c}$ a redundancy (e.g, covariance or correlation) vector.

(5) $|\mathbf{C}_N^{-1}| = 1/|\mathbf{C}_N|$

(6) The determinant of a product of matrices is the product of their determinants.

(NON)REDUNDANCY MEASURE

Consider a random function $\mathbf{Z}(\mathbf{u})$ sampled at N locations $\mathbf{u}_\alpha$, the data values being $z(\mathbf{u}_\alpha)$, $\alpha = 1, \ldots, N$. The random function is second-order stationary with unit variance, $\sigma^2 = 1$, and correlogram $\rho(\mathbf{h})$. The correlation matrix, $\mathbf{C}_N$

$= [\rho(\mathbf{h}_{\alpha,\beta})]_{N\times N}$, models the data location-to-data location redundancy. The degree of nonredundancy of random variable at location $\mathbf{u}_\alpha$ is defined as:

$$dnr(\mathbf{u}_\alpha) = \frac{|\mathbf{C}_N|}{|\mathbf{C}_{N-1}|} = \frac{\begin{vmatrix} 1 & \mathbf{c}'_\alpha \\ \mathbf{c}_\alpha & \mathbf{C}_{N-1} \end{vmatrix}}{|\mathbf{C}_{N-1}|} \tag{1}$$

where $\mathbf{C}_N$ is the data location-to-data location correlation matrix, $\mathbf{C}_{N-1}$ is the data location-to-data location correlation matrix excluding the random variable $Z(\mathbf{u}_\alpha)$ at location $\mathbf{u}_\alpha$, and $\mathbf{c}_\alpha$ is the correlation vector between random variable $Z(\mathbf{u}_\alpha)$ and the $(N-1)$ remainder random variables. The degree of nonredundancy, as defined in (1), measures the relative change (increase) in redundancy when adding the random variable $Z(\mathbf{u}_\alpha)$ to the $(N-1)$ other random variables.

As long as the redundancy matrix is modeled by a positive definite function, it can be seen as a matrix of scalar products. Then, the projection theorem can be applied and the minimum squared distance of any vector α to the subspace spanned by the $(N-1)$ others vectors can be computed as in Equation (1) (Luenberger, 1969, p. 57). As shown by Journel (1977), when the matrix of scalar products is a covariance matrix, this minimum squared distance is the simple kriging variance. A demonstration can be given as follows:

Let $\mathbf{A}$ and $\mathbf{B}$ be:

$$\mathbf{A} = \mathbf{C}_N = \begin{bmatrix} 1 & \mathbf{c}'_\alpha \\ \mathbf{c}_\alpha & \mathbf{C}_{N-1} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{C}^{-1}_{N-1} \end{bmatrix}$$

and their product:

$$\mathbf{B}\cdot\mathbf{A} = \begin{bmatrix} 1 & \mathbf{c}'_\alpha \\ (\mathbf{C}^{-1}_{N-1}\cdot\mathbf{c}_\alpha) & \mathbf{I}_{N-1} \end{bmatrix}$$

$\mathbf{A}$ is the redundancy matrix (e.g., correlation) and $\mathbf{B}$ is a matrix needed for the sole sake of the demonstration. Using the determinant properties given in the previous section:

$$|\mathbf{B}| = |\mathbf{C}^{-1}_{N-1}| = \frac{1}{|\mathbf{C}_{N-1}|}$$

$$|\mathbf{B}\cdot\mathbf{A}| = |\mathbf{B}|\cdot|\mathbf{A}| = \frac{|\mathbf{A}|}{|\mathbf{C}_{N-1}|} = dnr(\mathbf{u}_\alpha)$$

Now consider another matrix:

$$\mathbf{M} = \begin{bmatrix} 1 & \mathbf{0}' \\ (-\mathbf{C}^{-1}_{N-1}\cdot\mathbf{c}_\alpha) & \mathbf{I}_{N-1} \end{bmatrix}$$

such that $|\mathbf{M}| = 1$, hence:

$$|\mathbf{B} \cdot \mathbf{A} \cdot \mathbf{M}| = |\mathbf{B} \cdot \mathbf{A}|$$

because:

$$\mathbf{B} \cdot \mathbf{A} \cdot \mathbf{M} = \begin{bmatrix} (1 - \mathbf{c}'_\alpha \cdot \mathbf{C}^{-1}_{N-1} \cdot \mathbf{c}_\alpha) & \mathbf{c}'_\alpha \\ \mathbf{0} & \mathbf{I}_{N-1} \end{bmatrix}$$

then,

$$|\mathbf{B} \cdot \mathbf{A}| = |\mathbf{B} \cdot \mathbf{A} \cdot \mathbf{M}| = 1 - \mathbf{c}'_\alpha \cdot \mathbf{C}^{-1}_{N-1} \cdot \mathbf{c}_\alpha$$

Thus,

$$dnr(\mathbf{u}_\alpha) = \frac{|\mathbf{A}|}{|\mathbf{C}_{N-1}|} = |\mathbf{B} \cdot \mathbf{A}| = 1 - \mathbf{c}'_\alpha \cdot \mathbf{C}^{-1}_{N-1} \cdot \mathbf{c}_\alpha = \sigma^2_{SK}(\mathbf{u}_\alpha) \quad (2)$$

which is known to be the simple kriging variance when estimating the value at location $\mathbf{u}_\alpha$ from a linear combination of the $(N - 1)$ other data. The degree of nonredundancy is thus the crossvalidation kriging variance. The definition (1) is more general than the kriging variance definition (2) in that any similarity (redundancy) measure verifying the axioms of the scalar product can be used. Journel (1988) has proposed many such alternative spatial continuity measures, some of them being more robust with respect to outliers and data clustering than the traditional covariance (variogram) estimator.

In the particular situation of no spatial correlation, $\mathbf{C}_N$ resumes to the identity matrix, $\mathbf{c}_\alpha = \mathbf{0}$, thus $dnr(\mathbf{u}_\alpha) = 1$ for any data location $\mathbf{u}_\alpha$. If the random variable $Z(\mathbf{u}_\alpha)$ is perfectly redundant with some others, then $|\mathbf{A}| = |\mathbf{C}_N| = 0$ and its degree of nonredundancy is zero.

SPATIAL DECLUSTERING

As mentioned in the introduction, spatial declustering weights must account for specific spatial continuity observed in the sample dataset. Determination of the degree of nonredundancy (1) calls for a correlation function, $\rho(\mathbf{h})$, robust with respect to data clustering. Spatial continuity is modeled usually from the sample variogram (see for example Isaaks and Srivastava, 1989, chap. 7). One should use a declustered variogram estimator when handling spatially clustered datasets. The problem is that variograms are computed from data and declustering data require variogram. There seems to be a circular problem. To overcome this problem, it is suggested to use a spatial continuity measure robust against data clustering or to apply a data transform that reduces the impact of clustering on the variogram computation, or a combination of both. Examples of variogram estimators robust with respect to outliers and data clustering are

proposed in David (1977), Cressie and Hawkins (1980), Alfaro (1984), Armstrong (1984), Cressie (1984), Dowd (1984), Omre (1984), Isaaks and Srivastava (1988), and Journel (1988). The measure proposed here is based on the correlogram of the uniform scores of the clustered data values. The uniform transform rescales the data values in the interval [0, 1] which reduces the large fluctuations seen in many traditional sample variograms. The correlogram (variogram) of the uniform scores then is rescaled to have a unit variance (sill). Because the uniform transform is a rank-preserving transform, it preserves the major features of the original spatial continuity.

An important implementation aspect relates to the size of the search window centered at location $\mathbf{u}_\alpha$, when computing the correlation matrices $\mathbf{C}_N$ and $\mathbf{C}_{N-1}$. In other words, how large should N be? Ideally, N should be the total number of available data. However, when that number is too large (e.g., $N > 100$), the computation of $|\mathbf{C}_{N-1}|$ and $|\mathbf{C}_N|$ may be affected by numerical instability. Using a moving data neighborhood as in ordinary kriging, N is limited to the $n(\mathbf{u}_\alpha)$ data occurring within a window centered on $\mathbf{u}_\alpha$ with (directional) radii equal to the larger (directional) correlation lengths of the correlation model used. Indeed, the random variable (and its outcome) at data location $\mathbf{u}_\alpha$ has zero redundancy (correlation) with any random variable (or data location) outside this correlation window. If no other data occur within the correlation window, $n(\mathbf{u}_\alpha) = 0$, then datum α receives a full degree of nonredundancy (declustering weight equal to 1).

THE GSLIB DATASET

Figure 1A presents the exhaustive (reference) image (50 × 50 grid) from which the GSLIB sample data set **cluster.dat** was taken. Figure 1B shows the exhaustive (reference) omnidirectional experimental variogram. The exhaustive histogram is shown on Figure 1C. Figure 2A shows the location map for the data of the clustered sample data set (taken from Fig. 1A). Figure 2B shows the corresponding omdirectional sample variogram and Figure 2C presents the sample histogram. Note the impact on global statistics (histogram and variogram) of preferential sampling in the high-grade zones. The clustered data are the result of a two-step sampling (Deutsch and Journel, 1992, p. 34): a first set of 97 data is taken on a pseudoregular 5 × 5 grid, then a second set of 43 data was taken in areas detected as high-valued by the first stage of sampling. Figure 3A compares the exhaustive (reference) variogram, the traditional sample variogram of the clustered data and the sample variogram of the uniform scores of the same clustered data, all variograms (samples and exhaustive) are standardized to a unit variance (sill). The sample uniform score variogram reflects the underlying spatial continuity seen on the exhaustive experimental variogram, whereas the traditional sample variogram does not reveal any structure. The

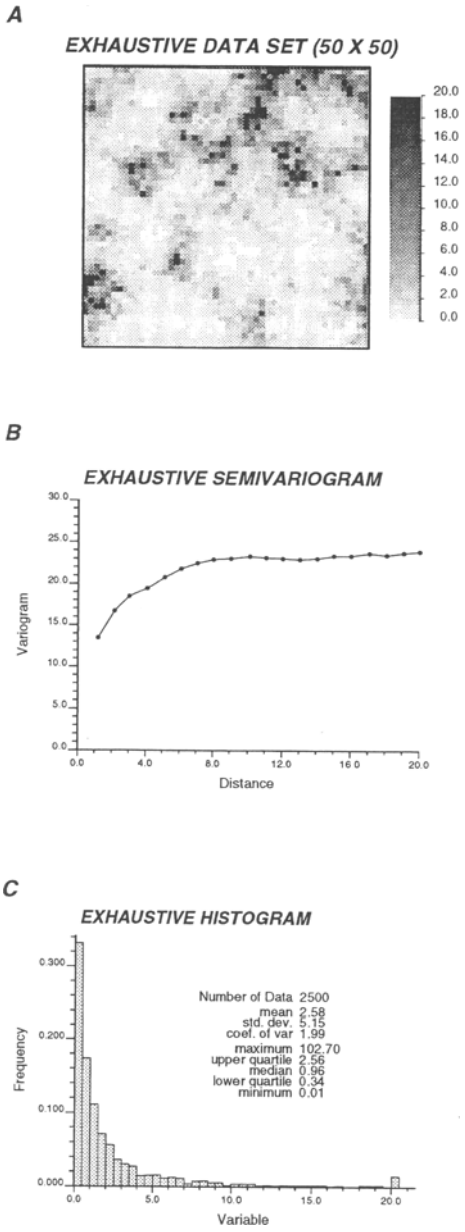

Figure 1. Reference map, exhaustive variogram and exhaustive histogram of GSLIB dataset.

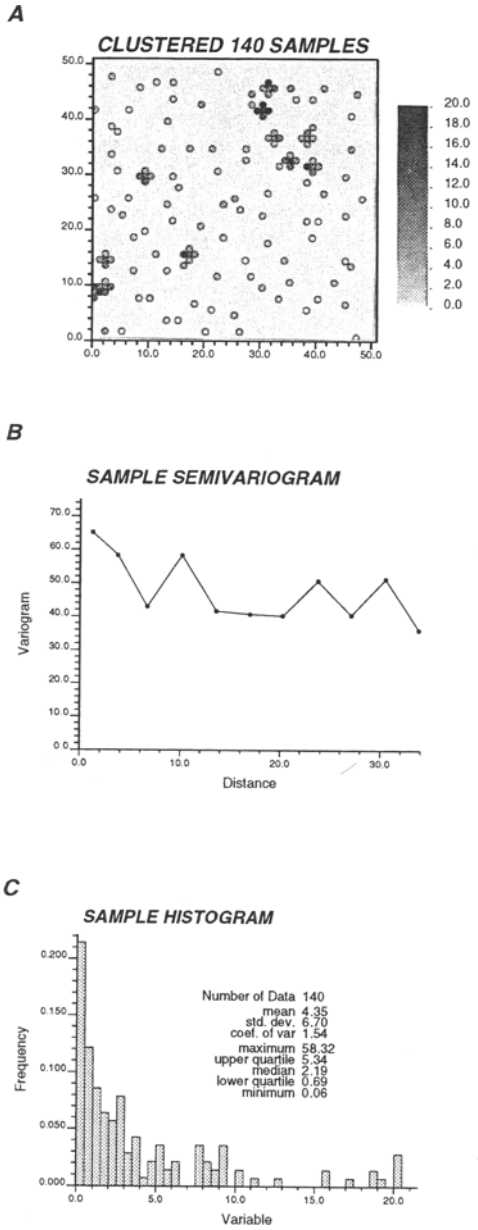

Figure 2. Location map, sample variogram histogram for 140 spatially clustered data.

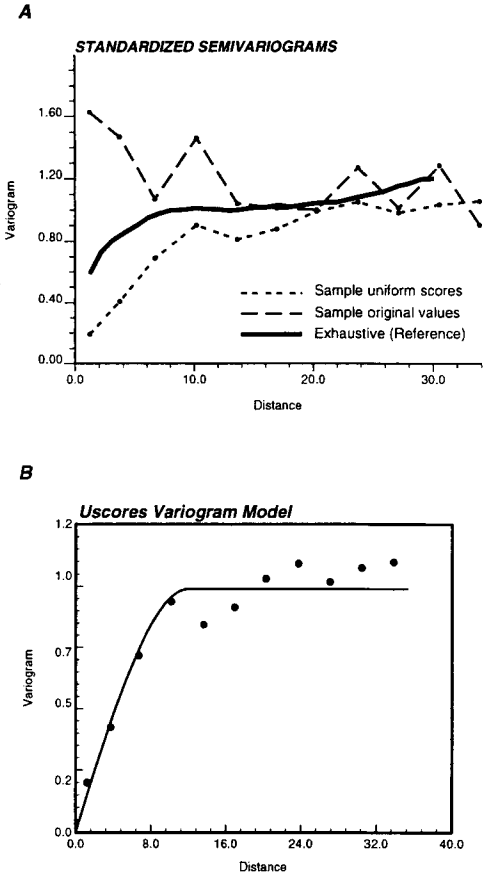

Figure 3. A, Comparison of exhaustive (reference) variogram (continuous line), sample variogram computed on original 140 data values (long dashed line) and sample variogram computed on uniform scores of 140 data values (short dashed). All variograms are standardized to unit sill. B, Modeling sample uniform scores variogram using spherical scheme with range of 10 units and zero nugget effect.

experimental variogram computed in using the uniform scores appears robust with respect to the data clustering. A standardized spherical variogram with range 10 units and zero nugget effect was used to model the spatial continuity from that sample uniform score variogram (Figure 3B). Figure 4 compares the declustered histogram obtained with the proposed approach and that obtained with program **Declus** of GSLIB (Deutsch and Journel, 1992, p. 207). The

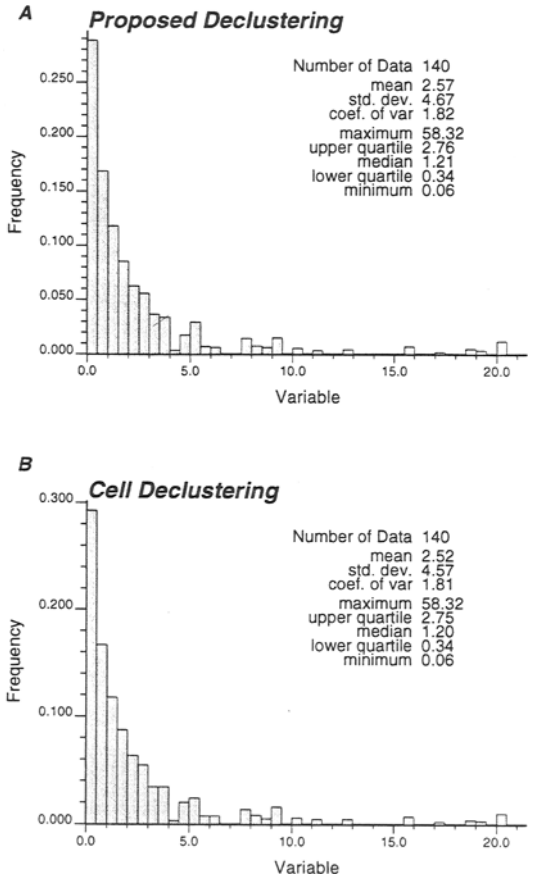

Figure 4. A, Declustered sample histogram obtained using proposed approach and local correlation window. B, Declustered sample histogram obtained using cell declustering technique.

program **Declus** applies the cell declustering technique initially proposed by Journel (1983). When compared to the reference statistics of Figure 1C, both techniques produce good results, with the proposed approach providing slightly more accurate mean and standard deviation. This is confirmed by the quantile–quantile plots of Figures 5, the proposed approach provides more accurate estimates of the higher reference quantiles (the most important in many applications). The search window used to compute matrix $\mathbf{C}_{n(\mathbf{u}_\alpha)}$ contains an average of 17 data values. To appreciate the sensitivity of the proposed approach to the size of the search neighborhood, that is against the number of data $n(\mathbf{u}_\alpha)$ retained

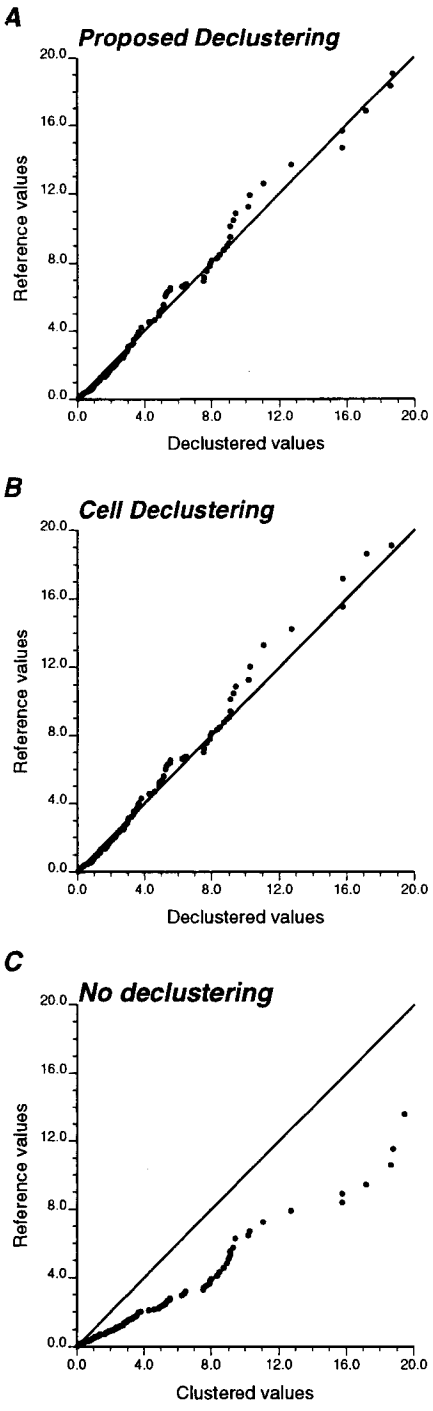

Figure 5. Quantile–Quantile plots comparing exhaustive (reference) distribution and sample distribution A, corrected with proposed degree of nonredundancy; B, corrected with cell declustering approach; and C, without correction.

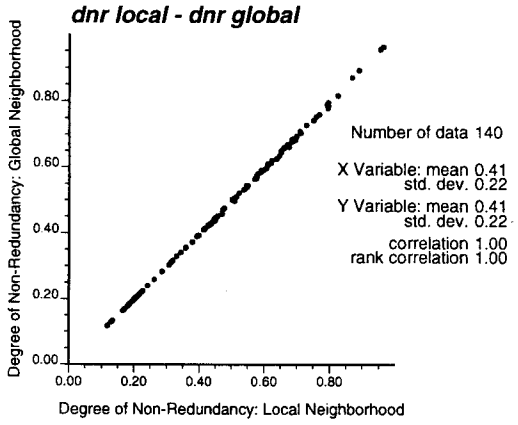

Figure 6. Cross plot between degrees of nonredundancy computed using local correlation window and using global neighborhood.

to build the correlation matrices $\mathbf{C}_{n(\mathbf{u}_\alpha)}$ and $\mathbf{C}_{n(\mathbf{u}_\alpha)-1}$, the degrees of nonredundancy, $dnr(\mathbf{u}_\alpha)$, were computed anew considering now a global neighborhood $n(\mathbf{u}_\alpha) = N = 140$. Figure 6 presents the cross plot between the degrees of nonredundancy computed using a moving correlation window vs. using a global neighborhood. For all practical purposes, the values are identical.

The sum of the 140 degrees of nonredundancy corresponding to the 140 clustered data is equal to 57.8, this suggests that the 140 clustered data provide the same information about global statistics as 58 nonredundant data, that is 58 data with interdistances larger than 10 units (correlation range).

CONCLUSIONS

The proposed declustering weight is defined as the ratio of determinants of correlation matrices, one including and the other excluding the regionalized random variable at the considered data location. This weight measures the increase in global redundancy when adding the regionalized random variable involved to the other regionalized random variables at data locations which occur within a correlation window centered at the location being declustered. When the redundancy matrices are covariance matrices, these declustering weights are equivalent to crossvalidation simple kriging variances. Similarly to the kriging variance, the proposed declustering technique accounts for both the sampling geometry and the spatial continuity of the dataset. The experimental covariance (or variogram) of the uniform scores of the original values, rescaled to a unit sill, is shown to be a correlation measure robust with respect to data clustering.

For the GSLIB dataset **cluster.dat,** the results of the proposed approach seem slightly better than those obtained with the traditional cell declustering technique. Both approaches should be compared on more case studies before drawing general conclusions. A definite advantage of the proposed approach over the cell declustering technique, is that it does not call for knowing *a priori* if clusters are in high values or low values areas. Another advantage is that the declustering weights all lie in the [0, 1] interval which makes them easy to interpret, 0 for complete redundancy and 1 for complete nonredundancy. The main limitation of the proposed approach is related to the need for a prior variogram estimator robust with respect to data clustering. In some situations, the variogram might be difficult to estimate. However, many robust measures of spatial continuity have been proposed in the geostatistical literature. Using those measures in combination with a rank preserving transform may be helpful in the characterization of the spatial continuity of clustered spatial datasets. One has to specially take care of the relative proportion of the nugget effect when modeling the spatial continuity. Indeed, being similar to a kriging variance, the proposed degree of nonredundancy is sensitive to that parameter.

REFERENCES

Alfaro, M., 1984, Statistical inference of the semivariogram and the quadratic model, *in* Verly, G., and others, eds., Geostatistics for natural resources characterization: Reidel, Dordrecht, The Netherlands, p. 45–53.

Armstrong, M., 1984, Improving the estimation and modeling of the variogram, *in* Verly, G., and others, eds., Geostatistics for natural resource characterization: Reidel, Dordrecht, The Netherlands, p. 1–20.

Cressie, N., 1984, Towards resistant geostatistics, *in* Verly, G., and others, eds., Geostatistics for natural resources characterization: Reidel, Dordrecht, The Netherlands, p. 21–44.

Cressie, N., and Hawkins, D., 1980, Robust estimation of the variogram: Math. Geology, v. 12, no. 2, p. 115–126.

Crozel, D., and David, M., 1983, The combination of sampling and kriging in the regional estimation of coal resources: comment: Math. Geology, v. 15, no. 4, p. 571–574.

David, M., 1977, Geostatistical ore reserve estimation: Elsevier, Amsterdam, 364 p.

Deutsch, C. V., 1989, DECLUS: A FORTRAN 77 program for determining optimum spatial declustering weight: Computers & Geosciences, v. 15, no. 3, p. 325–332.

Deutsch, C. V., and Journel, A. G., 1992, GSLIB: Geostatistical Software Library and user's guide, Oxford Univ. Press, Oxford, 340 p.

Dowd, P. A., 1984, The variogram and kriging: robust and resistant estimators, *in* Verly, G., and others, eds., Geostatistics for natural resources characterization: Reidel, Dordrecht, The Netherlands, p. 91–106.

Isaaks, E. H., and Srivastava, R. M., 1988, Spatial continuity measure for probabilistic and deterministic geostatistics: Math. Geology, v. 20, no. 4, p. 313–341.

Isaaks, E. H., and Srivastava, R. M., 1989, Introduction to applied geostatistics: Oxford Univ. Press, Oxford, 561 p.

Journel, A. G., 1977, Kriging in terms of projections: Math. Geology, v. 9, no. 6, p. 563–586.

Journel, A. G., 1983, Non-parametric estimation of spatial distributions: Math. Geology, v. 15, no. 3, p. 445–468.

Journel, A. G., 1988, New distances measures: the route towards truly non-Gaussian geostatistics: Math. Geology, v. 20, no. 4, p. 459–475.

Luenberger, D. G., 1969, Optimization by vector space methods: John Wiley & Sons, New York, 326 p.

Matheron, G., 1971, The theory of regionalized variables and its applications: Fascicule 5, Les Cahiers du Centre de Morphologie Mathématique, Ecole des Mines de Paris, Fontainebleau, 212 p.

Omre, H., 1984, The variogram and its estimation, *in* Verly, G., and others, eds., Geostatistics for natural resources characterization: Reidel, Dordrecht, The Netherlands, p. 107–125.

Schofield, N., 1993, Using the entropy statistic to infer population parameter from spatially clustered sampling, *in* Soares, A., ed., Geostatistics Tróia '92, Volumes 1-2, Quantitative Geology and Geostatistics, Volume 5: Kluwer Acad. Publ. Dordrecht, The Netherlands, p. 109–119.

Strang, G., 1988, Linear algebra and its applications (3rd ed.): Saunders College Publ., Harcourt Brace Jovanovich College Publ., Fort Worth, Texas, 505 p.