# Spatial Variability: Error in Natural Resource Maps?

F. CSILLAG and M. KERTÉSZ

Research Institute for Soil Science and Agricultural Chemistry of the Hungarian Academy of Sciences, Budapest /HUNGARY/

## Introduction

Natural resource surveys have long been engaged in characterizing the spatial /and temporal/ variability of Earth /surface/ features. Since data collection and the handling of the spatial information inherent in the data are cumbersome and particularly problematic in large area resource assessment, methods of interpolation and classification have come into the focus of interest. This process has naturally increased the role of statistics. However, the validation of models that provide a description of error leads out of the area of statistics and /re-/defines the problem of capturing measures of spatial information /sampling/ in an environmental or geoscientific context.

The fundamental principles applied in this paper to determine the spatial characteristics of variables related to natural phenomena have been labelled raster-oriented in the GIS and quantitative modelling community. The importance of this distinction is primarily due to the fact that the last two decades have provided a number of new data sources and processing techniques in spatial data handling, where even the assumption of having contours, polygons and the like is not necessary. Geographers, cartographers, remote sensing scientists and others have been working on the exploitation of tools and methods in quantitative resource mapping, which have had a significant impact on cartography /BURROUGH, 1987; CSILLAG, 1987; MORRISON, 1986/, as well as on a number of geosciences /BURROUGH, 1986; HAWKINS and MERRIAM, 1974; KITANDIS and VONVORIS, 1983; WEBSTER, 1985/.

A number of authors have proved the potential of the sampling theory applied for mapping, in particular in experiment design /McBRATNEY et al., 1981/, the calibration of remotely sensed data /CURRAN and WILLIAMSON, 1986/, the performing and testing of interpolation /RIPLEY, 1981/ and/or classification /GORDON, 1987; ROSENFIELD et al., 1982/. These results provide means of determining strategies for better sampling to achieve higher accuracy for - in general - a better /i.e. quantitative/ description and understanding of spatial phenomena leading to more reliable maps.

One of the most important and popular keywords used by those involved in this type of research is spatial variability. This paper reviews some statistical models for capturing information on spatial variability, outlines a method to deal with it quantitatively in terms of resolution and estimation variance, and finally raises the issue of incorporating this knowledge of spatial variability in a geographical information system /GIS/. The approach is illustrated with examples and is thought to be appropriate for a wide variety of applications.

In the first part of the paper sampling as such is reviewed comparing ideal vs. field or remote data collection and signal reconstruction with given constraints. Next, a concise summary of nested sampling is given, by which the most informative sampling distance can be effectively selected in terms of variance in a pilot study prior to the detailed sampling. The second part presents optimal interpolation techniques with special reference to spatial resolution and estimation variance. Finally, in the third part a prototype soil mapping experiment in Hungary demonstrates the potential of the approach.

## Sampling revisited

In general there is no a priori information basis for the optimum sampling distance. A convenient way to determine a reasonable sampling distance is the nested model of variance, recently described by OLIVER and WEBSTER /1986a/, which partitions the total variance by distance-stages in a hierarchical manner. Thus, the distance most sensitive to variability can be considered as characteristic of the spatial phenomena being sampled.

Many properties of the Earth's surface features can be treated as continuous signals: however, scientists need to describe their patterns using generally sparse point observations. Field sampling in this respect can be represented as seen in Fig. 1.

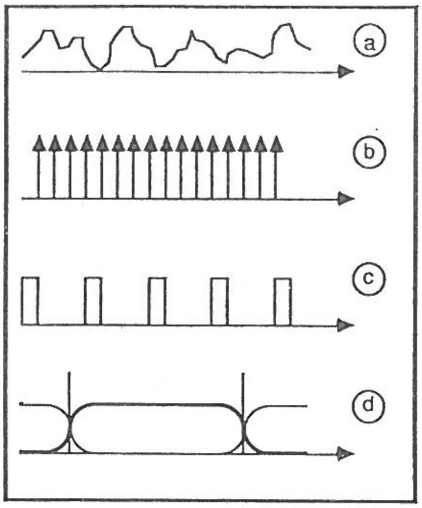


*Fig. 1*

Schematic representation of: a/ a variable as a continuous signal; b/ its ideal; c/ real in-situ, and d/ remote sampling

Having a continuous signal $g(x)$, where x denotes spatial coordinate/s/, point sampling can be approximated with a Dirac- $\delta$ series:

$$g(x)\Sigma_k \delta(x-kd) = \Sigma_k g(x)\, \delta\, (x-kd) \qquad /1/$$

where d denotes the sampling distance. From a number of illustrative trials one receives the impression that the shorter the sampling interval, the better the reconstruction of the signal, although the right hand side of equation /1/ is equal to zero at every $x \neq kd$.

The exact relationship between the sampled and original signal is defined by the sampling theorem /e.g. TOBLER, 1969; MESKO, 1984/. Recalling the Fourier-transformation of the Dirac-$\delta$ series, the Fourier-transformation of equation /1/ can be written as

$$G(f)*(1/d)\Sigma_k \delta(f-[k/d]) = (1/d)\Sigma_k G(f-[k/d]) \qquad /2/$$

where $*$ denotes convolution, f stands for spatial frequency and the expression gives the sampled spectrum $G_s(f)$. Sampling leads to a spectrum, which is periodic by the Nyquist interval, i.e. by its $[-1/2d,\ 1/2d]$ primary part. The summed terms of the right hand side may overlap, however:

$$\text{if } |f| > f_{up} \leq f_{Nyquist} = 1/2d \qquad /3/$$

$$\text{then } G_s(f) = 0 \text{ for } |f| > f_{up}$$

meaning that the original signal can be reconstructed without any loss of information if the period of the highest spatial frequency is sampled at least twice. If the sampling distance is selected arbitrarily and is greater than half the distance corresponding to the Nyquist interval, aliasing will significantly distort the spectrum.

Once having the Fourier-transformation of $g(x)$, the autocovariance function can be derived as follows:

$$\gamma(h) = F^{-1}\{|G(f)|^2\} \qquad /4/$$

from which spatial patterns /e.g. dominant frequency/ can be computed.

## Optimal interpolation

The interpolation of stationary data can be optimized, for instance, in terms of RMS-error with Wiener-filters /KENDALL and STUART, 1966; CLEARBOUT, 1976/. This leads to the solution of a set of linear equations, and allows us to determine the minimum error in the following form:

$$E_{min} = \phi(0)-\Sigma_k c_k \phi_+(k) \qquad /5/$$

where $\phi$ denotes the autocovariance of the measured data /i.e. $\phi(0)$ is the power/, $\phi_+$ is the cross-covariance of the signal-component and the measured data, while $c_k$ is the filter-coefficient. The practical limitation of this estimate is that it is heavily dependent on the covariance function estimate based on a limited sample of realizations as well as on a priori information about the form of signal and noise. /an example may serve to illustrate the

the importance of the "meaningful" assumptions: considering non-correlated signals and measurements, the $c_k = 0$ values for all k values are the optimum interpolator Wiener-filter coefficients./

Besides the practical limitations, there are theoretical ones, too. Equation /4/ implies that the process for which the g(x) signal is recorded should be stationary in both mean and covariance. As it happens, many properties of the land appear not to be stationary in this sense /OLIVER and WEBSTER, 1986b/. This led MATHERON /1965/ to consider the somewhat weaker assumptions of stationarity:

$$E\{g(x) - g(x+h)\} = 0 \qquad /6/$$

and

$$E\{[g(x)-g(x+h)]^2\} = 2\cdot\gamma h) \qquad /7/$$

where E denotes expectation and the function $\gamma(h)$ is called the semi-variogram. Its existence is based on the stationarity of the differences between samples.

If the process is second-order stationary, then the semi-variogram is related to the autocovariance function as: $\gamma(h) = \phi(0) - \phi(h)$, and either $\gamma(h)$ or $\phi(h)$ can be used to describe the spatial process. If, however, only the so-called intrinsic hypothesis holds (c.f. equations /6/ and /7/), then the covariance is undefined and g(x) is called a regionalized variable. The semi-variogram can be estimated without bias according to the definition implicit in equation /6/, or to the formulae for two or more dimensions and irregular sampling /WEBSTER, 1985/.

The method of estimation embodied in regionalized variable theory is known in earth sciences as kriging /JOURNEL and HUIJBREGTS, 1978/. It is essentially a means of weighted averaging:

$$g'(y) = \Sigma_k L_k g(x_k) \qquad /8/$$

in which the weights ($L_k$) are chosen so as to give unbiased estimates at y (g(y')).

It is optimal in the sense that it minimizes the estimation variance:

$$\sigma_E^2\ (y) = E\{[g(y)-g'(y)]\ \} = \Sigma_k L_k \gamma^*(x_k,y) - \gamma^*(y) \qquad /9/$$

where $\sigma_E$ (y) is the estimation variance at y /which can be either a point or a block/, E denotes expectation, $\gamma^*(x_k,y)$ stands for the semi-variance of the property estimated by all point-pairs between $x_k$ and y, taking into account both the distance and angle, while $\gamma^*(y)$ denotes the average semi-variance within the block. The latter term can be omitted, being zero for points. For formulae to obtain the estimation variance, see e.g. JOURNEL and HUIJBREGTS /1978/. Kriging also requires the estimation of a set of linear equations. Besides its advantage of weaker assumptions, it should be noted, from a practical point of view that given the semi-variogram, the kriging matrix is dependent only on the location of the samples /SZIDAROVSKY and YAKOWITZ, 1985/.

## Quantitative soil moisture mapping — An example

Large area surveys of natural resources, in general, require many samples when applying systematic sampling. Remote sensing techniques, however, confine ground sampling to training /and test/ areas. Nevertheless, neither of these approaches provide "automatically" reasonable tools for adjusting errors /in terms of estimation variances/ to equally reliable mapping of natural variables. The above-outlined geostatistical approach, however, assures sufficient information for such a purpose while constructing the map, and may also increase the efficiency of data processing in ways which have been implemented in the following experimental project.

The problem of defining optimum data content in attributes and space and suitable data structure is crucial for soil information systems in particular. The Hungarian Soil Information System /TIR=HunSIS/ /CSILLAG et al., 1986; VÁRALLYAY et al., 1985/ is at present capable of handling point and polygon data on a quadtree-based data base with a finest resolution of 25 m. These data contain conventional basic /static/ soil property and soil type descriptions. The study from which soil moisture data are partly presented below has been focused on spatial and temporal variability and interrelationships.

First of all, measurement accuracy was estimated for the given sampling area and equipment /detailed methodology in CSILLAG and SZABÓ, 1987/. From the point of view of spatial uncertainties, first the optimal sampling distance had to be determined /i.e. the distance to which most of the variance can be assigned over several orders of magnitude/ based on 3- and 4-stage nested sampling from 1 to 100 metres. Based on the results of nested analysis of variance, sampling was carried out based on a regular grid. This data set could be used for computing the semi-variogram or the power-spectrum, which can serve as the kernel-function in interpolation. Once varying size block-estimates are made, the estimation variances can be plotted against block-size /resolution/, and the evaluation of such functions can determine the maximum permissible resolution of the data in terms of acceptable error. This information can also be used to determine the optimal pixel-size for remotely sensed data as well as to be incorporated in a GIS.

### *Measurement accuracy*

The determination of measurement accuracy for natural variables provides an additive term in the final variance estimates as a contribution to the zero-lag /i.e. nugget/ variance. The reasons why it is briefly mentioned here are twofold:

- one should never neglect the definite limitations of a map derived from measurement /i.e. quantized/ data, especially in the case of new types of devices /BERNARD et al., 1984/, and
- if spatial and non-spatial covariances are to be used in further inferences, the uncertainties related to each variable must be taken into consideration /McBRATNEY and WEBSTER, 1983/.

In our case a novel, fast, capacitance-type soil moisture meter /VÁRALLYAY and RAJKAI, 1987/ was tested in the laboratory and under a range of field conditions against conventional methods and reproducibility. Without going into details of soil moisture measurements, it should be noted that field methods generally require more samples than laboratory measurements for the same error level /VACHAUD, 1978; HAWLEY et al., 1982/. This is particularly true when one compares moisture per weight /laboratory/ with moisture per volume /field/ data with conversion using bulk density values. As expected, the higher variances obtained with the volumetric method still pay off in the field for the instantaneous measurement capability. With a rise of an order of magnitude in the number of samples per point in the field /~50/, an acceptable error level /~10%/ could be achieved under several soil conditions.

*Nested sampling and analysis of variance*

Two study areas were selected for test purposes, where spatial interrelationships were to be evaluated cartographically and in terms of soils. At both study areas, samples were collected at 72 locations and variance was analyzed according to 3-stage and 4-stage /unbalanced/ nested design /as described by OLIVER and WEBSTER, 1986a/.

For illustration purposes let us take just one of the soil moisture data sets /neglecting temporal variations, mechanical composition and vegetation data/. Nested design for analysis of variance is an efficient way to obtain rough estimates for the proportions of variance that can be assigned for a given distance /lag/. There is no way to predict a priori decomposition of variance /except through field experience/, so the way of accepting models is based on the significances of the F-tests between stages. /Note that the highest significance values do not necessarily coincide with the highest variances./ The 3-stage model provided acceptable /i.e. significant/ differences between variances, so the highest proportion of variance assigned to a 5 m lag was taken into consideration for detailed sampling. Figure 2 demonstrates proportional cumulative variance and significance in nested analysis of variance for soil moisture measurements.

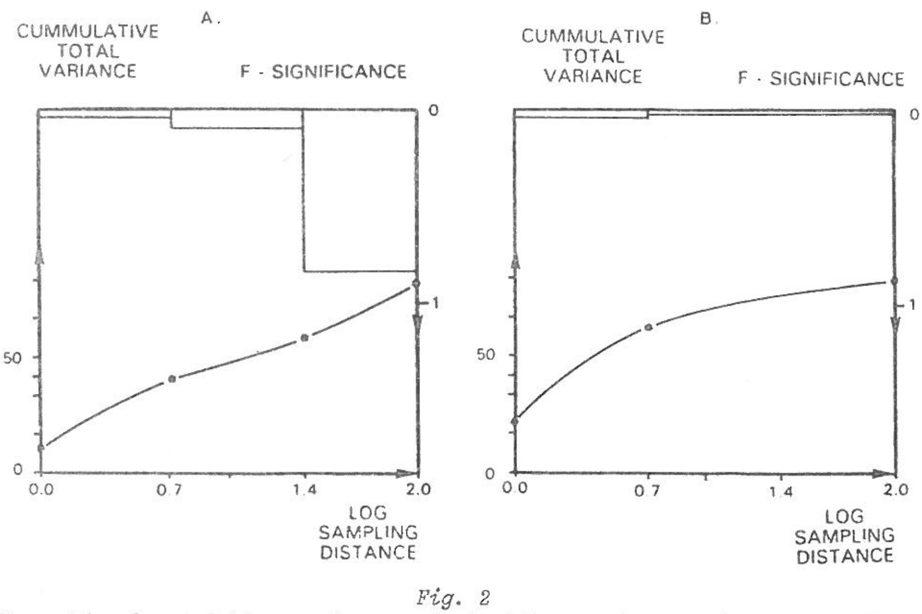


*Fig. 2*

Proportional cumulative variance and significance in nested analysis of variance for soil moisture measurements. A. 4-stage model /distances were 1, 5, 25 and 100 m, respectively, and the model was rejected due to the significance values of F-ratios between stages/. B. 3-stage model with the omission of the 25 m-apart stage /suggesting that the most characteristic distance lies between 5 and 25 m/ /the curves are just to guide the eye/

Detailed sampling for defining functions describing spatial characteristics and testing the accuracy of interpolation was carried out on a 4 metre 32*64 grid /again at different depths and for various variables, too/, and soil moisture data is presented in Fig. 3. The selection of 4 metres as the sampling distance was based on the sampling theorem and the nested analysis of variance, since the latter suggests that the characteristic distance is around 5 m /possibly a little more, according to the 4-stage model/, while the former suggests to sample that twice.

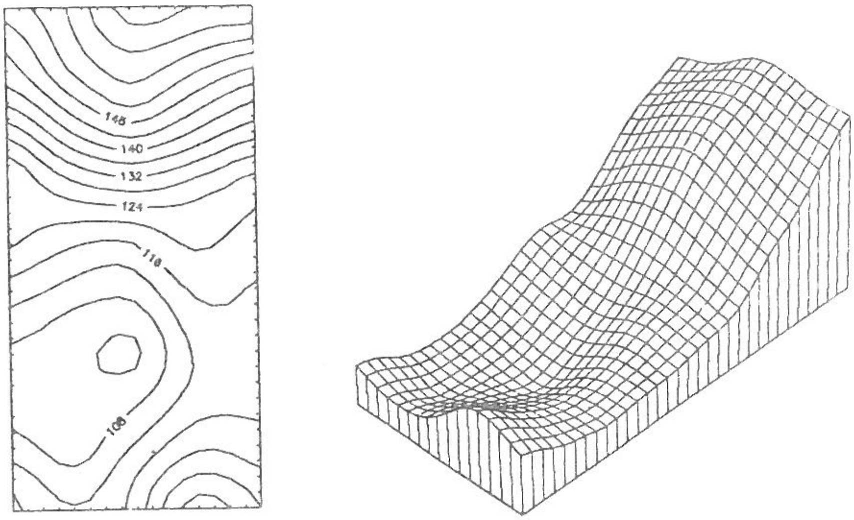


*Fig. 3*

Raw data set on a 4-metre grid /moisture in thousands/

*Defining functions of spatial characteristics*

Based on the detailed data set, the semi-variogram was estimated assuming a spherical model. In the case of moisture, there are certain implications of this model /e.g. superimposition of point sources; McBRATNEY and WEBSTER, 1986/, nevertheless, our interest is not focused on the "exact" model, but rather on using it for interpolation. Fig. 4 agrees with previous assumptions: the range of the semi-variogram is 2.5-times the sampling distance.

*Interpolation - block-variance estimates*

To study the effects of spatial decomposition on the estimation of variance, the scheme shown in Fig. 5 was used.

The estimates obtained by different resolutions make it possible to examine functions relating resolution to estimation variance. Fig. 6 contains some examples.

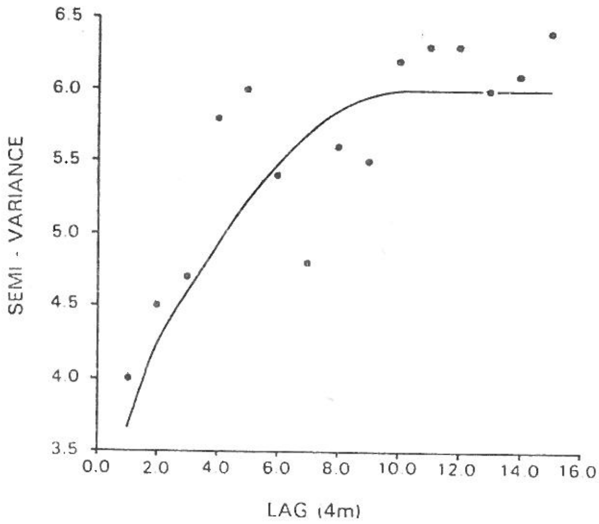


*Fig. 4*

Semi-variogram determined for soil moisture data measured on 4 metre grid /the curve is the result of fitting a spherical model with the following parameters: nugget variance = 3.0 [%moisture$^2$], sill = 6.0, range = 10, where the last term means a characteristic distance of 2.5 lags /sampling intervals/

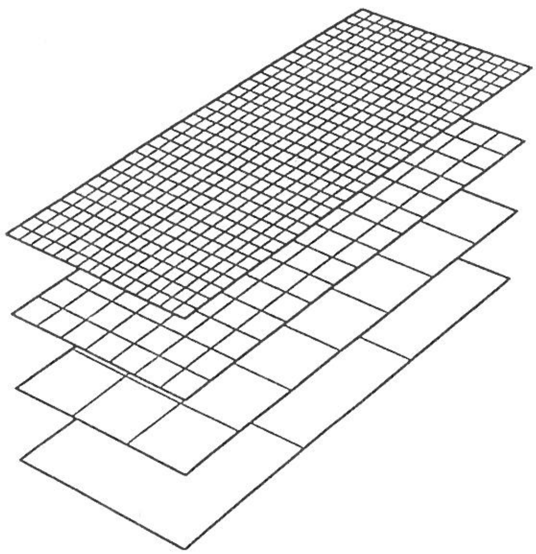

*Fig. 5*

Scheme of spatial decomposition for a study area /original resolution [32*16 points] decreasing from top to bottom/

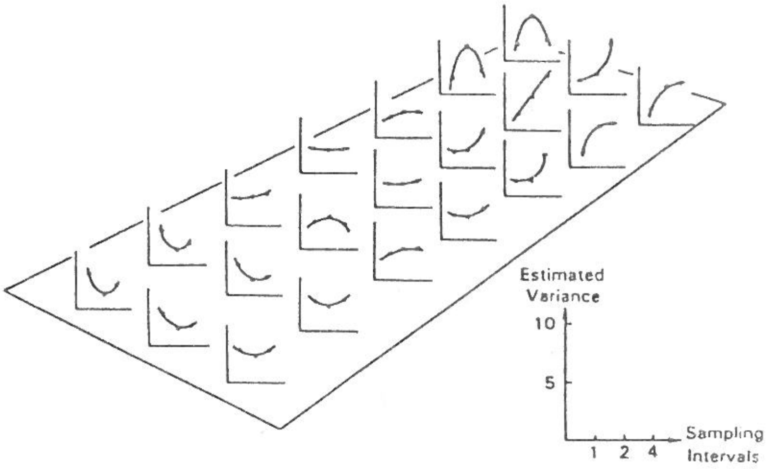


*Fig. 6*
Some functions relating resolution to variance /the functions are positioned to their approximate location and the curves are drawn based on 3 points to guide the eye/

*Defining optimal resolution based on functions relating resolution to variance*

Some publications seem to suggest monotonically decreasing block-variance /e.g. BURROUGH, 1986/. However, this is not the case. Some intuitive ideas about these functions were presented by the author, but further theoretical work is still needed to interpret and predict their different shapes. Nevertheless, as one might expect, in some instances the curves have well-defined minima, but not at the finest resolution. This enables us to produce maps with either locally minimum uncertainty, or maps with equal variances, but with varying resolution. As Fig. 7 suggests, this can be incorporated with the internal data structure /say, quadtree/ of a geographical information system even in such a way that no query is accepted for data with less confidence than a given limit.

## Conclusions and suggestions

The solution of the problem of making reliable maps has to be accompanied by the recognition that "accurate" and "erroneous" are not disjunct sets, but can rather be viewed as a continuum. The methodology presented here is suitable for the determination of spatial characteristics of variables related to natural phenomena, and applying that information for areal interpolation and variance estimate can lead to homogeneous variance maps with varying resolution, or vica versa.

The contradiction between the requirements of constant accuracy and constant resolution will necessitate significant contributions from different branches of science for modelling co-regionalized variables and/or elaborating classification strategies that can lead from rasters to vectors,

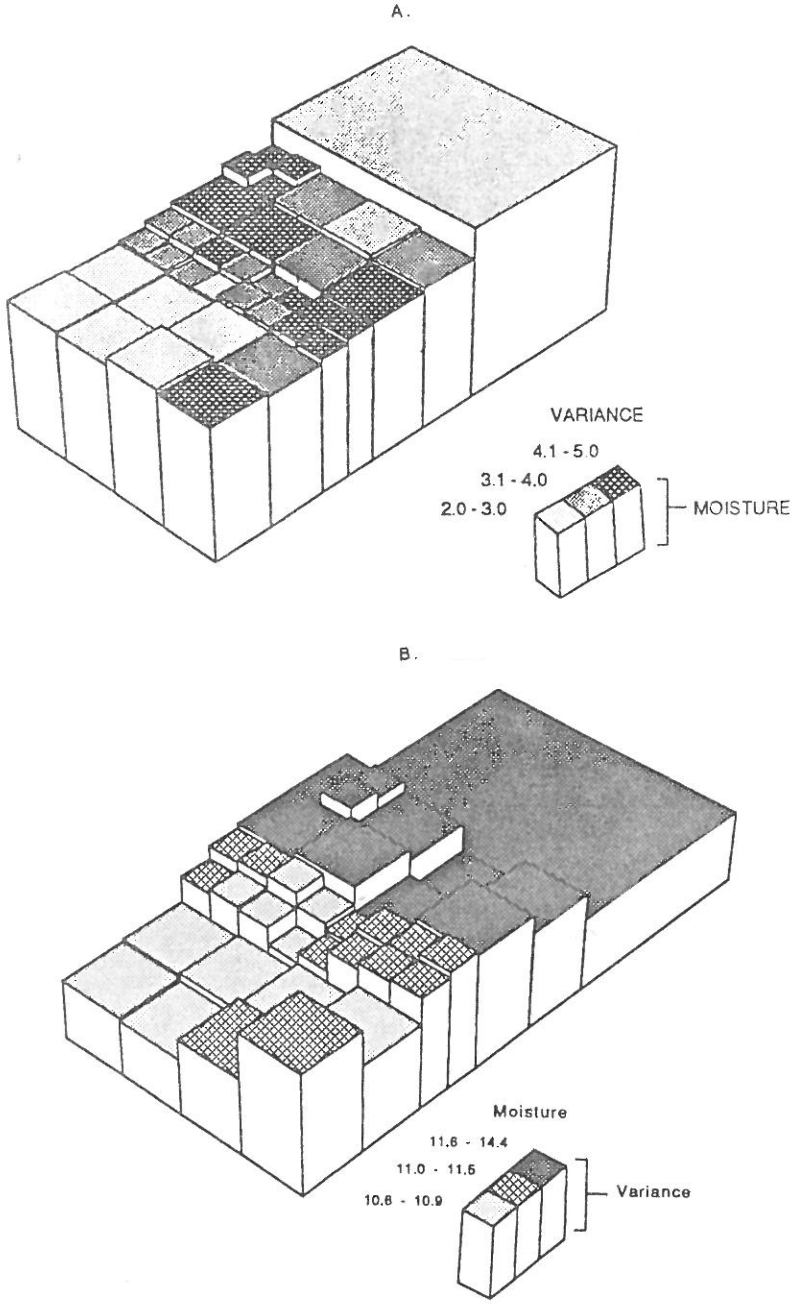


*Fig. 7*
Variable resolution sample map of soil moisture where all variances are below 5%. A. Shades represent variance; B. Shades represent moisture

or from vectors to rasters. It is also understood that at present, whenever a choice is made concerning data representation, it implies a particular model which is further used to estimate accuracy. A unified approach to spatial error modelling seems to be inevitable, to combine the advantages of vector- and raster-based models. Furthermore, once errors are determined their role in further processes /cognition or error absorption - BEDARD, 1987/ should be studied in more detail. These studies can link sampling methodology with error modelling, variance estimates with data structures, and consequently provide a basis for upgrading geographical information systems to geographical expert systems.

## References


BEDARD, Y., 1987. Uncertainties in land information system databases. Auto-Carto-8. /Ed.: CHRISMAN, N./ 175-184. ACSM. Falls Church.

BERNARD, R. et al., 1984. Comparison of three in-situ surface soil moisture measurements and application to C-band scatterometer calibration. IEEE Transactions on Geosciences and Remote Sensing, GE-22. 388-393.

BURROUGH, P. A., 1986. Principles of Geographical Information Systems for Land Resources Assessment. Oxford University Press. Oxford.

BURROUGH, P. A., 1987. Mapping and map analysis: New tools for land evaluation. Soil Use and Management. 3. 20-25.

CLEARBOUT, J. F., 1976. Fundamentals of Geophysical Data Processing. McGraw-Hill. New York.

CSILLAG, F., 1987. A cartographer's approach to quantitative mapping of spatial variability: from ground sampling to remote sensing of soils. AutoCarto-8 /Ed.: CHRISMAN, N./ 166-176. ACSM. Falls Church.

CSILLAG, F. and SZABÓ, J., 1987. Experimental results and theoretical considerations of ground calibration for remote sensing of soil moisture. Manuscript.

CSILLAG, F. et al., 1986. TIR: A computerized cartographic soil information system. AutoCarto London /Ed.: BLAKEMORE, M./. Vol. II. 52-60.

CURRAN, P. J. and WILLIAMSON, H. D., 1986. Sample size for ground and remotely sensed data. Remote Sensing of Environment. 20. 31-43.

GORDON, A. D., 1987. Classification and assignment in soil science. Soil Use and Management. 3 3-8.

HAWKINS, D. M. and MERRIAM, D. F., 1974. Zonation of multivariate sequences of digitized geologic data. Mathematical Geology. 6. 263-269.

HAWLEY, M. E., McCUEN, R. H. and JACKSON, T. J., 1982. Volume-accuracy relationship in soil moisture sampling. Proc. ASCE. 108. IR1-11.

JOURNEL, A. G. and HUIJBREGTS, Ch. J., 1978. Mining Geostatistics. Academic Press. London.

KENDALL, M. G. and STUART, A., 1966. The Advanced Theory of Statistics. Griffin. London.

KITANDIS, P. K. and VONVORIS, E. G., 1983. A geostatistical approach to the inverse problem in groundwater modelling /steady state/ and one dimensional simulations. Water Resources Research. 19. 677-690.

MATHERON, G., 1965. Les Variables Regionalisees et Leur Estimation. Masson. Paris.

McBRATNEY, A. B. and WEBSTER, R., 1983. Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. J. Soil Sci. 34. 137-162.

McBRATNEY, A. B. and WEBSTER, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. J. Soil Sci. 37. 617-639.

McBRATNEY, A. B., WEBSTER, R. and BURGESS, T. M., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. Computers and Geosciences. 7. 331-334.
MESKO, A., 1984. Digital Filtering: Applications in Geophysical Exploration. Akadémiai Kiadó - Pitman - Halsted. Budapest - London - New York.
MORRISON, J., 1986. Cartography: A milestone and its future. AutoCarto London. /Ed.: BLAKEMORE, M./ Vol. I. 1-13. AutoCarto London Ltd.
OLIVER, M. A. and WEBSTER, R., 1986a. Combining nested and linear sampling for determining the scale of spatial variation of regionalized variables. Geographical Analysis. 18. 227-242.
OLIVER, M. A. and WEBSTER, R., 1986b. Semi-variogram for modelling the spatial pattern of landform and soil properties. Earth Surface Processes and Landforms. 11. 491-504.
RIPLEY, B. D., 1981. Spatial Statistics. John Wiley. New York.
ROSENFIELD, G. H., FITZPATRICK-LINS, K. and LING, H. S., 1982. Sampling for thematic map accuracy testing. Photogrammetric Engineering and Remote Sensing. 48. 131-137.
SZIDAROVSKY,F. and YAKOWITZ, S.J., 1985. A comparison of kriging with nonparametric regression methods. Journal of Multivariate Analysis. 16. 21-53.
TOBLER, W. R., 1969. Geographical filters and their inverses. Geographical Analysis. 1. 234-253.
VACHAUD, G., 1978. Methods of field characterization of the hydrodynamic properties of an unsaturated soil. Annales Agronomiques. 29. 1-36.
VÁRALLYAY, Gy., KABOS, S. and ZILAHY, P., 1985. Potential uses of the Hungarian Soil Information System. 3rd GATE-UMASS Symposium on "Computer-Assisted Landuse/Land-Cover Mapping, June 24-26, Gödöllő, Hungary. 50-63.
VÁRALLYAY, Gy. and RAJKAI, K., 1987. Soil moisture content and moisture potential measuring techniques. International Conference on Soil Moisture Measurements, July 6-9, Logan, USA.
WEBSTER, R., 1985. Quantitative spatial analysis of soil in the field. Advances in Soil Science. 3. 1-70.