1410

GEOSTATISTICAL ANALYSIS OF A SOIL SALINITY DATA SET

G. Bourgault,[1] A. G. Journel,[1] J. D. Rhoades,[2] D. L. Corwin,[2] and S. M. Lesch[2]

[1]Geological and Environmental Sciences Department, Stanford University, Stanford, California 94305

[2]USDA-ARS, U.S. Salinity Laboratory, 450 Big Springs Road, Riverside, California 92507

I. INTRODUCTION

Rather than relating a specific case study with specific goals, this study aims at presenting a range of potential applications of modem geostatistics to soil survey problems using a real data set. Typical of the development of geostatistics, a discipline led by engineers, many new algorithms, although well published in

Advances in *Agronomy, Volume 58*

their field of inception (mining and petroleum), have not yet found their way into mainline statistical books; hence, they may not be readily accessible to professionals outside the extractive industry (Deutsch and Joumel, 1992; Dimitrakopoulos, 1993; Isaaks and Srivastava, 1989; Soares, 1993). It is hoped that this presentation may raise enough interest among soil scientists so that they find it worth their time to learn more about modem geostatistical concepts of data analysis, estimation, uncertainty assessment, and stochastic imaging. All results presented in this study were obtained using standard mapping routines and the public-domain GSLIB software (Deutsch and Joumel, 1992). FORTRAN source code of the latter is public domain; hence, it is available to whomever wishes to understand the details of any particular algorithm and/or to modify it to fit any particular problem at hand.

A common denominator of many soil sciences data sets is the sparsity of "hard" or direct measurements of the primary variable of interest, usually balanced by the prevalence of "soft" or indirect information related to the primary variable. Examples of hard data are core measurements and more generally expensive field-based data as opposed to soft data obtained, e.g., from remote sensors. Geographical information systems (GIS) and geostatistics pursue a similar objective-that of providing tools for the integration of different objectives, and that of providing tools for the integration of different information sources with varying relevance/reliability to build maps that summarize and expand the original hard data set. Geostatistics propose to add to the GIS toolbox various spatial data analysis tools to explore and model patterns of space/time dependence between the data available. The resulting numerical models, e.g., variograms or conditional distributions, can then be put to use for various mapping purposes and an assessment of the reliability of such maps.

Just as there is no unique or optimal sequence in using GIS tools, such as concatenation, intersection, or interpolation, there is no unique geostatistical approach to spatially distributed data. Many alternative covariance/variogram models can be fitted to the same data set depending on ancillary information available to the operator (including his or her own prior experience); there are many different kriging algorithms (generalized leastsquares regression) that can be used toward the same mapping goal depending on which particular aspect of the data one wishes to capitalize on. What may be lost to statistical objectivity (an extremely debatable concept) is gained in flexibility and ability to handle soft, yet essential information of various types. It is better to have a somewhat subjective but accurate assessment that accounts for all relevant information than a supposedly objective assessment that misses critical aspects of the problem. This chapter will illustrate the toolbox aspect of geostatistics, presenting several alternative ways to reach the same goal and proposing cross-validation exercises to help the operator in his or her decision.

A. BROADVIEW SALINITY DATA SET

As mentioned previously, the data set used in the following study is more a support for demonstration of geostatistical algorithms than the data base of an actual case study. For such demonstrative purposes, the name of the location involved and even the measurement units of the data could have been omitted, and coordinate values could have been changed by any one-to-one monotonic transform leaving unchanged the relative patterns of spatial variability of the various attributes.

The actual and complete Broadview salinity data base and its statistical analysis are presented in various papers and reports of the U.S. Salinity Laboratory (Lesch et al., 1995a,b). The reader is referred to these papers for any question related to sampling and salinity assessment in the Broadview water district. The results of the present study are based on a limited data set ignoring, in particular, such critical variable as soil water saturation: they should only be used to assess the worth of adding geostatistical tools to GIS and other toolboxes available to the soil scientist.

The Broadview data set covers approximately 6000 acres and comprises the following:

- a soil map digitized into 7 soil types (see Fig. la).
- 315 soil core measurements of electrical conductivity (EC,), taken at four depths (O-l, 1-2, 2-3, and 3-4 ft) (see Fig. lb). Unit is *dS/m.*
- *2385* measurements of soil vertical (EM,) and horizontal (EM,) electromagnetic response (see Fig. lc). Unit is *dS/m.* Each measurement is deemed representative of the vadose zone (upper 4 ft of soil). The extent of that electromagnetic information delineates the study area, as shown in Figs. la-1c.

For the purpose of this study, core *EC,* values are considered hard data directly related to soil salinity and represent the primary variable to be evaluated throughout the vadose zone. The electromagnetic induction readings represent a secondary variable less directly related to soil salinity; they are considered soft data used to complement the hard *EC,* data. With little loss of location accuracy, electromagnetic data have been relocated to the nodes of a 2D regular grid 100 X 100 *m.* The few grid nodes with no electromagnetic sample within a radius of 50 m were left uninformed. Another option could have been to interpolate the few missing nodal values. The grid includes 2385 *EM,* measurements. To allow standardization of the grayscales, the *EM,* data plotted in Fig. lc have been rescaled by a factor equal to the ratio of the standard deviations of original *EC,* and *EM,* data.

Although measured in the same unit *(dS/m), EC,* and *EM,* data have different

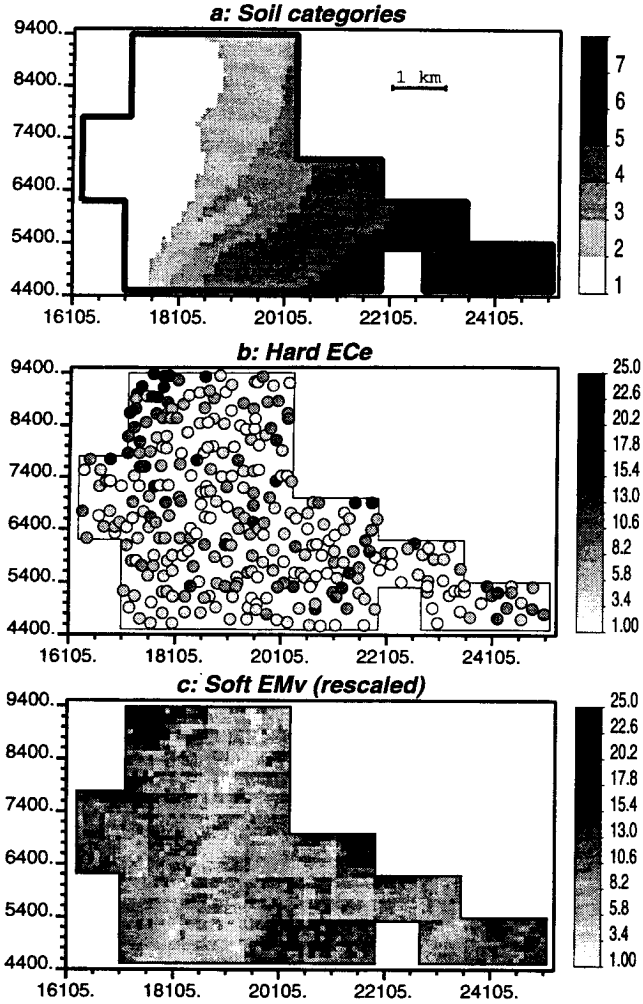

Figure 1 Broadview salinity data sets. (a) Seven soil types, (b) 315 core measurements of electrical conductivity (hard EC_e data), and (c) gridded electromagnetic data (soft EM_V data). To allow easier comparison of spatial structures, the EM_V variance has been rescaled to that of the hard EC_e data.

mean values (5.24 and 1.46, respectively). Such differences in mean values between hard and soft data are not uncommon in the earth sciences and are filtered out by various unbiasedness constraints in the algorithms used.

The soil map in Fig. 1a was digitized to allow marking each EC_e and EM_V sample location with a specific soil type denoted 1–7.

II. EXPLORATORY DATA ANALYSIS

The aim of an exploratory data analysis (EDA) is to acquire an overall familiarity with the data, their interrelations, statistical grouping, spatial distribution, clustering, etc. At this stage, the operator should not be constrained by any specific goal but rather he or she should be attentive to any clue the data may give that may prove useful in later interpretations. Because geostatistics deals with spatial data, extensive use should be made of isopleth maps and GIS-related routines depicting the relations between data values and their space/time coordinates. Beware that random sampling (random drawing of sample coordinates) does not make the data values independent inasmuch as it is the physical generating process that makes the data dependent and not the human decision about where samples are taken.

A. ELECTRICAL CONDUCTIVITY DATA

Figure 2 shows the succession of four grayscale ***EC,*** maps corresponding to the four measurement depths. The vertically averaged map is that shown in Fig. lb. There appears to be a gradual increase in soil salinity with depth, corroborated by the histograms shown on the right in Fig. 2.

A diagonal transect $N120°E$ crosscutting the $N30°E$ elongation of the seven soil categories shown in Fig. la was defined, then ***EC,*** data values were plotted against their coordinate value along that transect (see Fig. 3). At each depth level, the $n = 315$ ***EC,*** data were ranked from $r^{(1)} = 1$ to $r^{(n)} = 315$ and their standardized ranks $v^{(i)} = r^{(i)}/n$, or uniform scores distributed in [0,l], are grayscale plotted in Fig. 3. This uniform score transform allows identifying each level-specific ***EC,*** data set to the same uniform [O,l] distribution. This transform thus filters out the vertical trend previously observed and allows comparison of the strictly horizontal structures. The four $N120°E$ grayscale transects of uniform scores shown on the left in Fig. 3 show similarity of the horizontal variability of ***EC,*** data over the four depth levels. This is confirmed by the ***EC,*** uniform score (uscore) semivariograms calculated along the $N120°E$ direction and given on the right in Fig. 3. Therefore, these uscores-standardized variograms can be pooled together into a single model valid for all four depth levels.

The rank (or uscores) correlations between two vertically consecutive EC, data (thus with the same horizontal coordinates) are 0.64 for l-2 ft, 0.80 for 2-3 ft, and 0.89 for 3-4 ft. Therefore, except for the first transition from 1 to 2 ft, the ***EC,*** data are quite redundant from one level to the next one: there is little gain to be expected from a 3D interpolation versus a much simpler 2D exercise using only data from the level being estimated.

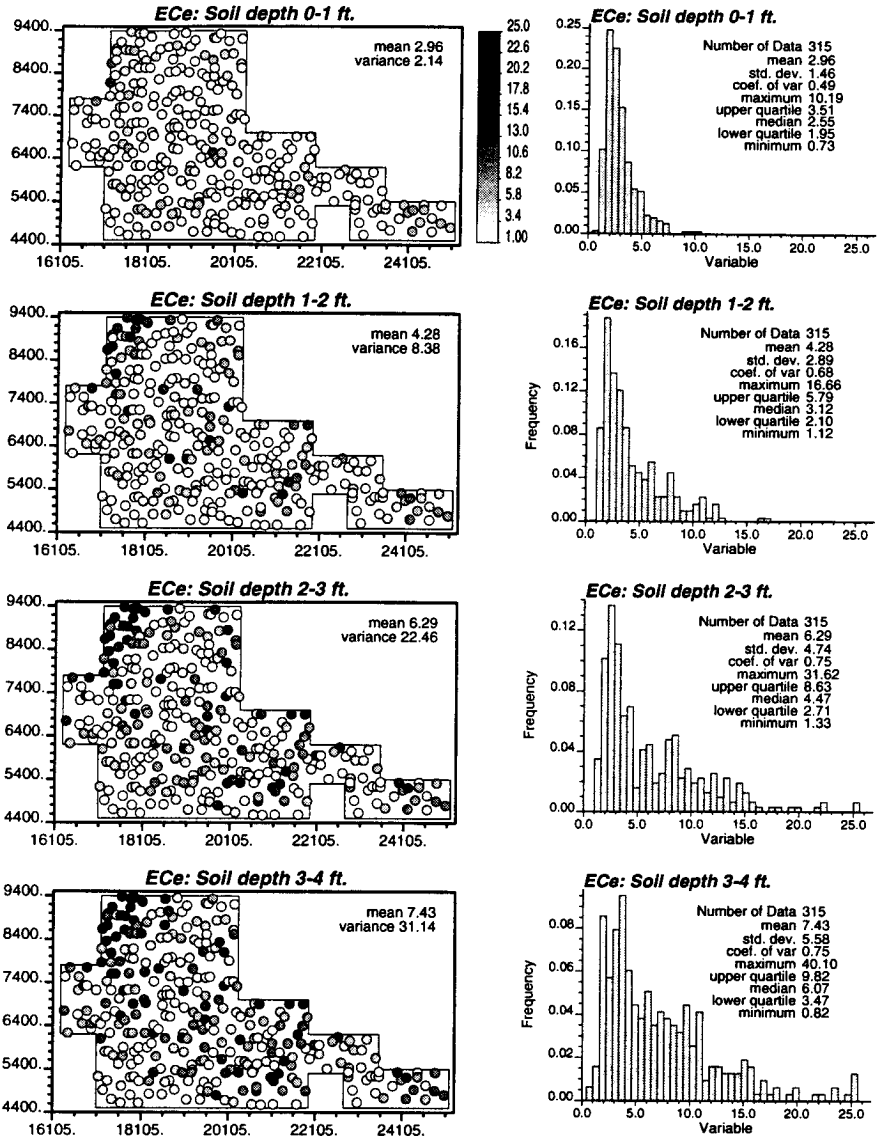

Figure 2 Grayscale plots of EC_e data and corresponding histograms showing increase with depth of soil salinity.

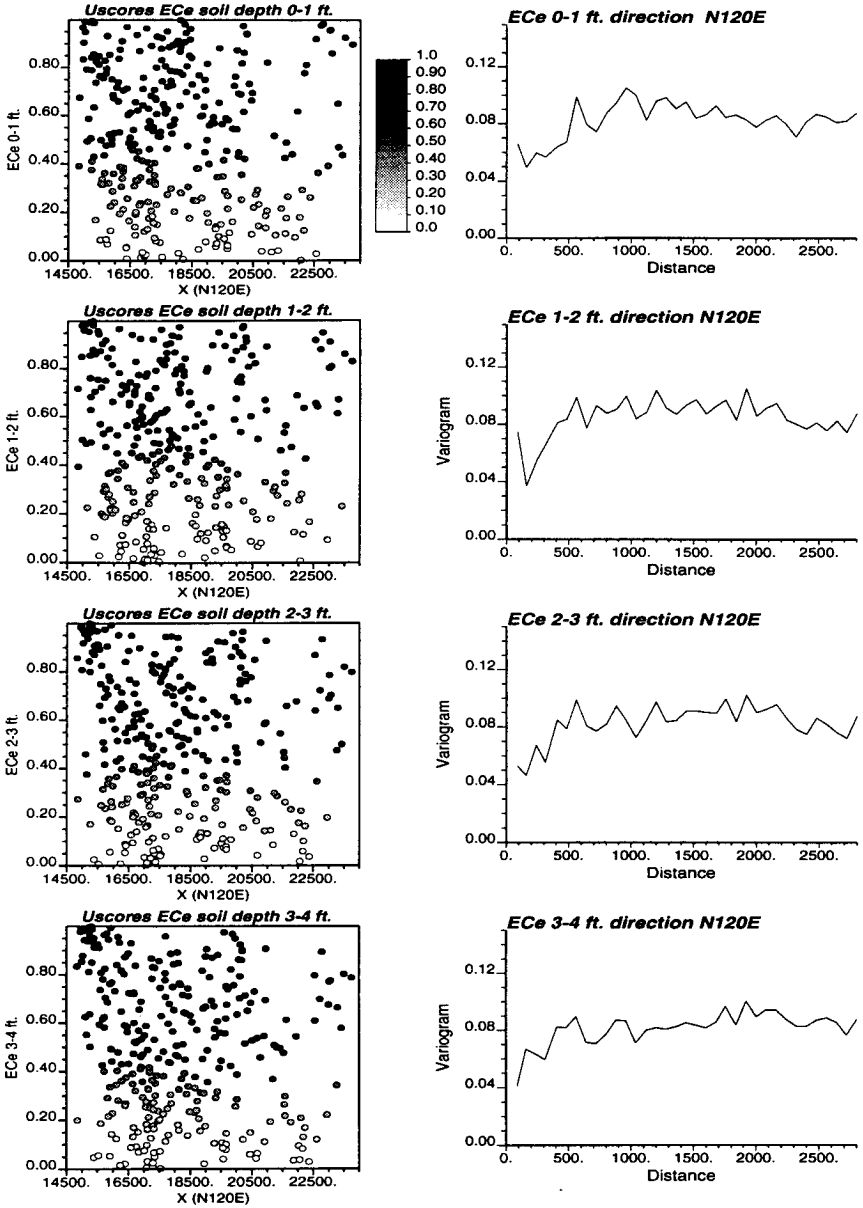

Figure 3 Grayscale plots of standardized ranks (uniform scores) of level-specific EC_e data projected along a $N120°E$ transect. The corresponding $N120°E$ semivariograms are given to the right.

For the remainder of that study and for reason of conciseness, only the vertically averaged ***EC,*** data (see Fig. 1 b) were considered together with the corresponding 2D-distributed electromagnetic data (see Fig. lc).

Note that the uniform score transform $x \rightarrow F_X(x)$, where $F_X(\cdot)$ is the cumulative distribution function (cdf) of random variable X, is the first step of a normal score transform (Deutsch and Joumel, 1992, p. 138). Unless properties specific to the Gaussian distribution are to be called for, there is no need for going beyond the standardized rank transform $F_X(\cdot)$. This rank transform, by definition, preserves the rank of the data as does the commonly used, albeit somewhat arbitrary, log transform. From the histograms of Fig. 2, the ***EC,*** data appear neither normal nor lognormal distributed; this was confirmed by probability graph plots (Deutsch and Joumel, 1992, p. 201) not shown here.

B. Electromagnetic Data

Figure 4a shows an extreme redundancy between the two secondary data, vertical ***(EM,)*** and horizontal ***(EM,)*** electromagnetic measurements. This redundancy was confirmed by maps and variograms analysis (not shown here). Because ***EM,*** has slightly better correlation with colocated vertically averaged ***EC,*** ***(see*** Figs. 4b and 4c), only ***EM,*** was retained as a source of secondary data for the rest of the study. The grayscale map of this ***EM,*** data was shown in Fig. lc.

Observe on Fig. 4b the nonlinear relation ***EC,-EM,*** To linearize that relation and capitalize on linear regression tools (such as kriging), a transform of the variables is necessary. If the two variables were to be made Gaussian distributed, a normal score transform (Isaaks and Srivastava, 1989, p. 138) would be necessary. Because the histograms of the original ***EC,*** and ***EM,*** values are not lognormal, the log transform does not identify the normal score transform. In any case, there is currently no need for any Gaussian assumption; hence, the rank transform (uniform scores) is enough.

Figure 4d shows the scattergram of the uniform score transforms of ***EC,*** and ***EM,*** data. Note how the rank transform has succeeded in linearizing the original regression between ***EC,*** and ***EM,*** data (see Fig. 4b). The linear rank regression remains though heteroscedastic, in the sense that higher ranks of ***EC,*** are better predicted by corresponding high ***EM,*** ranks than are lower ranks. These rank regressions will be fine-tuned later using soil type information.

C. Soil Type Differentiation

The 3 15 vertically averaged ***EC,*** data values were plotted against their coordinates along the $N120°E$ transect and grayscale coded for soil type (see Fig. 5a). Figure 5b provides a similar profile for the 2385 ***EM,*** data.

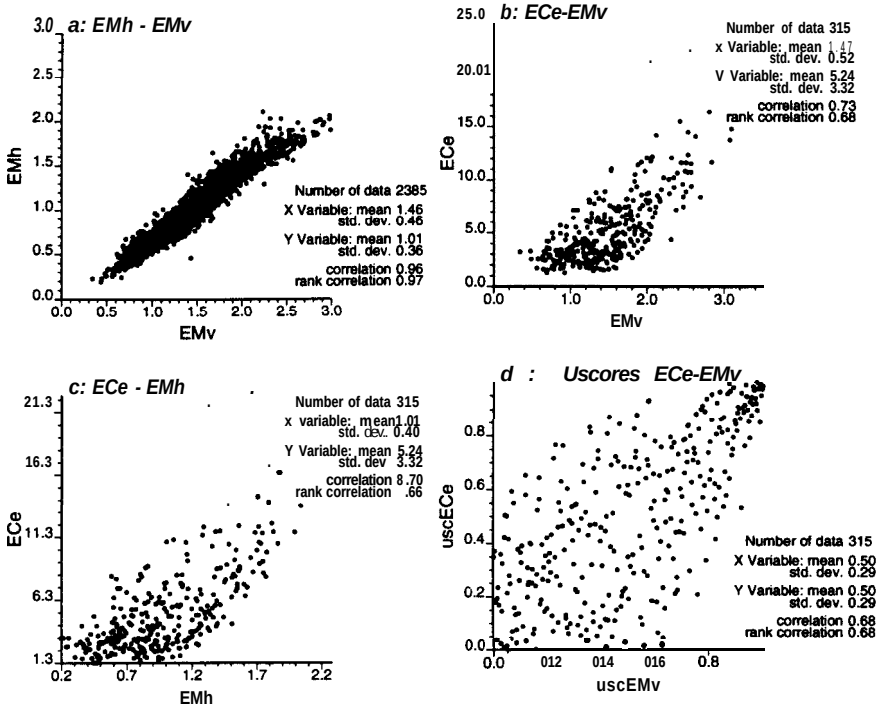

Figure 4 Scattergrams. (a) Soft EM, versus soft EM_V data; (b) hard EC, versus soft ***EM,*** data; (c) hard ***EC,*** versus soft EM_H data; and (d) uniform scores of ECe versus ***EMv*** data.

Except for soil types 1 and 6 (the latter being nonrepresentative because of lack of data), the ranges of ***EC,*** values appear homogeneous across soil types. This is confirmed by the ***EC,*** histograms per soil type (not shown).

The histograms of EM_V data per soil type (Fig. 6) would lead one to differentiate the following two groups based on mean ***EM,*** value:

- A first group including soil types 1, 4, 5, and 7 with a mean ***EM,*** value around 1.6
- A second group including soil types 2, 3, and 6 with a lower mean ***EM,*** value around 1.2

Note that these two groups are intermingled in space.

EM, data considered to be exhaustively sampled, are used to inform unsampled primary ***EC,*** values. To investigate how soil type influences the relation, ***EC,-EM,*** the seven soil-specific rank scattergrams of uscores of colocated ***EC,*** and ***EM,*** data are shown in Fig. 7. It appears that the linear rank regression

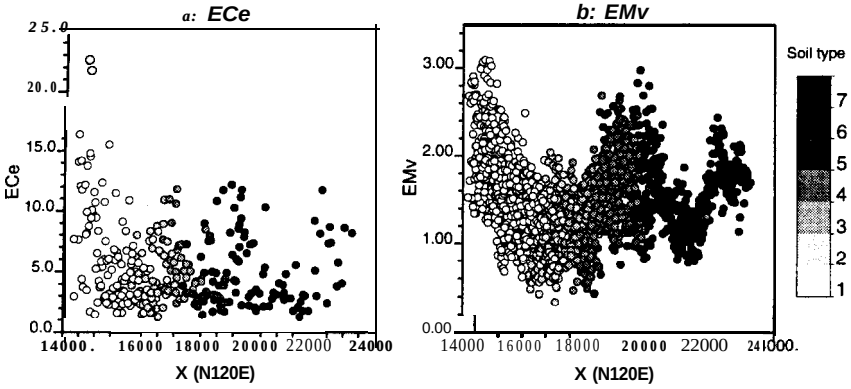

Figure 5 Profiles of EC, and EM, data along a $N120°E$ transect. The grayscale code indicates soil type.

observed when data are pooled across all soil types (Fig. 4d) is in fact constituted by several different regressions better fitted by power models of the type,

$$v_{EC_e} = [v_{EM_v}]^{\omega_{st}} \quad (1)$$

where v_{Ec} and v_{EM} are the uniform scores valued in [0,1], and ω_{st} is a power exponent specific to each soil type st.

The regression fits shown in Fig. 7 correspond to

$\omega_{st} = 1$ (linear rank regression) for soil types overall and type 6.
$\omega_{st} > 1$ for soil types 1, 4, 5, and 7.
$\omega_{st} < 1$ soil types 2 and 3.

These power regression models are used later to incorporate the secondary ***EM,*** information while accounting for soil type.

Were the electromagnetic data ***EM,*** a variable of primary interest, further considerations would be given to splitting its spatial distribution into soil type groups. However, because EM, represents only secondary information destined to supplement the hard ***EC,*** data, it was decided to model its spatial distribution across all soil types. The only soil type differentiation kept is that of the previous rank regression power models.

D. SPATIAL AND VARIOGRAM ANALYSIS

In preparation for spatial interpolation of the ***EC,*** values, joint variogram analysis of the EC,-EM, data was performed across all soil types.

Figure 8 shows the sample (cross)semivariograms for the original ***EC,*** and

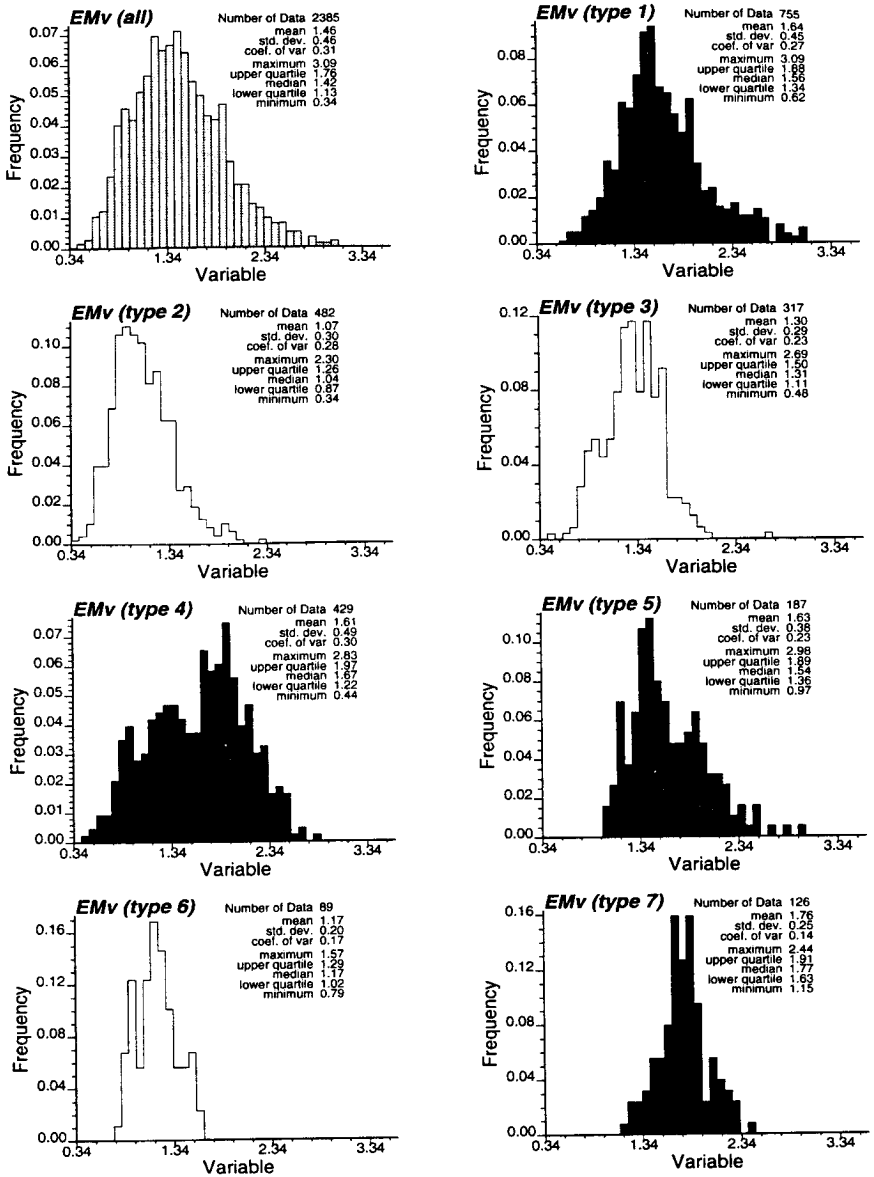

Figure 6 Histograms of EM_V data across all soil types and per soil type. The white histograms correspond to soil type groups 2, 3, and 6 with lower mean, the dark histograms correspond to groups 1, 4, 5, and 7 with higher mean.

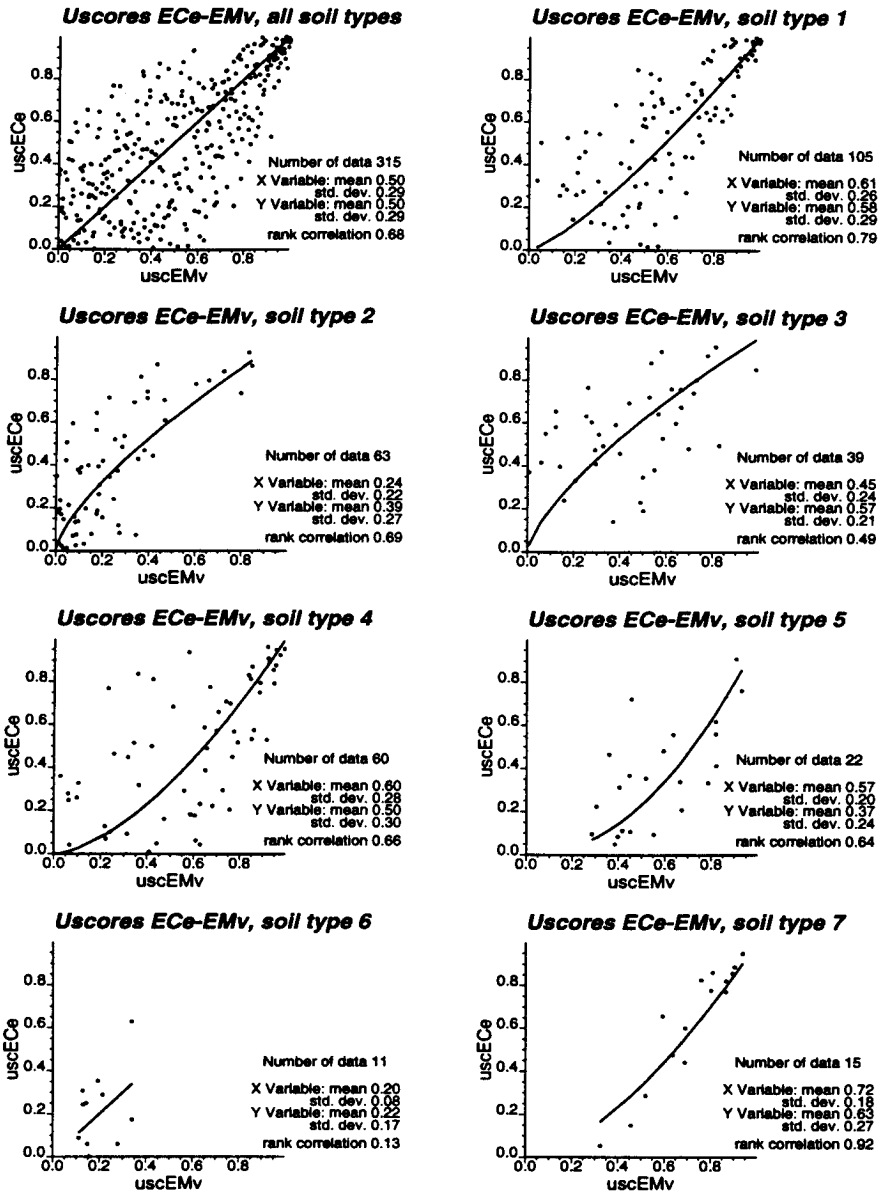

Figure 7 Rank scattergrams of EC_e–EM_v data pooled across all soil types and per soil type. The regression curve shown is fitted using a power model.

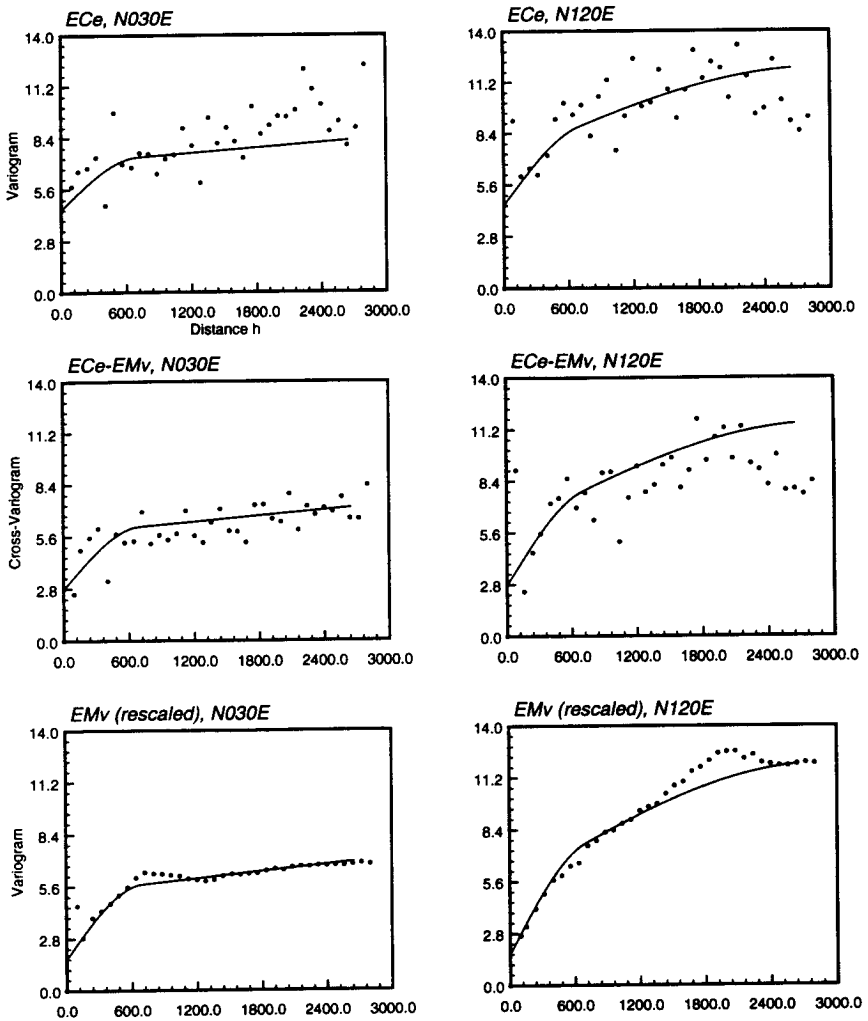

Figure 8 Sample (cross)semivariograms of EC_e–EM_V data and the model of coregionalization fitted.

EM_V data in the two main directions, $N30°E$ along soil types elongation and $N120°E$ across it. The EC_e and EM_V data used are those given in Figs. 1b and 1c. For reason of stability of future cokriging matrices, the secondary EM_V data have been rescaled by a factor allowing identification of their variance to that of the primary data EC_e; equivalently, one could have worked on (cross)correlograms instead of variograms.

The solid lines in Fig. 8 show the fit by a model of coregionalization (Joumel and Huijbregts, 1978, p. 172; Isaaks and Srivastava, 1989, p. 390). That model features

- An isotropic nugget effect accounting for about one-third of the total spatial variance of the ***EC,*** data
- A first isotropic structure of range 700 m accounting for another third of the ***EC,*** variance
- A second anisotropic structure of range 3000 m in the $N120°E$ direction across soil types and 16,000 m in the $N30°E$ direction along soil continuity

The model for ***EM,*** is similar although with lesser nugget effect due to the larger definition volume (averaging effect) of the electromagnetic data.

The original ***EC,*** and ***EM,*** data were then normal score transformed (Deutsch and Joumel, 1992, p. 138) so that both histograms identify a standard Gaussian distribution, and the corresponding sample (cross)semivariograms were calculated and modeled (see Fig. 9). The coregionalization model features the same characteristics as those fitted to the original data. Note that sampling fluctuations have not been significantly reduced by the normal score transform; this would have also been true had a log-transform been used.

The uniform scores of the EM, and ***EC,*** data used are shown in Figs. 10a and l0b. Compare these scores to the data in Fig;. lc and lb, respectively: except for the different grayscales, they are essentially the same. Again, we prefer comparing data through the uniform standardization in [0,l] provided by the standardized ranks (uniform scores). Figure 10c shows the location of 26 *EC,* random samples taken from the 3 15 original ***EC,*** data; this subsample is used later in the cross-validation exercises.

III. MAPPING THE EC, DISTRIBUTION

To demonstrate the various geostatistical mapping algorithms, four different approaches and two sampling cases are considered. ***EC,*** estimation is performed at each node of the 100 x 100-m grid covering the study area as defined by the template of electromagnetic data (see Fig. lc).

The three different approaches are

1. Simple kriging (SK) (Deutsch and Joumel, 1992, p. 62): ***EC,*** is estimated by a linear combination of the neighboring ***EC,*** data plus the overall ***EC,*** sample mean, $m = 5.24$. No secondary information is used; thus, this approach represents a base case.
2. Simple cokriging (coSK) (Deutsch and Joumel, 1992, p. 71): ***EC,*** is esti-

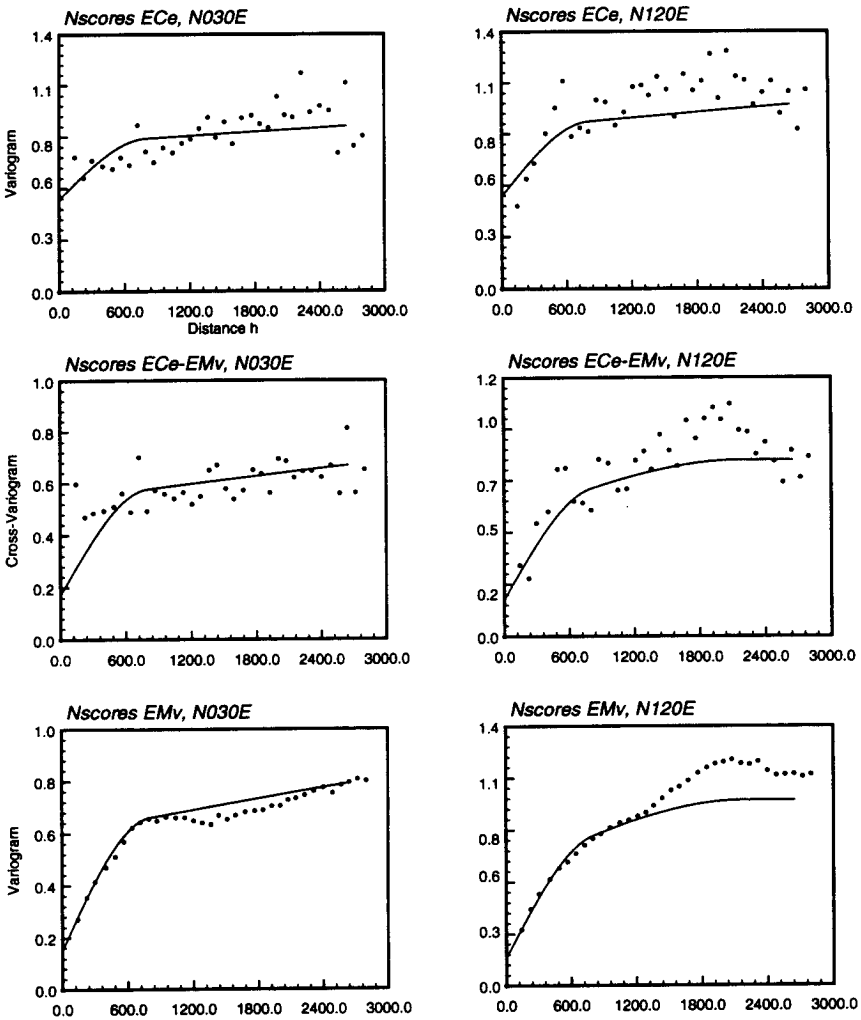

Figure 9 Sample (cross)semivariograms of normal score EC_e–EM_V data and the model of coregionalization fitted.

mated by a linear combination of neighboring EC_e data and EM_V data rescaled to the EC_e variance, $\sigma^2 = 11.0$. The coregionalization model shown in Fig. 8 is used. This is the most straightforward cokriging; it does not correct for the nonlinear relation observed between EC_e and EM_V data (see Fig. 4b).

3. Probability-field estimate (P-field): This approach is a variant of the p-field simulation algorithm introduced by Srivastava (1992); it accounts for the soil

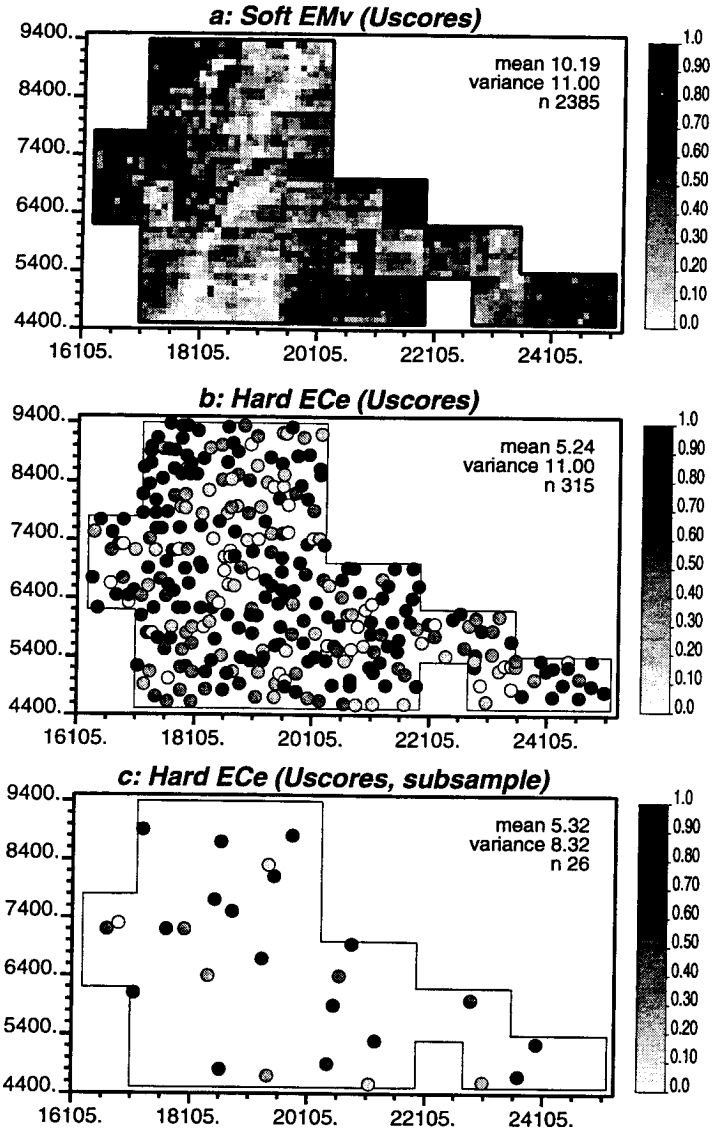

Figure 10 Grayscale maps of uniforms scores of EM_V and EC_e data.

type information available at each location being estimated. A simple kriging of the EC_e normal score is performed using only EC_e normal score data. Adopting a Gaussian random function model for these normal scores (Deutsch and Journel, 1992, p. 136), the conditional distribution of any unsampled EC_e normal score

value is Gaussian with mean and variance identified to the simple kriging mean and variance. The Gaussian conditional cumulative distribution function (cdf) can be denoted by

$$\text{Prob}\ \{Y(u) \leq y \mid \text{neighboring } y(\mathbf{u}_\alpha) \text{ data}\} = G\left(\frac{Y - y^*_{SK}(\mathbf{u})}{\sigma_{SK}(\mathbf{u})}\right) \qquad (2)$$

where Y(u) is the normal score transform of ***EC,*** at grid location u, $y^*_{SK}(\mathbf{u})$ and $\sigma^2_{SK}(\mathbf{u})$ are the simple kriging estimate and variance using neighboring normal score ***EC,*** data ($y(\mathbf{u}_\alpha)$); and the function $G(\cdot)$ is the standard normal cdf.

Let q(u;p) be the corresponding conditional quantile function or inverse of the previous conditional ***cdf:***

$$q(u; p) \text{ such that } G\left(\frac{q(\mathbf{u}; p) - y^*_{SK}(\mathbf{u})}{\sigma_{SK}(\mathbf{u})}\right) = \quad p \in [0, 1]. \qquad (3)$$

Srivastava's p-field approach (Srivastava, 1992) consists of simulating a p-field, that is, a set of spatially correlated uniformly distributed p,(u) values, then transforming them through the previous quantile function into simulated normal score values $y_s(\mathbf{u})$ for ***EC,:***

$$y_s(\mathbf{u}) = q(\mathbf{u}; p_s(\mathbf{u})) \qquad (4)$$

Because there may be several realizations (outcomes) for $p_s(\mathbf{u})$ at any location u, there may be several "simulated" realizations y,(u), hence the subscript notation *s* for simulation.

The variant proposed here consists of determining, at each location u, a single value p(u) resulting in a single estimate y*(u) for the ***EC,*** normal score value:

$$y^*(\mathbf{u}) = q(\mathbf{u}; p(\mathbf{u})) \qquad (5)$$

with $p(u) = [v_{EM_v}(\mathbf{u})]^{\omega_{st}}$, as given by the power model in Eq. [I], and *st* is the soil type prevailing at location **u.**

In words, the p-field value to be plugged into the conditional quantile function q(u;p) is the p value obtained by the regression model [Eq. (1) and Fig. (7)] specific to the soil type prevailing at u and to the uniform score $v_{EM_v}(\mathbf{u})$ of the electromagnetic datum at u.

This p-field approach requires the Gaussian random function model to determine the conditional cdf Eq. (2) from the only two parameters, mean and variance, provided by simple kriging.

A final step back-transforms the normal score estimate y*(u) into ***EC,*** estimates expressed in the original ***EC,*** units.

Note: The three different approaches proposed here to interpolate ***EC,*** values do not cover the range of different geostatistical algorithms that could be used for

this purpose. The first two approaches proposed are the most straightforward and are likely to be familiar to many readers. The latter approach is a bit more involved; it is intended to give the reader a glimpse of the forefront of applied geostatistics in which new variants are constantly proposed to better match the problem at hand and the specific data available.

For each of the previous three approaches, two sampling cases are considered:

1. All 315 hard ***EC,*** data are used together with the (exhaustive) EM_V and soil type information present at all nodes being estimated
2. A subsample of only 26 hard ***EC,*** data is used (see Fig. 10c) in addition to the previous ***EM,*** and soil type information

This latter sampling case allows a model-validation exercise using the remainder 289 hard EC, data. The problem with using such a small sample size (26) is the difficulty of doing any reliable statistical inference. We have decided to set apart the two problems of statistical inference and model validation of the estimation approaches proposed. More precisely, for the latter sampling case, although only 26 ***EC,*** data were retained for the various krigings, the statistics needed (histogram, variograms, and regression) are those established using all 3 15 hard data, i.e., the same statistics used for the first sampling case. This decision corresponds to the extractive industry practice of borrowing statistics from a similar and better sampled field but using only field-specific data for local estimation.

The objective of this specific model validation exercise is twofold:

1. Observe the performance of each model or algorithm under data sparsity
2. Evaluate the worth of the secondary information ***(EM,)*** under the same conditions of data sparsity

Three approaches times two sampling cases result in six sets of results. Each set of results given hereafter includes

- An estimated ***EC,*** map in the original ***EC',*** unit.
- The corresponding estimated ***EC,*** uniform score map, unit free and valued in **[0,** 1]. These uniform scores are the standardized ranks of the previous ***EC,*** estimates. Again, this standardization allows a visual comparison that is unit free and free of color or grayscale effect.
- The semivariograms of the ***EC,*** estimated values plotted against the model fitted to the (3 15) sample semivariograms. That model is the one depicted by the continuous curves in the two top graphs in Fig. 8. This comparison allows for the evaluation of the smoothing effect (Deutsch and Joumel, 1992, pp. 17, 61) of the estimation algorithm considered.
- For the three sets of results corresponding to the second sampling case, the

Table 1

Summary of Results[a]

	Full sample (315)		Crow. sample (289)	
	In	σ^2	m	σ^2
Reference	5.24	11.00	5.30	11.02
Simple kriging	5.15	4.54	5.34 (ρ = 0.21)	0.49
Simple cokriging	5.21	8.29	5.43 (p = 0.76)	8.45
P-field	5.31	13.32	5.57 (p = 0.75)	13.03

[a] The first two columns give the mean and variance of EC, estimates to be compared to the reference EC, sample used (size 315). The third and fourth columns give the mean and variance of 289 reestimated EC, values and their linear correlation with the 289 actual values. For the latter, the EC, subsample size is 26.

cross-validation scattergram of the 289 "true" ***EC,*** values versus the corresponding estimated values

Table I provides a summary of the major results.

A. COMMENTS ON RESULTS

1. Simple Kriging

The results of the base case, simple kriging using only the full sample of hard ***EC,*** data (3 15), are shown in Fig. 11. The grayscale map of the ***EC,*** estimates (top map) reveals a severe smoothing effect: the variance of the estimates is only 4.54 versus the 315 hard data variance of 11.00. This smoothing effect is a well-known shortcoming of all linear weighted average-type estimators including kriging (Joumel and Huijbregts, 1978, p. 450): typically, the distribution of estimates understates the actual proportions of extreme values, whether high or low values. If detection of spatial patterns of extreme values is the goal of the study, then kriging is not an appropriate mapping algorithm (Joumel and Alabert, 1988). Instead, one should consider one of the stochastic imaging algorithms, also known as conditional simulations (Deutsch and Joumel, 1992, p. 117), which aim to reproduce the patterns of spatial variability seen from the sample and modeled through the variogram. Conditional simulations are the topic of the latter part of this chapter (see Section VII).

The uniform score transform (middle map in Fig. 11) filters the effect of smoothing on the global variance and reveals $N30°E$ structures clearly associated

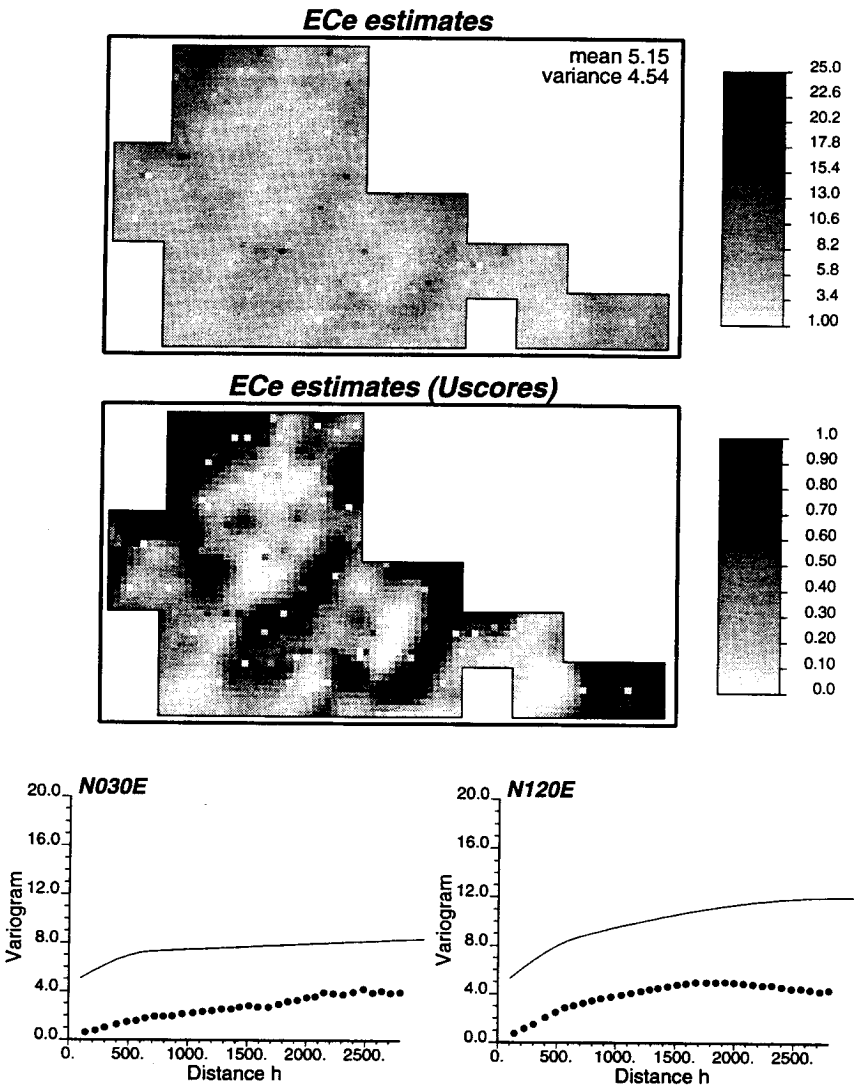

Figure 11 Simple kriging results; full EC_e sample (315). The semivariograms (dots) of the estimated values versus the model fitted from the full sample (continuous curve). Note the severe smoothing effect (lesser variance of the estimates).

to the soil type direction of continuity. These structures are somewhat similar to those seen on the rescaled EM_V (soft data) map in Fig. 1c.

The lower graphs in Fig. 11 confirm the variance deficiency of the simple

kriging estimates. Note that the shape of the variogram models is reasonably well reproduced.

The smoothing effect of kriging is particularly dramatic when only 26 EC_e data are available (see Fig. 12): the variance drops to 0.49 versus 11.0 for the original

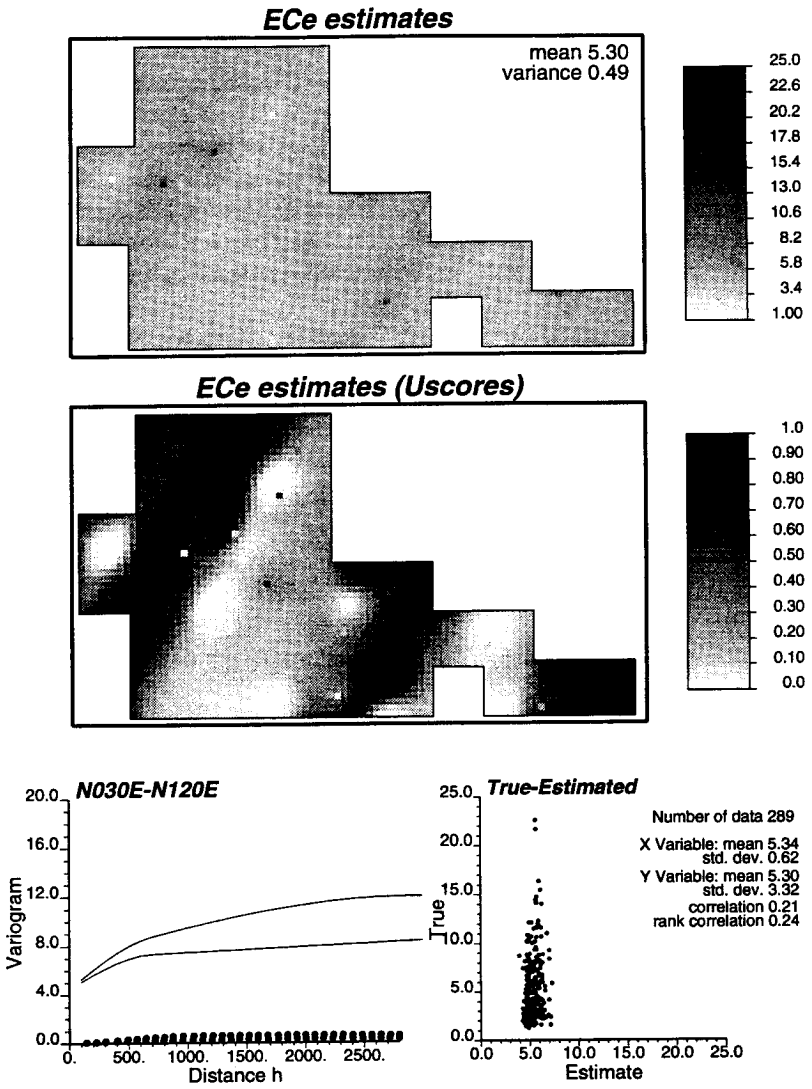

Figure 12 Simple kriging cross-validation results; EC_e subsample (26). The lower right graph gives the scattergram of actual EC_e values versus the reestimated values.

sample data. Although globally unbiased, the distribution of the 289 reestimated values fails badly in reflecting the actual proportions of nonmedian EC_v data values outside the interval [4.0, 8.0] (see scattergram at the lower right in Fig. 12). This is known, in geostatistical jargon, as conditional bias (Joumel and Huijbregts, 1978, p. 458).

2. Simple Cokriging

When using the densely sampled secondary (EM_v) information, the smoothing of the EC_v estimates is partially corrected to 8.29, a value still less than the original sample variance of 11 .OO. The variograms of the cokriging estimates approach those of the model much better. The spatial structures of the cokriging estimates closely reproduce those of the EM_v data (compare the uscores maps of Figs. 13 and 10a); this is as expected given the large EM_v sample size and the relatively strong correlation (0.73) between colocated EC_v and EM_v data (see Fig. 4b).

The contribution of the dense EM_v secondary information is more dramatic when only 26 hard EC_v data are available (compare Figs. 12 and 14). The scattergram of true versus reestimated EC_v values (lower right graph in Fig. 14) indicates a substantial correction of the smoothing effect and related conditional bias of the simple kriging estimates. In this cross-validation exercise, cokriging (using the secondary information) has raised the true-versus-estimate correlation from a low 0.21 to a reasonable 0.76.

3. P-Field Estimates

In addition to the secondary EM_v data, the p-field approach implemented here accounts for the soil type information.

From the results of Fig. 15, it appears that the smoothing effect seen on the simple kriging and cokriging estimates in Figs. 11 and 13 has been overcorrected. The variance of the p-field estimates, 13.30, is now larger than that of the original EC_v sample values, 11 .00; the overcorrection takes place in the $N120°E$ direction across soil continuity (see the lower right variograms in Fig. 15). The soil type information appears to have imposed too much of the soil discontinuity along that direction.

The cross-validation results of Fig. 16 confirm the correction of the smoothing effect: the variogram model is well reproduced in both $N30°E$ and $N120°E$ directions. The correlation true-versus-reestimated values is not significantly improved from the results of cokriging (Fig. 14). Note that the two dots departing most from the 45" line of the scattergram in Fig. 16 are the same as those in Figs. 12 and 14: no estimation algorithm can improve the estimation of outlier values, i.e., values departing significantly from the statistics of the data used. At best,

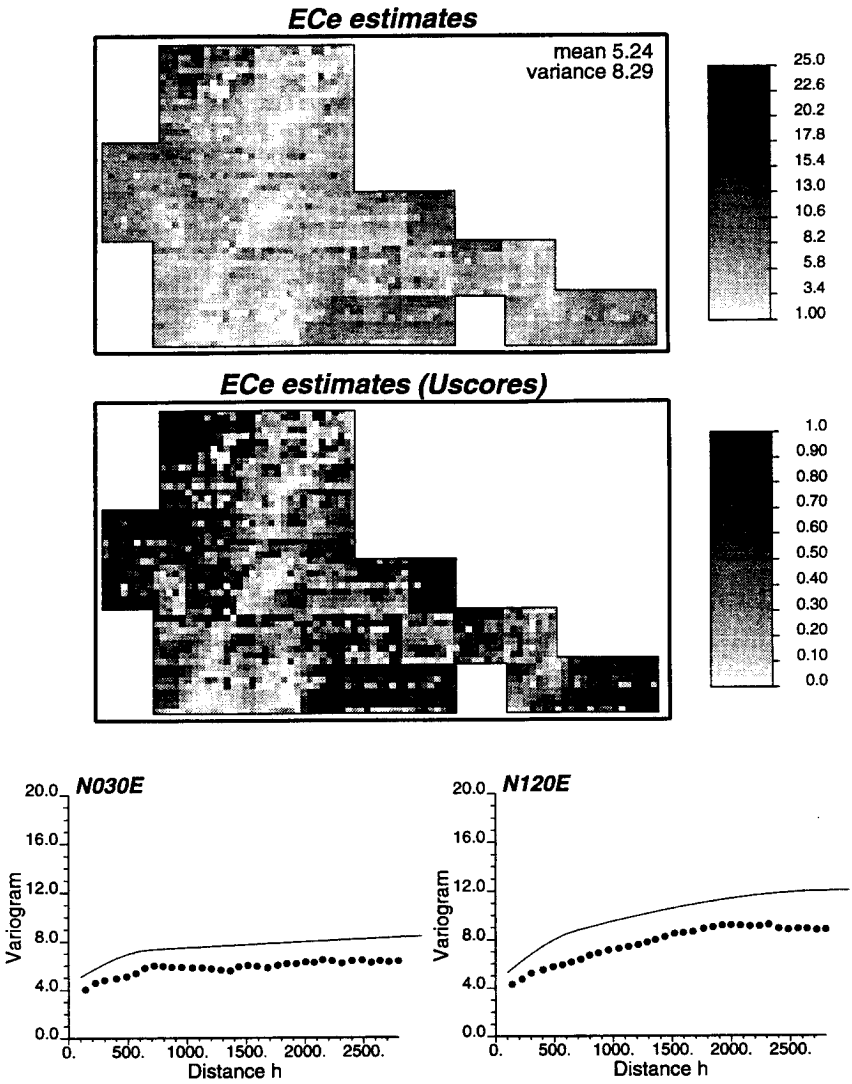

Figure 13 Simple cokriging results; full EC_e sample (315).

the occurrence of such outlier values can be simulated, as is done in stochastic simulation.

In summary, from the results in Figs. 11–16, it appears that

1. Accounting for a dense secondary information, such as EM_V data, does improve the resolution of the estimated EC_e map more so if the primary

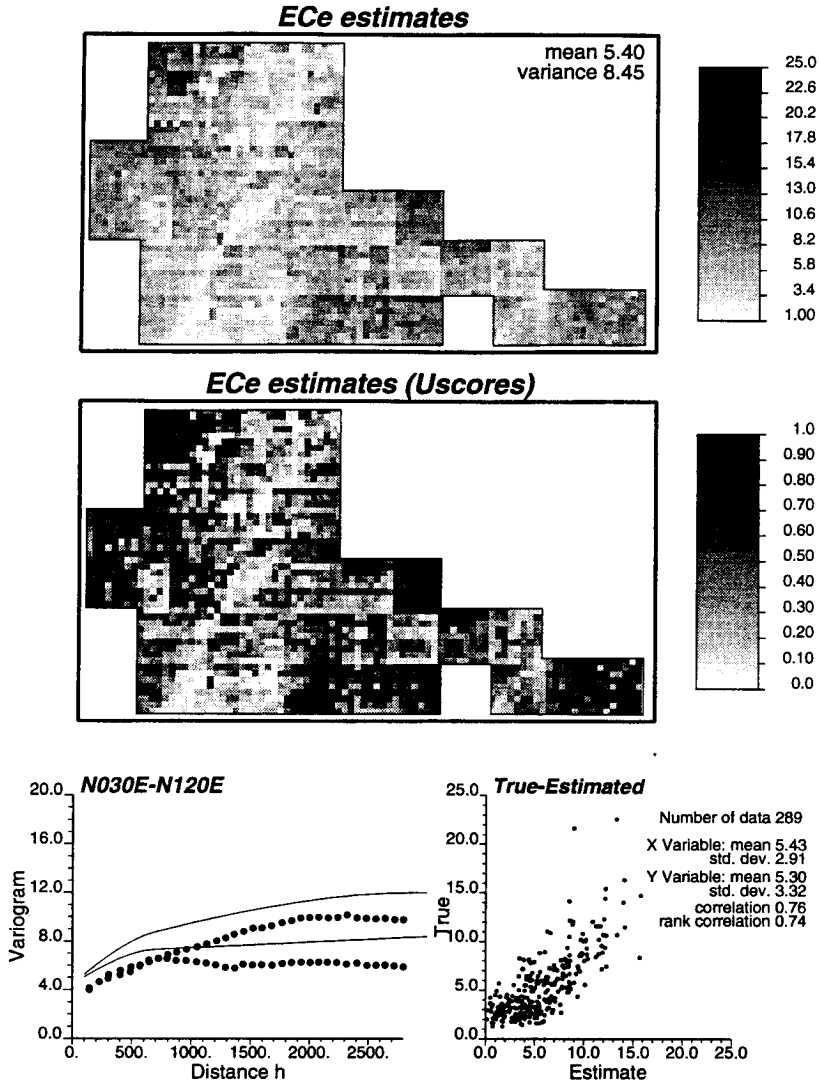

Figure 14 Simple cokriging cross-validation results; EC_e subsample (26).

EC_e data set is sparse. However, one should question whether the additional resolution borrowed from the secondary information (EM_V) reflects actual patterns of variability of the primary variable (this is more thoroughly discussed in the following sections).

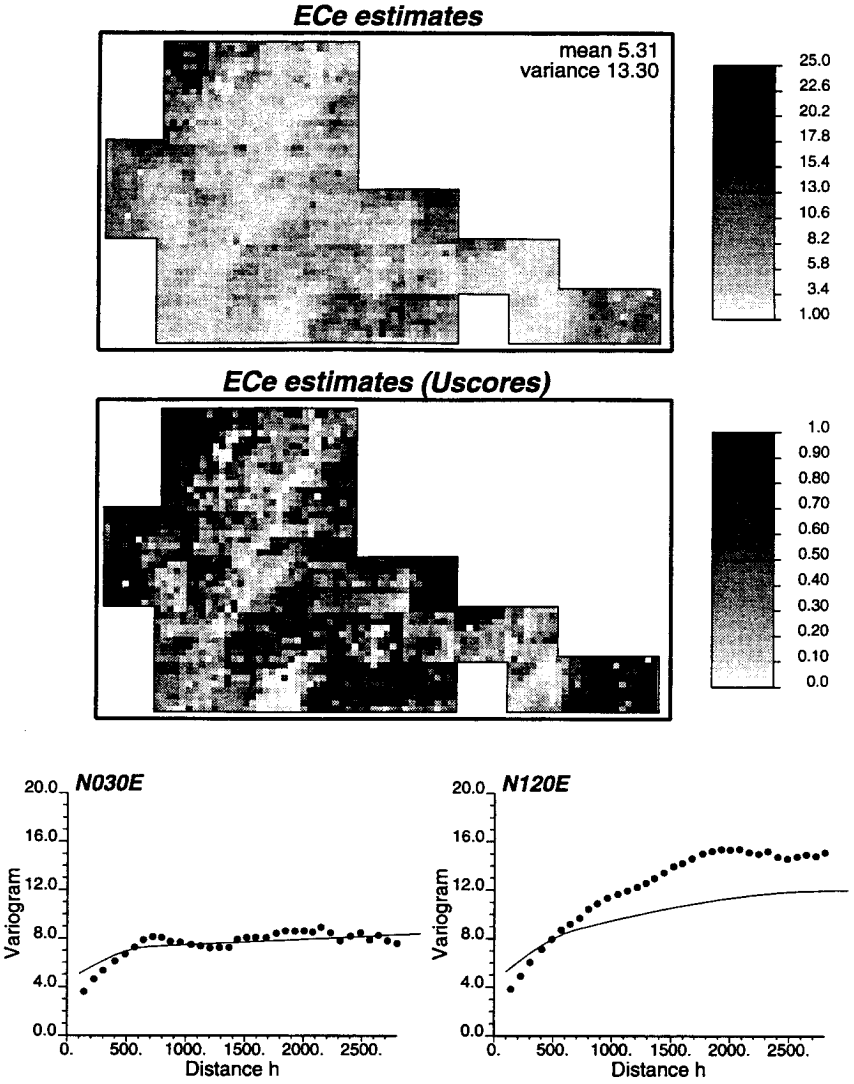

Figure 15 P-field results; full EC_e sample (315).

2. Utilization of a second soft information—in this case, soil type—does not bring significant improvement if that information is partially redundant with the first soft information used (in this case, exhaustive EM_V data). The p-field approach utilizing soil type data only performs marginally better

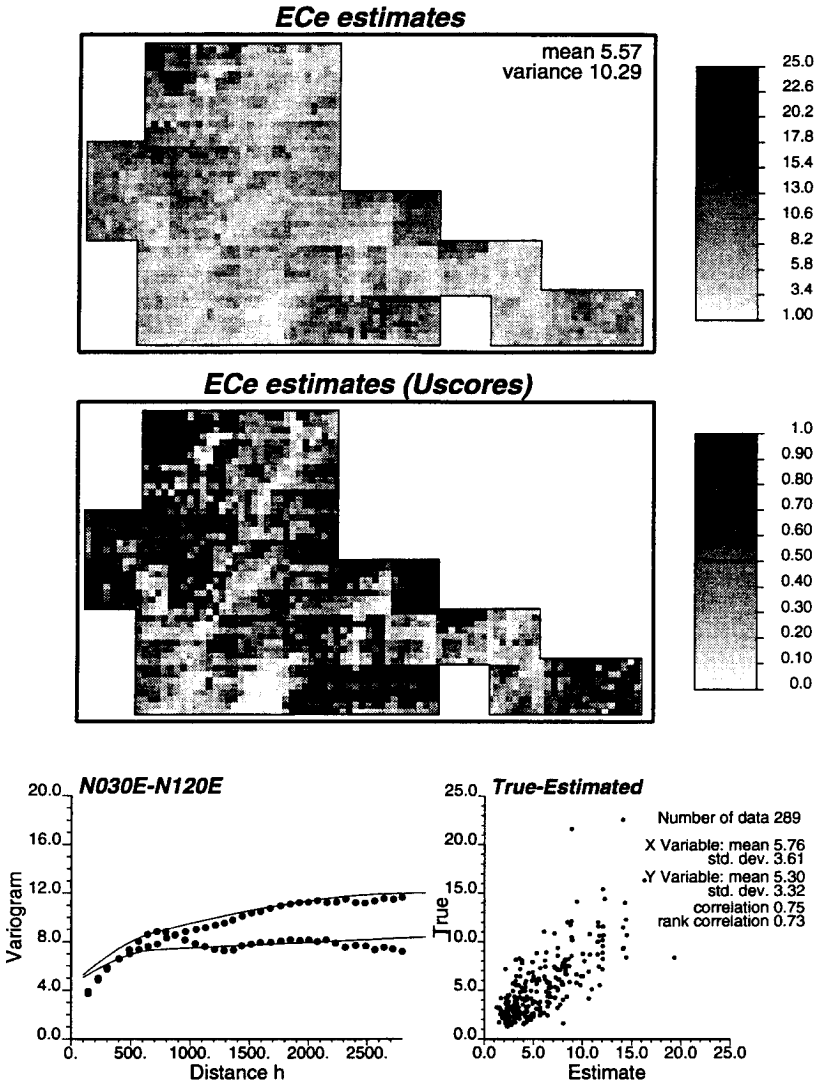

Figure 16 P-field cross-validation results; EC_e subsample (26).

than the cokriging approach that accounts only for EM_V as soft information (cf. Fig. 16 to Fig. 14). Nevertheless, for the subsample case, the p-field approach provides a better variogram reproduction.

B. Influence of *EMv* Data

In the two latter estimation algorithms, the sheer density of soft information (one EM, datum at each location being estimated) coupled with the good *EC, - EM,* correlation (0.73) tend to overwhelm the few hard data available. One should then wonder how much of the structures seen on the *EC,* estimated maps in Figs. 13 and 15 pertain to *EC,* and how much is mere *EM,* import.

Figure 17 recalls the sample *EC, -EM,* scattergram as given in Fig. 4b, then gives the four scattergrams of *EC,* estimates versus colocated *EM,* values. The correlation *EC, -EM,* is lowest (0.59) for simple kriging estimates, as expected. Accounting for *EM,* data increases that correlation to a level (0.86, 0.90) higher than that of the original sample (0.73). This higher correlation indicates that indeed there may be too much import of the *EM,* structures into the *EC,* mapping exercise. Note that the cokriging estimates show a linear relationship when plotted against the secondary *EM,* data (Fig. 17c). The p-field estimates (Fig. 17d) reproduce better the nonlinear relationship seen in the sample *EC, -EM,* scattergram (Fig. 17a).

At the limit, one may think of forfeiting altogether the *EC,* data and use the *EM,* map after proper rescaling to identify the sample *EC,* histogram (Joumel and Xu, 1994). The geostatistical toolbox offers one such algorithm that allows transforming any data set, e.g., the grid of *EM,* values, with any given histogram *H,* into another set of values identifying a target histogram *H,,* with H_2 possibly quite different from H_1. In addition to approximating the target histogram, this algorithm allows reproducing (freezing) a few original data values at their specific locations. This algorithm is a generalization of the well-known normal score transform whose target histogram is the standard Gaussian distribution (Deutsch and Joumel, 1992, p. 138).

Figure 18a shows the histogram of the 315 *EC,* data (the target histogram). Figure 18b shows the histogram of the 2385 *EM,* data transformed to match the target histogram: note the excellent histogram reproduction. These transformed *EM,* data, expressed in *EC,* units, are taken as estimates of *EC,* with their map shown in Fig. 18c. Per definition of the transformation algorithm, the uscores of these *EC,* estimates identify exactly the *EM,* uscores (Fig. 10a). Figure 18e shows the scattergram of *EC,* estimates (actually transformed *EM,* values) versus the original *EM,* values: this scattergram has a rank correlation 1 .00 reflecting the rank-preserving algorithm underlying the transform used. Recalling the scattergrams of Fig. 17, we have, indeed, gone all the way into importing all *EM,* structures into the *EC,* mapping exercise. This time, although there is a good linear correlation coefficient of 0.89 between the *EC,* estimates *(EM,* transformed) and the *EM,* data, the nonlinear relationship *EC, -EM,* is over-reproduced (cf. Figs. 18e and 17a). The *EC,* variogram model shape is not

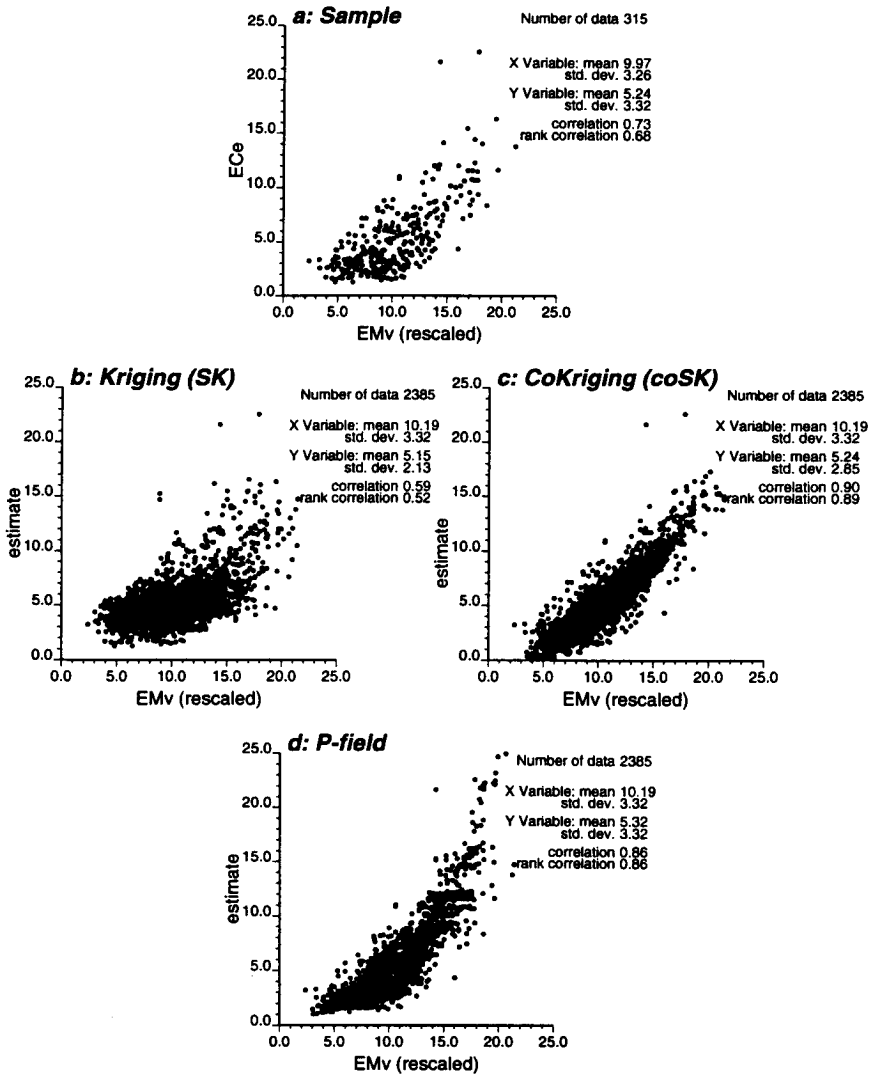

Figure 17 Scattergrams EC_e–EM_V (original sample and EC_e estimated values).

perfectly reproduced. This indicates that EC_e and EM_V spatial structures are not self-similar and that perfect identification of the two histograms is not enough to identify variograms (see Fig. 18d).

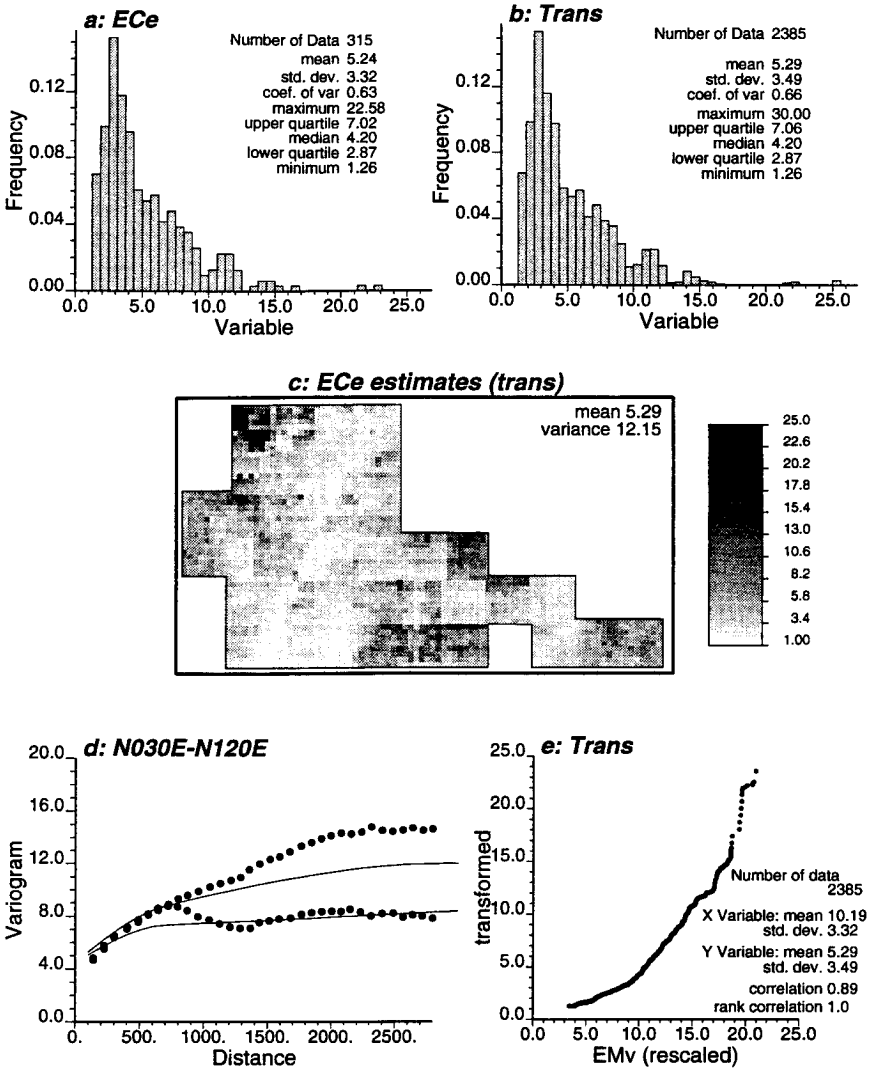

Figure 18 Transforming EM_V data to identify the EC_e sample histogram.

IV. FILTERING STRUCTURES

Geostatistics provides kriging algorithms for filtering from a spatial distribution any structure present in its variogram model. Consider, for example, the variogram model fitted from the 315 EC_e data and shown in the top row of Fig. 8.

This model includes a large nugget effect, a first isotropic structure with range of 700 m, and a second anisotropic structure with long range of 16,000 m in the N30°E direction of soil continuity and short range of 3000 m in the N120°E direction across continuity. The factorial kriging algorithm (Deutsch and Joumel, 1992, p. 68) allows filtering out from the simple kriging estimated map at the top of Fig. 11 (reproduced in Fig. 19a) the influence of both nugget effect and short-scale (700 m) structure, leaving the large-scale anisotropic structures (see Fig. 19b). Alternatively, one can filter out the influence of the large-scale variogram component leaving the short-scale structure (see Fig. 19c). The "sum" of Figs. 19b and 19c plus the nugget effect values at sample locations (not shown) add up to the original simple kriging map of Fig. 19a.

Because the nugget effect and short-scale structure account for such a large proportion of the EC_v spatial variance, the impact of the previous filtering is better seen on the corresponding uscore maps (see Fig. 20). Recall that the uniform score transform standardizes all distributions (hence variances) to a uniform distribution in [0,1].

Figures 19b and 20b depict the clear anisotropy of the large-scale structure associated to the soil type distribution (cf. Fig. la). Conversely, Figs. 19c and 20c zoom on shorter scale patterns of soil salinity possibly related to human activities: note the appearance of 1×1-km quadrats delineating different soil usage (Lesch et *al.*, 1995b). Recall that these maps are based on simple kriging-that is, ignoring the EM_v information.

Instead of being integrated in the simple kriging system (Deutsch and Joumel, 1992, p. 68), the factorial kriging algorithm can be used on any already available map such as the cokriging EC_v map shown at the top of Fig. 13. Figures 21 and 22 give for simple cokriging the same series of maps as given in Figs. 19 and 20 for simple kriging, the difference being utilization of the secondary EM_v information. As noted in the previous section, accounting for the dense EM_v information adds considerable local resolution to the estimated EC_v maps (cf. Figs. 21a and 19a).

After filtering, the EC_v cokriging maps (Figs. 21b and 22b) depict the same large-scale structure (related to soil type) seen on the filtered simple kriging maps (Figs. 19b and 20b). However, the short-scales structures seen on the cokriging maps (Figs. 21c and 22c) differ markedly from those seen on the simple kriging maps (Figs. 19c and 20c). To further investigate that difference in short-scale structures, the EM_v data rescaled to the EC_v variance 11.0 were directly filtered (see Figs. 23 and 24, whose layouts are the same as those in Figs. 21 and 22). It appears that the large-scale structure of the EM_v data is indeed similar to that observed on both the simple kriging and the cokriging EC_v estimates (cf. Figs. 23b and 24b to Figs. 19b and 20b and to Figs. 21b and 22b). However, the short-scale structures of the EM_v data (Figs. 23c and 24c) are clearly different from those of the simple kriging EC_v estimates (Figs. 19c and 20c). The short-scale EM_v structures reveal a NS-EW lattice possibly linked to the layout of the

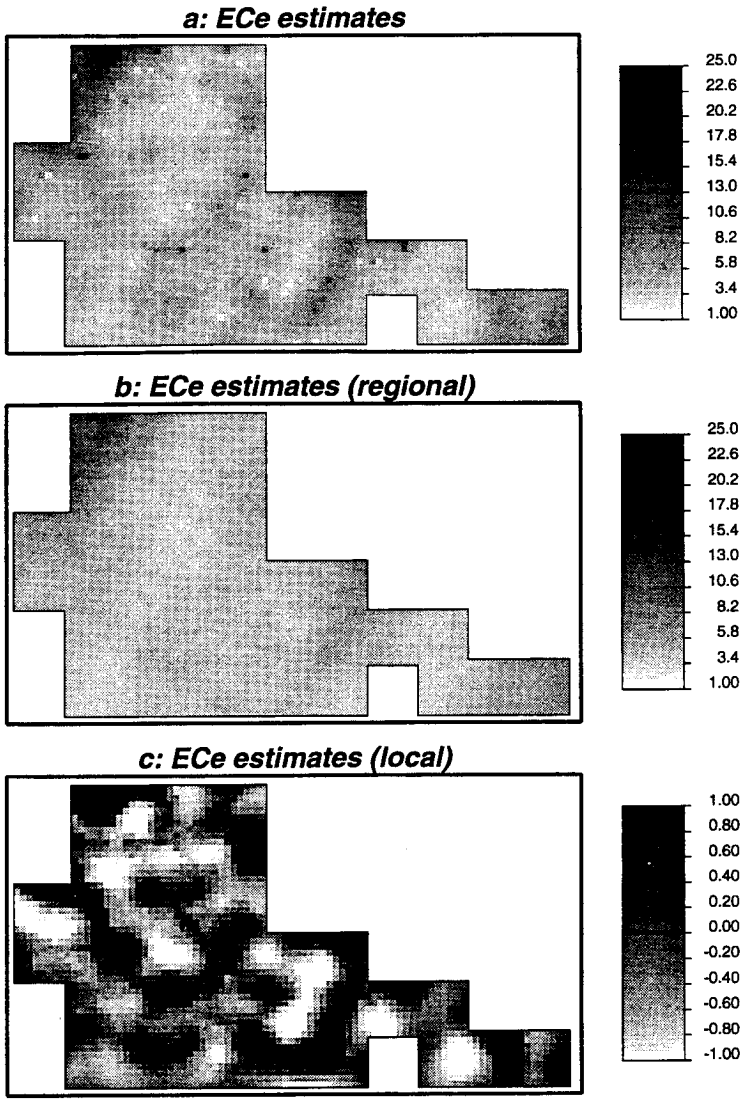

Figure 19 Filtering applied to the simple kriging map. (a) Simple kriging EC_e estimates, as in Fig. 11a, (b) large-scale EC_e patterns after filtering, and (c) short-scale EC_e patterns after filtering.

electromagnetic measures,[1] whereas the short-scale EC_e structures (Figs. 19c and 20c) are more curvilinear, possibly related to human activity.

Figure 25a shows the scattergram of the regional components of EM_V and EC_e

[1]Posterior to this comment, discussion with Scott Lesch (Lesch *et al.*, 1995a,b) pointed to an interpretation of the mostly EW short-scale lattice sturcute seen on the EM_v map (Fig. 24c). This EW

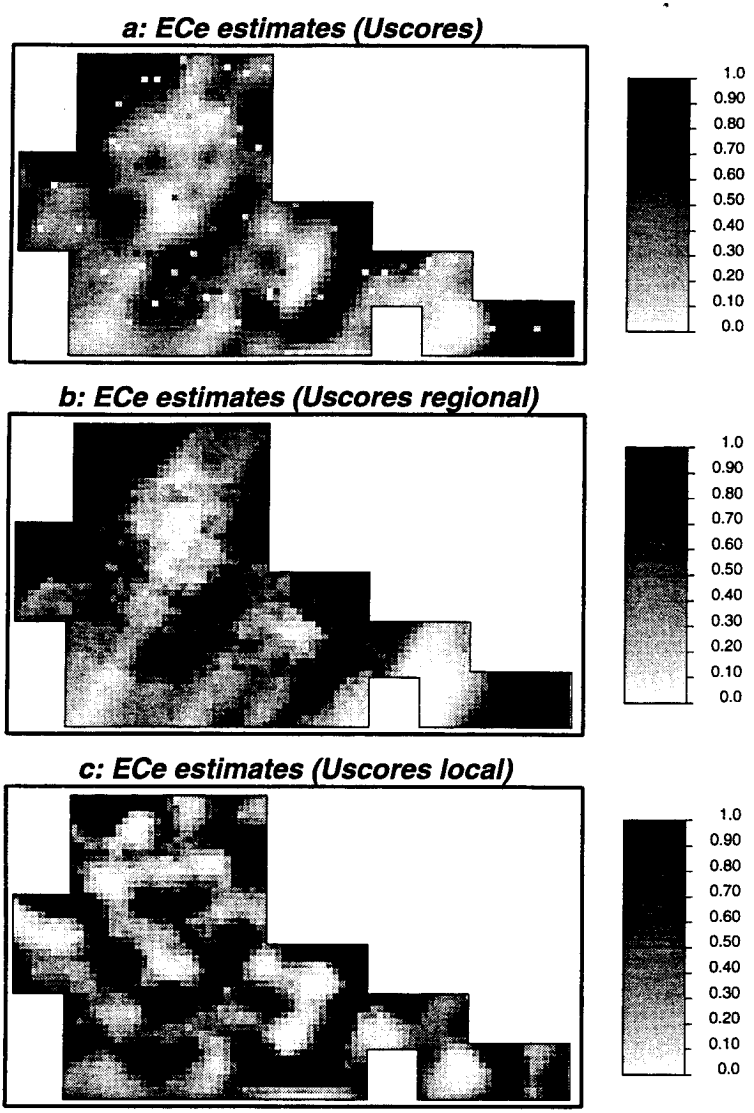

Figure 20 Uscore transform of simple kriging estimated maps. (a) Prior to filtering, (b) large-scale EC_e patterns after filtering, and (c) short-scale EC_e patterns after filtering.

---

structure may be the result of repeated directional flood irrigation causing, over the years, a certain amount of soil salinity to redistribute along furrows from head to tail in each field. There are not enough EC_e hard data within each field to reveal that structure.

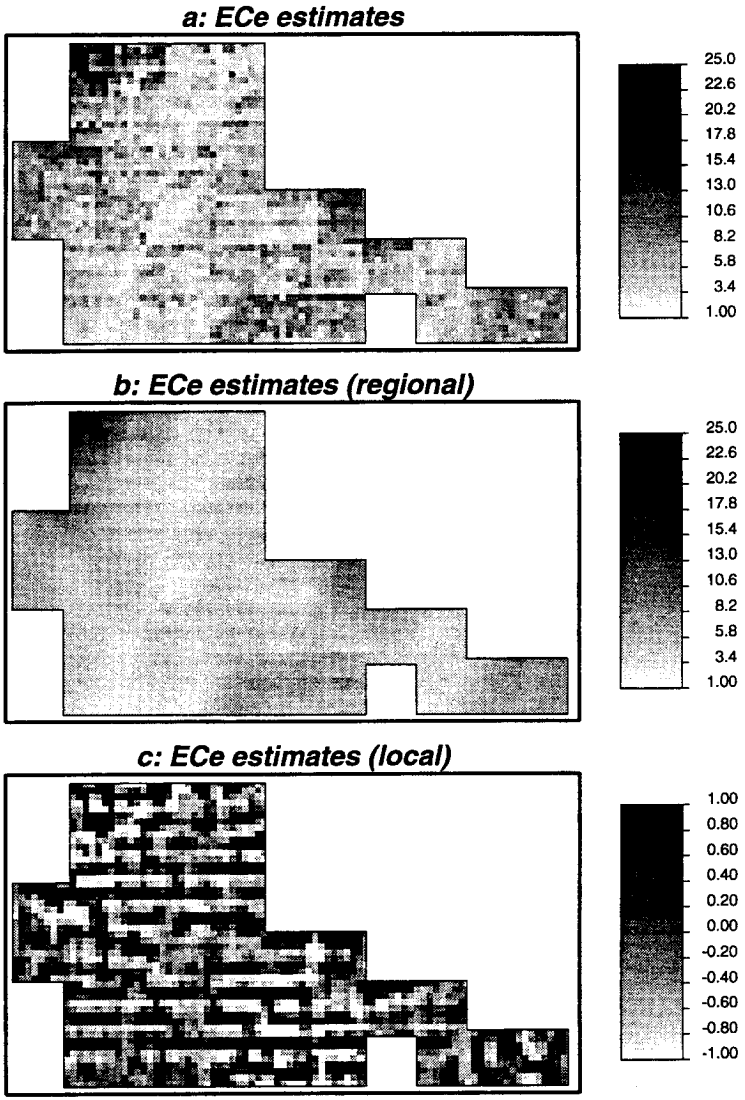

Figure 21 Filtering applied to the cokriging map. (a) Simple cokriging EC_e estimates, (b) large-scale EC_e patterns after filtering, and (c) short-scale EC_e patterns after filtering.

as read from Figs. 19b and 23b, respectively. Figure 25b shows the scattergram of the corresponding short-scale components as read from Figs. 19c and 23c. Note the poor correlation of the short-scale structures ($\rho = 0.30$) as opposed to

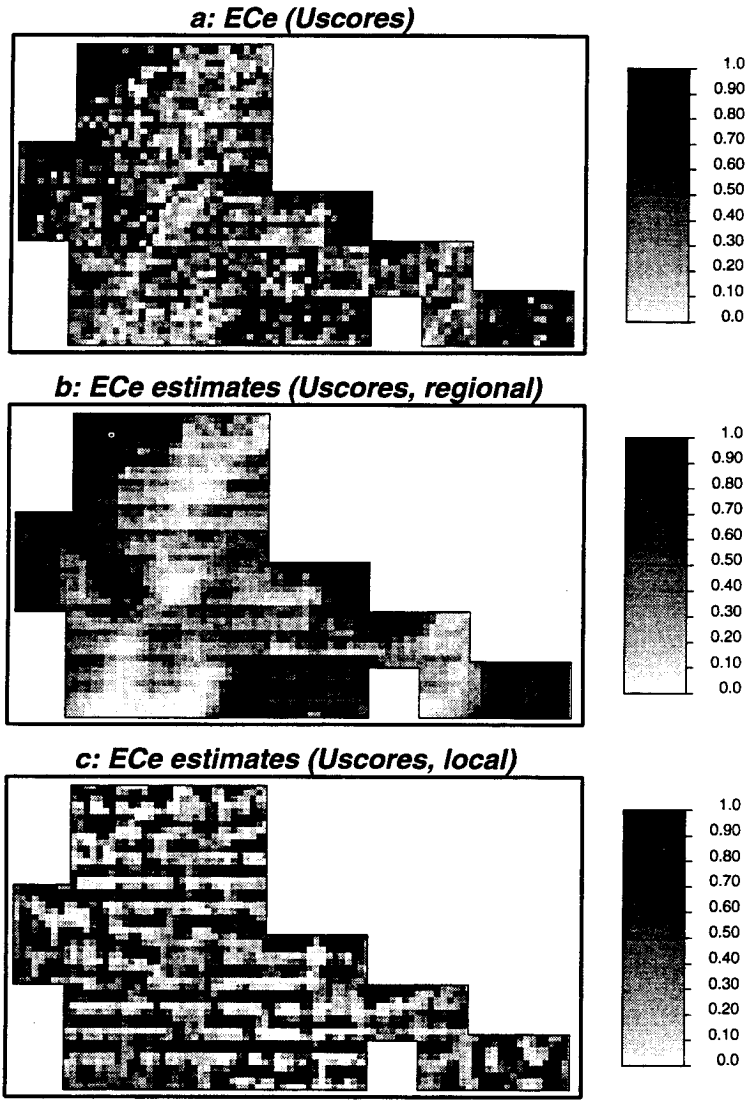

Figure 22 Uscore transform of simple cokriging estimated maps. (a) Prior to filtering, (b) large-scale EC_e patterns after filtering, and (c) short-scale EC_e patterns after filtering.

the better correlation of the regional, large-scale, components ($\rho = 0.63$). On the regional scale scattergram, there appears to be two populations most likely related to the two groups of soil types previously distinguished from EM_V histo-

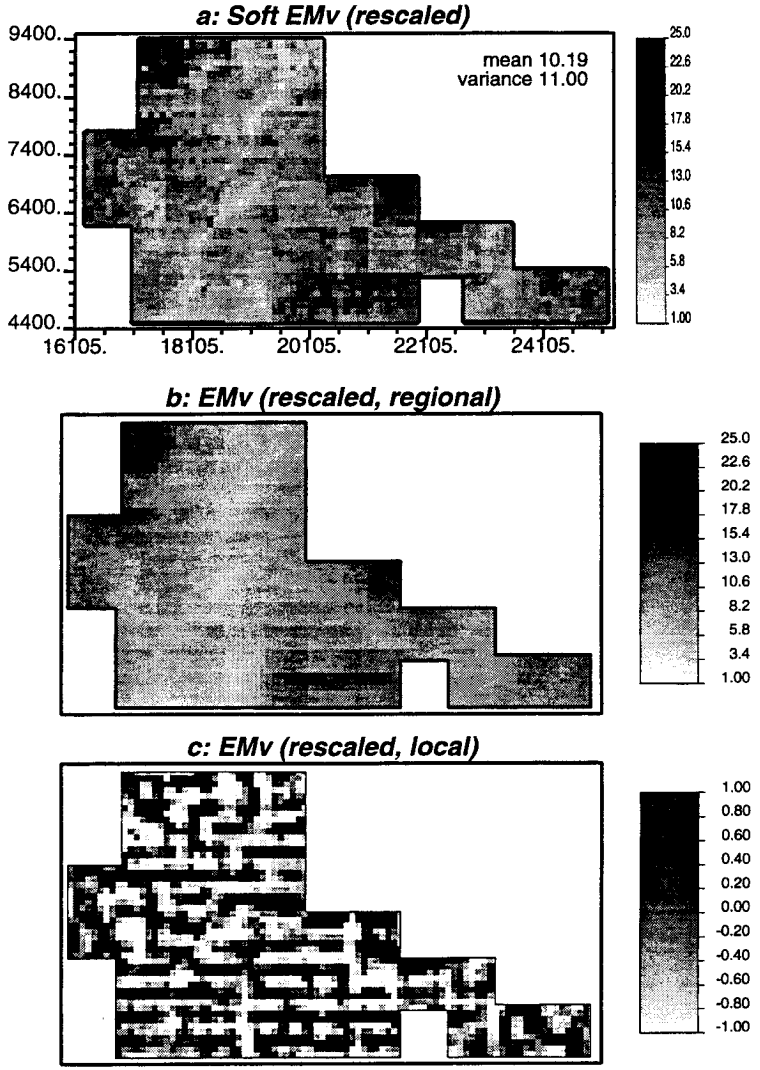

Figure 23 Filtering applied to the EM_V data. (a) EM_V data rescaled to EC_e variance, prior to filtering; (b) large-scale EM_V patterns after filtering; and (c) short-scale EM_V patterns after filtering.

grams per soil type (see Fig. 6). Were these two groups separated, the regional structures correlation observed on Fig. 25a would be even higher.

This analysis would suggest filtering the short-scale structures from the EM_V data before using them in the cokriging of EC_e values.

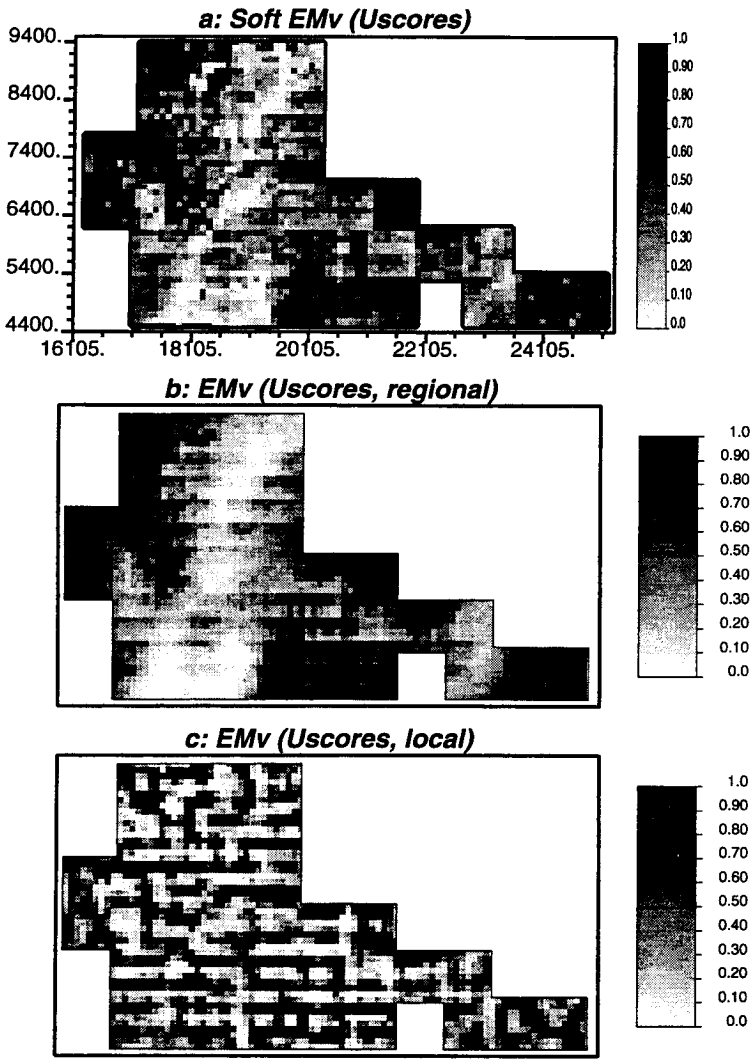

Figure 24 Uscore transform of EM_V data. (a) Prior to filtering, (b) large-scale EM_V patterns after filtering, and (c) short-scale EM_V patterns after filtering.

V. SPATIAL CLUSTER ANALYSIS

Spatial clustering allows to delineate in space relatively contiguous zones with similar attribute values. When dealing with multiple attribute values, e.g., EM_V,

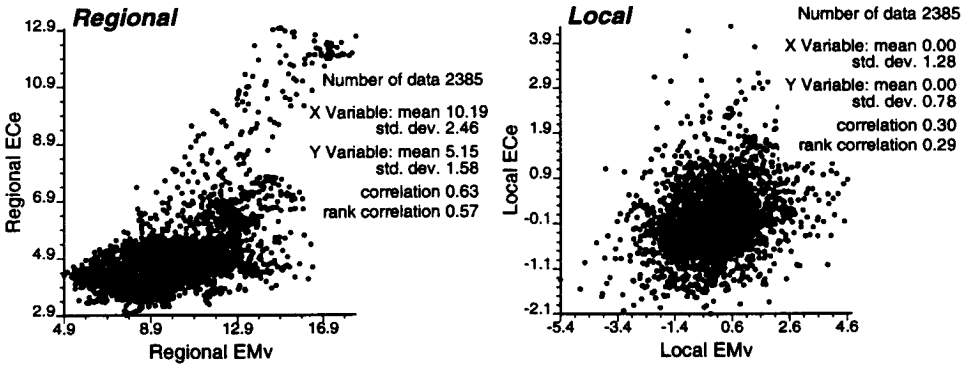

Figure 25 Scattergrams of EM_V–EC_e scale-specific components of spatial variability. (a) Large-scale (regional) and (b) short-scale (local).

EM_H, and EC_e values, definition of such zones from multiple contour maps would be fastidious and highly subjective. Standard clustering techniques (Sneath and Sokal, 1973) typically consider only the correlation matrix $[\rho_{kk'}(0)]$ of colocated attribute values, thus ignoring the spatial information provided by the full correlation matrix $[\rho_{kk'}(\mathbf{h})]$, $\forall$ lag distance $\mathbf{h}$ between samples. Such standard clustering often results in groups homogeneous in terms of attribute values but dispersed in space; more precisely, any particular group may comprise several disjoint areas/volumes in space (see Fig. 26b).

Bourgault *et al.* (1992) have proposed to weigh the traditional similarity measure (s_{ij}) between any two samples i,j by a function $r(\mathbf{h}_{ij})$ of their separation vector $\mathbf{h}_{ij}$:

$$S_{ij} = s_{ij} \times r(\mathbf{h}_{ij}) \tag{6}$$

where

$s_{ij} = \mathbf{Z}_i^T \cdot \Sigma^{-1} \cdot \mathbf{Z}_j$ is the traditional similarity measure, $\mathbf{Z}_i$ is the K vector of standardized attribute values at location $\mathbf{u}_i$, Σ is the $K \times K$ correlation matrix of colocated attribute values, and Σ^{-1} is the traditional Mahalanobis distance, e.g., for $K = 2$ attribute values (EM_V, EM_H),

$$\mathbf{Z}_i = \left[\frac{EM_V(\mathbf{u}_i) - m_V}{\sigma_V}, \frac{EM_H(\mathbf{u}_i) - m_H}{\sigma_H} \right]$$

where m_V, σ_V^2 is the mean and variance of EM_V data, m_H, σ_H^2 is the mean and variance of EM_H data, $\Sigma = \begin{pmatrix} 1 & \rho_{VH} \\ \rho_{VH} & 1 \end{pmatrix}$ with ρ_{VH} being the linear correlation coefficient between colocated EM_V and EM_H data, and $r(\mathbf{h}) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \omega_{kk'} \rho_{kk'}(\mathbf{h})$ is the multivariate covariance defined as a linear combination of the

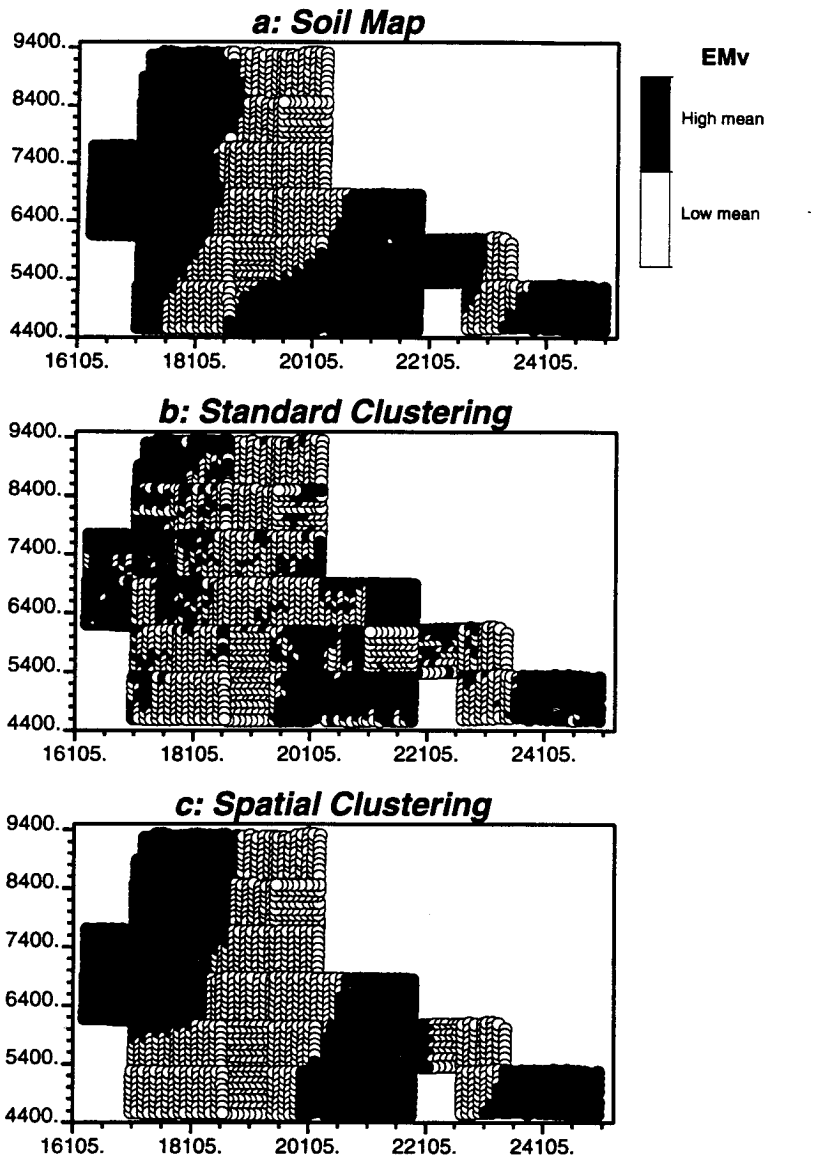

Figure 26 Comparison of standard versus spatial clustering of electromagnetic data in reproducing soil type groups.

K X K standardized (cross)covariance functions $\rho_{kk'}(\mathbf{h})$ between any two RV's $Z_k(\mathbf{u})$ and $Z_{k'}(\mathbf{u}+\mathbf{h})$ separated by vector h. $\mathbf{h}_{ij} = \mathbf{u}_i - \mathbf{u}_j$ is the vector separating the two samples i, j at locations $\mathbf{u}_i, \mathbf{u}_j$. Typically, the $(K \times K)$ weight matrix $[\omega_{kk'}]$ is identified to the Mahalanobis distance: $[\omega_{kk'}] = \Sigma^{-1}$.

Spatial continuity, anisotropy, and cross-correlation are accounted for through the term $r(\mathbf{h}_{ij})$ in the definition [Eq. (6)] of the similarity measure S,. Cluster analysis using this measure will result in grouping of samples with similar attribute values (term s_{ij}) but also spatially close together [term $r(\mathbf{h}_{ij})$].

The spatial cluster analysis algorithm is demonstrated using only the two densely sampled electromagnetic attributes, EM_V and EM_H. No soil data were used in order to check that cluster analysis using only electromagnetic data does result in groups consistent with soil type differentiation.

The algorithm progresses as follows:

1. An initial number of groups is chosen arbitrarily, not too large to allow statistical characterization of each group. Here, seven groups were retained according to the actual number of soil types.
2. All 2385 electromagnetic (EM_V and EM_H) samples are randomly assigned, with equal probability $\frac{1}{7}$, to one of the seven groups.
3. For each sample i, with $i = 1, \cdots, 2385$,

• Calculate its average similarity with group (g) defined as

$$\bar{S}_{i,(g)} = \frac{1}{|(g)|} \sum_{j \in (g)} S_{ij}, \ \forall\, g = 1, \cdots, G = 7 \text{ groups} \qquad (7)$$

where $|(g)|$ is the number of samples j currently classified in group (g).

. Assign sample i to the group with which it has the highest similarity, then update the constitution of all groups.

4. Step 3 is repeated until no change is observed in the constitution of the G groups.
5. The targeted number G_0 with $G_0 \leq$ G, of groups is obtained by concatenating groups having similar characteristics; for example, in the case of Fig. 29, $G_0 = 2$ corresponding to the two super groups having mean EM_H values greater (lesser) than the overall mean $EM_H = 1.47$.

Figures 26b and 26c show the $G_0 = 2$ super groups resulting from a standard clustering algorithm using similarity measure s_{ij} and the proposed spatial clustering algorithm using measure S,. Figure 26a shows the reference binary soil type map regrouping the seven original soil types into only two super groups depending on whether the mean EM_H values exceed the overall mean 1.47. Utilization

of spatial information results in a much cleaner image closer to the reference image obtained from soil type data.

VI. STOCHASTIC IMAGING

In the previous sections, various estimated ***EC,*** maps have been presented but their accuracy was not assessed. As opposed to mere interpolation algorithms, the main contribution of a geostatistical approach is to provide an assessment of the reliability of any given estimated value. What is the reliability of the ***EC,*** estimated value at any specific location u? What is the reliability of any cluster of, e.g., high ***EC,*** estimated values as seen on the estimated map of Fig. 13? Can there be alternative estimated maps using the same information?

Besides the kriging estimated value, the solution of any kriging system yields a kriging variance-that is, the minimized error variance (Isaaks and Srivastava, 1989, p. 286). Unfortunately, because this kriging variance is data values independent, it is a poor measure of estimation accuracy; instead, it is only a ranking index of data configuration-the data configuration corresponding to a lesser kriging variance would yield on average (over all possible data values for that configuration) a more accurate estimate.

Even if the kriging variance $\sigma_K^2(\mathbf{u})$ was a measure of accuracy of the estimated ***EC,*** value at location u, the two kriging variances $\sigma_K^2(\mathbf{u})$ and $\sigma_K^2(\mathbf{u}')$ would not provide assessment of joint accuracy at the two locations u and u'. For example, these two kriging variances would not allow assessing the probability that the two unsampled values Z(u), Z(u') be jointly above a given threshold z_0.

The concept of stochastic simulation (stochastic imaging) was developed to answer this need for a joint spatial measure of uncertainty (Deutsch and Joumel, 1992, p. 17). As opposed to kriging or any other interpolation algorithm, stochastic simulation yields not one but many alternative **equiprobable**[2] images of the distribution in space of the attribute under study (in this case, ***EC,; see*** Fig. 29). The difference between these alternative realizations, or stochastic images, provides a visual and numerical measure of uncertainty, whether involving a single location **u** or many locations jointly.

Similar to kriging, there are many stochastic simulation algorithms (Deutsch and Joumel, 1992, p. 117) depending on which particular feature (statistics) of the data ought to be reproduced. The first goal of geostatistical simulations is to correct for the smoothing effect observed in any (co)kriging estimated map.

[2]These stochastic images are equiprobable in the sense that, for a given simulation algorithm with its specific computer code and choice of statistics, each image is uniquely indexed by a seed number that starts the algorithm. The seed numbers are drawn from a probability distribution uniform in [0,1]; hence. each image is equal likely to be drawn.

Hence, the simulated values have a similar spatial continuity (variogram) to that of the sample data set used. In the following section, we present an indicator simulation algorithm modified to account for the soft information provided by electromagnetic (EM,) data. The indicator simulation algorithm (Deutsch and Joumel, 1992, p. 146) allows the simulated values to display different spatial continuities (variograms) for different classes of values.

A. SIMULATION ALGORITHM

Any unsampled ***EC,*** value at location u is interpreted as a random variable Z(u). This random variable (RV) can be seen as a set of possible outcome values or realizations, $z^{(l)}(\mathbf{u}), l = 1, 2 \cdots$ characterized by a probability distribution, denoted Prob $\{Z(u) \leq z](n)\}$, where the notation $|(n)$ is read as "conditional to the information set (n)." In the approach adopted here, this probability distribution is modeled by a weighted linear combination of neighboring indicator data $i(\mathbf{u}_\alpha;z)$, which is set to 1 if the ***EC,*** datum value $z(\mathbf{u}_\alpha)$ at sample location $\mathbf{u}_\alpha$ does not exceed threshold z and set to zero otherwise:

$$\text{Prob}\ \{Z(u) \leq z|(n)\} = \sum_{\alpha=1}^{n} \text{h,(z)} \cdot i(\mathbf{u}_\alpha; z) \tag{8}$$

The weights h,(z) are given by an indicator kriging system specific to each threshold value z (Deutsch and Joumel, 1992, p. 150). The ***n*** indicators retained correspond to the hard ***EC,*** sample values found in the neighborhood of location u. Nine threshold values z corresponding to the nine deciles of the ***EC,*** sample histogram (sample size is 3 15) were retained to discretize the range of variability of Z. In this case, the indicator simulation algorithm accounts for the spatial continuity specific to each decile of the ***EC,*** data values.

The model [Eq. (8)] accounts only for the hard data. Introduction of the soft ***EM,*** information was done through the "external drift" concept (Deutsch and Joumel, 1992, p. 67) whereby the set of ***n*** weights h,(z) is constrained such as to ensure that the expected value of the estimator [Eq. (8)] identifies a prior probability deduced from the ***EM,*** information. More precisely, the constraint is

$$\sum_{\alpha=1}^{n} \lambda_\alpha(z) \cdot p(\mathbf{u}_\alpha; z) = p(\mathbf{u}; z) \tag{9}$$

where p(u;z) = Prob $\{Z(u) \leq z|Y(\mathbf{u}) = y(u)\}$ is the prior probability of ***EC,*** value Z(u) given the colocated ***EM,*** sample value y(u). The qualifier, "prior," indicates that this probability value is obtained prior to using the neighboring values.

These prior probabilities [Eq. (8)] are read from the scattergram of colocated EC_e and EM_V values [see Figs. 27b–27d for three such prior distributions corresponding to three specific conditioning $y(\mathbf{u})$ value (EM_V)]. Following the suggestion made at the end of Section IV, only the regional (large-scale) component of the EM_V data was retained for the calibration scattergram of Fig. 27a. The map of this EM_V regional component was shown in Fig. 23b.

Figures 28a and 28b show the greyscale maps of the nine sets of prior probability values $p(\mathbf{u};z_k)$, with one probability value per location $\mathbf{u}$, and one map for each of the nine decile threshold values z_k, $k = 1,\cdots, 9$. Figs. 28a and 28b actually map the residual value $p(\mathbf{u};z_k) - p(z_k)$, with $p(z_k) = k/10$ being the marginal probability values. Dark areas are areas where the prior probability

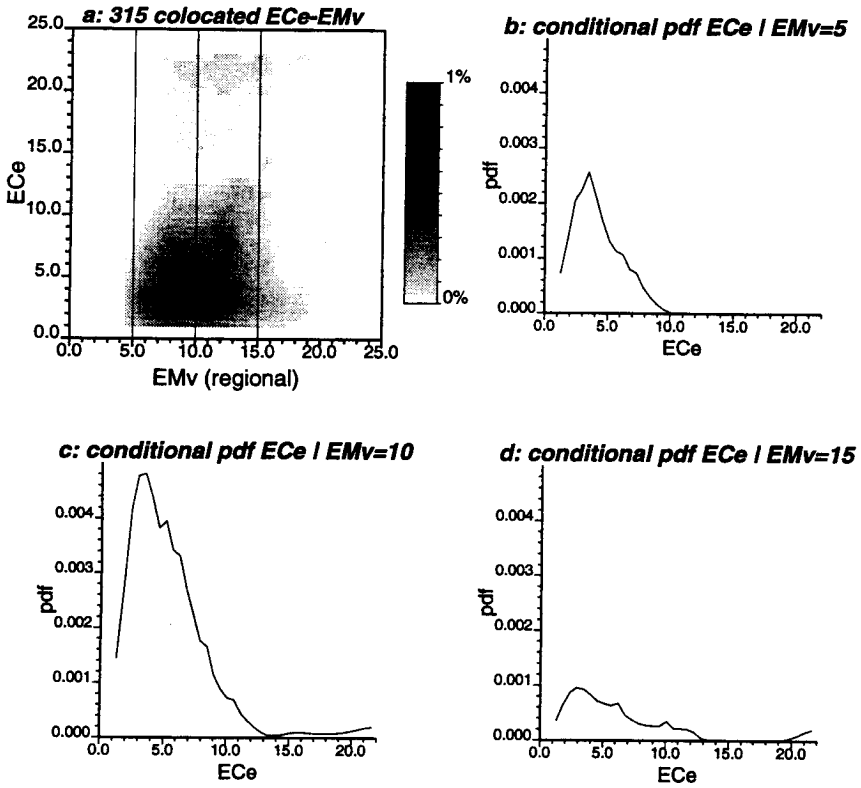

Figure 27 Scattergram of colocated EC_e and EM_V (regional component) values. The scattergram is intersected by three vertical lines; EM_V = constant providing three EC_e conditional probability distributions.

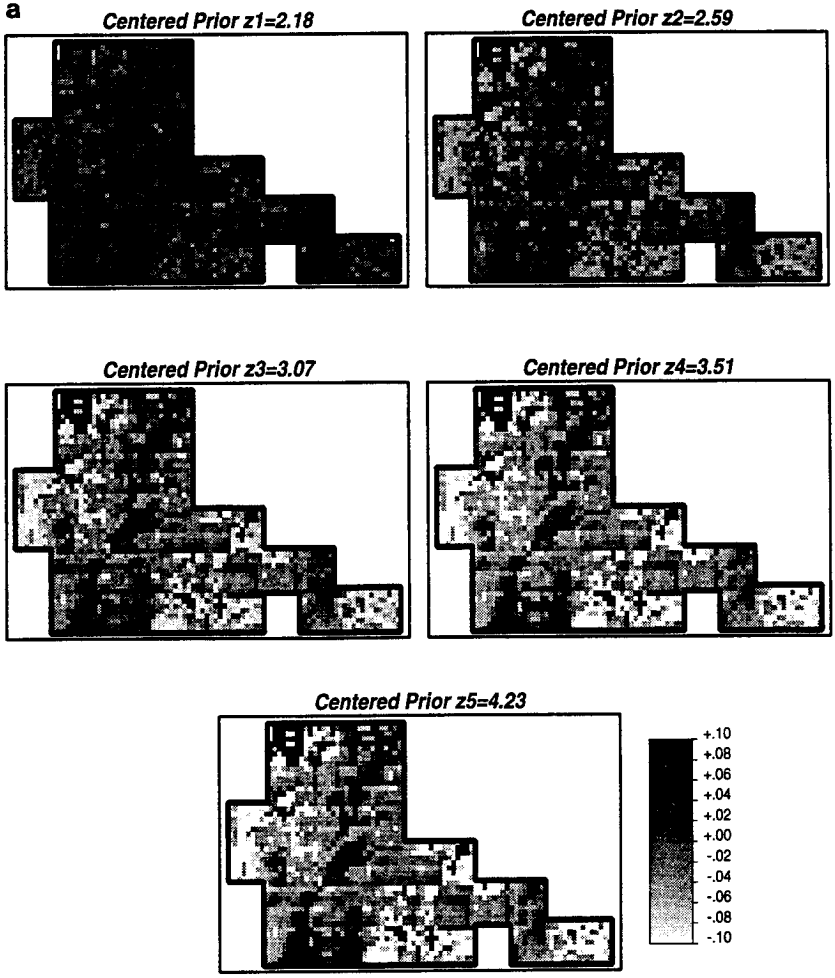

Figure 28 Maps of centered prior probability values for EC_e as obtained from the colocated EM_v data values.

$p(\mathbf{u};z_k)$ exceeds its marginal or mean value $p(z_k)$; conversely, light areas are where that prior probability is lesser than the marginal value.

The sequential simulation paradigm calls for visiting along a random sequence all nodes $\mathbf{u}$ of the simulation grid. At any such node $\mathbf{u}$, the prior probability values $p(\mathbf{u};z_k)$, $k = 1,\cdots, 9$ are updated through the indicator kriging process [Eqs. (8) and (9)] to account for neighboring original values and EC_e values

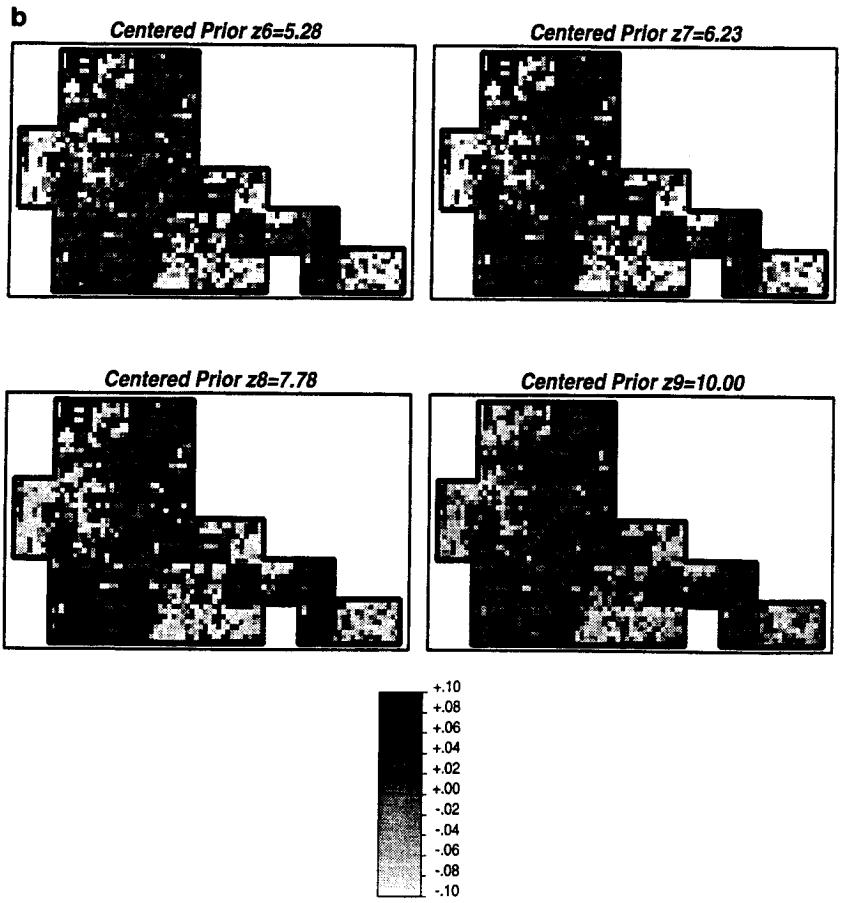

Figure 28 (*continued*)

simulated at previously visited nodes (Deutsch and Journel, 1992, p. 123). A simulated EC_e value, $z^{(l)}(\mathbf{u})$, is then drawn from that updated probability distribution. A stochastic image, e.g., the *lth*, is completed when all nodes $\mathbf{u}$ of the simulation grid have been visited and filled in with a simulated value. Iteration of the entire process starting from another random seed number provides another equiprobable stochastic image. Fifty such stochastic images were generated; the three first realizations are shown in Fig. 29. The pointwise average of the 50 realizations is shown in Fig. 30a. This average, also called *E*-type map for expected value map, is similar although not identical to the direct simple cokriging and p-field maps respectively shown at the top of Figs. 13 and 15.

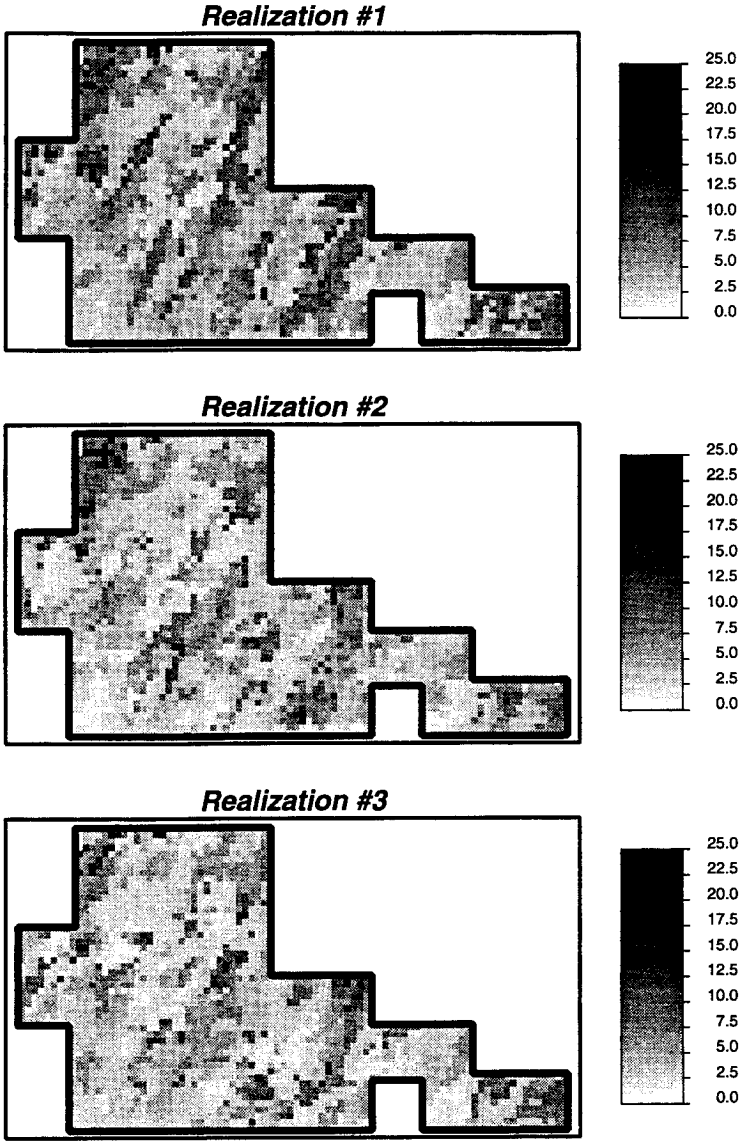

Figure 29 Three stochastic images of the distribution in space of EC_e values.

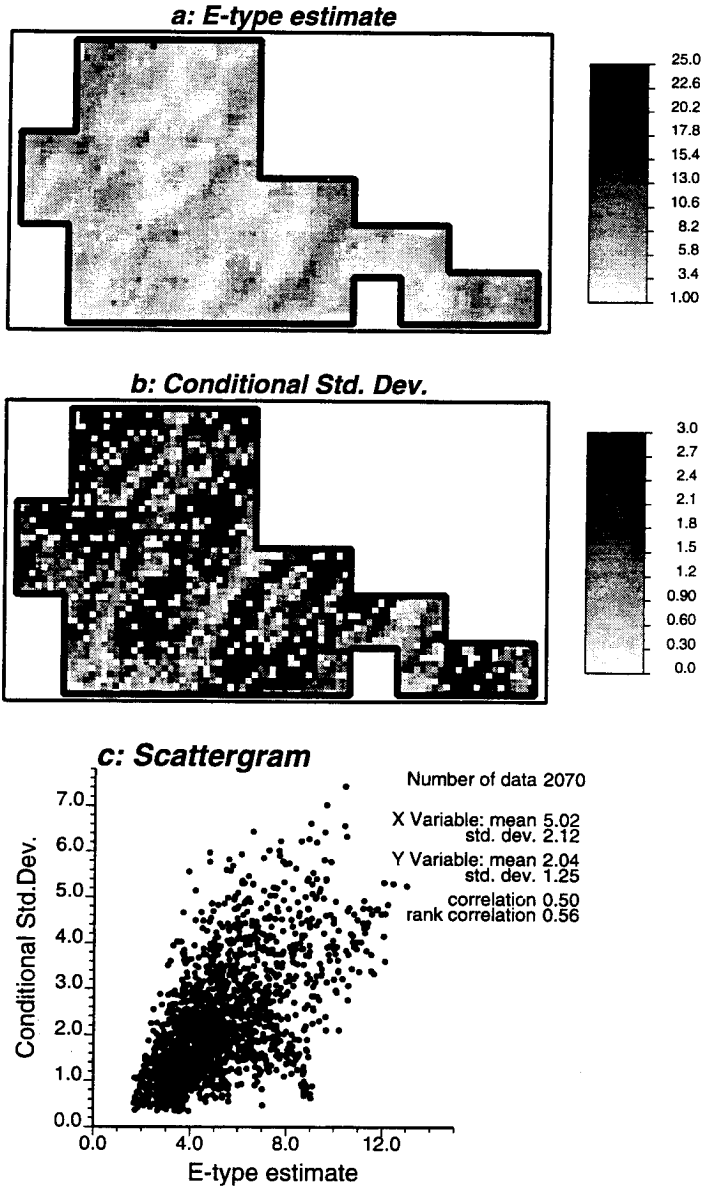

Figure 30 Pointwise average and standard deviation of the 50 EC_e stochastic images. The bottom graph gives their scattergram showing that high E-type estimated values are also the most uncertain.

VII. ASSESSMENT OF SPATIAL UNCERTAINTY

The availability of the 50 equiprobable stochastic images of the distribution in space of ***EC,*** values allows derivation of multiple measures of uncertainty beyond a mere visual inspection of these images.

- Local uncertainty: The uncertainty about ***EC,*** at any location u can be assessed by any measure of spread of the 50 simulated ***EC,*** values at that location, $z^{(l)}(\mathbf{u}), l = 1, \cdots, 50$. For example, one could consider the standard deviation of these 50 values. The corresponding grayscale map is shown in Fig. 30b. Note that as opposed to the kriging variance, a variance of simulated values is an estimation variance *conditional* to the data values retained to simulate these values. At ***EC,*** sample locations, that conditional estimation variance is zero (white pixels in Fig. 30b). Elsewhere, that estimation variance depends on the data values and not only on the data configuration; this property is known in statistics as heteroscedasticity. Here, the estimated high E-type values are also the most uncertain in that the corresponding variance between simulated values is larger (see the scattergram in Fig. 3Oc).
- Probability maps: At each location u one can count the proportion of simulated values $z^{(l)}(\mathbf{u})$ lesser (or greater) than any given threshold value z, then map these proportions. Figure 31 shows two such probability maps: (i) the probability that ***EC,*** is no greater than the first decile $z_1 = 2.18$ of the sample ***EC,*** histogram. Dark areas (high probability) on this map are areas where ***EC,*** is surely low valued; and (ii) the probability that ***EC,*** exceeds the ninth decile $z_9 = 10.0$ of the sample ***EC,*** histogram. Dark areas (high probability) on this map point to areas where ***EC,*** is surely high valued. Note that probability maps are unit free, valued in [0,l].
- Quantile maps: For some applications it is convenient to merge in a single map an "estimate" of the attribute value and the assessment of the accuracy of that estimate. Quantile maps provide such joint assessment.

Figure 32a provides a low (0.1) quantile map; more precisely, the map of the ***EC,*** value that is exceeded by 90% of the simulated values at the same location u. Therefore, a location appearing high (dark) in Fig. 32a has a high probability (90%) to be actually higher. Dark areas on a low-quantile map are areas that are surely high valued.

Conversely, Fig. 32b shows a high (0.9) quantile map; more precisely, the map of the ***EC,*** value that is higher than 90% of the simulated values. Therefore, a location appearing low (light gray) in Fig. 32b has a high probability (90%) to be actually even lower. Light gray spots on a high quantile map point to areas that are surely low valued.

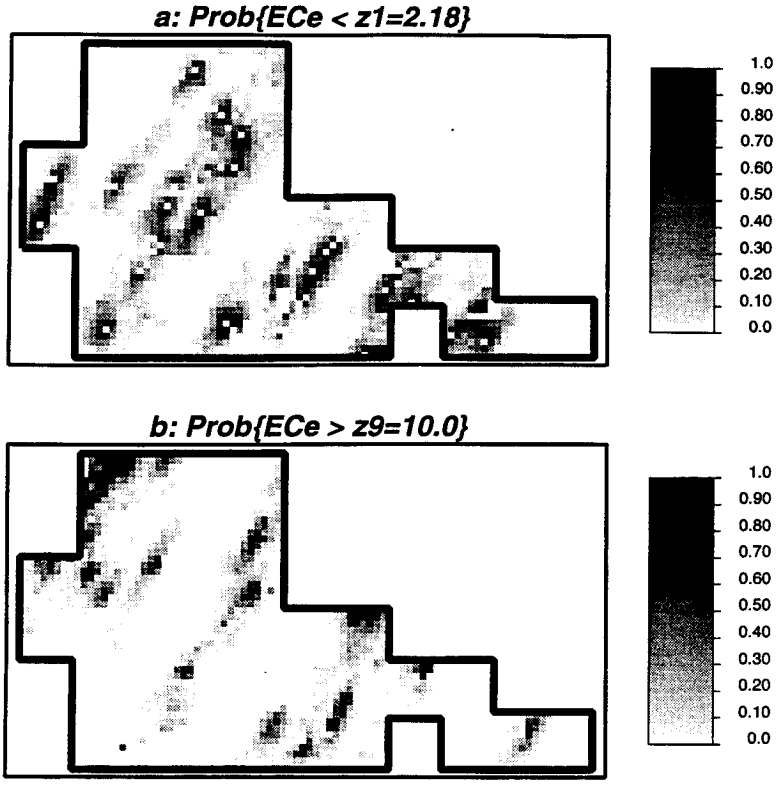

Figure 31 Probability maps. (a) EC_e does not exceed $z_1 = 2.18$, and (b) EC_e exceeds $z_9 = 10.0$.

Note that quantile maps are in the unit of the attribute itself. Figure 32c shows the median or 0.5 quantile map. The median is the value that has equal probability to be exceeded or not exceeded by the simulated values. It can be used in lieu of the *E*-type (average) map in Fig. 30a. As opposed to the *E*-type map, the median map carries a measure of its uncertainty.

Probability and quantile maps are numerical and visual aids to decision making. They allow spotting areas that are, e.g.,

1. Critical and surely so; hence, where remediation work can start without delay
2. Potentially critical (e.g., high estimated pollution but with large uncertainty); such areas should be earmarked for additional data
3. Safe and surely so; hence, where no further action is warranted

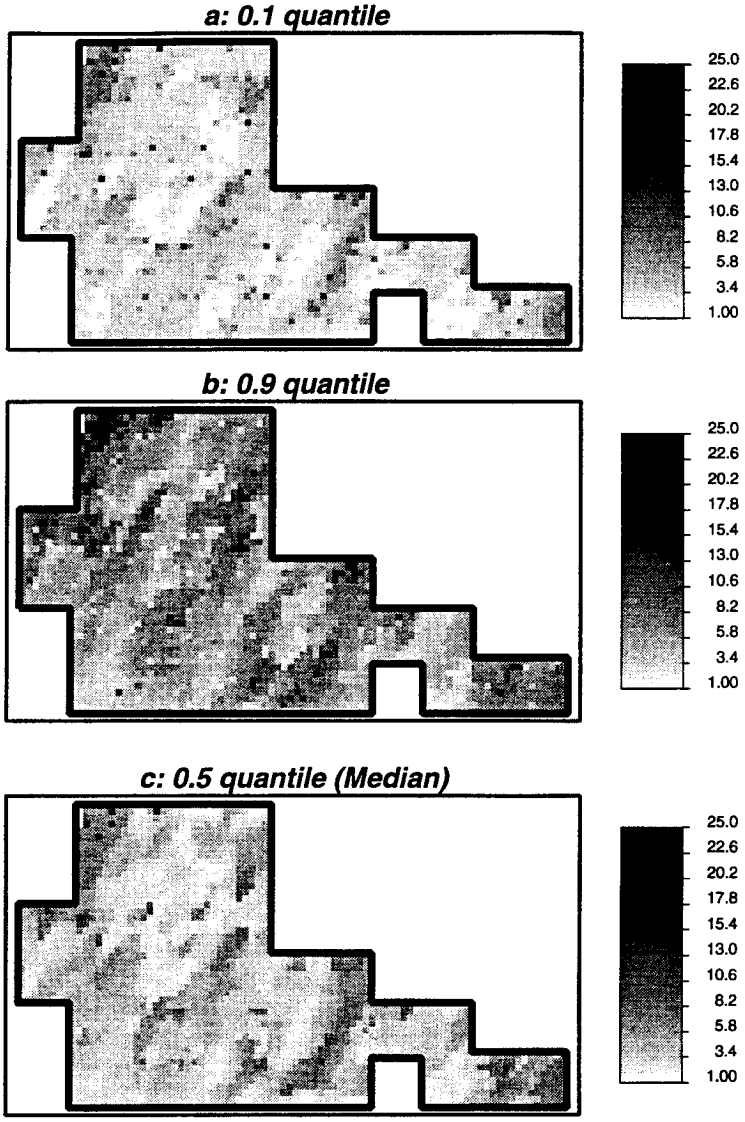

Figure 32 Three quantile maps of the EC_e spatial distribution.

As more or upgraded information becomes available, the probability distributions of Eq. (8) should be updated again and corresponding stochastic images should be generated leading to updated probability and quantile maps.

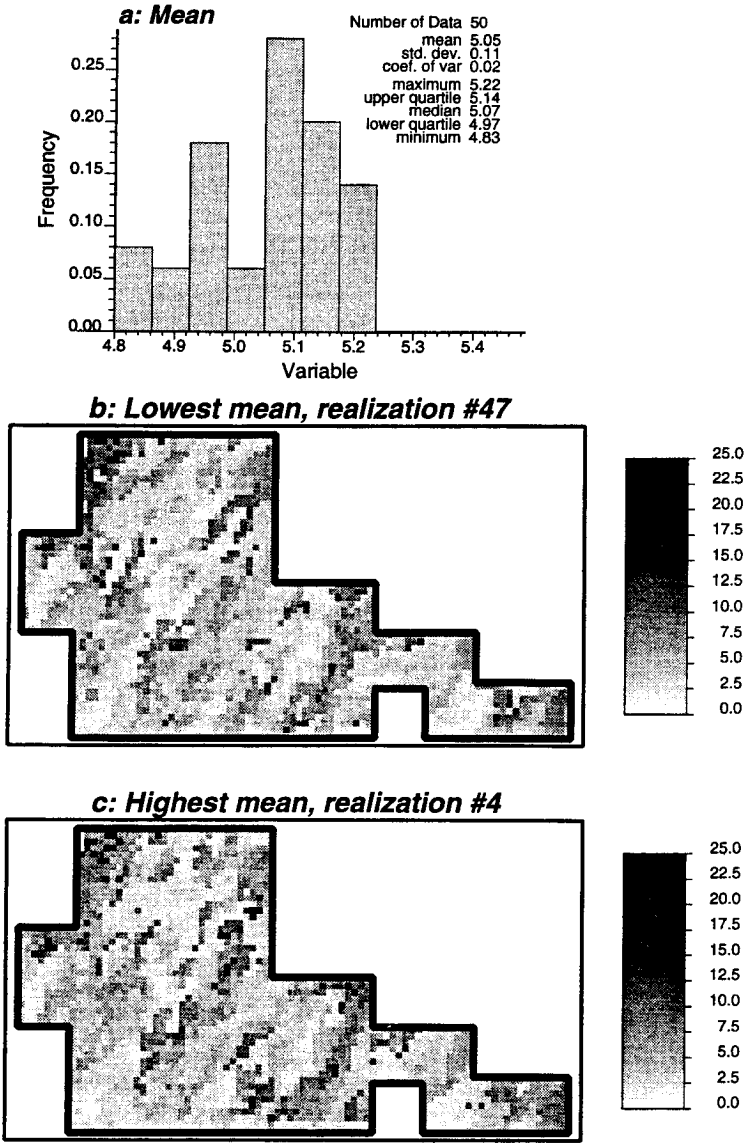

Figure 33 Ranking of stochastic images. (a) Histogram of simulated overall mean values, (b) realization with lowest mean, and (c) realization with highest mean.

VIII. RANKING OF STOCHASTIC IMAGES

The various stochastic images can be ranked according to a criterion relevant to their usage. If the attribute value is a pollutant concentration, one may want to rank the stochastic images according to their global mean concentration. If the concentration z(u) is weighted by a "criticality" factor c(u), with c(u) high in critical zones such as playgrounds and c(u) low in less critical zones such as fenced industrial yards, one may consider the stochastic images of the new variable c(u) **x z(u)** and rank them according to their global mean. Typically, such operation can be achieved with the help of GIS tools.

Figure 33a shows the histogram of the 50 simulated global mean ***EC,*** values. Figs. 33b and 33c show the corresponding two realizations with, respectively, the lowest and highest global ***EC,*** mean value.

IX. CONCLUSIONS

The aim of this study is not so much assessment of soil salinity but rather to present a typical geostatistical analysis of a data set representative of the diversity and complexity of data sets handled through GIS. There is much more to geographical (spatial) data analysis than performing elementary operations of overlay, merge, and split and then merely mapping data with somewhat arbitrary, eye-pleasing, spline algorithms. The data talk when their geographic interdependence is revealed; there is an essential third component to any two data values taken at two different locations in space or time-their relation is seen as a function of the separation vector linking these two locations. Pictorial and numerical models of patterns of space/time dependence allow us to go far beyond data locations into alternative (stochastic equiprobable) maps that depict the true complexity of the data while always preserving an assessment of uncertainty. Present GIS essentially fail to read between the lines of data.

When statistics is used, it is elementary statistics, which ignores data locations and the relation of data with space and/or time. It is suggested that the most robust geostatistical tools, as presented in this study, be made available to soil scientists and users of geographical information systems. There cannot be efficient data utilization without data interpretation and modeling. When data are distributed in space, such interpretation and modeling necessarily call for geostatistics.

REFERENCES

Bourgault, G., Marcotte, D., and Legendre, P. (1992). The multivariate (co)variogram as a spatial weighting function in classification methods. Math. Geol. 24, 463-478.

Deutsch, C. V., and Joumel, A. G. (1992). "GSLIB: Geostatistical Software Library and User's Guide," pp. 340. Oxford Univ. Press, London.

Dimitrakopoulos, R. (ed.) (1993). "Geostatistics for the Next Century," pp. 497. Kluwer, Academic, Dordrecht.

Isaaks, E. H., and Srivastava, R. M. (1989). "Introduction to Applied Geostatistics," pp. 561. Oxford Univ. Press, London.

Joumel, A. G. (1989). "Fundamentals of Geostatistics in Five Lessons. Short Course in Geology," Vol. 8, pp. 40. American Geophys. Union Press, Washington DC.

Joumel, A. G., and Alabert, F. (1988). "Focusing on Spatial Connectivity of Extreme-Valued Attributes: Stochastic Indicator Models of Reservoir Heterogeneities," SPE paper No. 18324. Soc. of Pet. Eng.

Joumel. A. G., and Huijbregts, Ch. J. (1978). "Mining Geostatistics," pp. 600. Academic Press, San Diego.

Joumel, A. G., and Xu, W. (1994). Posterior identification of histograms conditional to local data. *Math. Geol. 26, 323-360.*

Lesch, S. M., Strauss, D. J., and Rhoades, J. D. (1995a). Spatial prediction of soil salinity using electromagnetic induction techniques. I. Statistical prediction models: A Comparison of multiple linear regression and cokriging. *Water Resour. Res.* 31, 373-386.

Lesch, S. M., Strauss, D. J.. and Rhoades, J. D. (1995b). Spatial prediction of soil salinity using electromagnetic induction techniques. 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. *Water Resour. Res.* **31,387-398.**

Sneath, P. H. A., and Sokal, R. R. (1973). "Numerical Taxonomy-The Principles and Practice of Numerical Classification," pp. 573. Freeman, San Francisco.

Soares, A. (ed.) (1993). Geostat Troia 1992. In "Proceedings of the 4th Geostatistical Congress," Vols. l-2, pp. 1088. Kluwer, Academic, Dordrecht.

Srivastava, R. M. (1992). "Reservoir Characterization with Probability Field Simulation," SPE paper No. 24753. Soc. of Pet. Eng.