Comparing Stochastic Simulation Algorithms by Measures of Local Accuracy and Precision

A THESIS
SUBMITTED TO THE DEPARTMENT OF PETROLEUM ENGINEERING
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

by
Yexiang Mo
May 1997

Approved for the Department:

Clayton V. Deutsch
(Advisor)

Acknowledgments

The author wishes to express his gratitude to Professor Clayton V. Deutsch for suggesting this research topic and for his continuous guidance, encouragement throughout this study. I would also like to thank Dr. Xian-Huan Wen for his constant help during the course of this study.

Financial support from the Stanford Center for Reservoir Forecasting (**SCRF**) is gratefully knowledged

Abstract

There are a large number of geostatistical approaches to model the spatial variability of petrophysical properties. For a given problem, different approaches may be appropriate depending on the goals of the study. An important question occurs: among many geostatistical methods, which one is the "best" for a particular problem at hand?

The conventional approach to test an estimation algorithm is to consider cross validation and look at summary statistics such as mean absolute error, mean squared error, or the correlation coefficient between the estimated values and the true values. The cross plot of the true values versus the estimated values reveals the "goodness" of the estimation technique.

In the context of stochastic simulation, these measures are on longer relevant since we have a distribution of simulated values. So, an approach that can measure the "goodness" of the multiple realizations is needed. This study presents a new approach based on the "leave-one-out" cross validation method. The first step is to build multiple realizations at the location where a true value is temporarily removed. Secondly, a local conditional cumulative distribution function (**ccdf**) model is built and four measures, i.e., accuracy(A), precision (P), goodness(G) and uncertainty(U) are specifically designed to establish the "goodness" of this local **ccdf** model. The advantages and weaknesses of these new measures are demonstrated by comparing them to the conventional cross validation tools. Furthermore, this study describes the methodology to understand strengths and shortcomings of different geostatistical simulation approaches and to assess them in relative accuracy, precision, and uncertainty.

Contents

List of Figures

Chapter 1

Introduction

The primary objective of the application of geostatistical tools for reservoir modeling is to build numerical geological models of lithofacies, porosity and permeability. There are many geostatistical methods that could be used to realize this objective. Estimation techniques such as kriging and cokriging can generate estimates with minimum local error variance. These estimates, however, may not be suitable when taken all together, that is:

- Local details tend to be smoothed out; small values are overestimated while large values are underestimated. This is a serious shortcoming when the aim is to detect patterns of extreme attribute values such as zones of high permeability.
- The smoothing effect varies spatially; it is less near the data locations and increases as the location being estimated gets farther away from data locations. As a result, a kriged map appears more variable in densely sampled areas than in sparsely sampled areas.

Compared to any interpolation algorithm, simulation has two major advantages[1].

- Estimation techniques provide unique local estimates of the attribute without considering the spatial statistics of the estimated values. Simulation provides alternative global representations where reproduction of patterns of spatial continuity prevails, i.e., reproduction of global features and statistics such as the histogram and variogram take precedence over local accuracy.

- Kriging cannot provide information on the joint uncertainty of several locations considered together. Simulation is specifically designed to provide such measures of uncertainty. These measures are given by the differences between L alternative simulated realizations.

Different simulation algorithms impart different global statistics and spatial features to the realizations. The practitioner must choose from the variety of available algorithms. A further complication is that there is no " best-for-all-cases " simulation algorithm but rather a toolbox of alternative algorithms from which to choose the method best suited to the problem at hand. According to Deutsch [7], three criteria can be used to select an appropriate simulation algorithm:

- the human and CPU time required to generate a set of realizations,
- the amount of relevant information that can be accounted for, and
- the precision and accuracy of generated distributions of uncertainty

Indicator-based (**SISIM**) and simulated annealing algorithms (**SASIM**) require greater CPU cost than Gaussian-based algorithms(**SGSIM**); however, this may be offset by their greater flexibility to incorporate additional types of data. Sequential Gaussian simulation is fast and straightforward, in that the modeling of the distribution of uncertainty at a location calls for solving a single kriging system. However, the implicit assumption of a multiGaussian RF model may be inappropriate if the structural analysis indicates that extreme values have a high degree of correlation.

In reservoir modeling practice, a hybird approach is considered to generate realizations that reflect widely different of features. The typical steps to generate a reservoir model may be descibed as[4]:

- Start with an object-based modeling approach or cell-based indicator simulation method to generate the geometric architecture of the various lithofacies.
- Apply cell-based geostatistical algorithms such as **SGSIM, SISIM**, or **SASIM** to simulate the distribution of continuous petrophysical properties within each lithofacies.

- Finally, iterative or simulated annealing methods can be used to modify local petrophysical properties to match additional well and production data.

The focus in this report will be on the second step, that is, the modeling of continuous petrophysical properties. In particular, the focus will be on the third criteria, i.e. assessing of the accuracy and precision of the different algorithms that could be used to model continuous petrophysical properties.

To compare different geostatistical algorithms, we use the "leave-one-out" cross validation approach. In a conventional cross validation exercise, the estimation method is tested at the locations of existing sample, that is, the sample value at a particular location is temporarily discarded from the sample data set. The value at that location is estimated with the remaining samples. This procedure is then repeated for all available samples. The estimates are then compared to the true sample values. Such a comparison typically falls short of clearly indicating the best alternative and does not provide any information on uncertainty. Therefore, we need to consider additional methods to predict uncertainty at each data location and check the "goodness"of those estimates of uncertainty.

A 2D exhaustively sampled porosity data set will be used to test different stochastic simulation algorithms. Based on cross validation, four new measures: **Accuracy, Precision, Goodness, Uncertainty** are considered to measure and compare the local accuracy and precision of different methods. Given L stochastic realizations at each left-out data location, we generate a probability distribution and define symmetric probability intervals of varying width p. The acuracy and precision are based on the actual fraction of true values falling within those intervals. In general, accuracy refers to the ultimate excellence of the data or computed results. Precision is a measure of the narrowness of a probability distribution and refers to the repeatability or refinement of a measurement or computed result. Uncertainty represents the spread of the distribution[3]. Normally, we expect a geostatistical algorithm to be both accurate, precise, and with smallest possible uncertainty.

In addition to performing the comparison with simulation algorithms, we also use estimation techniques (inverse distance, kriging and cokriging) where the traditional

statistics such as mean absolute error, mean squared error are available to measure the goodness of the estimates.

The variogram model for the comparative study is taken from the reference data. The comparison is fair because the reference variogram model has been used by all algorithms. The purpose of this report is to compare different geostatistical methods rather than address problems of variogram inference in the presence of sparse data.

Also, geostatistical simulation methods should integrate all available data into the geological model. In practice, primary data or "hard" data are usually limited, but they are supplemented by a relatively large number of soft data such as seismic data. A cosimulation approach could be used to integrate these soft data. Sequential Gaussian simulation with collocated cokriging (**COLSGSIM**) or sequential indicator simulation with Markov-Byes model for coregionalization (**MBSIM**) are two alternatives.

In general, no single simulation or estimation algorithm is flexible enough to handle the reproduction of the wide variety of features and statistics encountered in all practical problems. Nevertheless, such comparison, will provide information that is necessary to help practitioners select the appropriate algorithm for a particular situation.

Chapter 2

Exploratory Geostatistical Analysis

2.1 Well Data and Seismic Data

The data used in this case study are from a synthetic eolian sandstone constructed by U.S. Wind Tunnel Laboratory[3]. The area of interest is a 2-D rectangular area 100 dimensionless units long in the horizontal direction and 40 dimensionless units high in the vertical direction. The porosity and seismic data are available at every pixel. While the usual objective of geostatistics is to create petrophysical property models that honor the histogram and variogram of well data together with seismic data, the objective here is to compare different estimation and simulation algorithms. This is possible with the exhaustively sampled data set. To compare the "goodness" of simulation techniques by using a secondary variable, a 2-D "seismic" data was generated by using non-conditional sequential Gaussian simulation with the collocated cokriging alternative. For simplicity, the synthetic seismic data has same resolution as porosity data. Figure 2.1 shows the spatial images of porosity and seismic data. As we can see, most of the low values are distributed among the top and bottom of the layer. It is unnecessary to decluster these data because they are on a regular grid. The cross-plot between porosity and seismic data is given on the bottom of Figure 2.1. The correlation coefficient is about 0.60, which is typical of practice. The histogram of the porosity data is shown on the left side of Figure 2.2. No extremely large or low values are found. Basic statistics of porosity data are listed below:

Number of Data	Mean	Median	Variance	Coef. of variance
4000	0.2111	0.2020	0.1212	0.5742

2.2 Variography

2.2.1 Variogram

The reference porosity data were transformed to a standard normal distribution. The variogram map of the transformed porosity data in Figure 2.3 clearly indicates that the horizontal direction has the greatest continuity and the vertical has the least. The experimental variogram in these two directions are shown on Figure 2.3. The variogram was fitted with the following anisotropic model:

$$\gamma(\mathbf{h}) = 0.02 + 0.76Sph\sqrt{(\frac{h_x}{25.0})^2 + (\frac{h_z}{3.0})^2} + 0.22Sph\sqrt{(\frac{h_x}{\infty})^2 + (\frac{h_z}{6.0})^2} \qquad (2.1)$$

The fitted variogram model is shown by solid lines in Figure 2.3. From the Figure, we find that the experimental points in the horizontal direction are matched very well while the vertical variogram displays more noise. Horizontal stratification causes the variogram in the horizontal direction not to reach the theoretical sill of 1.

2.2.2 Indicator Variograms

Figure 2.5 and Figure 2.6 show the experimental vertical and horizontal indicator variograms based on the 5 thresholds listed below:

Probability Values ($F(z)$)	0.1	0.3	0.5	0.7	0.9
Thresholds (z)	0.055	0.125	0.202	0.282	0.396

The indicator variogram models are :

$F(z) = 0.1$;

$$\gamma(\mathbf{h}) = 0.02 + 0.056Sph\sqrt{(\frac{h_x}{18})^2 + (\frac{h_z}{2})^2} + 0.022Sph\sqrt{(\frac{h_x}{60})^2 + (\frac{h_z}{50})^2} \qquad (2.2)$$

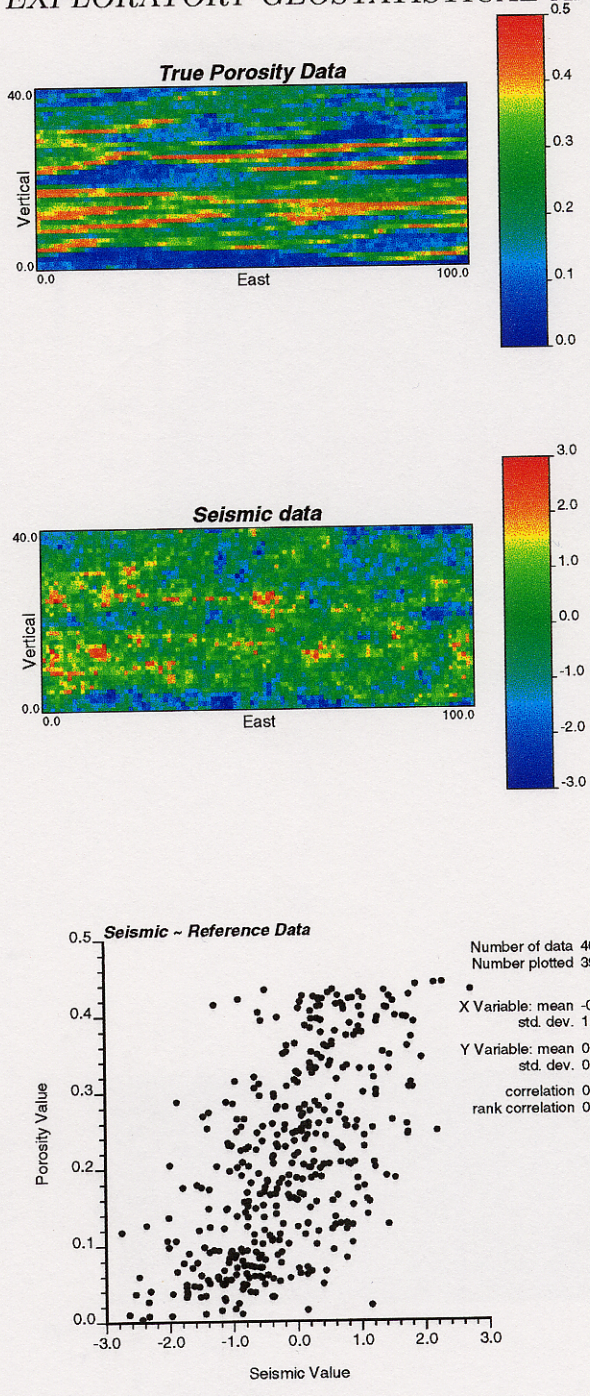


Figure 2.1: Location map of the reference porosity data, seismic data, and a scaterplot of the porosity and seismic data(only 10 percent of the pixels are shown so that the plot is not obscured by the bullet size)

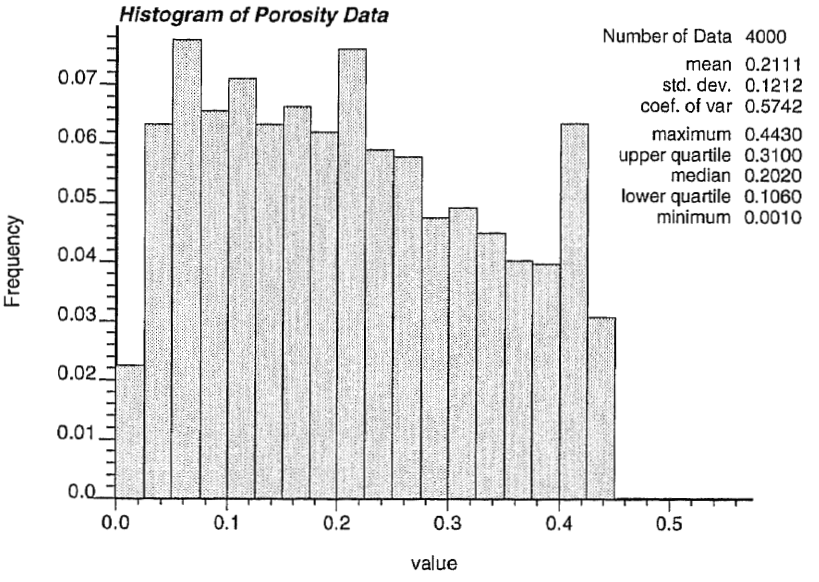


Figure 2.2: Histogram of the reference data. Note no extremely large or low values are found

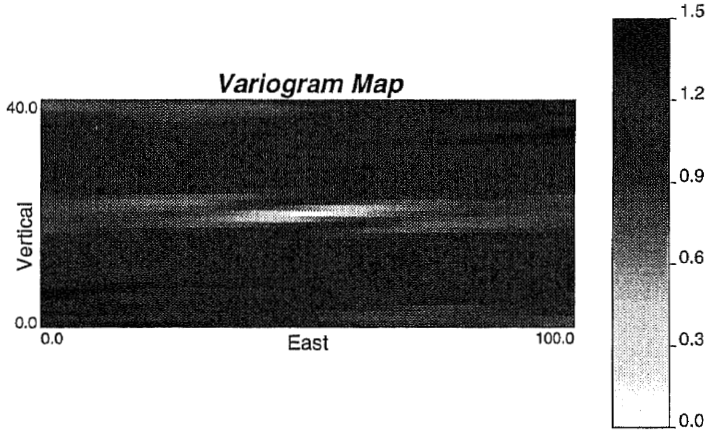


Figure 2.3: Variogram map of the reference data. Note the maximun and minmun continuity along the horizontal and vertical direction, respectively

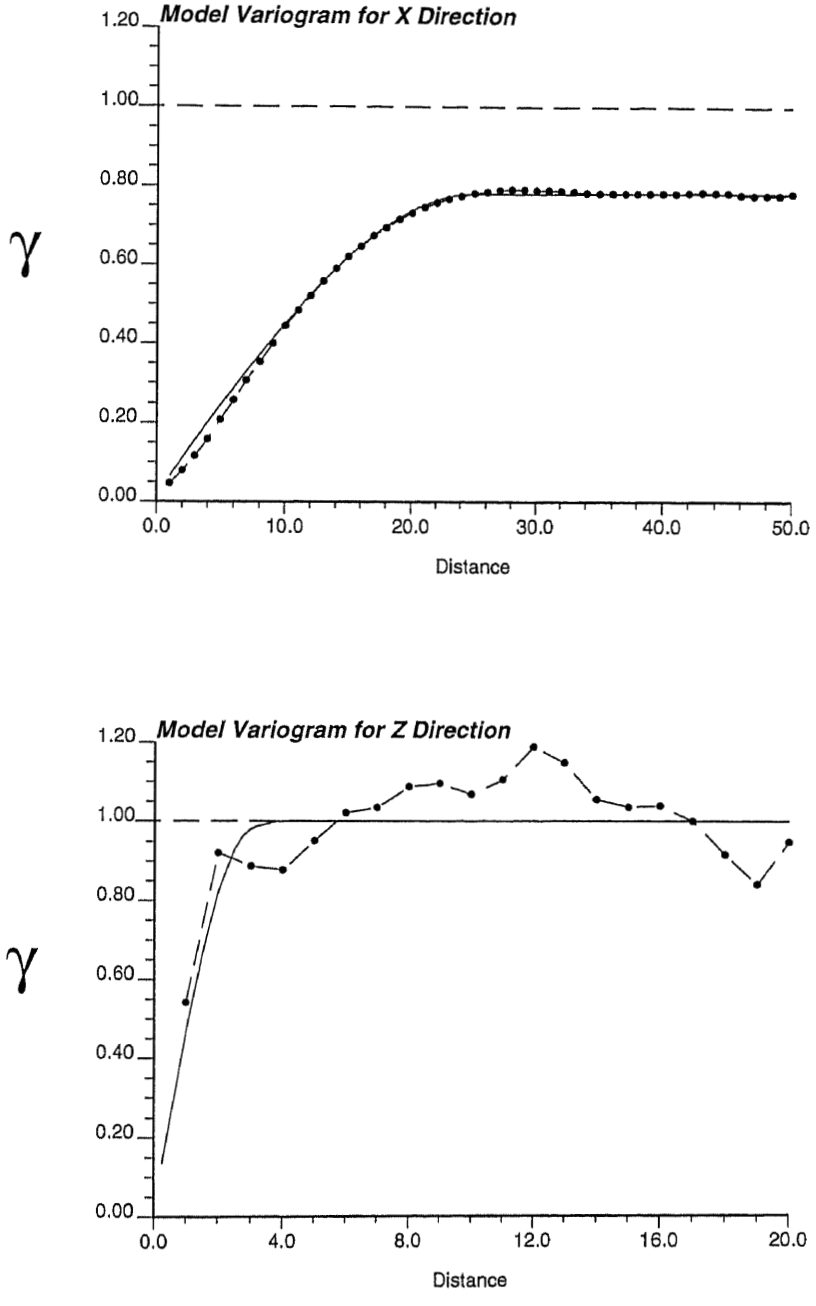


Figure 2.4: Horizontal and vertical normal score variograms and anisotropic model fitted to the experimental points

$F(z) = 0.3$;

$$\gamma(\mathbf{h}) = 0.03 + 0.12Sph\sqrt{(\frac{h_x}{22})^2 + (\frac{h_z}{2})^2} + 0.06Sph\sqrt{(\frac{h_x}{80})^2 + (\frac{h_z}{11})^2} \quad (2.3)$$

$F(z) = 0.5$;

$$\gamma(\mathbf{h}) = 0.04 + 0.19Sph\sqrt{(\frac{h_x}{24})^2 + (\frac{h_z}{2})^2} + 0.02Sph\sqrt{(\frac{h_x}{\infty})^2 + (\frac{h_z}{10})^2} \quad (2.4)$$

$F(z) = 0.7$;

$$\gamma(\mathbf{h}) = 0.02 + 0.16Sph\sqrt{(\frac{h_x}{20})^2 + (\frac{h_z}{4})^2} + 0.03Sph\sqrt{(\frac{h_x}{\infty})^2 + (\frac{h_z}{5})^2} \quad (2.5)$$

$F(z) = 0.9$;

$$\gamma(\mathbf{h}) = 0.01 + 0.05Sph\sqrt{(\frac{h_x}{11})^2 + (\frac{h_z}{1})^2} + 0.03Sph\sqrt{(\frac{h_x}{50})^2 + (\frac{h_z}{4})^2} \quad (2.6)$$

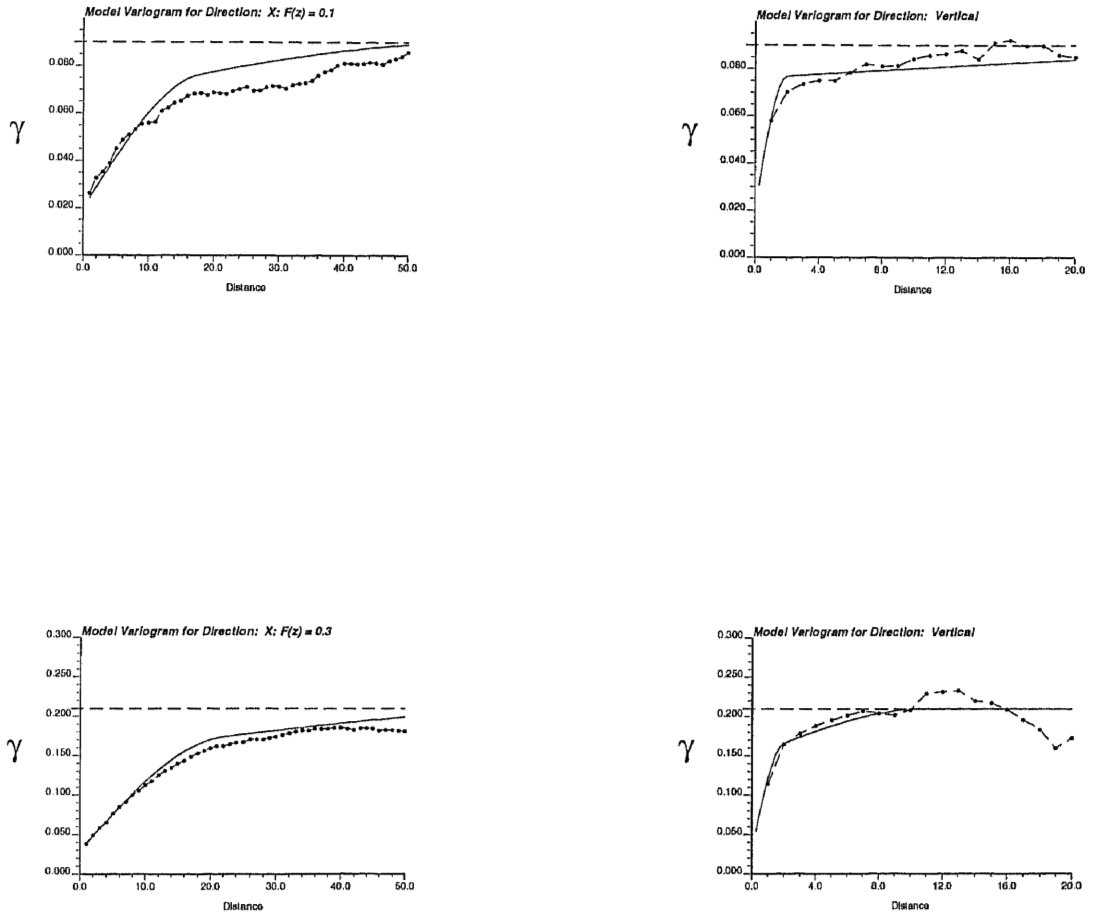


Figure 2.5: Vertical and horizontal indicator variogram of reference data and their fits by anisotropic models.

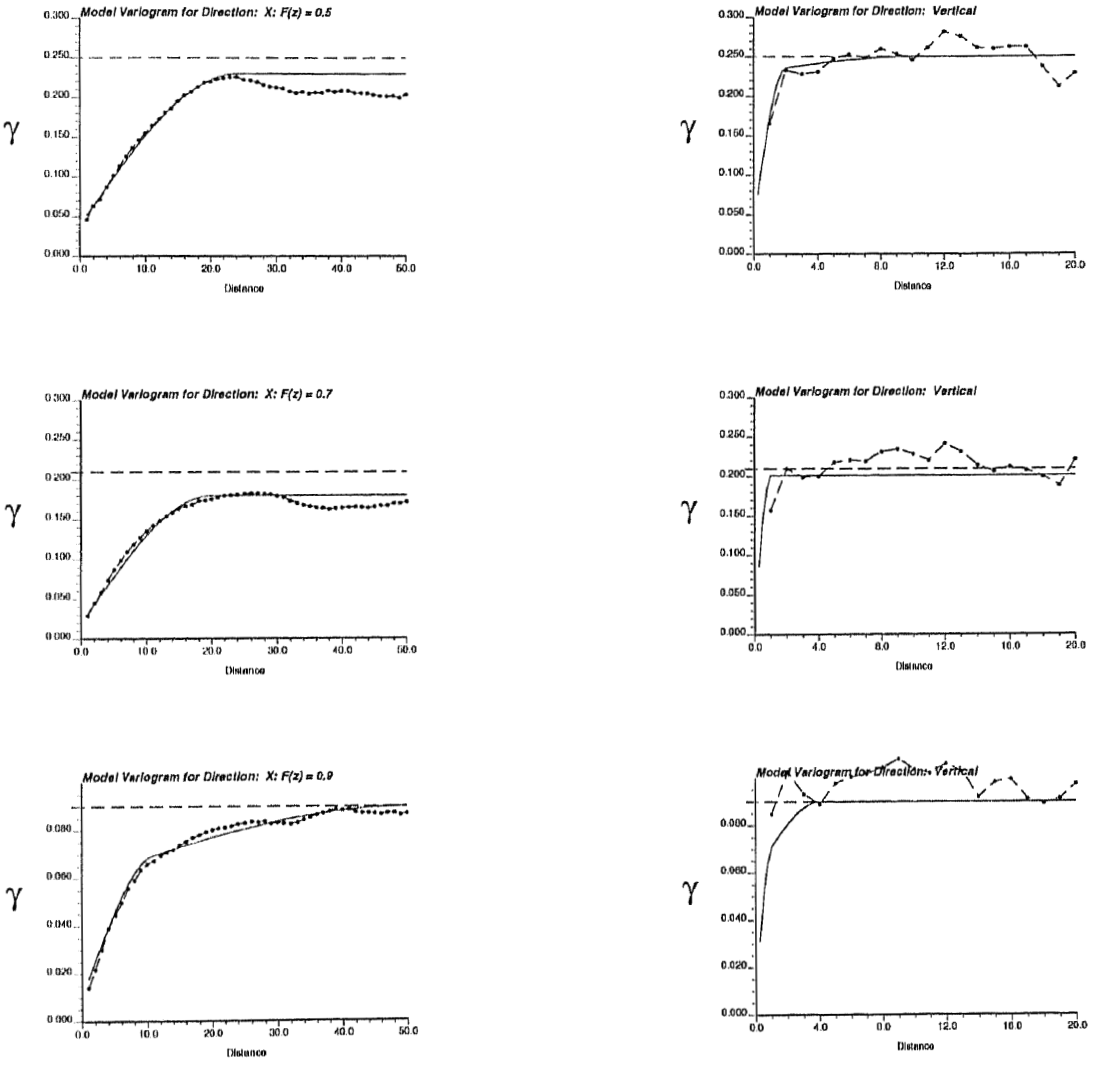


Figure 2.6: Vertical and horizontal indicator variogram of reference data and their fits by anisotropic models(continued).

Table 1 lists the parameters values in above indicator variogram models. Note that all five models are of the same type:

first structure + second structure								
k	F(z)	$C_0(z_k)$	$C_1^1(z_k)$	a_x^1	a_z^1	$C_1^2(z_k)$	a_x^2	a_z^2
1	0.1	0.220	0.622	18	2	0.222	60	50
2	0.3	0.143	0.571	22	2	0.286	80	11
3	0.5	0.160	0.760	24	2	0.080	∞	10
4	0.7	0.095	0.762	20	4	0.143	∞	5
5	0.9	0.111	0.555	11	1	0.333	50	4

2.2.3 Checking BiGaussianity

Bivariate normality is an important assumption in Gaussian related algorithms. In fact, **SGSIM** assumes all multipoint distributions are Gaussian. We can check the bivariate Gaussian assumption, unfortunately, the Gaussianity of three-point,..., multipoint experimental distributions are difficult to check in practice. To check bivariate Gaussian model, we compare the two-point experimental indicator variograms and theoretical model at different indicator thresholds given by the program **BIGAUS**[1]. Figure 2.7 and Figure 2.8 show that the experimental indicator variograms do not invalidate the BiGaussian assumption. Another simple way to check for bivariate Gaussian model is to check for the symmetric "destructuration" of the extreme porosity values: The practical ranges of the indicator variogram should decrease symmetrically and continuously as quantile $F(z)$ tends toward its bound values of 0 and 1. This fact was approximately shown by the ranges listed in Table 1. Although the fits may be considered reasonable, we should notice the indicator variograms are systematically slightly lower than the theoretical indicator variogram model. This may be a reason to consider the indicator formalism.

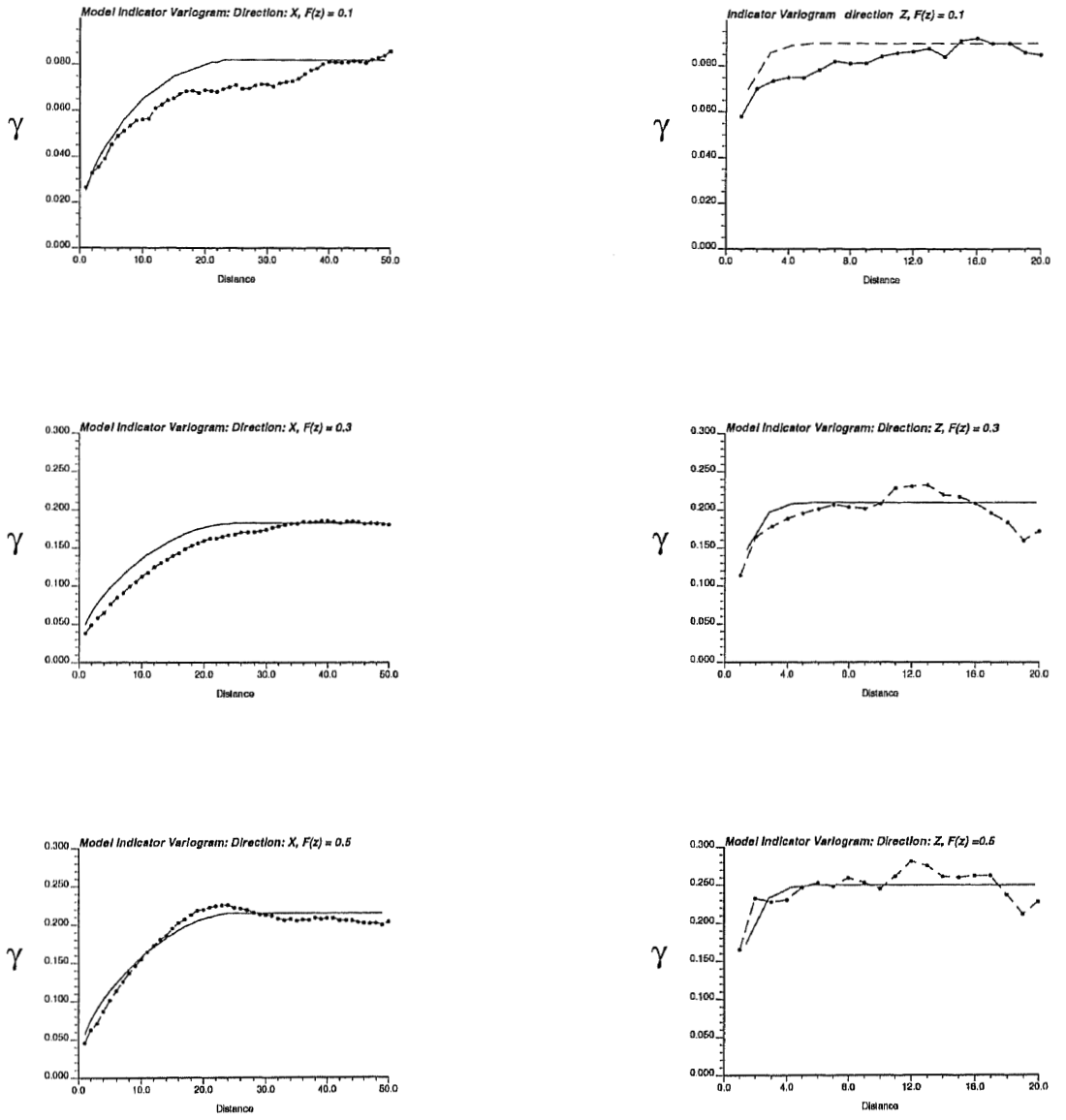


Figure 2.7: Experimental indicator semivariogram and their fits by Gaussian RF model-derived theoretical curves.

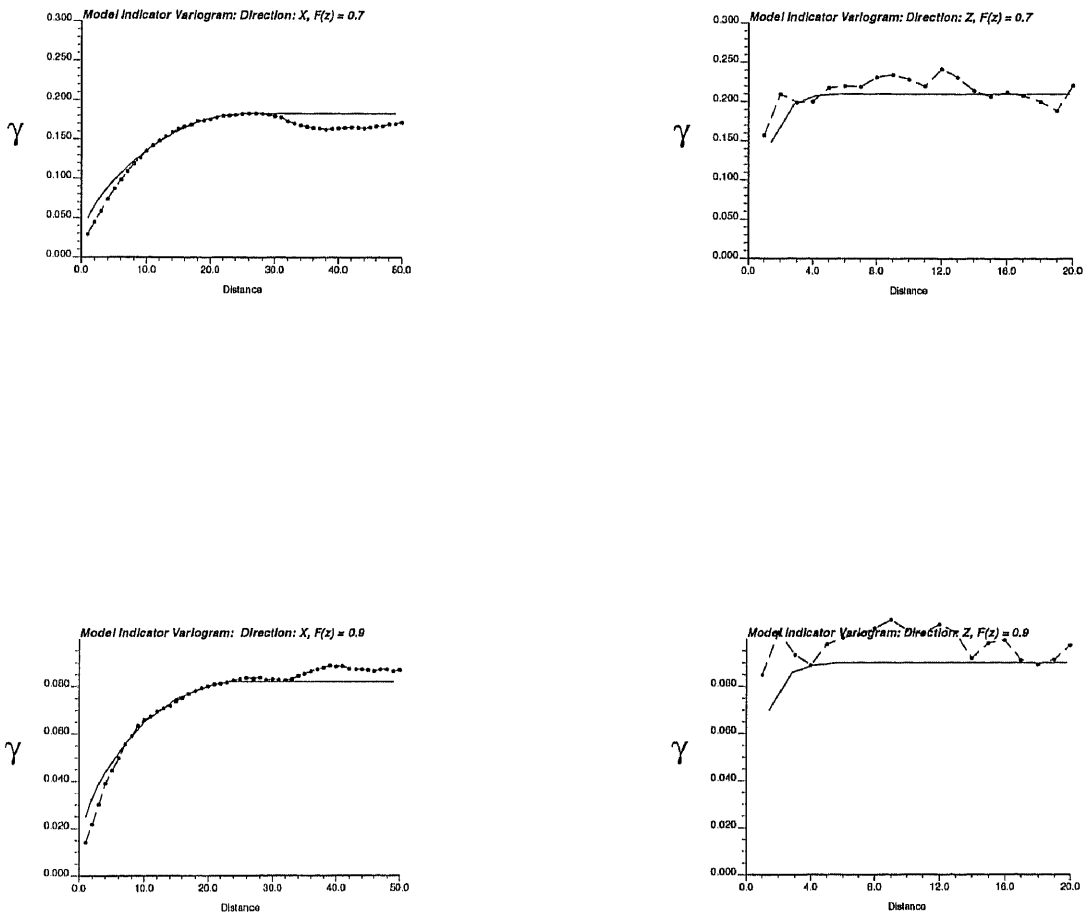


Figure 2.8: Experimental indicator semivariogram and their fits by Gaussian RF model-derived theoretical curves.

2.2.4 Cross Variogram

Figure 2.9 gives the cross-variogram between the porosity and seismic data. The variogram was fitted with the following anisotropic model:

$$\gamma(\mathbf{h}) = 0.01 + 0.40Sph\sqrt{(\frac{h_x}{25})^2 + (\frac{h_z}{3.0})^2} + 0.22Sph\sqrt{(\frac{h_x}{\infty})^2 + (\frac{h_z}{6.0})^2} \quad (2.7)$$

This linear coregionalization model is only used in the cokriging estimation technique. To handle the seismic data in simulation, a full indicator cokriging approach would be quite demanding in that a matrix of cross variogram models must be inferred. In later chapters, we use the Markov hypothesis to avoid this tedious inference.

The cross variogram model is constructed using the same basic variogram models as auto-variogram models. In this case, both auto-variogram and cross variogram are constructed using spherical model with two structures. The coefficients were chosen to ensure a licit linear coregionalization model.

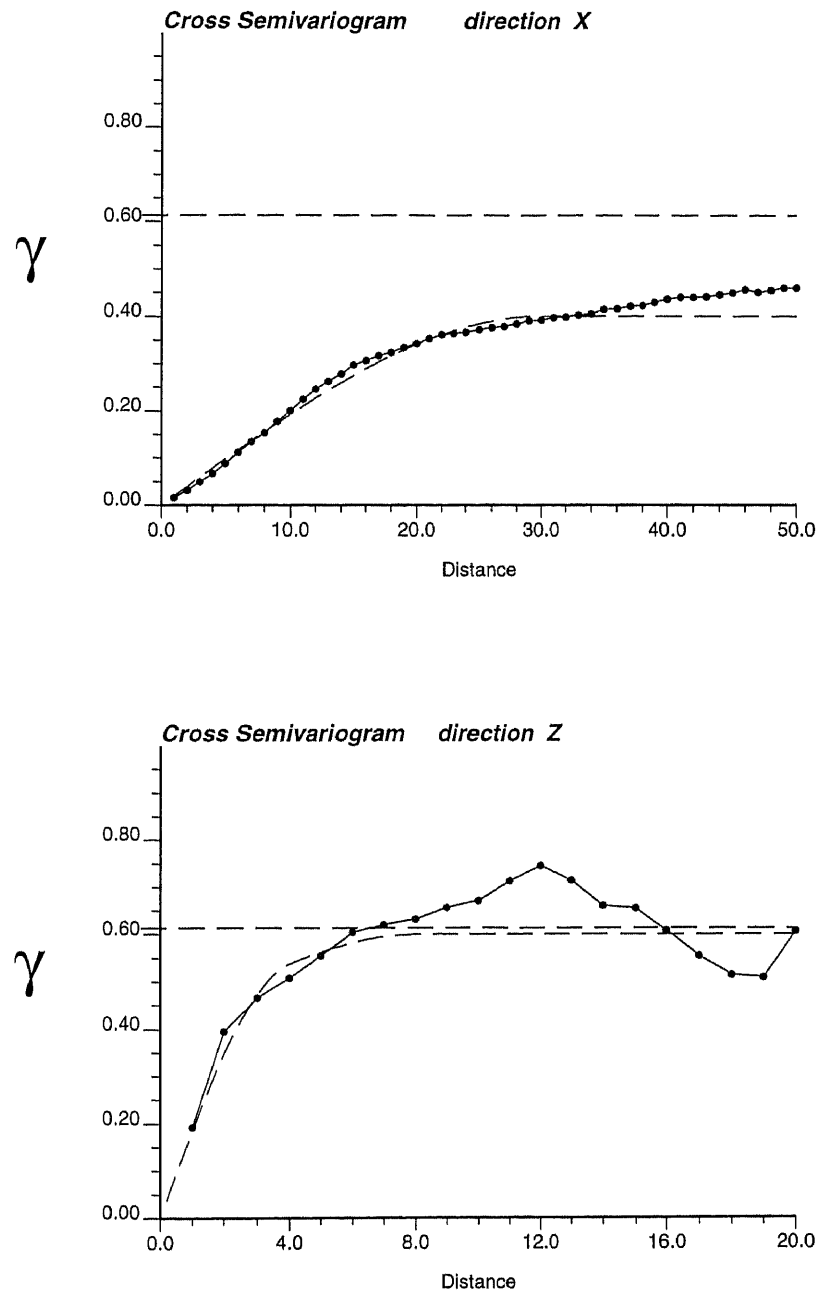


Figure 2.9: Cross variogram between porosity and seismic data and their model fits

Chapter 3

Estimation Techniques

Most estimation techniques consider a linear weighted combination of data within a local neighborhood, that is:

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})[Z(\mathbf{u_i}) - m(\mathbf{u_i})] \quad (3.1)$$

where $Z^*(\mathbf{u})$ is the estimated value at location $\mathbf{u}$, $\mathrm{n}(\mathbf{u})$ is the number of data in the neighborhood of $\mathbf{u}$, and $\lambda_i(\mathbf{u})$ are the weights applied to the data values $Z(\mathbf{u_i})$, $i = 1, ..., n(\mathbf{u})$. $m(\mathbf{u})$ is the expected value of **RV** $Z(\mathbf{u})$. In the presence of large scale trends, the trend is modeled first and the Z data are residual values from that mean surface. Here, we only present two traditional estimation techniques, i.e, inverse distance estimatè and kriging estimate.

3.1 Inverse Distance Method

The principle of inverse distance methods is straightforward, i.e., make the weights assigned to sample data inversely proportional to the distance between the data and the location being estimated. The distance is raised to an exponent, p, to impart different levels of smoothness to the estimates. The estimator:

$$Z^*(\mathbf{u}) = \frac{\sum_{i=1}^{n(\mathbf{u})} \frac{Z(\mathbf{u_i})}{d_i^p}}{\sum_{i=1}^{n(\mathbf{u})} \frac{1}{d_i^p}} \quad (3.2)$$

A common choice for the inverse distance exponent is 2(used in this thesis). As we decrease the exponent, the weights given to the samples become more similar. On the other hand, if we increase the exponent, the individual weights become more dissimilar, i.e. the farthest samples receive a smaller proportion of the total weight, while the nearest samples become more influential. Another aspect we should mention here is a 10:1 horizontal to vertical anisotropy ratio was used in the distance calculations. This is clearly displayed in the map of **ISD** estimate.

In Figure 3.1c we show the estimated point values using the 80 conditioning data, which are located at left and right side of the reservoir. The estimated values at the middle locations remain almost constant. This is because all the conditioning data get almost equal weights as the distance between the conditioning data and estimated location increases. The histogram of the **ISD** errors in Figure 3.1a shows a symmetric distribution. A slight positive mean (0.01) may reflect a general tendency toward overestimation. Another feature we need to check is the spread of the error distribution. Ideally, we would like the error distribution to have minimum spread and a mean close to 0. But, in practice, these two objectives are not independent and we have to trade one off against the other. Algorithms for incorporating both characteristics are introduced in Section 3.4.

Figure 3.1b shows the scatterplot of the true values and **ISD** estimates. Conditional biasedness, i.e., overestimation of low values and underestimation of high values, can be observed from this scatterplot. Figure 3.1d is the absolute error map of **ISD**. As we can see, the errors are most significant near the high and low values. Another important issue is to check the errors generated by any procedure have no spatial correlation. The variogram of the error values confirms this fact, i.e., variogram shows a pure nugget effect(not shown here).

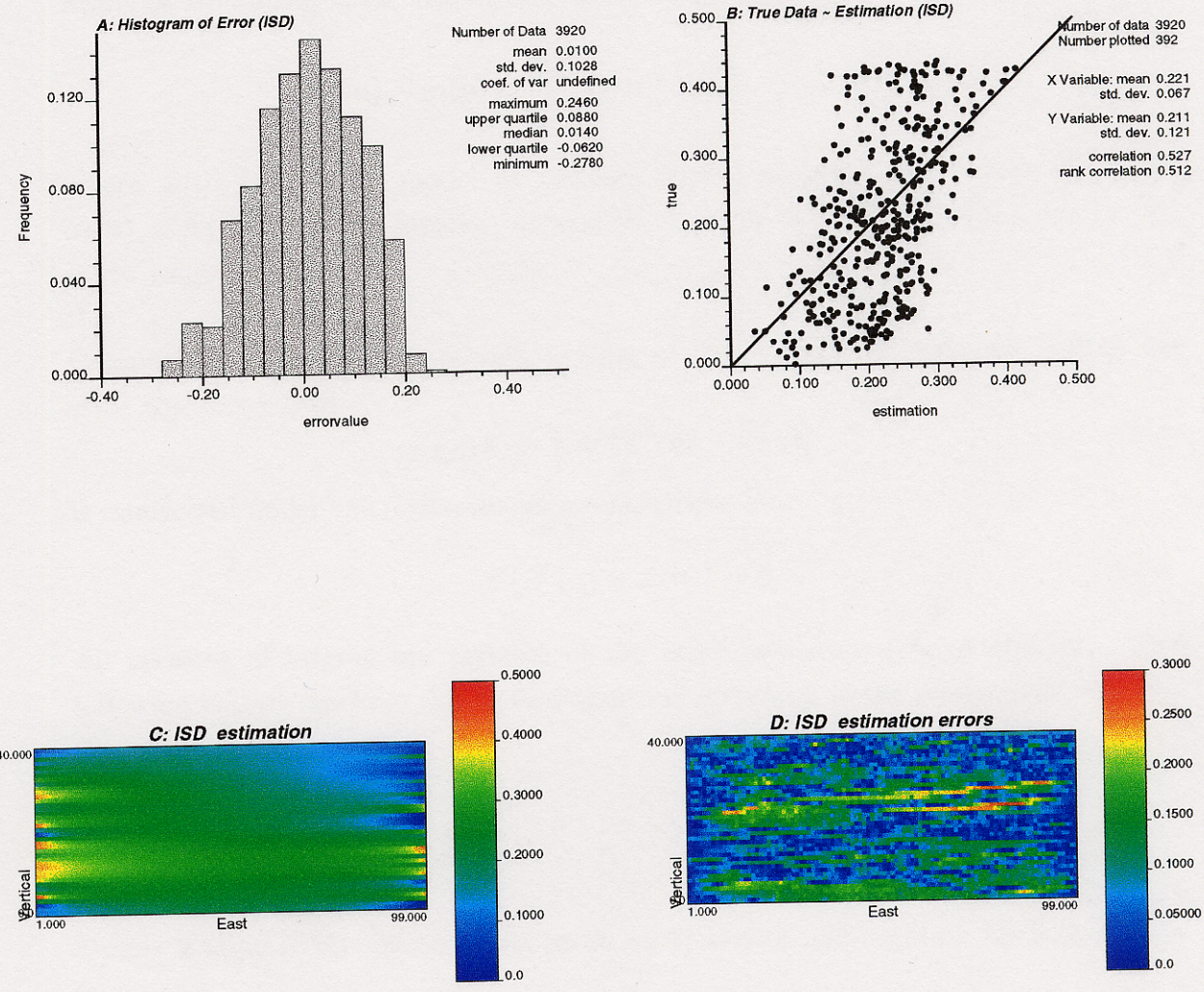


Figure 3.1: Cross validation of inverse square distance methods

3.2 Kriging

To estimate the value of an attribute z at any unsampled location $\mathbf{u}$ using only data available over study area, kriging is a favorable method:

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u})[Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)] \quad (3.3)$$

where $Z(\mathbf{u})$ is the **RV** model at location $\mathbf{u}$, the $\mathbf{u}_\alpha$'s are the n data locations, $m(\mathbf{u}) = E\{Z(\mathbf{u})\}$ is the location-dependent expected value of **RV** $Z(\mathbf{u})$, and $Z^*(\mathbf{u})$ is the kriging estimator.

Compared to above **ISD** methods, kriging is usually a better methods in that the error variance $\sigma_E(\mathbf{u})$

$$\sigma_E(\mathbf{u}) = Var\{Z^*(\mathbf{u}) - Z(\mathbf{u})\} \quad (3.4)$$

is minimized under the constraint of unbiasedness, i.e

$$E\{Z^*(\mathbf{u}) - Z(\mathbf{u})\} = 0 \quad (3.5)$$

All versions of kriging are variants of the basic estimator defined above[1]. The following kriging methods can be distinguished according to the model considered for the mean $m(\mathbf{u})$.

3.2.1 Simple Kriging

Simple kriging (**SK**) considers the mean $m(\mathbf{u})$ known and constant throughout the study area.

$$Z^*_{\mathbf{SK}}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{\mathbf{SK}}(\mathbf{u}) Z(\mathbf{u}_\alpha) + [1 - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{\mathbf{SK}}(\mathbf{u})]m \quad (3.6)$$

where $\lambda_\alpha^{\mathbf{SK}}(\mathbf{u})$ is the weight associated to the data in the **SK** estimates. m is the global mean.

Figure 3.2c shows a map of the **SK** estimates. Note that the **SK** estimates:

- Identify the conditioning data (exactitude property).

- Become closer to the mean when the estimated location is far away from the conditioning data. At locations beyond the correlation range (25 units), the estimated values are equal to the mean (0.2111).

- The map shows the typical smoothing effect of kriging. However, as mentioned before, the smoothing effect is not uniform, i.e., smoothing is less close the conditioning data locations and increases as the location estimated gets farther away from conditioning data. As a result, the kriged map in Figure 3.2c appears more variable at locations next to the wells.

The scatterplot of the true values and **SK**estimates exhibits conditional biasedness. This is also seen in map of absolute error map in Figure 3.2d. The histogram of the **SK** errors in Figure 3.2a shows a symmetric distribution. Compared to the **ISD** error distribution, the mean is slightly lower. Pure nugget effect in error variogram confirms the fact that there is no any general trend in error distribution, i.e., the errors generated by **SK** at different locations are independent from each other.

3.2.2 Ordinary Kriging

Ordinary kriging (**OK**) allows one to account for local fluctuation of the mean by considering the mean constant within local neighborhood. The unknown local mean is filtered from the linear estimator by constraining the kriging weights to sum to 1[1].

$$Z^*_{\mathbf{OK}}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{\mathbf{OK}}(\mathbf{u}) Z(\mathbf{u}_\alpha) \text{ with } \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{\mathbf{OK}}(\mathbf{u}) = 1 \tag{3.7}$$

where $\lambda_\alpha^{\mathbf{OK}}(\mathbf{u})$ is the weight associated to the data in the **OK** estimates.

Figure 3.3c shows **OK** estimates of the porosity. Note that:

- like **SK**, **OK** identifies the conditioning data.

- According to [5], one can deduces the following relation between the **SK** and

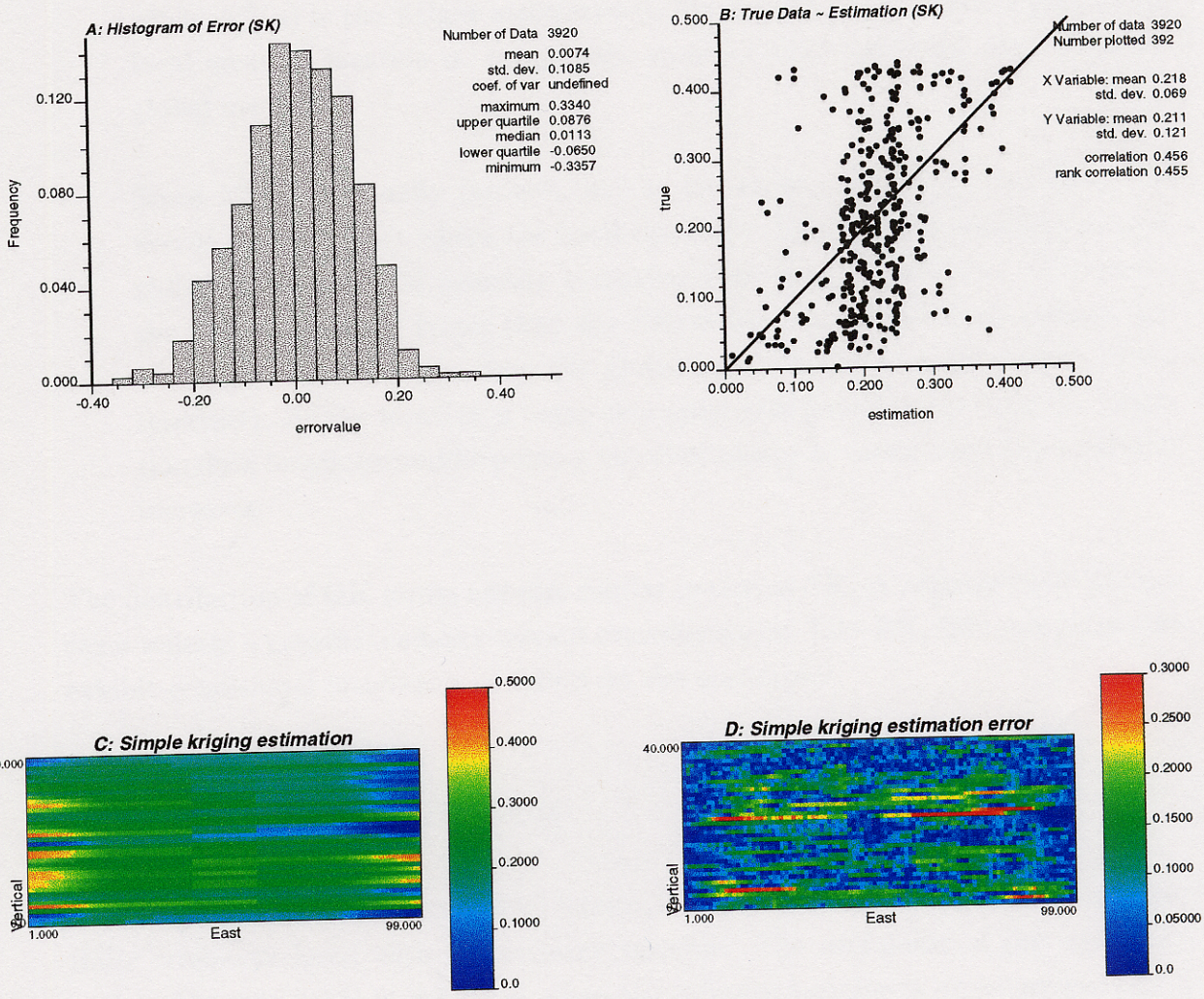


Figure 3.2: Cross validation of simple kriging estimation

OK estimators:

$$Z^*_{\mathbf{OK}}(\mathbf{u}) = Z^*_{\mathbf{SK}}(\mathbf{u}) + \lambda^{\mathbf{sk}}_m(u)[m^*_{\mathbf{OK}}(\mathbf{u}) - m] \tag{3.8}$$

where $\lambda^{sk}_m(u)$ is the weight associated to the data in the **OK** estimation of the local mean at location $\mathbf{u}$. $m^*_{\mathbf{OK}}$ is the mean of local condition data. m is the global mean

Since $\lambda^{\mathbf{sk}}_m(u)$ is usually positive, at the low-valued areas (the top right corner of the reservoir) where the local condition data mean is smaller than the global mean, the **OK** estimate is smaller than the **SK** estimate. Conversely, the **OK** estimate is larger than the **SK** estimate in high-valued areas where the local mean is larger than the global mean(the bottom left corner of the reservoir). As the location $\mathbf{u}$ being estimated gets farther away from the condition data locations(middle points), the discrepancy between these two estimates increases.

The distribution of **OK** errors appears similar to that of **SK**. A positive mean in both cases reflects a general tendency toward overestimation. Like **SK**, **OK** estimates also exhibit conditional biasedness as shown on the scatterplot.

3.3 Simple Cokriging

A cokriging approach is required to integrate the secondary variable which are spatially cross-correlated with the primary variable. The usefulness of the secondary variable(seismic data) is enhanced when the primary variable is undersampled(as is the case here). Simple cokriging **SCK** considers the local mean of every attribute known and constant within the study area[2]. Consider only one single secondary attribute z_2:

$$Z^{(1)*}_{\mathbf{SCK}}(\mathbf{u}) - m_1 = \sum_{\alpha_1=1}^{n_1(\mathbf{u})} \lambda^{\mathbf{SCK}}_{\alpha_1}(\mathbf{u})[Z_1(\mathbf{u}_{\alpha_1}) - m_1] + \sum_{\alpha_2=1}^{n_2(\mathbf{u})} \lambda^{\mathbf{SCK}}_{\alpha_2}(\mathbf{u})[Z_2(\mathbf{u}_{\alpha_2}) - m_2] \tag{3.9}$$

where λ_{α_1} 's are the weights applied to the n_1 primary data and λ_{α_2} 's are the weights applied to the n_2 secondary data. m_1 and m_2 are the global mean for primary data

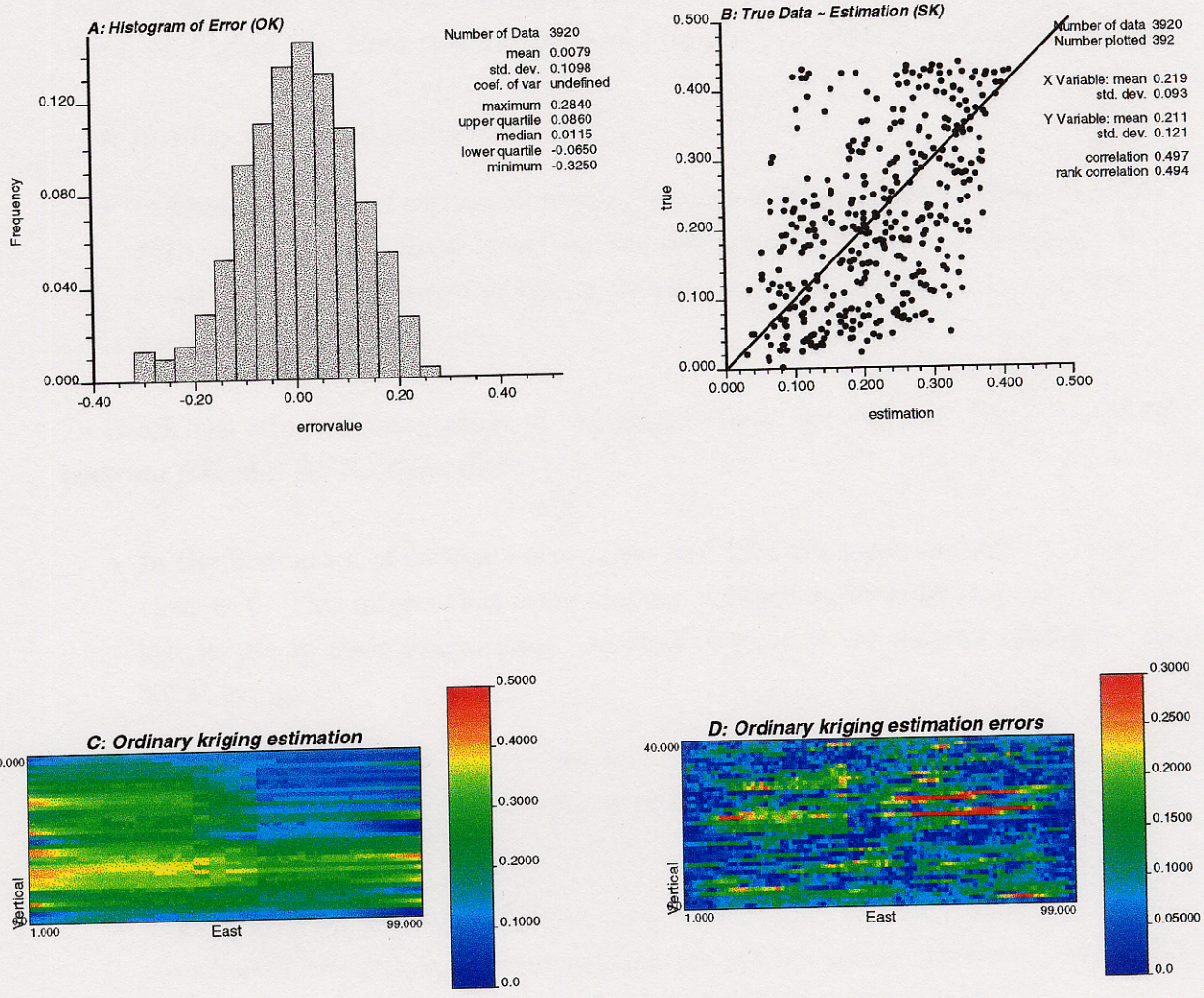


Figure 3.3: Cross validation of ordinary kriging estimation

and secondary data, respectively. $Z^{(1)*}_{\mathbf{SCK}}(\mathbf{u})$ is the **SCK** estimator of the primary attribute z_1.

Figure 3.4c shows the **SCK** estimates using the linear model of coregionalization shown on Figure 2.9. Note this model should be positive definite and the estimates:

- Identify the conditioning data
- In most locations of estimated points, because of the limit of conditioning hard data, estimation is done based on the secondary variable(seismic data) As a result, the spatial image in cokriging estimation is very close to the seismic data.

To check the effect of integration of the seismic data, we compare the estimations between **SK** and **SCK**. Note that:

- In the histogram of estimate errors, we find that the mean and the median are close to 0. This impies that overestimates balance underestimates and they are symmetric in their magnitudes. Another feature is the smaller spread in the **SCK** errors.
- The scatterplot of the true values and **SCK**, like **SK**, displays conditional biasedness. However, in **SCK**, the correlation coefficient(0.665) is higher than **SK**(0.455).
- Comparison between the absolute error maps shows similar result, i.e., overestimation of low values and underestimation of high values. However, we should notify the magnitudes of error is decreased significantly in **SCK**.
- no spatial correlation found in **SCK** error map.

In general, if there is a good correlation between primary and secondary variable, like the situation in here, the secondary variable plays a very important role in estimation, especially at the location when sparse/no neighborhood conditioning data available.

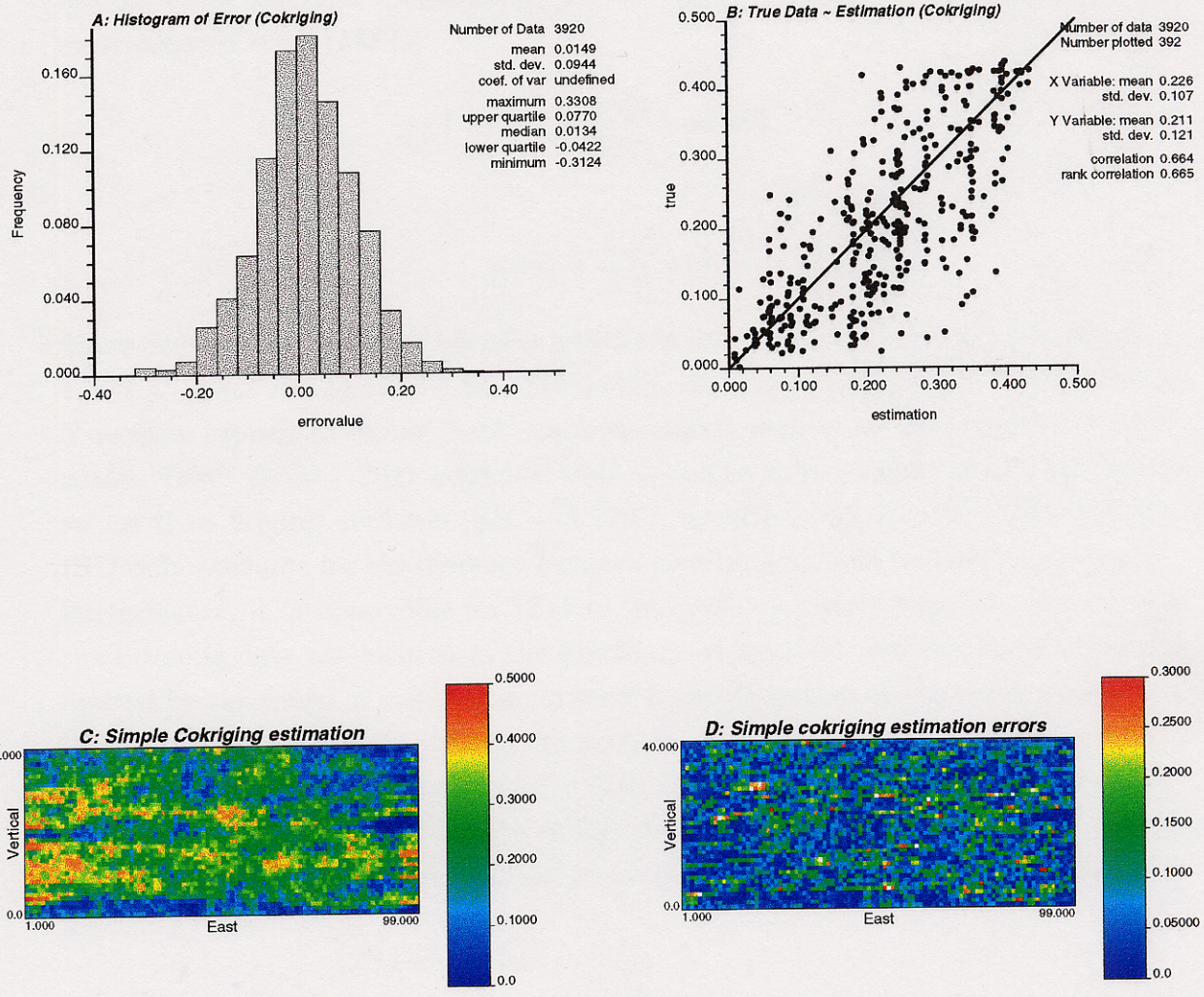


Figure 3.4: Cross validation of simple cokriging estimation

3.4 Comparison of Estimation Techniques

Traditionally, there are two summary statistics that incorporate both the bias and the spread of the error distribution, i.e., the mean absolute error (**MAE**) and the mean squared error (**MSE**):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid error \mid \tag{3.10}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} error^2 \tag{3.11}$$

where *error* is the difference between estimated value and true value at same location. Table 3.1 gives the **MSE**, **MAE** values and other statisticsfor above four methods. If decision process to choose 'best" methods mainly depend on these variables, **SCK** is the "best" choice. **ISD** methods, even seems to have smaller **MAE** and **MSE** values than kriging methods (**SK** and **OK**), usually is not a good choice because **ISD** only accounts for the distance between conditioning data and estimated points. Furthermore, it is impossible for **ISD** to integrate secondary variable. Because all conditioning data are located at the boundary of reservoir and the search radii are limited by the range of the variogram model, the departure of **OK** local mean from **SK** global mean is not significant. As a result, the difference between **SK** and **OK**, in this particular case, is quite small. Often **OK** is preferred to **SK** since it requires neither knowledge nor stationarity of the mean over the entire area. Finally, **SCK** is considered as the best estimation technique among these four methods due to following facts:

- primary variable(porosity data) is undersampled relative to secondary variable(seismic data). At the location where estimated points are far away from the conditioning porosity data, seismic data is the only information to decide the estimation values. In this case, the result form **SCF** almost reproduced all of the spatial characteristics in seismic data.

- There are two possible reasons that **SCK** may get identical result as **SK**: (1) primary and secondary variables are uncorrelated (2) the linear model of coregionalization is proportional to the primary variogram model. Fortunately, in

this case, we do not have such problems. So, the improvement brought by the large amount of seismic data is seen to be significant.

Methods	Error mean	Error median	Coef. of Correlation	MSE	MAD
ISD	0.0100	0.0140	0.527	0.011	0.0839
SK	0.0074	0.0113	0.456	0.011	0.0849
OK	0.0079	0.0115	0.494	0.012	0.0871
SCK	0.0149	0.0134	0.665	0.009	0.0736

3.5 Conclusions

There are two main disadvantages in above traditional cross-validation tools:

- If we need to use simulation to generate different realizations, no information about the "goodness" of uncertainty models can be assessed by these summary statistics.
- Even with estimation techniques, we find the above comparison falls short of clear indicating which method is best, i.e., a diagnostic tool that can give qualitatively information of comparison will be appreciated whenever the results from two methods are quite similar.
- If we increase the smoothness of the estimate, the relative **MSE** and **MAE** values should be smaller. Because , at this particular case, the estimate from **ISD** seems to be more smoother than **SK** and **OK**, it is not surprise to find **MSE** and **MAE** values in **ISD** are smaller than that of the **SK** and **OK**.

in next chapter, we discuss simulation algorithms and four new measures that is specifically designed to solve the problems in traditional cross-validation methods.

Chapter 4

Local Accuracy and Precision of Simulation

Unlike kriging, which only provides a unique estimate $z^*(\mathbf{u})$, the goal in simulation is to build alternative, equally probable models of the spatial distribution of $z^{(l)}(\mathbf{u}), \mathbf{u} \in A$,l= $1, ..., L$. In this chapter, we will introduce four new concepts to directly assess the accuracy and precision of the local distributions of uncertainty provided by simulation techniques. Then, based on these new concepts, we compare different simulation techniques using the same data as in chapter 3.

Once again, consider a cross validation exercise with the "leave-one-out" approach, the detail steps:

- Except the 80 values at left and right side of reservoir, all values(3920) in Figure 4.1a are temporarily discarded from the true data-set.
- Generate 100 realizations at each of these 3920 locations for each stochastic simulation approach(**SGSIM, COLSGSIM, SISIM and MBSIM**). These realizations only consider the 80 data values.
- Each set of 100 realizations provides a model of the conditional cumulative distribution function (**ccdf**) at that location[6]. Figure 4.1 shows some examples of the distribution functions.

The series of single-point ccdfs $F(\mathbf{u_i}; z|(n)), i = 1, ..., N$ do not provide any measure of "multiple-point" or spatial uncertainty such as the probability that z-values over the N locations $\mathbf{u_i}$ are jointly no greater than a critical threshold z_c, i.e.,:

$$Prob\{Z(\mathbf{u_i}) \leq z_c, i = 1, ..., N|(n)\} \neq \prod_{i=1}^{I} F(\mathbf{u_i}; z_c|(n)) \tag{4.1}$$

Where (n) is the set of data at the 2 conditioning well locations. This is why we refer to "local" uncertainty assessment rather than "spatial" uncertainty assessment.

Once we have the probability distribution at every location, we calculate the probabilities of the true values $z(\mathbf{u_i}), i = 1, ..., 3920$ using the predicted distributions of uncertainty[6];

$$F(\mathbf{u_i}; z(\mathbf{u_i}) \mid n(\mathbf{u_i})), i = 1, ..., 3920 \tag{4.2}$$

Define a symmetric p-probability interval(PI) by corresponding lower and upper probability values:

$$p_{lower} = \frac{(1-p)}{2} \tag{4.3}$$

$$p_{upper} = \frac{(1+p)}{2} \tag{4.4}$$

where p is probability value. For example, for $p = 0.5$, $P_{lower} = 0.25$ and $P_{upper} = 0.75$.

Next, for each of these intervals, check if the probability associated to the true value is located inside this interval or not(Figure 4.1):

Next, define an indicator function $\xi(\mathbf{u_i}; p)$ at each location $_i$ as:

$$\xi(\mathbf{u_i}; p) = \begin{cases} 1, & \text{if } F(\mathbf{u_i}; z(\mathbf{u_i}) \mid n(\mathbf{u_i})) \in (p_{lower}, p_{upper}] \\ 0, & \text{otherwise} \end{cases} \tag{4.5}$$

The average of $\xi(\mathbf{u_i}; p)$ over the n = 3920 locations $\mathbf{u_i}$;

$$\overline{\xi(p)} = \frac{1}{n} \sum_{i=1}^{n} \xi(\mathbf{u_i}; p) \tag{4.6}$$

is the proportion of locations where the true value falls within the symmetric $p - PI$.

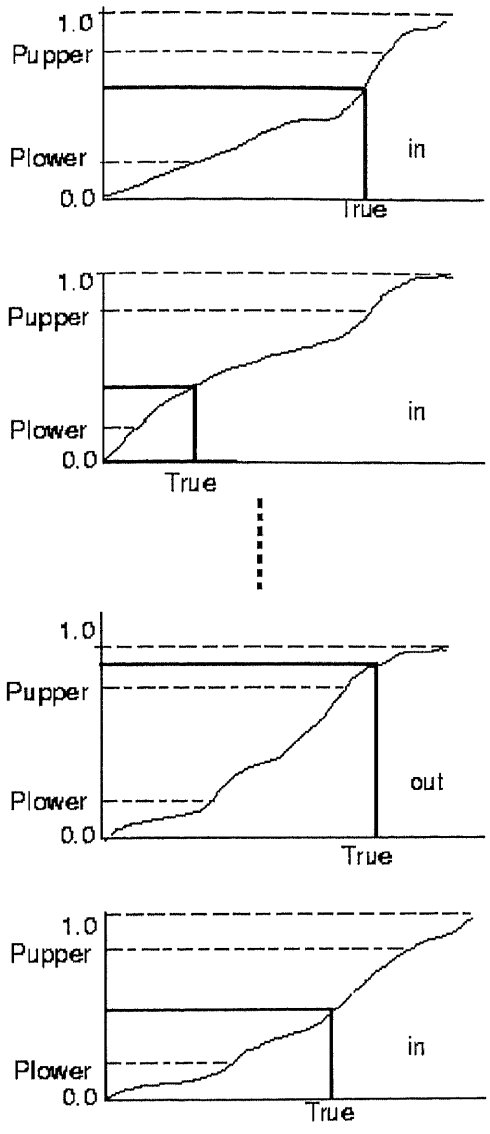


Figure 4.1: Definition of accuracy and precision. By counting the proportion of true values fall within symmetric $p - PI$, we know probability distribution is accurate if the fraction of true values falling in the p interval exceeds p for any p in [0,1]

4.1 Accuracy and Precision

In the above formula, if the fraction of true values falling in the p interval exceeds p for every p in [0,1], a probability distribution is said be accurate. For example, given a probabilistic model, if the fraction of the true values falling in the 10 percent interval exceeds or equals 10 percent, the fraction of the true values falling in the 20 percent interval exceeds or equals 20 percent, and so on for every p in [0,1], this probability model is accurate.

In practice, to check the accuracy of a probability model, we plot $\overline{\xi(p)}$ versus p and see that all of the points fall above or on the 45^o line. Ideally, we hope all of the points are located above the 45^o. This plot is referred to as an **accuracy plot.** Figure 4.4 show the accuracy plot of **SGSIM and COLSGSIM.** Note that most of points are above the 45^o line.

Sometime we find that two simulation approaches are accurate, i.e., all points are above the 45^o line. But the distance the points depart from the 45^o line in each accuracy plot is different. The precision of an accurate probability distribution is a measure of the closeness of the fraction $\overline{\xi(p)}$ of the true value to p for all p in [0,1]. An ideal case is where all points fall on-the-line, i.e., the probability distributions are accurate and precise.

The accuracy plot is one way to check the uniformity of the distribution of $F(\mathbf{u_i}; z(\mathbf{u_i}) \mid n(\mathbf{u_i})), i = 1, ..., 3920$. Equivalently, we could directly check that histogram. A good simulation approach should generate a uniform distribution over all unsampled locations. For a probabilistic model to be accurate and precise, the marginal distribution of $F(\mathbf{u_i}; z(\mathbf{u_i}) \mid n(\mathbf{u_i})), i = 1, ..., 3920$ must be uniform in [0,1] [6]. So, if the interval [0,1] is divided into N equal width classes , one should get approximately the same number of simulated values in each class.

4.2 Quantitative Measures

There are two main reasons to consider use quantitative measures of accuracy and precision:

- If the probability distributions generated by different simulation algorithms are similar, the accuracy plot may not give a clear indication of which algorithm is "best".
- In practice, the accuracy plot of a particular probabilistic model may show only few points below the 45^o line, it may not be revealing to simply declare that this probabilistic model is not accurate, especially in the case where most of points are very close to the 45^o line.

4.2.1 Accuracy

To give a quantitative measure of accuracy, define an indicator function $a(p)$ for each probability interval p, $p(0,1]$:

$$a(p) = \begin{cases} 1, & \text{if } \overline{\xi(p)} \geq p \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

Then for all $p \in (0,1]$, we have values of $a(p)$. The accuracy value is the summation of $a(p)$ over K probability values:

$$A = \int_0^1 a(p)dp = \sum_{k=1}^{K} a(p_k) \triangle p_k \tag{4.8}$$

where $k = 1, .., K$, K is the number of the probability intervals. If A =1.0, we get maximum accuracy, i.e., for all $p \in (0,1]$, $\overline{\xi(p)} \geq p$. A = 0.0 is the worst situation in which fraction of true values are contained in any probability intervals is less then the width of the interval.

4.2.2 Precision

Precision is a measure of the narrowness of the distribution and only defined for accurate probability distributions.

$$P = 1 - 2\int_0^1 a(p)[\overline{\xi(p)} - p]dp = 1 - 2\sum_{k=1}^{K} a(p_k)(\overline{\xi(p_k)} - p_k) \triangle p_k \tag{4.9}$$

where P = 1 represents maximum precision. On an accuracy plot, precision equal to 1 means all the points are exactly located on the 45^o line.

4.2.3 Goodness

In order to account for the inaccurate case $\overline{\xi(p)} \leq p$, which is not considered in definitions of accuracy and precision. We define a measure of the goodness as the departure of the points from the 45^o line on accuracy plot

$$G = 1 - \int_0^1 [3a(p) - 2][\overline{\xi(p)} - p]dp \tag{4.10}$$

Where $G = 1$ for maximum goodness and $G = 0$ is the worst case. Note that inaccurate values get more penalty, i.e., in this case, are weighted twice for inaccurate values($\overline{\xi(p)} \leq p$)

4.2.4 Uncertainty

In practice, two different probabilistic models may be equally accurate and precise($G = 1.0$). However, the spreads or uncertainties of the distributions may be different. Ideally, without losing accuracy and precision, the uncertainty generated by the best simulation algorithm is smallest. There are many measures to quantify the uncertainty of a particular probability model. Here, the uncertainty is simply defined as the average conditional variance at all locations in the simulated area.

$$U = \frac{1}{N} \sum_{i=1}^{N} \sigma^2(\mathbf{u_i}) \tag{4.11}$$

where the variance of location $\sigma(\mathbf{u_i})$is calculated from the local ccdf $F(\mathbf{u_i}; z \mid n(\mathbf{u_i}))$.

4.3 Simulation

To further illustrate the direct assessment of accuracy and precision and the advantages of using new measures, we do stochastic simulations using the same data in Chapter 2. In order to compare the difference between simulation and estimates techniques, all of the simulation approaches using the same conditioning data and variogram models listed in Chapter 2 and 3. For each approach, 100 realizations are generated and the corresponding accuracy plots, A, P, G, and U values are given by a program named **accplt**. A typical parameter file for **accplt** program is listed below.

```
                    Parameters for ACCPLT
                    *********************

START OF PARAMETERS:
true.data                       \file with true values
4                               \column number for true value
0                               \ccdf type (0=sim, 1=IK-type, 2=Gaussian)
100                             \   number of realizations/thresholds
.1 .2 .3 .4 .5 .6 .7 .8 .9      \   thresholds              (ccdf type 1)
0.0    1.0                      \   zmin, zmax              (ccdf type 1)
4   5                           \   column for mean and var (ccdf type 2)
100sgsim.out                    \file with realizations/distributions
sgsim.acc                       \summary file (accuracy)
sgsim.pro                       \output file for true quantiles
0.01                            \probability increment
```

4.3.1 Gaussian Simulation

As discussed in Chapter 1, Gaussian simulation(**SGSIM**) is the most straightforward algorithm for simulation of continuous variables. Since checking for bivariate normality in chapter 2 did not invalidate the multiGaussian assumption, we applied the **SGSIM** algorithm to generate 100 realizations of the porosity field[1]. Figure 4.2 lists the first 4 realizations. Note these realizations were made to honor:

- 2 wells with conditioning data,
- the reference histogram, and the
- variogram model.

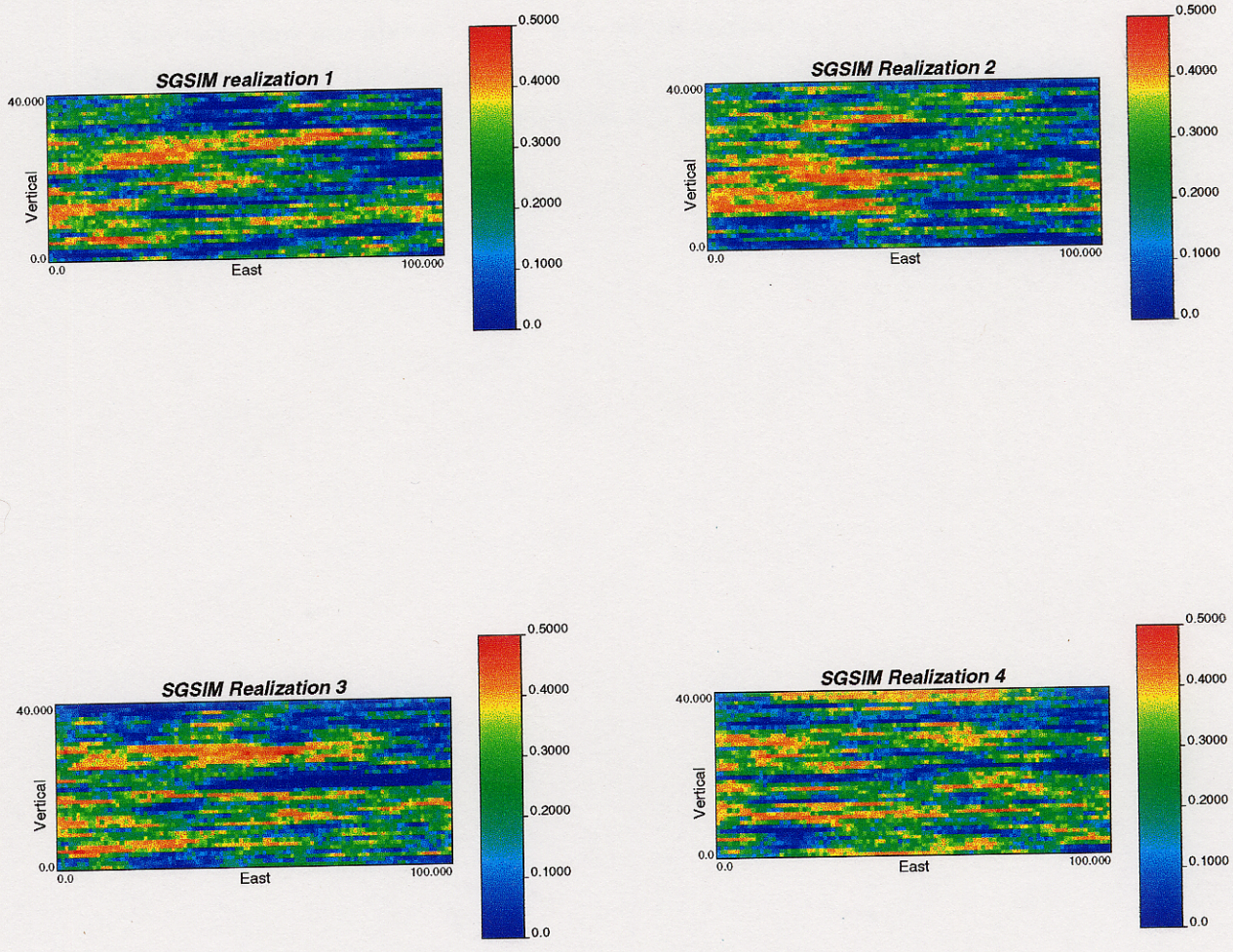


Figure 4.2: Four realizations generated by SGSIM method

The most important spatial character in the true porosity data, i.e., low values among the bottom and right top corner of the reservoir, are reproduced. A typical check for a simulation algorithm is variogram reproduction which in this case, is also reproduced well(Figure 4.3). Figure 4.4 shows the corresponding accuracy plots, A, P, G, and U scores (also summarized on Table 2). Note that most of the points in accuracy plot are above the 45^o line which indicates the ccdf models are accurate. The narrowness of the distributions, i.e., precision, however, is not good because the departure of the accuracy plot from the 45^o line is significant.

4.3.2 Gaussian Simulation with Co-located Cokriging

The above **SGSIM** does not account for the seismic data, which are useful supplements to the sparse conditioning data. To integrate seismic information, a co-located cokriging method(**COLSGSIM**) is used[3][4], i.e., only the seismic data co-located with the node being simulated is retained in the cokriging system. Thus, the computational time is significantly decreased. However, since only one seismic value is used at each grid, the trade-off costs of such approach are:

- The simulation can only reproduce the correlation between primary and secondary variable at co-located locations, i.e., no control on reproduction of correlation between primary and secondary variable at lags $||h|| \neq 0$.
- the secondary variable must be available at all simulated grid nodes, which is satisfied in this case.

Figure 4.5 shows four realizations generated by **COLSGSIM**. The variogram reproduction is checked in Figure 4.6. Compared to **SGSIM**, **COLSGSIM** improves the simulation in that, at locations near the center where no primary variable not available, seismic data are the critical for the simulation values. Figure 4.7 shows the cross plot of the simulated porosity with the seismic attribute. It reproduces the input model, i.e., correlation coefficient of about 0.6. The character of the seismic data, e.g., low values streak near the left-bottom and right-top corner of the field, are reproduced in the **COLSGSIM**.

The bottom plot of Figure 4.3 shows the accuracy plot of **COLSGSIM** and corresponding summary values. Even though there are more points below the 45^o line, we still consider this method as accurate because they are few and the departure is small. But, in this case, **COLSGSIM** leads better ccdf models in that the precision value in **COLSGSIM** is much better than **SGSIM**. Also, another important measure, i.e., the uncertainty U in **COLSGSIM** is lesser than **SGSIM**. So, using seismic data as soft information allows further reduction of uncertainty while maintaining accuracy and precision.

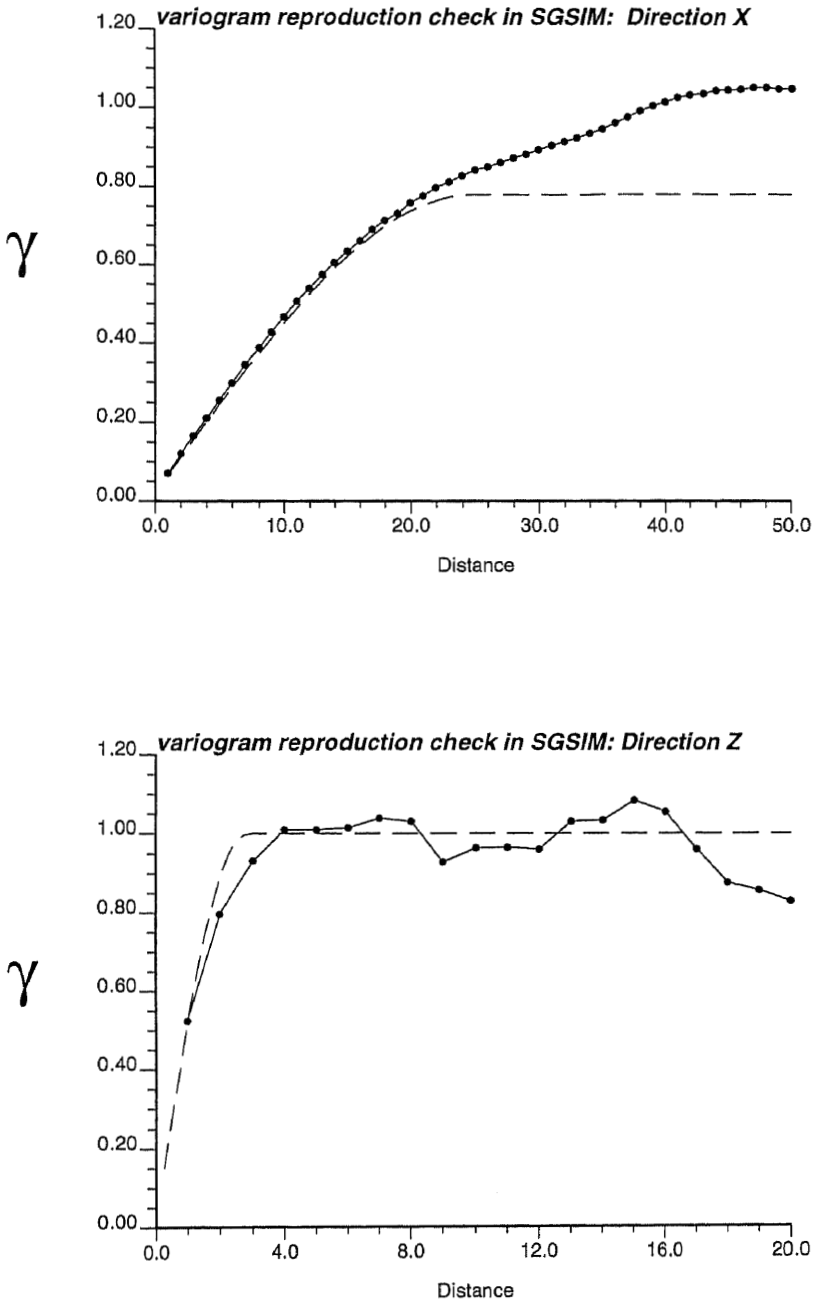


Figure 4.3: Variogram reproduction in SGSIM

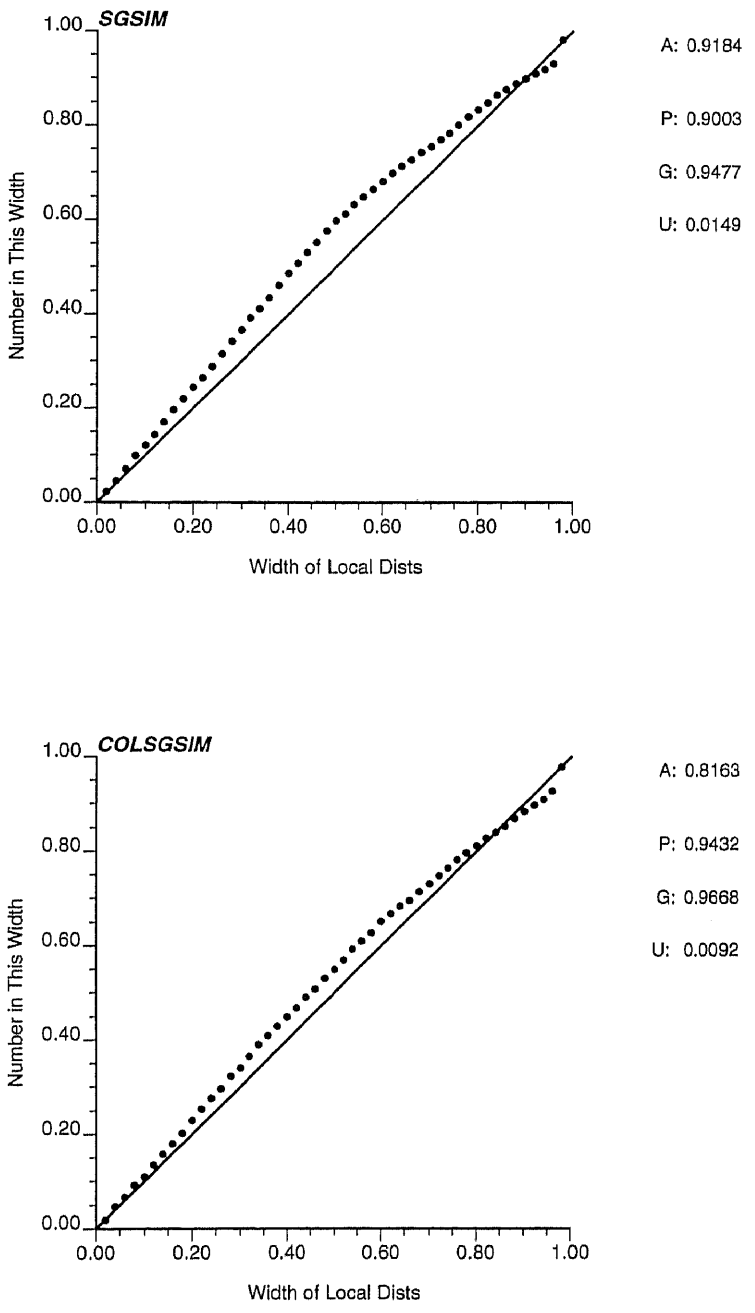


Figure 4.4: accuracy plot in SGSIM and COLSGSIM

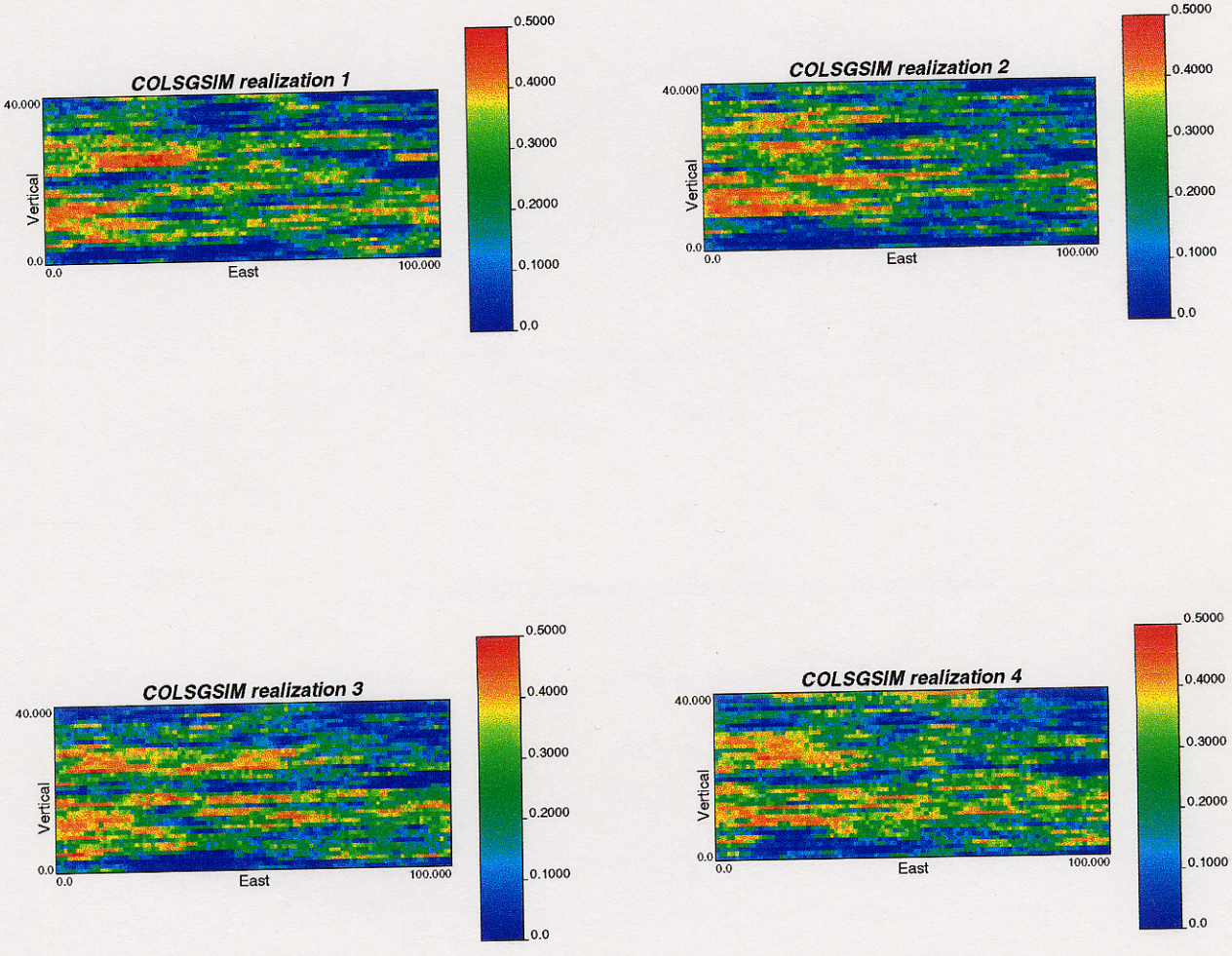


Figure 4.5: Four realizations in COLSGSIM

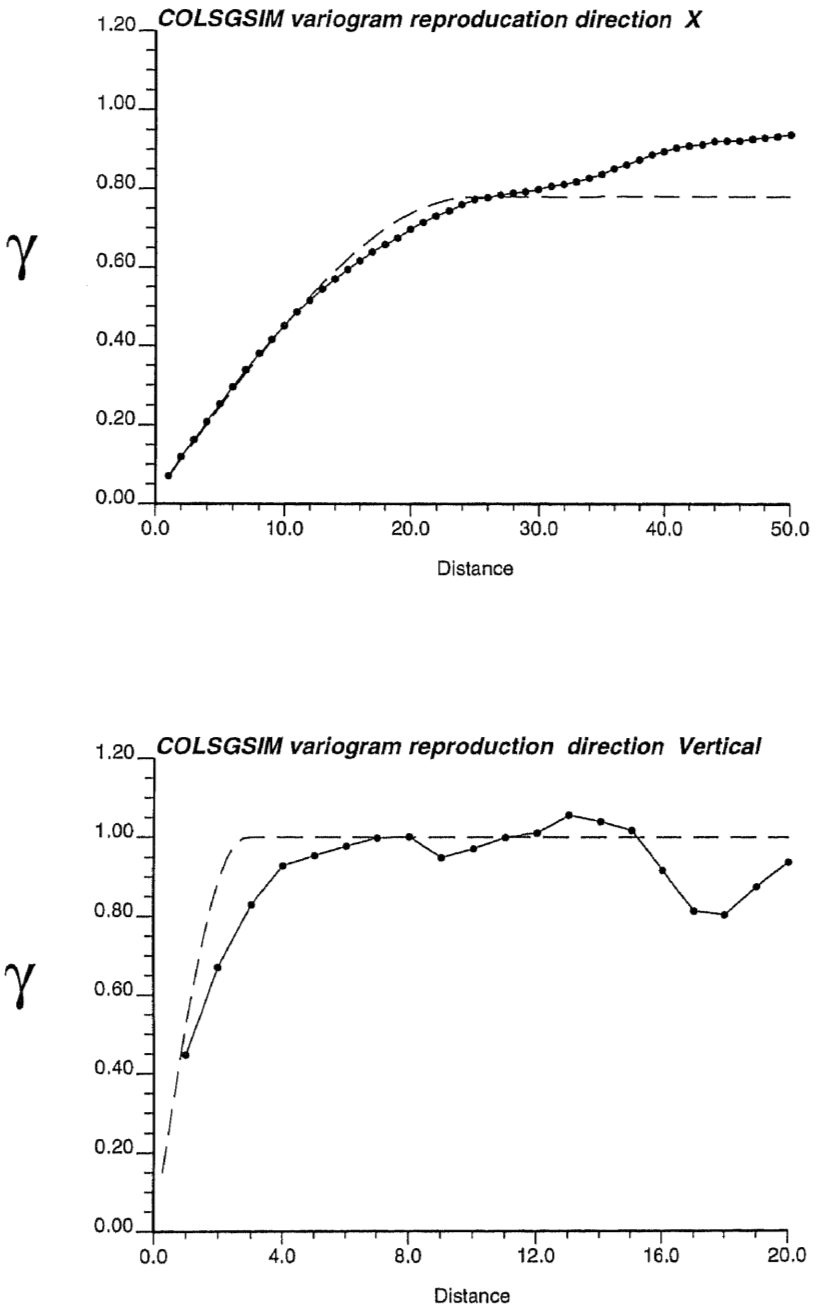


Figure 4.6: Variogram reproduction in COLSGSIM

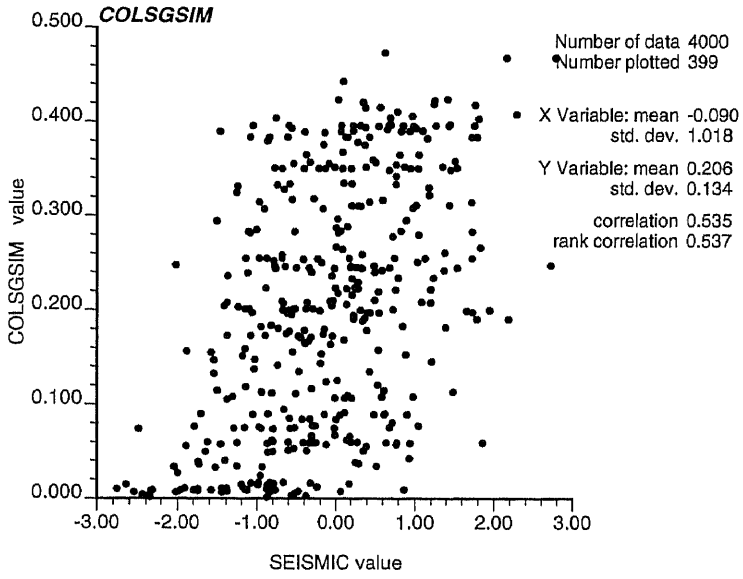


Figure 4.7: Correlation reproduction

4.3.3 Indicator Simulation

SGSIM and **COLSGSIM** require the spatial variation of porosity data to be multi-Gaussian distributed. The simplicity and congeniality of these approachs are offset by several shortcomings:

- If the porosity data are not multiGaussian distributed, the multiGaussian approach will lead to suboptimal distributions of uncertainty.
- Under the multiGaussuan model, extreme high and low values are spatially uncorrelated, an assumption often invalidated in practical situations.
- If the soft information is a categorical variable or provides constraint intervals, it is difficulty to incorporate in Gaussian type models.

An alternative algorithm that handles these shortcomings is the indicator based approach: **SISIM**. The main difference between **SISIM** and **SGSIM** is that **SISIM** does not assume any particular shape or analytical expression for the conditional distribution function $F(\mathbf{u}; z|(n))$. Instead, $F(\mathbf{u}; z|(n))$ is modeled through a series of K threshold values z_k by discretizing the range of variation of z:

$$F(\mathbf{u}; z|(n)) = Prob\{Z(\mathbf{u}) \leq z_k|z_k(n)\}, k = 1, ..., K \tag{4.12}$$

Where K is the cutoff number. Figure 4.8 shows four realizations generated by **SISIM**. The apparent discontinuity of the very high simulated values in middle locations is due to:

- Only 5 cutoffs were used, which causes artificially large within-class noise. As a result, the discrepancies between the model and experimental variogram for these five quantiles cannot be ignored (Figure 4.9). One may reduce the within-class noise by increasing the number of cutoffs. However, because of the tedious calculation of indicator variograms and problems with order relations in **SISIM**, $K = 5$ is considered adequate[1].

- The apparent discontinuity of the very high simulated values distributes at middle locations. These locations are beyond the correlation range from the conditioning data and no soft information available on these locations. As a result, the realizations generated by **SISIM** without seismic data gives poor reproduction at these locations, i.e., discontinuity of very high simulated values.

The accuracy plot and corresponding quantitative values are listed in Figure 4.11. Note that the uncertainty of **SISIM** is 0.015, which is larger than the values for the Gaussian method.

4.3.4 Markov-Bayes Indicator Simulation

As mentioned before, incorporating soft data is easy with indicator-based methods using indicator cokriging. However, traditional indicator full cokriging method requires to model cross-indicator variogram of all the cutoff values. One way to reduce this tedious calculation is to use Markov-Bayes assumption. The Markov-Bayes method establishes the covariance and cross-covariances of soft indicator data by calibration with the hard indicator covariance models. Calibration parameters $B(z_k)$ from the calibration scatterplot provide the linear rescaling parameters needed to establish this cross covariance. The $B(z_k)$ values for the five thresholds are listed below:

K	z_k	B(z_k)
1	0.055	0.3231
2	0.125	0.3746
3	0.202	0.3545
4	0.282	0.3234
5	0.396	0.0741

Figure 4.11 and Figure 4.12 display the four realizations and variogram reproduction in **MBSIM** method, respectively. Another important character in **MBSIM** method is shown on accuracy plot. Note all of the points are above the 45^o line, i.e, the accuracy values of **MBSIM** is 1.0.

4.3.5 Comparison in Simulations

Table 2 summaries all the A, P, G and U values of ccdf generated by simulation approaches. In this case, we find that COLSGSIM is best simulation algorithm due to following facts:

- The attribute we simulated here is porosity values, which, as we can see in chapter 2, has no extremely large or low values. Furthermore, the small number of large and low values are spatially uncorrelated, i.e., multiGaussian distribution is an appropriate assumption. So, Gaussian related methods are better choices

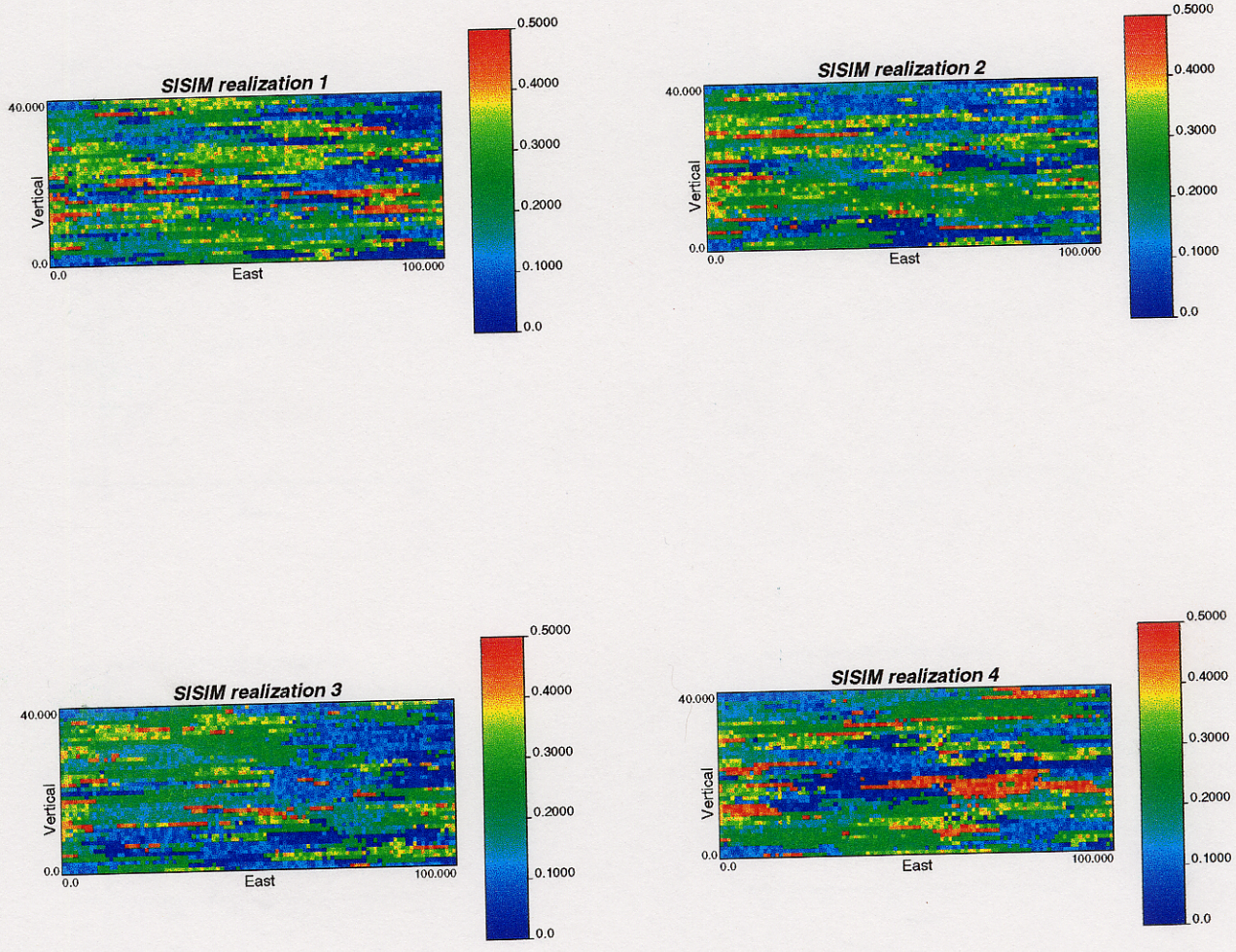


Figure 4.8: Four realizations in SISIM

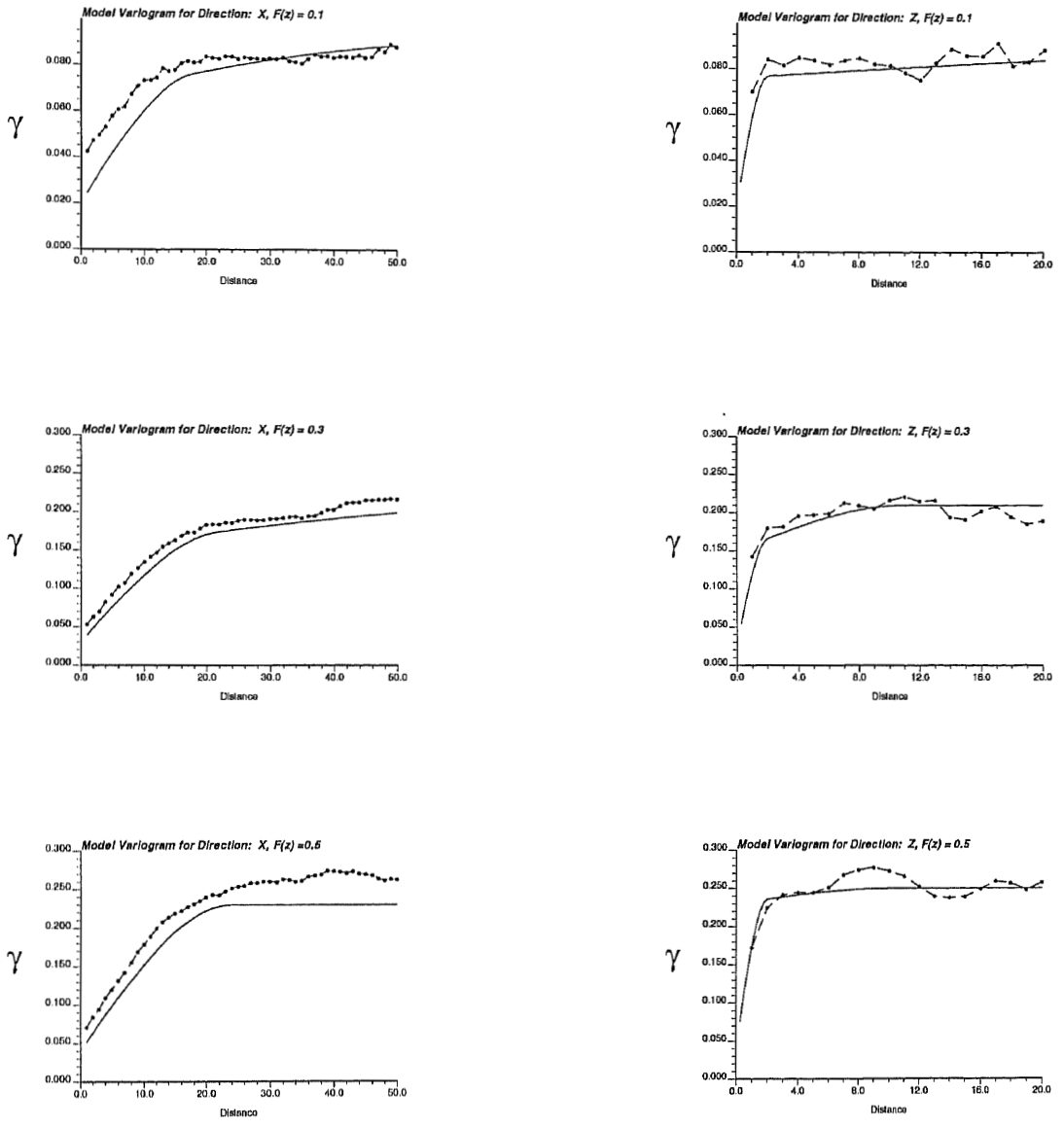


Figure 4.9: Variogram reproduction in SISIM

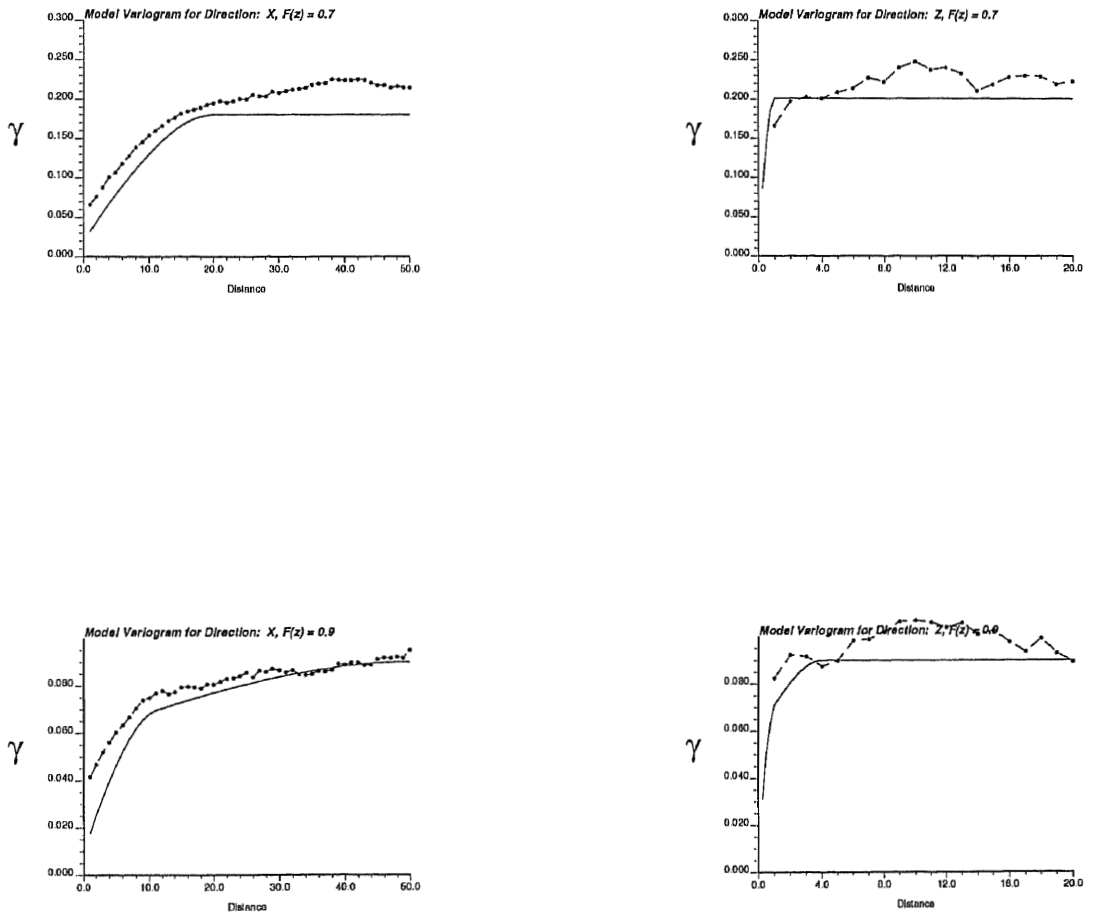


Figure 4.10: Variogram reproduction in SISIM(continued)

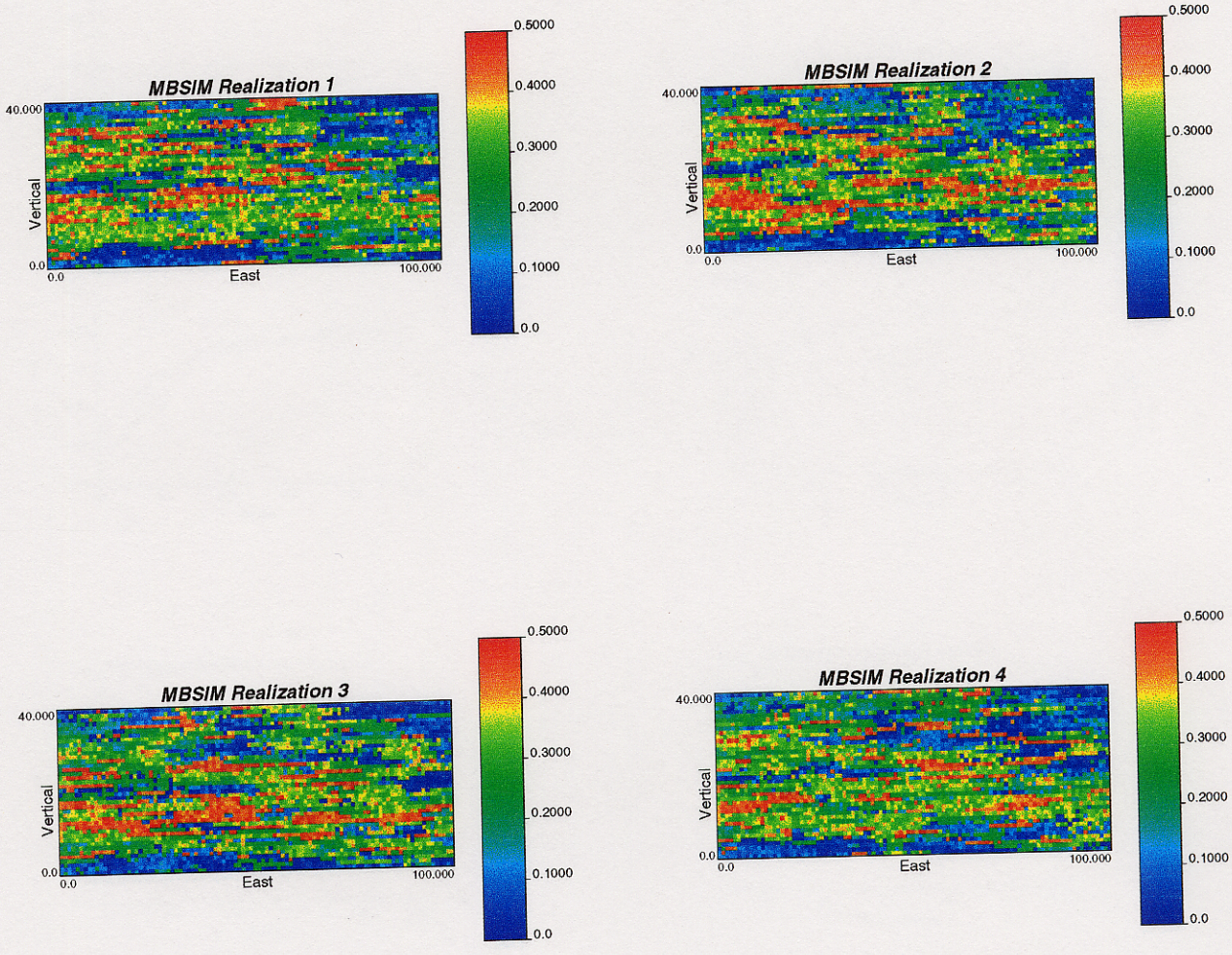
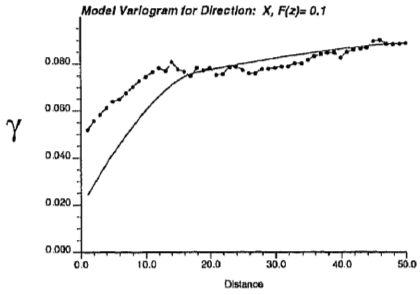
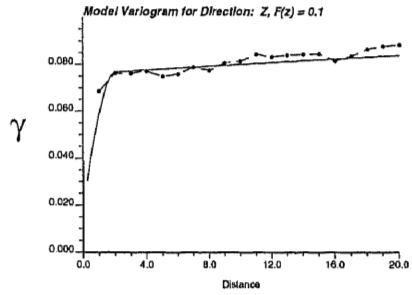
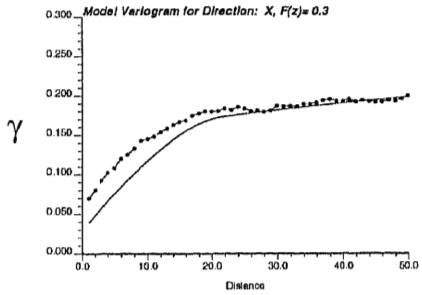
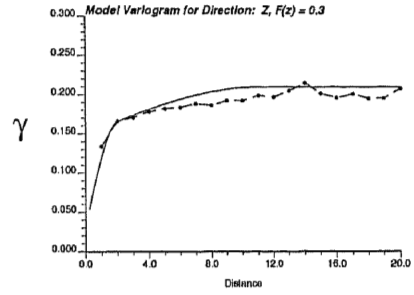


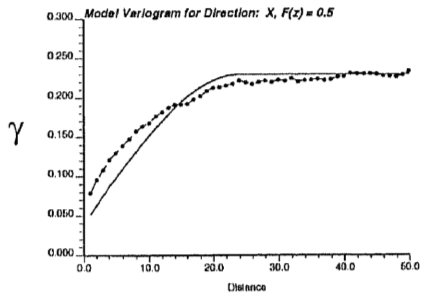
Figure 4.11: Four MBSIM realizations











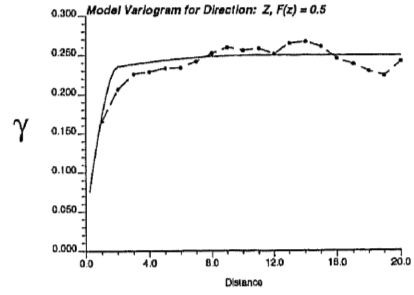


Figure 4.12: Variogram reproduction in MBSIM

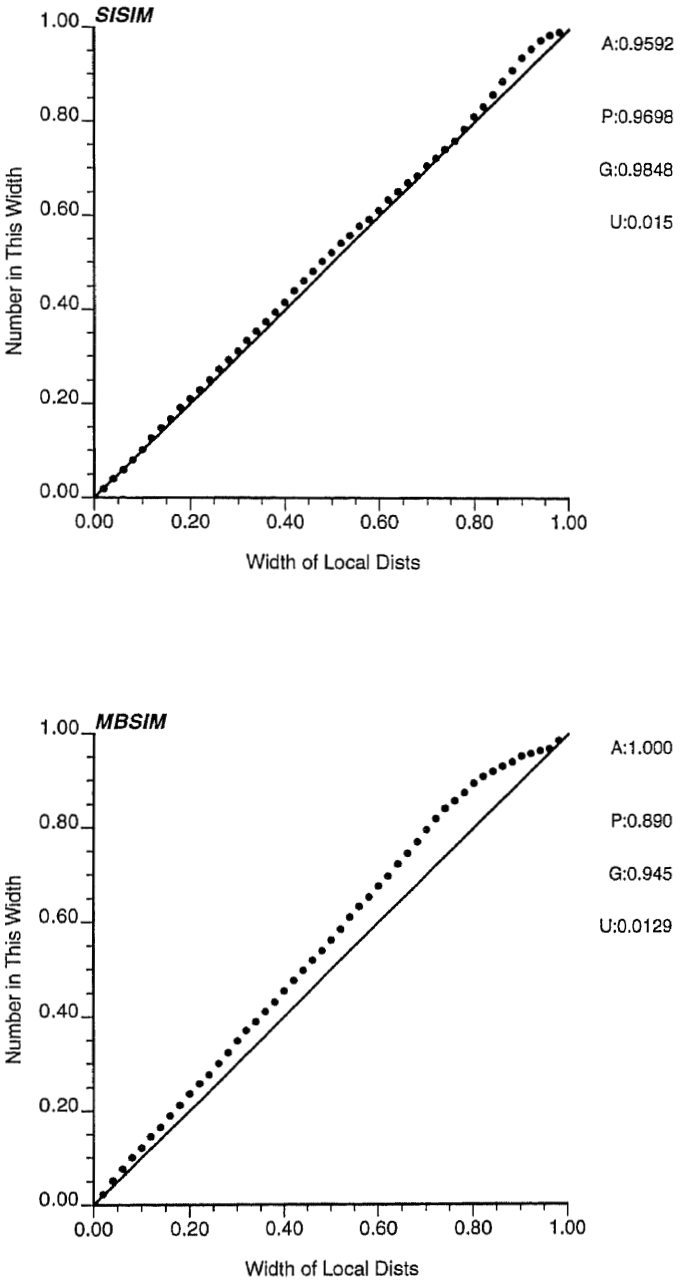


Figure 4.13: Accuracy plot for SISIM and MBSIM

for this problem. The advantages of using **SISIM** and **MBSIM**, such as allowing one to build a more complete specification of the bivariate distribution than Gaussian model, is not needed here. The problem with **SISIM**, the discontinuity of very large simulated values, may be avoided with more cutoffs and, hence, the results may appear better.

- Seismic data provides useful information in **COLSGSIM**. At the location where no conditioning hard data available(limited by search neighborhood) or beyond the correlation range (indicated by normal score variogram), there is no updating of the global model of uncertainty. So, at these locations, if we use **SGSIM** method without seismic data, the local ccdf model is the standard normal distribution, i.e, stationary zero mean and unit variance(no reduction in uncertainty).

Of all the four measures, U is the most important factor since it gives the final information about uncertainty. As we see, the uncertainty value of **COLSGSIM** is the smallest.

Simulation algorithm	Accuracy	Precision	Goodness	Uncertainty
SGSIM	0.9184	0.9003	0.9477	0.0149
COLSGSIM	0.8163	0.9432	0.9668	0.0092
SISIM	0.9592	0.9698	0.9848	0.0150
MBSIM	1.0000	0.8900	0.9451	0.0129

4.4 Comparison Between Traditional Cross Validation Techniques and New Measures

As we know, traditional techniques, such as mean absolute error and mean squared error cannot measure the uncertainty information of any stochastic simulations. They only apply for the techniques such as kriging that provide a unique estimate. In order to compare traditional cross validation techniques and new measures introduced in this chapter, we pick up one realization that generated by **COLSGSIM** and use

the same tools introduced in chapter 3 to do cross validation study on this realization.

Figure 4.14 shows the histogram of error, scatterplot of true values versus simulated value and absolute error map. Note that:

- In the histogram of estimate errors, the mean and the median are very close to 0. The spread of the error distribution is larger than for estimation because the simulated values have larger variance.
- The scatterplot of the true values and **COLSGSIM** displays conditional biasedness, but a good correlation coefficient. In general, this correlation will be less than with estimation methods because of the additional variance in the simulated values.
- no spatial correlation is found in the **COLSGSIM** error map.

The mean square error and mean absolute deviation are 0.015 and 0.095, respectively. However, if another realization is selected for cross validation, one may get different results based on the same tools. Indeed, the difference between these realizations carries the uncertainty information, which cannot be measured by traditional cross validation tools. Uncertainty is a critical factor in geostatistics analysis in that it reflects our imperfect knowledge of the unsampled value $z(\mathbf{u}_i), i = 1, ..., n$ and more generally, of the distribution of $z(\mathbf{u}_i)$ within the area. Therefore, a stochastic model with uncertainty information is considered more reasonable than a deterministic value or single realization. Consequently, the tools for cross validation should be able to reflect and measure the uncertainty for different simulation algorithms. How to model this uncertainty at any partiular location? Here, the local uncertainty is modeled through a set of possible realizations of the random variable at that location. Based on this model, the probabilities associated to the true values are calculated. Figure 4.15 shows the map of true values, absolute error and probabilities associated to the true values $z(\mathbf{u}_i), i = 1, ..., n$. Note that:

- The absolute error map indicates the "goodness" of a realization. However, these error maps are different for different realizations, hence it is impossible to use a single error map to compare different simulation algorithms.

- Except at the locations where conditioning data are available, all probabilities values are less than 1.0. The local **ccdf** models for a given procedure are built to be dependent from each other. However, the probabilistic values associated to the true values should not have spatial correlation. Let's assume true porosity values are critical threshold for risk analysis. At any particular location **u**, the magnitude of the misclassification risk depends on the local **ccdf** model, not on a particular simulated value at this location.

- By checking the probability map, we see how many simulated values are lower than or equal to the true value at each particular location. However, our objective here is to know how "close" of the simulated values are to the true porosity values. The mean absolute deviation value is the traditional tool to evaluate this deviation. But it only applies to one realization at a time. The accuracy and precision values are useful measures to detect this "closeness". As we mentioned, for a probability distribution, these two values give the actual fraction of true values falling within symmetric probability intervals of varying width p. They implicitly represent the average deviation between simulated value and true data for all realizations.

- The accuracy plot can be considered a counterpart of scatterplot in traditional tools. Both of them represent the deviation between simulated(estimated) value and true data. However, conditional biasedness character is not illustrated in the accuracy plot.

- The mean squared error is a measure for assessing the spread of error distribution. However, like error map, this measure only applies to one realization. The uncertainty(variance) defined by equation 4.11 at this chapter, measures the average conditional variance of all locations. The variance at any particular location, is easy to calculate since we already have the **ccdf** model.

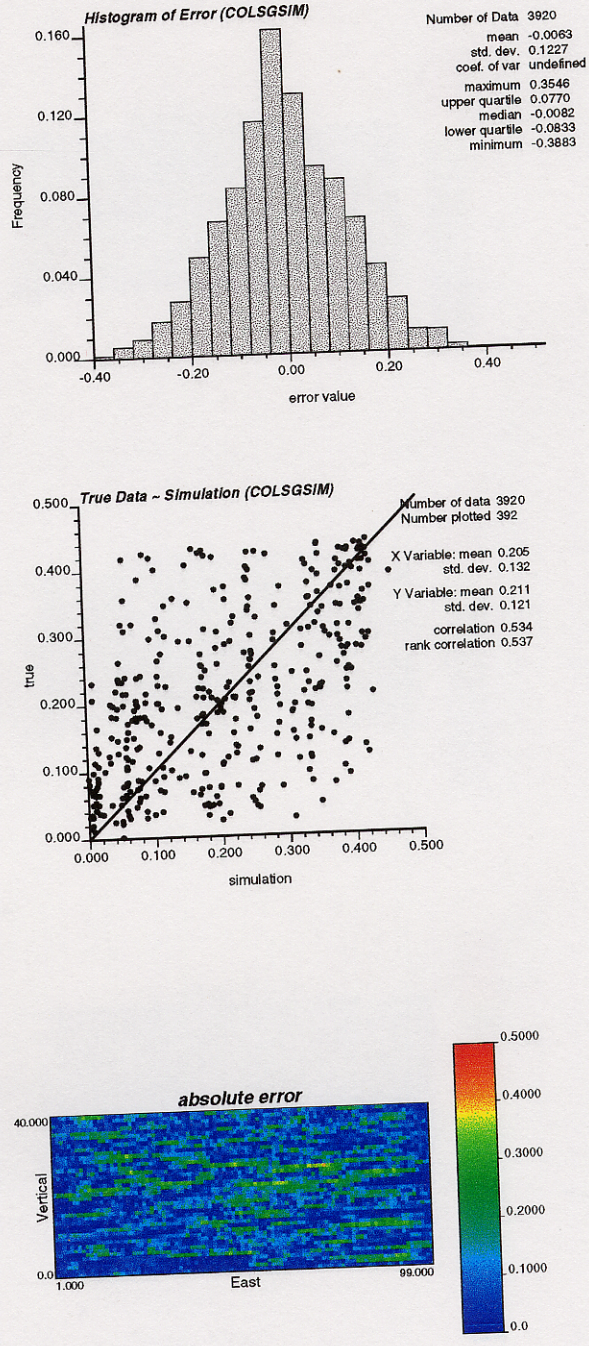


Figure 4.14: Traditional cross validation in one realization generated by COLSGSIM.

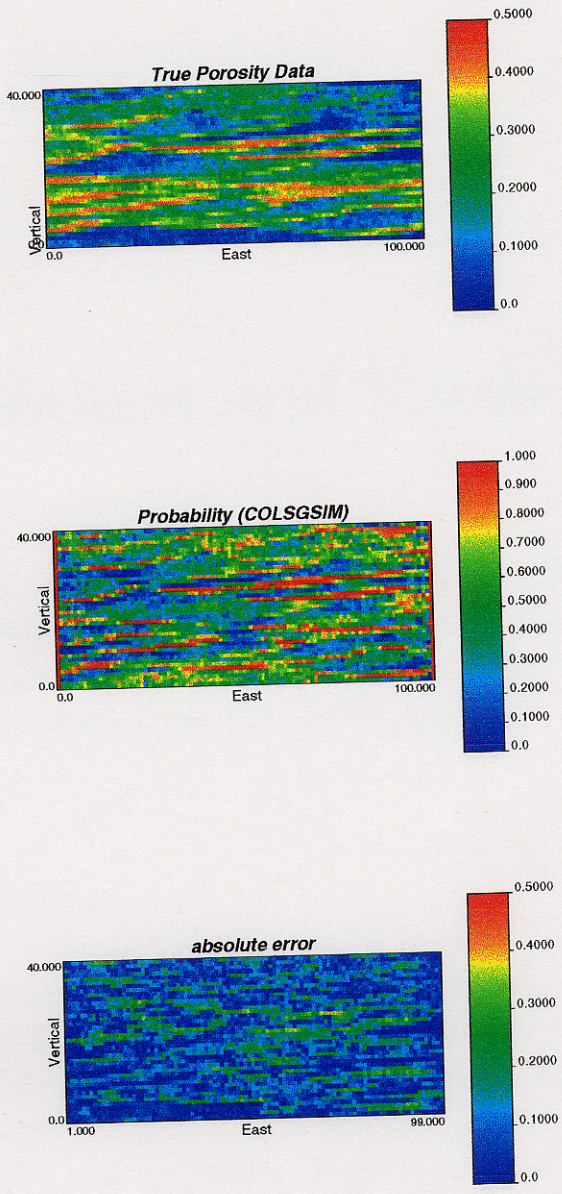


Figure 4.15: True value, probability and error map in COLSGSIM.

Chapter 5

Conclusions and Future Work

Two aspects of geostatistical reservoir modeling have been studied in this thesis. One is the comparison of traditional cross validation tools to the measures of accuracy(**A**), Precision(**P**), Goodness(**G**) and Uncertainty(**U**). The second one is criteria to choose an optimal simulation algorithm. Following are general comments based on these comparisons.

5.1 Conclusions

- The traditional cross validation tools apply to unique estimates. To handle the problem of multiple realizations and to account for the uncertainty information in stochastic simulations, measures of accuracy and precision are more suitable. A probabilistic model of uncertainty is "good" if it is both accurate and precise. In addition, the uncertainty should be smallest while preserving accuracy and precision.

- To quantify the "goodness"of a probabilistic model, we can use accuracy plot combined with quantitative measures of accuracy, precision, goodness, and uncertainty. In general, accuracy refers to the ultimate excellence of the data or computed results, e.g., conformity to truth or to a standard. Precision refers to the repeatability or refinement of a measurement or computed result[6]. Of all the four measures, uncertainty is the most important factor.

- One drawback of the accuracy and precision checks presented here is that they only consider one location at a time. An algorithm could have very good accuracy and precision scores; however, the simulated values, when taken all together, may not reproduce important patterns of spatial correlation such as the probability of occurrence of a string of large or small values. If we had enough data, we could cross validate the multivariate properties of a simulation algorithm[6].

5.2 Future Work

To simplify the problem, there were some assumptions made in this case study.

- The data we used here are porosity values where the connectivity of extreme values is less important. However, the problem of spatial connectivity of extreme high and low values is very common and should be considered in practice. For example, the indicator approach could prove more suitable for permeability where the connectivity of extreme values is a critical spatial feature.

- Another important assumption in seismic data is the identical volume support between primary and secondary attribute, which is unrealistic. The decision of choosing optimal simulation algorithm should also consider the issue of data with different volume supports[5].

- In this case study, the true values are exhaustively sampled and the vertical and horizontal variogram are easy to model. Indeed, all estimations and simulations use the same variogram models from the true values, a process to ensure the fairness in the comparisons. In practice, variogram models must be calculated from limited sample data and secondary information.

- In addition to quantifying uncertainty and comparing different simulation algorithms, another main use for these cross validation tools is the detection of implementation errors such as inappropriate search radii. Due to the time limit, this application is not tested in here.

As mentioned before, cross validation is a very useful method that gives us the ability to check the impact of our choices about the estimation and simulation methodology. However, the success of cross validation can never guarantee the success of the final performance in a particular estimation and simulation. The real advantage of a cross validation study is the early detection of major problems. One should focus on the negative aspects of the results. For example, a cross validation study shows some locations with worst errors or consistent bias. These problems lead us to check the reasons such as inappropriate kriging methods or clustered sample values. In general, The choice of a reasonable stochastic simulation algorithm by cross validation is necessary but, however, is not sufficient for making final decision. An optimal algorithm should also consider spatial characters of sample data, the goal of study and the difficulty to handle secondary information. A decision that blindly depends on the positive aspects of cross validation residuals may go wrong.

Chapter 6

Bibliography

1. C. Deutsch and A. Journel. *GSLIB: Geostatistical software library and user's guide.* Oxford University Press, New York, 1996

2. E. Isaaks and R. Srivastava. *An introduction to applied geostatistics.* Oxford University Press, New York, 1996

3. X. Wenlong and T. Tran. *Integrating seismic data in reservoir modeling: The collocated cokriging alternative.* SPE paper 27412, 1993

4. R. Behren and T. Tran. *Incorporating seismic attribute maps in 3D reservoir models.* SPE paper 36499, 1996

5. C. Deutsch, S.Srinivasan and Y.Mo. *Geostatistical reservoir modeling accounting for precision and scale of seismic data.* SPE paper 36497, 1996

6. C. Deutsch. *Direct assessment of local accuracy and precision.* SCRF report, Stanford University, 1996