DIRECT ASSESSMENT OF LOCAL ACCURACY AND PRECISION

C.V. DEUTSCH
Department of Petroleum Engineering
Stanford University, Stanford, CA 94305-2220

Abstract Geostatistical techniques are used to build probabilistic models of uncertainty about unknown true values. The *goodness* of these probabilistic models can be assessed by measures related to the two concepts of accuracy and precision.

A probability distribution is said to be accurate if the 10% symmetric probability interval (PI) contains the true value 10% (or more) of the time, the 20% PI contains the true value 20% (or more) of the time, and so on for increasingly wide probability intervals. To directly assess this measure of accuracy we can use the "leave-one-out" cross validation approach or the "keep-some-back" jackknife approach. For a given probabilistic model, the idea is to build distributions of uncertainty at multiple locations where the true values are known. Accuracy may then be judged by counting the number of times the true values actually fall within fixed probability intervals. Accuracy could be quantified for different probabilistic models (Gaussian, Indicator, Object-based, and Iterative/Annealing) and different implementation options.

Precision is a measure of the narrowness of the distribution. Precision is only defined for accurate probability distributions; without accuracy, a constant value or Dirac distribution would have the ultimate precision. A probability distribution where the 90% PI contains the true value 99% of the time is accurate but not precise. Optimal precision is when the 90% PI contains the true value exactly 90% of the time.

1. Introduction

The basic paradigm of geostatistics is to model the uncertainty about any unknown value z as a random variable (RV) Z characterized by a specific probability distribution. The unknown value z could be a global parameter such as the economic ultimate recovery of an oil field or a local attribute

E.Y. Baafi and N.A. Schofield (eds.), Geostatistics Wollongong '96, Volume 1, 115-125.
© 1997 *Kluwer Academic Publishers. Printed in the Netherlands.*

such as the porosity at a specific location **u**. Often, global parameters are the result of a complex non-linear process that may be simulated with relatively costly flow simulation of a high resolution numerical model. There are many ways of building these numerical models: (1) parametric approaches such as the multiGaussian model, (2) parameter-rich or distribution-free models such as provided by the indicator formalism, (3) object-based algorithms where objects are stochastically positioned in space, or (4) via simulated annealing. Moreover, each approach calls for many subjective implementation decisions related to specific computer coding, variogram measures of continuity, search strategies, size distributions, and convergence parameters.

This paper is concerned with checking the goodness of a probabilistic model, comparing it to alternative models, and perhaps fine-tuning the parameters of a chosen model. Before proceeding further, we must recognize that probabilistic models may only be checked by: (1) the data used for modeling, (2) some data held back from the beginning, or (3) additional knowledge of the physics of the phenomena, e.g., information that would allow classifying some realizations as implausible. In this paper, the "leave-one-out" cross validation and the "keep-some-back" jackknife approachs are considered. As for point (3), the goal is to incorporate such knowledge into the probabilistic model as soft or secondary data.

The goodness of a probabilistic model may be checked by its accuracy and precision. 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 (significant figures) of a measurement or computed result.

For clarity and in the context of evaluating the goodness of a probabilistic model, we propose specific definitions of accuracy and precision. For a probability distribution, accuracy and precision are based on the actual fraction of true values falling within symmetric probability intervals of varying width p:

- A probability distribution is accurate if the fraction of true values falling in the p interval exceeds p for all p in $[0, 1]$.
- The precision of an accurate probability distribution is measured by the closeness of the fraction of true values to p for all p in $[0, 1]$.

A procedure for the direct assessment of local accuracy and precision is now described.

2. Assessing Local Accuracy

2.1. DEFINITIONS

Consider the "leave-one-out" cross validation approach. Values at n data locations $\mathbf{u}_i, i = 1, \ldots, n$, are simulated one at a time using the remaining $n-1$ data values, i.e., leaving out the data value $z(\mathbf{u}_i)$. Stochastic simulation leads to L (L large) stochastic realizations $\{z^{(l)}(\mathbf{u}_i), l = 1, \ldots, L\}$ at each left out data location. These L realizations provide a model of the conditional cumulative distribution function (ccdf):

$$F(\mathbf{u}_i; z|n(\mathbf{u}_i)) = Prob\{Z(\mathbf{u}_i) \leq z|n(\mathbf{u}_i)\} \tag{1}$$

where $n(\mathbf{u}_i)$ is the set of n data minus the data at location $\mathbf{u}_i$. These local ccdf models may be (1) derived from a set of L realizations, (2) calculated directly from indicator-kriging, or (3) defined by a Gaussian mean, variance, and transformation.

The probabilities associated to the true values $z(\mathbf{u}_i), i = 1, \ldots, n$ are calculated from the previous ccdf as:

$$F(\mathbf{u}_i; z(\mathbf{u}_i)|n(\mathbf{u}_i)), \;\; i = 1, \ldots, n$$

For example, if the true value at location $\mathbf{u}_i$ is at the median of the simulated values then $F(\mathbf{u}_i; z(\mathbf{u}_i)|n(\mathbf{u}_i))$ would be 0.5.

Consider a range of symmetric p-probability intervals (PIs), say, the centiles 0.01 to 0.99 in increments of 0.01. The symmetric p-PI is defined by corresponding lower and upper probability values:

$$p_{low} = \frac{(1-p)}{2} \quad \text{and} \quad p_{upp} = \frac{(1+p)}{2}$$

For example, for $p = 0.9$, $p_{low} = 0.05$ and $p_{upp} = 0.95$.

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

$$\xi(\mathbf{u}_i; p) = \begin{cases} 1, & \text{if } F(\mathbf{u}_i; z(\mathbf{u}_i)|n(\mathbf{u}_i)) \in (p_{low}, p_{upp}] \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The average of $\xi(\mathbf{u}_i; p)$ over the n locations $\mathbf{u}_i$:

$$\overline{\xi(p)} = \frac{1}{n} \sum_{i=1}^{n} \xi(\mathbf{u}_i; p) \tag{3}$$

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

According to our earlier definition of accuracy, the simulation algorithm used to generate the ccdfs (1) is accurate when $\overline{\xi(p)} \geq p$, $\forall\, p$. A graphical way to check this assessment of accuracy is to cross plot $\overline{\xi(p)}$ versus p and see that all of the points fall above or on the 45^o line. This plot is referred to as an *accuracy plot*, see Figure 1 and example hereafter.

An equivalent way to calculate $\overline{\xi(p)}$ is to define the indicator $\xi(\mathbf{u};p) = 1$ if $z(\mathbf{u}) \in [F^{-1}(\mathbf{u};p_{low}|n(\mathbf{u})), F^{-1}(\mathbf{u};p_{upp}|n(\mathbf{u}))]$, otherwise $\xi(\mathbf{u};p) = 0$. That is, define the indicator on z values instead of p values. The advantage of defining the indicator on probability values as in relation (2) is that significantly fewer quantiles must be calculated because p_{low} and p_{upp} are global parameters independent of the location $\mathbf{u}_i$. For example, with $n = 1000$ and $n_p = 99$ centiles we calculate either 1000 probabilities or $99 \cdot 2 \cdot 1000 = 198000$ quantile values for the same result.

2.2. A FIRST EXAMPLE

Consider $n = 1000$ *independent* uniformly distributed random numbers $z_i, i = 1, \ldots, n$ drawn from the ACORN generator (Wikramaratna 1989). For each outcome, the correct corresponding ccdf of type (1) is a uniform distribution in $[0, 1]$:

$$F(i; z|(n(i))) = \begin{cases} 0.0, & \text{for } z \leq 0 \\ z, & \text{for } 0 < z < 1 \\ 1.0, & \text{for } z \geq 1 \end{cases}$$

The probability of each true value $z_i, i = 1, \ldots, n$ is that value itself, i.e.,

$$F(i; z_i|(n(i))) = z_i, \quad i = 1, \ldots, n \tag{4}$$

The average indicator function $\overline{\xi(p)}$ was calculated for each centile $p = \frac{j}{100}, j = 1, \ldots, 99$ and the resulting accuracy plot is shown on the top of Figure 1. Note that the plot is very close to a 45^o line indicating that this ccdf model is both accurate and precise.

The middle row graphs of Figure 1 illustrate the accuracy plot if the ccdfs were expected to be uniform between -0.5 and 1.5. The simulated distributions are still accurate, because the probability intervals are sufficiently wide, but not precise. The 0.5 PI of this ccdf model contains all of the true values, therefore, $\overline{\xi(0.5)} = 1.0$.

The lower graphs on Figure 1 show the results when the ccdfs are modeled to be triangular distributions between 0 and 1 with mode at 0.5. The simulated distributions are no longer accurate. In fact the 0.5 PI only contains 25% of the true values.

2.3. QUANTITATIVE MEASURES OF ACCURACY AND PRECISION

A distribution is accurate when $\overline{\xi(p)} \geq p$. To develop a measure of accuracy, an indicator function $a(p)$ is defined for each probability interval p, $p \in$

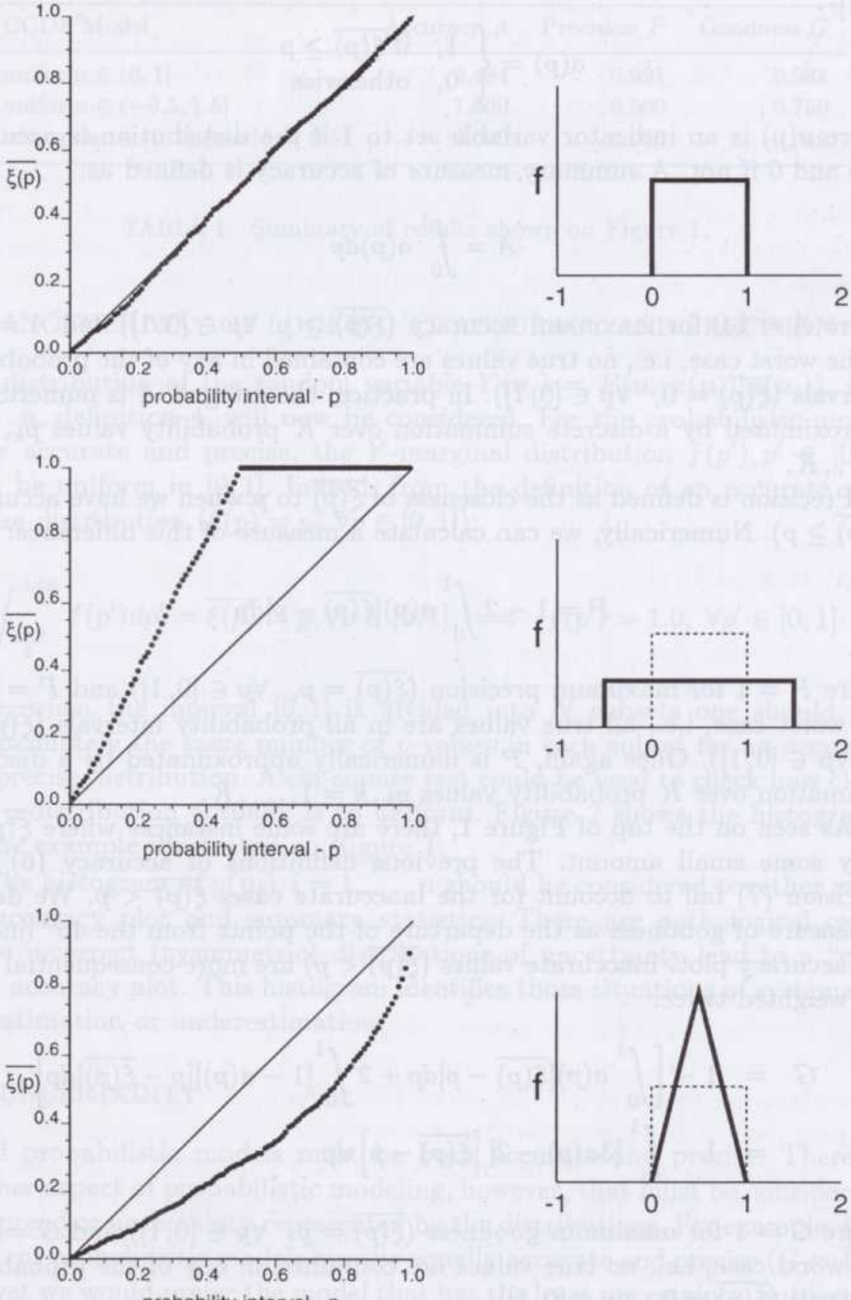

Figure 1. Accuracy plots for $n = 1000$ independent uniformly distributed in $[0, 1]$ random numbers $z_i, i = 1, \ldots, n$. The plots correspond to (1) a correct uniform distribution between 0 and 1, (2) a uniform distribution between -0.5 and 1.5, and (3) a triangular distributions between 0 and 1 with mode at 0.5.

(0, 1]:

$$a(p) = \begin{cases} 1, & \text{if } \overline{\xi(p)} \geq p \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $a(p)$ is an indicator variable set to 1 if the distribution is accurate at p and 0 if not. A summary measure of accuracy is defined as:

$$A = \int_0^1 a(p)dp \tag{6}$$

where $A = 1.0$ for maximum accuracy ($\overline{\xi(p)} \geq p \;\; \forall p \in [0,1]$) and $A = 0.0$ in the worst case, i.e., no true values are contained in any of the probability intervals ($\overline{\xi(p)} = 0, \;\; \forall p \in [0,1]$). In practice, the integral A is numerically approximated by a discrete summation over K probability values $p_k, k = 1, \ldots, K$.

Precision is defined as the closeness of $\overline{\xi(p)}$ to p when we have accuracy ($\overline{\xi(p)} \geq p$). Numerically, we can calculate a measure of this difference:

$$P = 1 - 2\int_0^1 a(p)[\overline{\xi(p)} - p]dp \tag{7}$$

where $P = 1$ for maximum precision ($\overline{\xi(p)} = p, \;\; \forall p \in [0,1]$) and $P = 0$ in the worst case, i.e., all true values are in all probability intervals ($\overline{\xi(p)} = 1.0 \; \forall p \in [0,1]$). Once again, P is numerically approximated by a discrete summation over K probability values $p_k, k = 1, \ldots, K$.

As seen on the top of Figure 1, there are some instances where $\overline{\xi(p)} < p$ by some small amount. The previous definitions of accuracy (6) and precision (7) fail to account for the inaccurate cases $\overline{\xi(p)} < p$. We define a measure of goodness as the departure of the points from the 45^o line on the accuracy plot. Inaccurate values ($\overline{\xi(p)} < p$) are more consequential and are weighted twice:

$$\begin{aligned} G &= 1 - \left[\int_0^1 a(p)[\overline{\xi(p)} - p]dp + 2\int_0^1 [1 - a(p)][p - \overline{\xi(p)}]dp\right] \qquad (8) \\ &= 1 - \int_0^1 [3a(p) - 2]\left[\overline{\xi(p)} - p\right]dp \end{aligned}$$

where $G = 1$ for maximum goodness ($\overline{\xi(p)} = p, \;\; \forall p \in [0,1]$) and $G = 0$ in the worst case, i.e., no true values are contained in any of the probability intervals ($\overline{\xi(p)} = 0, \;\; \forall p \in [0,1]$).

Table 1 gives the accuracy, precision, and goodness statistics for the results on Figure 1. Note that the triangular distribution is not as good as the too-wide uniform distribution because of the more severe penalty for inaccuracy ($\overline{\xi(p)} < p$).

CCDF Model	Accuracy A	Precision P	Goodness G
uniform $\in (0,1]$	0.484	0.991	0.985
uniform $\in (-0.5, 1.5]$	1.000	0.500	0.750
triangular $\in (0,1]$ (mode at 0.5)	0.000	0.000	0.649

TABLE 1. Summary of results shown on Figure 1.

2.4. ANOTHER WAY OF LOOKING AT ACCURACY AND PRECISION

The distribution of the random variable $Y(\mathbf{u}_i) = F(\mathbf{u}_i; z(\mathbf{u}_i)|n(\mathbf{u}_i))$, $i = 1, \ldots, n$, definition 4, will now be considered. For the probabilistic model to be accurate and precise, the Y-marginal distribution $f(p'), p' \in [0,1]$ must be uniform in $[0,1]$. Indeed, from the definition of an accurate and precise distribution ($\xi(p) = p, \; \forall p \in [0,1]$):

$$\int_{\frac{1-p}{2}}^{\frac{1+p}{2}} f(p')dp' = \overline{\xi(p)} = p, \forall p \in [0,1] \implies f(p') = 1.0, \; \forall p' \in [0,1]$$

Thus, when the interval $[0,1]$ is divided into N subsets one should get approximately the same number of y-values in each subset for an accurate and precise distribution. A chi-square test could be used to check how close that y-distribution actually is to uniform. Figure 2 shows the histograms for the example illustrated on Figure 1.

The histogram of $y(\mathbf{u}_i), i = 1, \ldots, n$ should be considered together with the accuracy plot and summary statistics. There are pathological cases where incorrect (aysmmetric) distributions of uncertainty lead to a "correct" accuracy plot. This histogram identifies those situations of systematic overestimation or underestimation.

2.5. UNCERTAINTY

Good probabilistic models must be both accurate and precise. There is another aspect of probabilistic modeling, however, that must be considered: the spread or uncertainty represented by the distributions. For example, two different probabilistic models may be equally accurate and precise ($G \approx 1.0$) and yet we would prefer the model that has the least uncertainty. Subject to the constraints of accuracy and precision we want to account for all relevant data to reduce uncertainty. For example, as new data becomes available our probabilistic model may remain equally good ($G = 1.0$) and yet there is less uncertainty.

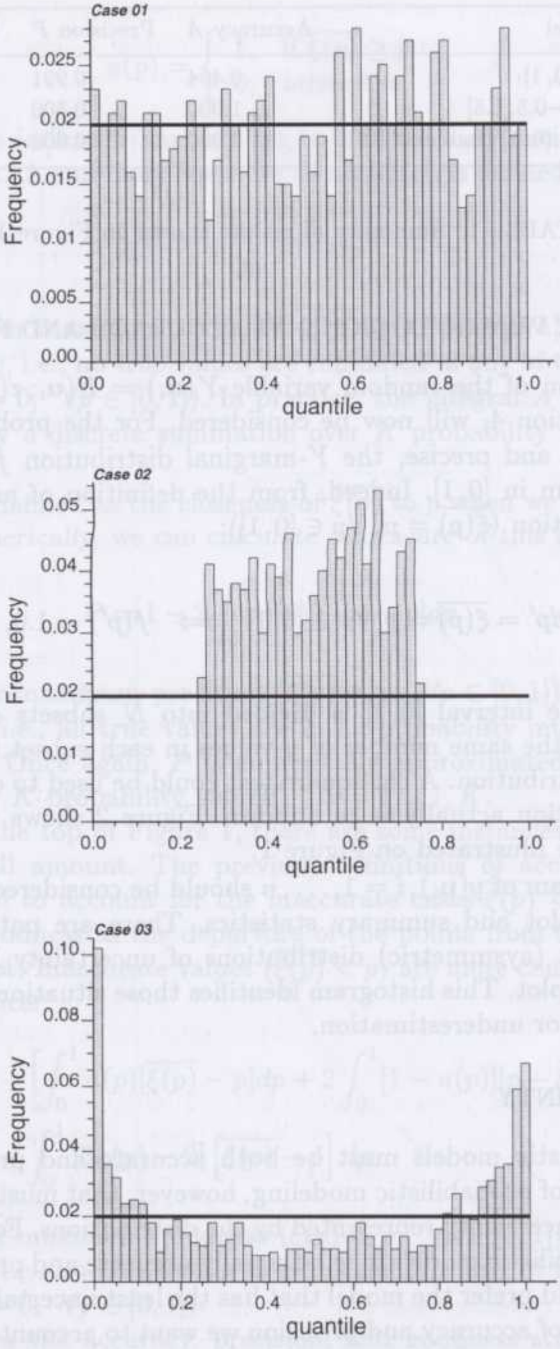

Figure 2. Histograms of $y(\mathbf{u}_i), i = 1, \ldots, n$ for the three examples illustrated on Figure 1.

If we systematically took the global histogram as the model of uncertainty at each location we would find that this conservative probabilistic model is both accurate and precise. The uncertainty, however, is large. Uncertainty increases as the spread of the probability distribution increases and could be quantified by measures such as entropy, interquartile range, or variance.

The measure of entropy has attractive features from the perspective of information theory or statistical mechanics. The interquartile range has attractive features from the perspective of robust statistics. The variance is preferred here because of its simplicity and wide acceptance as a measure of spread. The uncertainty of a probabilistic model may be defined as the average conditional variance of all locations in the area of interest:

$$U = \frac{1}{N} \sum_{i=1}^{N} \sigma^2(\mathbf{u}_i) \tag{9}$$

where there are N locations of interest $\mathbf{u}_i, i = 1, \ldots, N$ with each variance $\sigma^2(\mathbf{u}_i)$ calculated from the local ccdf $F(\mathbf{u}_i; z|n(\mathbf{u}_i))$.

The uncertainty statistics for the simple example presented in Figure 1 are 0.0833, 0.3333, 0.0424, respectively. Although the triangular distribution has the least uncertainty, it is not recommended since it is neither accurate nor precise, see Figure 1 and Table 1. To be legitimate, uncertainty can not be artificially reduced at the expense of accuracy.

3. Reservoir Case Study

To further illustrate the direct assessment of accuracy and precision, consider the "Amoco" data consisting of 74 well data related to a West Texas carbonate reservoir. Figure 3 shows a location map of the 74 well data and a histogram of the vertically averaged porosity for the main reservoir layer of interest.

Sequential Gaussian simulation (Isaaks 1990, Deutsch & Journel 1992) was performed at each of the 74 well locations using the "leave-one-out" cross-validation approach. Two alternative semivariogram models for the normal score transform of porosity were considered: (1) that built from the complete 3-D set of porosity data, and (2) that built from the 74 vertically averaged values. These two different variogram models result in two different sets of ccdfs. Figures 4 and 5 show the two semivariogram models and the corresponding accuracy plots, A, P, and G scores. The semivariogram built from the vertically averaged values leads to better ccdf models. Also, that semivariogram model leads to a lesser uncertainty measure $U = 0.433$ down from $U = 0.757$ for the first semivariogram model.

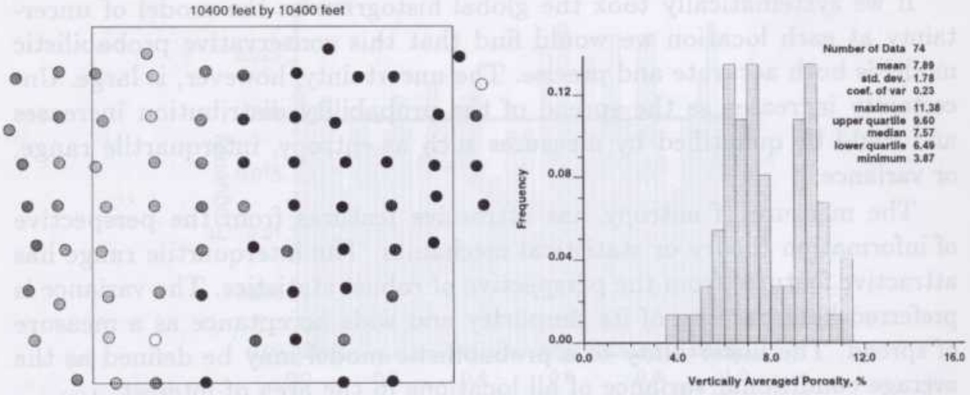

Figure 3. Location map and histogram of 74 well data.

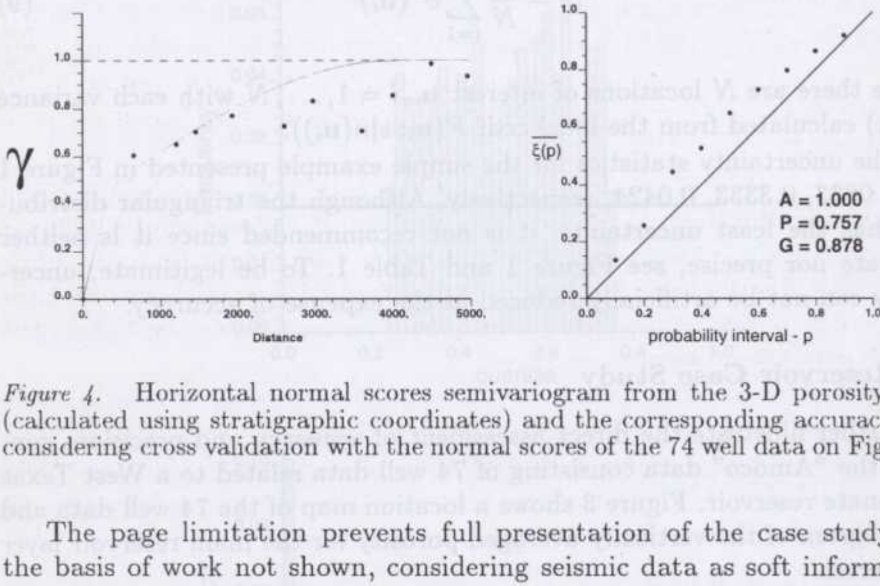

Figure 4. Horizontal normal scores semivariogram from the 3-D porosity data (calculated using stratigraphic coordinates) and the corresponding accuracy plot considering cross validation with the normal scores of the 74 well data on Figure 3.

The page limitation prevents full presentation of the case study. On the basis of work not shown, considering seismic data as soft information for simulation of porosity with a carefully fit linear model of coregionalization allows further reduction of uncertainty while maintaining accurate and precise distributions. Also, indicator simulation and annealing-based simulation were shown to perform slightly better than Gaussian-based simulation because more spatial information is considered through multiple indicator variograms.

4. Conclusions

We have developed the idea of directly checking local accuracy and precision through cross validation. A probabilistic model of uncertainty is "good" if it is both accurate and precise. In addition, the uncertainty should be as

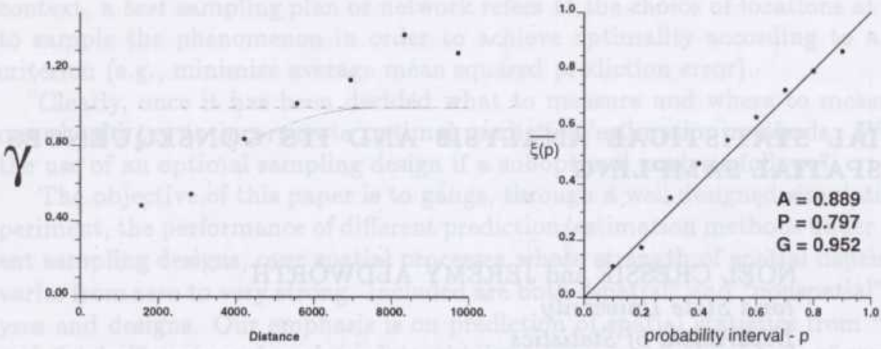

Figure 5. Horizontal normal scores semivariogram from the 74 vertically averaged porosity and the corresponding accuracy plot considering cross validation with the normal scores of the 74 well data on Figure 3.

small as possible while preserving accuracy and precision. The accuracy plot combined with measures of accuracy (A), precision (P), goodness (G), and uncertainty (U) are useful summaries to quantify the "goodness" of a probabilistic model.

The main uses for the diagnostic tools presented here are: (1) detecting implementation errors, (2) quantifying uncertainty, (3) comparing different simulation algorithms (e.g., Gaussian-based algorithms versus indicator-based algorithms versus simulated annealing-based algorithms), and (4) fine-tuning the parameters of any particular probabilistic model (e.g., the variogram model used).

These tools provide basic checks, i.e., necessary but not sufficient tests that any reasonable stochastic simulation algorithm should pass. They do not assess the multivariate properties of the simulation. Care is needed to ensure that features that impact the ultimate prediction and decision making, such as continuity of extreme values, are adequately represented in the model.

References

Deutsch, C. & Journel, A. (1992). *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York.

Isaaks, E. (1990). *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*, PhD thesis, Stanford University, Stanford, CA.

Wikramaratna, R. (1989). ACORN - a new method for generating sequences of uniformly distributed pseudo-random numbers, *Journal of Computational Physics* **83**: 16–31.