Computers & Geosciences Vol. 15, No. 4, pp. 593–598, 1989
Printed in Great Britain. All rights reserved

0098-3004/89 $3.00 + 0.00

DESCRIPTION OF A COMPUTER PROGRAM FOR ANALYZING MULTIVARIATE SPATIALLY DISTRIBUTED DATA

Hans Wackernagel

Centre de Géostatistique, Ecole des Mines de Paris, 35, rue Saint Honoré, 77305 Fontainebleau, France

(*Received* 1 *October* 1986; *received for publication* 27 *September* 1988)

Abstract—This paper gives a detailed description of a program for the factor analysis of multivariate data from samples taken in a physical environment. The spatial correlation of the samples is represented by a model of nested spatial structures. The correlations of the variables are summarized by performing a principal component analysis on the coefficients of the spatial structures. The result is a linear model of the coregionalization that can be used for factorial kriging, conditional simulation, and cokriging. The program could be written for a microcomputer connected to a mainframe.

Key Words: Geostatistics, Multivariate data, Variogram, Spatial data, Principal components.

INTRODUCTION

The computer program LINMOD has been developed for geostatistical factor interpretation of multivariate spatial information. This paper does not contain the source code of LINMOD. Rather it is intended to give a detailed description for the construction of such a program in any language and for any computer.

The probabilistic framework of geostatistics has been left aside in this paper. The ideas on which the algorithms are based on are described in the geostatistical literature (Matheron, 1965, 1982; Wackernagel, 1987).

Geostatistical factor interpretation is based on a linear model of the coregionalization. It assumes that a set of spatially correlated (regionalized) variables can be represented as a linear combination of uncorrelated factors (principal components).

As an illustration, data from a geochemical prospection campaign near Brilon (F.R.G.) is used. The content of three elements copper (Cu), lead (Pb), and zinc (Zn) has been measured on soil samples taken at irregular spacings in a region of $5 \times 6\,\mathrm{km}^2$.

The different steps for the interpretation of the data are the following:

(1) Calculation of the experimental variograms for all variables (simple variograms) and all pairs of variables (cross variograms).

(2) Definition of subsets of variables having simple variograms with a similar shape.

(3) Fitting a variogram model to the experimental variograms of a subset of variables. The model decomposes the variograms into several nested spatial structures.

(4) Calculation of principal components (factors) by an eigenvalue decomposition of the matrices of coefficients of each spatial structure.

(5) Interpretation of the spatial structures and of the factors using the nonnumerical (geological, biological, etc...) information related to the data.

(6) Estimation or simulation, in the sampled area, of values of the synthetic variables associated to structural and principal components of the original variables.

The program LINMOD only performs steps (3) and (4) of the data analysis and synthesis.

THE EXPERIMENTAL VARIOGRAM

The spatial increment of a variable z_i for a lag h shall be the difference between two values of the variable for two sample points x and $x + h$:

$$z_i(x) - z_i(x + h)$$

where

x is a vector containing the coordinates of a sample point,

$x + h$ is a vector giving the coordinates of another sample point,

i is an index for the different variables.

One-half of the average of increment products for different lags h belonging to a lag class h_k can be calculated. This is termed the experimental variogram:

$$\gamma_{ij}(h_k) = \frac{1}{2}\frac{1}{N_k}\sum_{\alpha=1}^{N_k}[z_i(x_\alpha + h) - z_i(x_\alpha)] \times [z_j(x_\alpha + h) - z_j x_\alpha)]$$

where

k is an index for the different lag classes,

h is a vector belonging to the lag class h_k,

N_k is the number of increment pairs for the lag class h_k,

j is another index for the different variables.

For $i = j$, the results for different lag classes h_k form the simple variogram of a variable z_i. For $i \neq j$, this is termed the cross variogram between a variable z_i and another variable z_j.

The cross variogram is symmetrical for any vector class h_k:

$$\gamma_{ij}(h_k) = \gamma_{ij}(-h_k).$$

Therefore it cannot detect a possible shift in the spatial correlation of a variable pair. In nature, the cross correlation between two variables may be shifted, especially if one variable has been displaced (e.g. a chemical element that has gone into solution and has migrated).

The experimental cross covariance can be calculated and be checked for shifted cross correlation. But as the covariance requires a more restrictive hypothesis of stationarity than the variogram, it is not a good criterion.

THE SHAPE OF THE VARIOGRAM

The experimental variogram of a variable may be connected to information about physical (geological, biological) processes in the area in which the variable was measured.

The shape of the variogram can be subdivided into several parts characterized by a sudden change of the slope. Each of these changes occurs at a certain distance termed the "range".

For example, there might be a discontinuity at the origin. It could be due to either measurement error (and then there is no range) or it could be due to the presence of geometrical objects of a size far below the size of the sampling grid. Then the corresponding range is almost zero and this is termed the 'nugget-effect".

There also might be a sudden change of slope, if the changes in the values of the variable are connected to the occurrence of objects such as mineralized lenses. Then, the distance at which a change of slope occurs will reflect the average diameter of the objects in the direction of calculation of the variogram. Serra (1968) has studied extensively the Lorraine iron ore deposit in this manner, and he identified six different ranges at different spatial scales.

It should be noted that clusters in the sampling pattern can lead to an artificial range that has nothing to do with the physical behavior of the variable.

THE FIT OF THE MODEL

In this procedure, the experimental variograms of a given variable set are modeled with a linear combination of variogram functions $g_u(h)$ with coefficients b^u_{ij}:

$$\gamma_{ij}(h) = \sum_{u=0}^{N_s} b^u_{ij} g_u(h)$$

where

u is the index of different spatial structures,

N_s is the number of spatial structures,

$g_u(h)$ is a variogram function with a specific range,

b^u_{ij} is the coefficient of a variogram function for a variable or a variable pair.

This variogram model should be used only for a set of variables that has the same number of spatial structures on its simple variograms. The variables therefore should be classified into different sets characterized by a particular variogram shape.

The data interpreter selects the number and the types of the variogram functions $g_u(h)$(Appendix 1) and specifies their ranges.

For a given set of N variables, there are $N(N + 1)$ 2 variograms to be fitted. The $(N_s + 1)N(N + 1)$ 2 coefficients b^u_{ij} are calculated by weighted least squares. We used the routine VE04AD from A.E.R.E. (1985) for this purpose. The weights are subjective and should be modified by the user so that the shape of each experimental variogram is reproduced satisfactorily by the variogram model (Appendix 2).

The routine VE04AD allows the programmer to define bounds for the values of the coefficients b^u_{ij}. As the matrices B_u of these coefficients have to be positive semidefinite, the following necessary (but not sufficient) conditions have to be respected:

$$b^u_{ii} \geqslant 0$$

and

$$|b^u_{ij}| \leqslant \sqrt{b^u_{ii} b^u_{jj}} \qquad \text{(Cauchy-Schwarz).}$$

This can be obtained easily by first fitting the direct variograms subject to the bounds:

$$0 \leqslant b^u_{ii} < \infty$$

where ∞ is the biggest number representable by the computer.

Afterwards, the cross variograms are fitted under the conditions:

$$-\sqrt{b^u_{ii} b^u_{jj}} \leqslant b^u_{ij} \leqslant +\sqrt{b^u_{ii} b^u_{jj}} \ .$$

The experimental variograms are stored on a binary file and are read into memory one at a time. Only the parameters of the model need to be retained in memory. After each fit, a plot of the experimental values and fitted model is realized on a graphic terminal.

For the cross variogram, the following curve is displayed:

$$\text{Hull}\,(\gamma_{ij}(h)) = \pm \sum_{u=0}^{N_s} \sqrt{b^u_{ii} b^u_{jj}}\, g_u(h) \ .$$

It is termed the "hull of perfect correlation", because it represents the ideal situation of a perfect positive or negative correlation in the frame of the model. It allows the user to judge the fit of the cross variogram in the context of the correlation between two variables.

THE BRILON DATA

The three variables Cu, Pb, Zn were measured on 2049 soil samples collected north of the town of Brilon (F.R.G.) by the Bundesanstalt fuer Geowissenschaften und Rohstoffe (B.G.R.) during the project "Rhenoherzynikum".

The fit of the six variograms of the Brilon data is shown on Figure 1.

The three simple variograms (Fig. 1, left-hand side) have a similar shape: a jump at the origin, a section with a steep slope up to a range of 130 m, a section with a smooth slope, that gets flatter at distances over 2300 m. Thus the following model was selected:

$$\gamma_{ij}(h) = b^0_{ij} + b^1_{ij}\,\mathrm{sph}\,(h,\ 130\,\mathrm{m}) + b^2_{ij}\,\mathrm{sph}$$
$$\times\ (h,\ 2300\,\mathrm{m})$$

where sph (h, a) is a spherical variogram function with a range parameter a (see Appendix 1).

The same model was fitted on the three cross variograms (Fig. 1, right-hand side), which show a different behavior of the structures. The degree of presence or absence of a spatial structure on the cross variogram reflects the degree of correlation of a pair of variables at a given spatial scale. The dotted curve on the graphs of the cross variograms represents the hull of perfect (positive or negative) correlation.

Table 1 shows the structural correlation coefficients calculated on the basis of the variogram model, together with the ordinary statistical correlation coefficient, which does not take into account the spatial nature of the data.

Cu and Pb seem to have a weak correlation at the regional scale (130–2300 m) and no correlation at the other scales. For Cu and Zn, the highest correlation is at the level of the scale below the smallest sample spacing (10 m). While for Pb and Zn there seems to be some correlation at the local level (10–130 m).

PRINCIPAL COMPONENTS

Principal components are an efficient tool to summarize the information contained in a variance–covariance matrix or a correlation matrix. Thus we can try to summarize the matrices B_u of coefficients b^u_{ij}, termed "coregionalization matrices", using a sum of uncorrelated factors.

Principal components establish a linear model of the variables composed of linear factors. The linear model of the coregionalization states that the original variables Z_i can be represented as a linear combination of uncorrelated variables Y^u_p with transformation coefficients a^i_{up}:

$$Z_i(x) = \sum_{u=0}^{N_s} \sum_{p=1}^{N_v} a^u_{ip}\, Y^u_p(x)$$

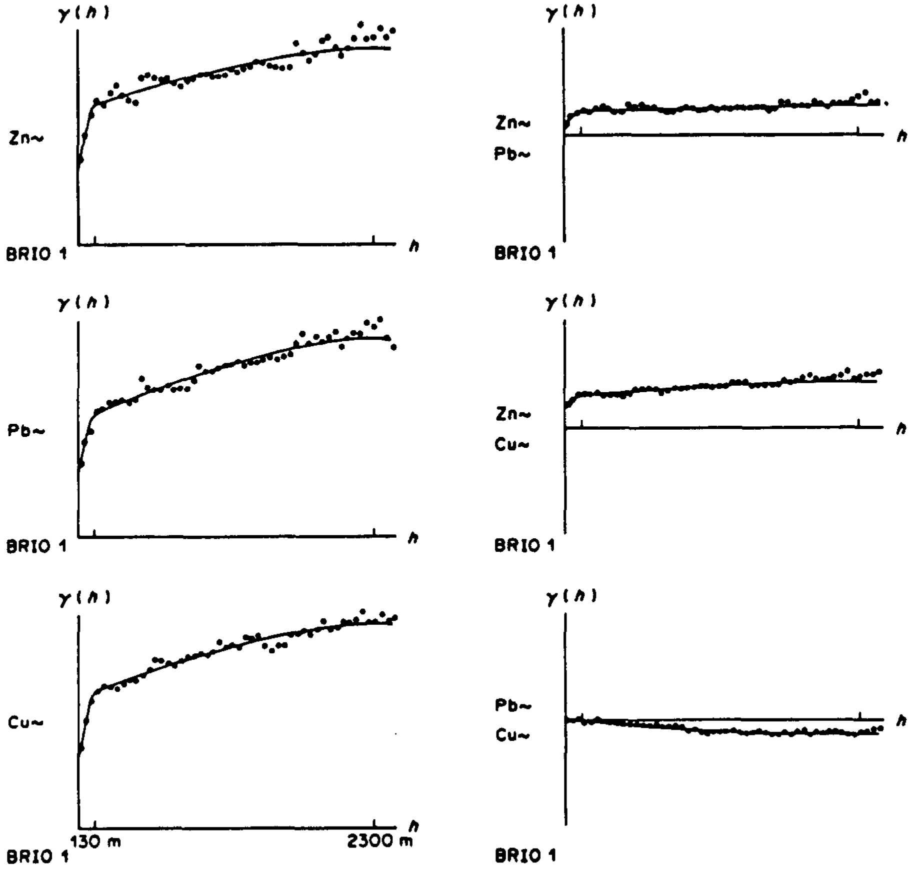

Figure 1. Six variograms of Brilon data with ranges of 130 and 2300 m.

Table 1. Statistical correlation coefficient and three structural correlation coefficients

	STATISTICAL CORRELATION	MICRO STRUCTURE 0 - 10M	LOCAL STRUCTURE 10 - 130M	REGIONAL STRUCTURE 130 - 2300M
CU-PB	-.08	0	-.04	-.36
CU-ZN	.42	.57	.31	.42
PB-ZN	.35	.23	46	11

where

p is the index of the principal components,

N_v is the number of variables, that is principal components.

The coregionalization matrices B_u are decomposed into a system of eigenvalues and eigenvectors:

$$B_u = Q_u^T \Lambda_u Q_u = (\sqrt{\Lambda_u} Q_u)^T \sqrt{\Lambda_u} Q_u = A_u^T A_u$$

where

T denotes matrix transposition,

Λ_u is a diagonal matrix of eigenvalues λ_p^u,

Q_u is an orthonormal matrix of eigenvectors q_p^u,

A_u is the matrix of transformation coefficients a^i_{up}.

The program LINMOD uses the routine EA06CD from A.E.R.E. (1985) to perform the eigenvalue decomposition.

The orthornormal matrices Q_u describe the position of the original variables on the unit hypersphere centered at the origin. Thus the projection of the variables on the plane spanned by two principal axes lies inside the unit circle around the origin. The importance of each axis is given by the coefficient:

$$\frac{\lambda_p^u \; 100\%}{\sum_u b_{ii}^u}.$$

It expresses the part of the total variance (contained in a coregionalization matrix) which is explained by a principal axis.

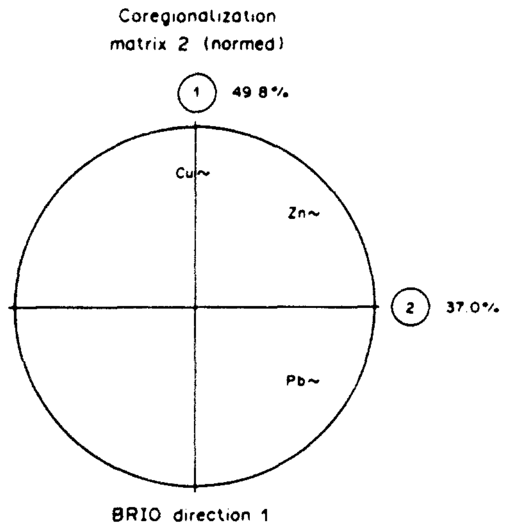

Figure 2. First two principal axes for normed coregionalization matrix of regional structure.

As an example, we show on Figure 2 the two most important principal axes (out of three) of the normed coregionalization matrix related to the regional structure of the Brilon data. These two axes explain 50 and 37% of the correlations at the level of the regional structure. The position of the variables on the graph is given by their coordinates on the principal axes, which are contained in the eigenvectors.

The graph summarizes the correlations between the three variables at the level of the second spatial structure. There is an opposition of Cu and Pb on the first axis because of their negative correlation. On the second axis, Zn lies nearer to Cu because its correlation with Cu is higher than that with Pb.

ORGANIZATION OF THE PROGRAM

A flowchart for the program LINMOD is given in Figure 3. It shows a simple vertical structure for the initial version of the program. The program, which should be interactive, could evolve later to a more horizontal structure with different modules for different purposes. For example, one section of the program could just fit simple variograms for determining an adequate model. Another section could fit the model to all variograms of a set of variables, and so on.

The program does not use much memory because it reads the experimental variograms one at a time from a binary file. It does not use much computing time either. A run of LINMOD including the fit of 120 variograms with 16 lag classes and the subsequent diagonalization of two 15 × 15 coregionalization matrices took 56 sec on a VAX 11/780. So LINMOD could be programmed on a small machine. But the prior step of a geostatistical analysis, the calculation of the experimental variograms, requires probably a bigger one. It takes about 20 min CPU time to calculate the 120 variograms of a set of 15 variables using 16 lag classes for 1200 samples.

CONCLUSION

The geostatistical factor analysis technique is a powerful tool for exploring the structure of multivariate spatial data. It also is a simple one, as the models for the variogram and for the coregionalization are linear. Furthermore it is rather economic,

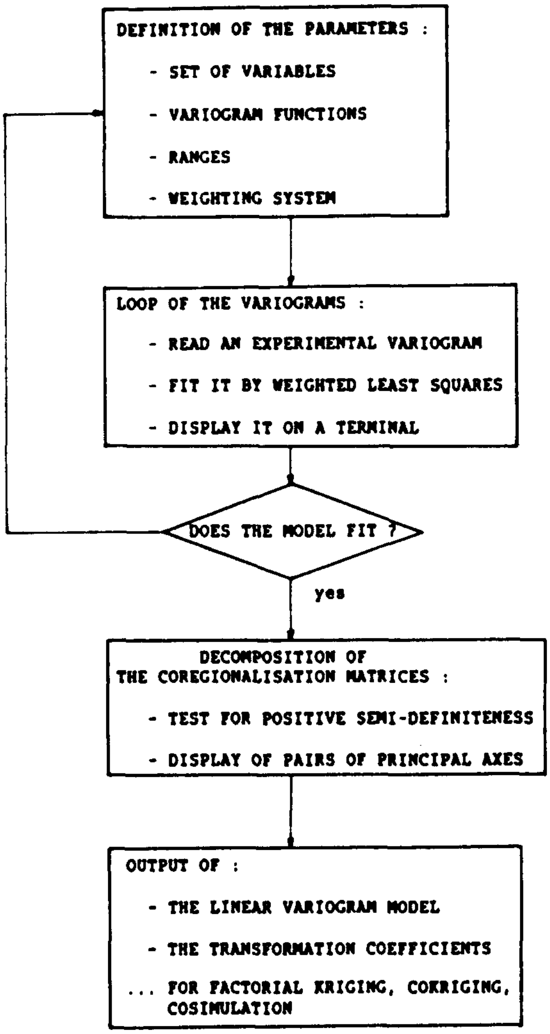
DEFINITION OF THE PARAMETERS :
- SET OF VARIABLES
- VARIOGRAM FUNCTIONS
- RANGES
- WEIGHTING SYSTEM
LOOP OF THE VARIOGRAMS :
- READ AN EXPERIMENTAL VARIOGRAM
- FIT IT BY WEIGHTED LEAST SQUARES
- DISPLAY IT ON A TERMINAL
DOES THE MODEL FIT ?
yes
DECOMPOSITION OF
THE COREGIONALISATION MATRICES :
- TEST FOR POSITIVE SEMI-DEFINITENESS
- DISPLAY OF PAIRS OF PRINCIPAL AXES
OUTPUT OF :
- THE LINEAR VARIOGRAM MODEL
- THE TRANSFORMATION COEFFICIENTS
... FOR FACTORIAL KRIGING, COKRIGING, COSIMULATION

Figure 3. Flowchart of program LINMOD.

because once the experimental variograms have been calculated, just a small computer is needed to analyze the coregionalization.

The linear model of the coregionalization is interesting for the scientist, because it provides a description of the behavior of natural phenomena at different spatial scales. This description, together with the original data, can then be used to obtain maps of these phenomena for the sampled area using kriging or conditional simulation techniques.

Acknowledgments—I wish to thank Ghislain de Marsily and Hazel Teil for helpful comments.

REFERENCES

A.E.R.E., 1985, Harwell subroutine library: Computer Science and Systems Division, Oxfordshire (U.K.), in FORTRAN.

Matheron, G., 1965, Les variables régionalisées et leur estimation: Masson, Paris, 305 p.

Matheron, G., 1982, Pour une analyse krigeante de données régionalisées: Centre de Géostatistique (Fontainebleau), Rept. N-732, 22 p.

Serra, J., 1968, Les structures gigognes—morphologie mathématique et interprétation métallogénique: Mineralium deposita, no. 3, p. 135–154.

Wackernagel, H., 1987, Geostatistical techniques for interpreting multivariate spatial information, *in* Chung, C.F., and others, eds., Quantitative analysis for mineral and energy resources: D. Reidel, NATO ASI Series C 223, Dordrecht, The Netherlands, 18 p.

OTHER REFERENCES

BLUEPACK-3D, General presentation of the package: Centre de Géostatistique, Fontainebleau, France, 209 p.

Touffait, Y., Renard, D., and Geoffroy, F., 1984, MAGMA—une structure de fichier adaptée aux applications géostatistiques: Sciences de la Terre, Ser. Inf., No. 20, p. 549-566.

Volle, M., 1985, Analyse des données (3rd ed.): Economica, Paris, 324 p.

Wackernagel, H., Webster, R., and Oliver, M. A., 1988, A geostatistical method for segmenting multivariate sequences of soil data, *in* Bock, H. H., ed., Proceedings of First Conference of International Federation of Classification Societies: North Holland, The Netherlands, 10 p.

APPENDIX 1

Variogram Functions

A few variogram functions $g(h)$ are listed:

(0) The nugget effect variogram function:

$$\text{nug}\,(h) = \begin{cases} 0 & \text{for } h = 0 \\ 1 & \text{for } h > 0. \end{cases}$$

(1) The exponential variogram function:

$$\exp\,(h, a) = 1 - \exp\left(-\frac{h}{a}\right).$$

(2) The spherical variogram function:

$$\text{sph}\,(h, a) = \begin{cases} \dfrac{3}{2}\dfrac{h}{a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3 & \text{for } 0 \leqslant h < a \\ 1 & \text{for } h \geqslant a. \end{cases}$$

(3) The Gaussian variogram function:

$$\text{Gauss}\,(h, a) = 1 - \exp\left[-\left(\frac{h}{a}\right)^2\right].$$

The Gaussian variogram function should not be used alone, as it is differentiable infinitely at the origin and corresponds to a phenomenon having the same property. A slight nugget-effect thus should be added always.

(4) The power variogram function:

$$\text{pow}\,(h, p) = h^p \text{ with } 0 < p < 2.$$

(5) The cubic variogram function:

$$\text{cub}\,(h, a) = \begin{cases} \left(\dfrac{h}{a}\right)^2 \left\langle 7 - \dfrac{h}{a}\left\{\dfrac{35}{4} - \left(\dfrac{h}{a}\right)^2\left[\dfrac{7}{2} - \dfrac{3}{4}\left(\dfrac{h}{a}\right)^2\right]\right\}\right\rangle & \text{for} \quad 0 \leqslant h < a \\ 1 & \text{for} \quad h \geqslant a. \end{cases}$$

APPENDIX 2

Weighting Systems

A few suggestions for setting the weights for the least-squares procedure are made:

(1) Set all weights equal to 1.
(2) Weight by the number N_k of couples of increments in a lag class.
(3) Weight by the inverse of the lag class number k (taken to some power).
(4) Give each lag class an individual weight.