Constrained Smoothing of Histograms and Scatterplots With Simulated Annealing

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

Histogram and scatterplot models are often required for statistical inference. In the field of petroleum engineering, stochastic simulation algorithms require, among other statistics, a model for the histogram of the petrophysical attribute (porosity/permeability) being simulated. Often this model is taken to be the declustered distribution of the sample data. When there are many data (say, greater than 1,000), this histogram may be reasonably informed. Most often, however, the sample histogram shows multiple sawtoothlike spikes that are not representative of the entire population; the sample histogram must be smoothed. A simulated annealing-based procedure is proposed for smoothing one-variable (univariate) histograms and two-variable scatterplots (bivariate histograms). The smoothed histograms are constrained to the sample mean, variance, specified quantiles, and a measure of smoothness. In the bivariate case, the distribution must be consistent with both marginal histograms and can be additionally constrained to a linear correlation coefficient. Several examples with real reservoir data are presented.

KEY WORDS: Bivariate histograms; Conditional simulation; Distribution modeling; Kernel functions; Nonparametric modeling.

1. INTRODUCTION

A model of the histogram or *univariate distribution* is required for stochastic or geostatistical simulation. When considering two variables, such as porosity and permeability, a model for the scatterplot or *bivariate distribution* is also needed. These distribution models must be inferred from the sample data. In practice, however, there are often too few sample data to describe the full range of values that may be encountered in the population (reservoir).

Some modeling including *smoothing* is necessary because assumptions must be made for extrapolation beyond the smallest datum value (lower tail) and the largest datum value (upper tail) and for interpolation between two consecutively ranked data. There is little advantage to smoothing if the goal is to inform quantiles near the center of the distribution (Sheather and Marron 1990; Silverman and Young 1987). A need to define extreme low and high quantiles provided added motivation for histogram and scatterplot smoothing.

A second motivation for smoothing is that sawtoothlike spikes appear when there are few sample data. If more data were available, these spikes would not likely appear; they are an artifact of data paucity and should be smoothed out.

The problems of lack of resolution and spikes in the univariate histogram become far worse in the case of a bivariate histogram (scatterplot) due to the number of class probabilities that must be informed. For example, if 100 classes are required for the porosity and permeability histograms, then 10,000 classes would be required for the bivariate histogram. There is rarely enough data to reliably inform the bivariate histogram without modeling or *smoothing*.

A first conventional approach to smooth univariate and bivariate distributions is to fit a parametric distribution (such as a normal, lognormal, or power-law distribution) to the sample data (Johnson and Kotz 1970; Scott 1992). The parametric model then overcomes all problems related to resolution and spikes in the sample histogram. The problem, however, is that real earth-science data can rarely be fitted with simple parametric distributions.

A second approach is to replace each datum with a kernel function (Scott 1992; Silverman 1986)—that is, a parametric probability distribution with a mean equal to the datum value and a small variance. The *smooth* distribution is obtained by integrating these kernel functions. The resulting distribution does not in general (Jones 1991, 1993) simultaneously honor the mean, variance, and quantiles of the sample data and may show negative values although the variable is positive.

This article documents a procedure to smooth univariate and bivariate histograms so that critical summary statistics, that are deemed reliably informed by the sample data, are reproduced. For example, data limits, the sample mean and variance, certain quantiles (such as the median), linear correlation coefficients, bivariate quantiles (for nonlinear behavior in the scatterplot), and measures of smoothness can be imposed. The proposed methodology has several advantages: (a) It is conceptually simple; (b) bounded variables, such as volumetric concentrations, are handled quite naturally; and (c) the approach does not introduce any bias in the mean, variance, or shape when considering a highly skewed distribution.

1.1 Notation

Although this article focuses on porosity ϕ and perme-

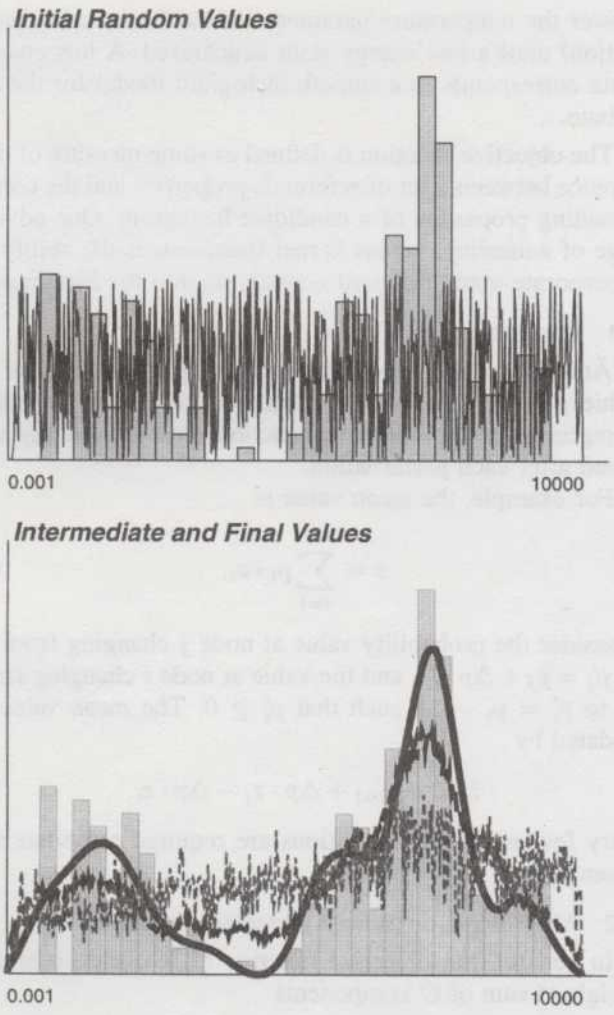

Figure 1. The Evolution of the Probabilities From Random to Smoothed Probabilities: - - -, Intermediate (one-third finished); – – –, Intermediate (two-thirds finished); ———, Final Values.

ability K, the approach is generally applicable to histograms of any attribute(s).

The ϕ and K values are each modeled by a random variable Z informed by n data values—$z(\mathbf{u}_\alpha), \alpha = 1, \ldots, n$, with $\mathbf{u}_\alpha$ being the location-coordinates vector for datum α. The data values are often correlated with each other in space (close values being more similar) and spatially clustered in areas of relative importance (i.e., zones of high permeability), yet we require a histogram that is representative of the entire area of interest. To obtain a representative distribution, one approach is to assign declustering weights whereby values in zones with more data receive less weight than those in sparsely sampled zones. The program DECLUS (Deutsch 1989; Deutsch and Journel 1992) provides an algorithm for determining such declustering weights. The declustering weight assigned to each datum location is denoted by ω_α with $\omega_\alpha \in [0, 1]$ and $\sum_{\alpha=1}^{n} \omega_\alpha = 1$. An equally weighted histogram would correspond to $\omega_\alpha = 1/n$, for all α.

A histogram $f(z)$ gives the probability of encountering any z value for the attribute being studied. A cumulative histogram or cumulative distribution function, cdf $F(z)$, is useful because Monte Carlo simulation proceeds by drawing from such cdf's. The cdf $F(z)$ can be modeled by, at most, n step functions if all n data values $z(\mathbf{u}_\alpha)$ are different:

$$F^*(z) = \sum_{\alpha=1}^{n} \omega_\alpha \cdot i(\mathbf{u}_\alpha; z), \tag{1}$$

with $\omega_\alpha \in [0, 1], \sum_{\alpha=1}^{n} \omega_\alpha = 1$, where $i(\mathbf{u}_\alpha; z)$ is the indicator datum, set to 1 if $z(\mathbf{u}_\alpha) \leq z$, to 0 otherwise, and ω_α is the declustering weight attached to datum location $\mathbf{u}_\alpha$.

The problem of modeling a histogram is one of assigning N probability values $p_i, i = 1, \ldots, N$, to evenly spaced z values between the minimum $z_{\min}$ and maximum $z_{\max}$. The minimum $z_{\min}$ and maximum $z_{\max}$ are chosen on the basis of physical limits (e.g., 0 to 100%) or, in the case of unbounded distributions, practical bounding limits.

In general, $N \gg n$; that is, the resolution is enhanced in the modeling. In cases in which n is acceptably large, there may still be a need to smooth the distribution to a more continuous function. The challenge is to find a set of $p_i, i = 1, \ldots, N$, values such that the resulting smooth distribution is close to the original declustered data distribution and reproduces summary statistics such as the sample mean, variance, and selected quantiles.

The problem of smoothing a bivariate histogram is one of assigning $N_1 \cdot N_2$ probability values $p_{i,j}, i = 1, \ldots, N_1; j = 1, \ldots, N_2$, to evenly spaced z_1 (say, porosity) and z_2 (say, permeability) values. The resulting smooth distribution should capture the important features of the original data distribution and honor summary statistics deemed reliable, such as the smoothed marginal distributions, sample mean(s), variance(s), correlation coefficient, and selected quantiles.

1.2 The Proposed Solution

The problem of finding the set of smooth univariate probabilities $p_i, i = 1, \ldots, N$, or bivariate probabilities $p_{i,j}, i = 1, \ldots, N_1; j = 1, \ldots, N_2$, is solved with simulated annealing (Journel 1993). An objective function may be constructed as the sum of multiple constraints such as reproduction of the sample mean, variance, selected quantiles, and minimization of a smoothness measure.

The initial set of N probabilities $p_i, i = 1, \ldots, N$, is assigned randomly and then perturbed according to a standard annealing approach [see Sec. 2 and Kirkpatrick, Gelatt, and Vecchi (1983); Press, Flannery, Teukolsky, and Vetterling (1986)] until an optimal set is found. Figure 1 shows an initial set of randomly assigned probability values ($N = 500$), two sets of intermediate probabilities, and the final *smoothed* probabilities. The computer program takes between 1 to 5 seconds on a Silicon Graphics Indy workstation to arrive at the final smooth probabilities while reproducing the original sample mean, median, variance, 11 evenly spaced quantiles, and imposed minimum and maximum values.

There are other optimization techniques such as genetic algorithms, threshold accepting, and steepest descent that should give comparable results to those presented here. The

remainder of this article documents the procedure in more detail and presents several real examples.

2. SIMULATED ANNEALING

The essential feature of the method is the formulation of the modeling problem as an optimization problem to be solved with simulated annealing.

The global optimization technique known as simulated annealing is based on an analogy with the physical process of annealing. Annealing is the process by which a material undergoes extended heating and is slowly cooled. Thermal vibrations permit a reordering of the atoms/molecules to a highly structured lattice—that is, a low energy state. In the context of smoothing histograms, the *annealing* process may be simulated through the following steps:

1. An initial histogram (analogous to the initial melt in true annealing) is created by assigning a random probability value for a series of regularly spaced z values.
2. An energy function (analogous to the Gibbs free energy in true annealing) is defined as a measure of difference between desired features and those of the realization. For example, one component of the energy or objective function could be the squared difference between the mean of the realization and a model mean. Another component could be a measure of smoothness.
3. The probability values are perturbed by choosing a pair of values (with indices i and j) and adding an incremental value Δp to one and subtracting it from the other (this mimics the thermal vibrations in true annealing).
4. The perturbation (thermal vibration) is accepted if the energy is decreased; it is accepted with a certain probability even if the energy is increased (the Boltzmann probability distribution of true annealing). Technically the name "simulated annealing" only applies when the acceptance probability is based on the Boltzmann distribution (Kirkpatrick et al. 1983). In common usage, however, the name "annealing" is used to describe the entire family of methods that are based on the principle of stochastic relaxation.
5. The perturbation procedure is continued while reducing the probability that unfavorable swaps are accepted (lower the temperature parameter of the Boltzmann distribution) until a low energy state is achieved. A low energy state corresponds to a smooth histogram model for the attribute.

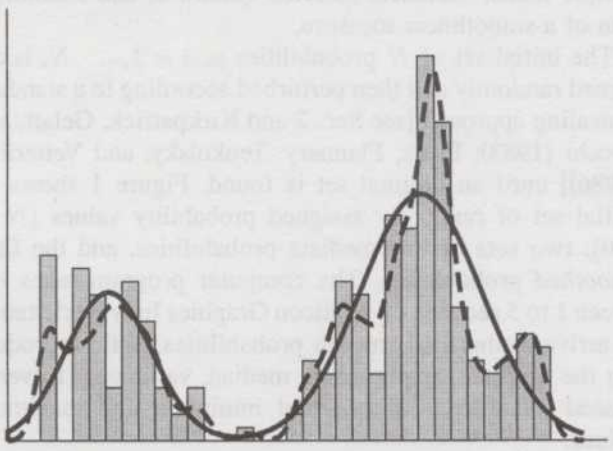

Figure 2. The Smoothed Distribution Obtained by Fitting Two Gaussian Distributions: ——, Two Normal Distributions; – – –, Kernel Smoothing.

The objective function is defined as some measure of difference between a set of reference properties and the corresponding properties of a candidate histogram. One advantage of annealing, versus kernel functions, is the ability to incorporate many different constraints into the histogram.

2.1 Updating

Annealing techniques rely on many perturbations to achieve a final acceptable realization. This implies that each component of the objective function must be quickly updated after each perturbation.

For example, the mean value is

$$\bar{z} = \sum_{i=1}^{n} p_i \cdot z_i. \tag{2}$$

Consider the probability value at node j changing from p_j to $p'_j = p_j + \Delta p \leq 1$ and the value at node i changing from p_i to $p'_i = p_i - \Delta p$ such that $p'_i \geq 0$. The mean value is updated by

$$\bar{z}_{\text{new}} = \bar{z}_{\text{old}} + \Delta p \cdot z_j - \Delta p \cdot z_i. \tag{3}$$

Very few arithmetic operations are required to update the mean value.

2.2 Weighting Component Objective Functions

In general, the objective function O is made up of the weighted sum of C components

$$O = \sum_{c=1}^{C} \nu_c O_c, \tag{4}$$

where ν_c and O_c are the weights and component objective functions, respectively. The component objective functions measure how certain features of the simulated image differ from the desired control or reference properties. For example, one component could be a measure of difference between the variance of the smoothed model and the variance of the sample data, a second component could measure reproduction of specific quantiles, and a third component could measure the smoothness of the histogram.

Each component objective function O_c could be expressed in widely different units of measurement. For example, a component measuring variance departure may be in units squared (>1,000), whereas a component measuring the reproduction of a correlation coefficient may be quite small (<.5).

The weights ν_c allow equalizing the contributions of each component in the global objective function. Decisions of whether to accept or reject a perturbation are based on the change to the objective function,

$$\Delta O = O_{\text{new}} - O_{\text{old}}, \quad \text{with}$$

$$\Delta O = \sum_{c=1}^{C} \nu_c [O_{c_{\text{new}}} - O_{c_{\text{old}}}] = \sum_{c=1}^{C} \nu_c \Delta O_c. \tag{5}$$

The weights $\nu_c, c = 1, \ldots, C$, are established so that, on average, each component contributes equally to the change in the objective function ΔO. That is, each weight ν_c is made inversely proportional to the average change in absolute value of the component objective function:

$$\nu_c = \frac{1}{\overline{|\Delta O_c|}}, \qquad c = 1, \ldots, C. \tag{6}$$

In practice, the average change of each component $\overline{|\Delta O_c|}$ may be numerically approximated by evaluating the average change due to a certain number M (say 1,000) of independent perturbations:

$$\overline{|\Delta O_c|} = \frac{1}{M} \sum_{m=1}^{M} |O_c^{(m)} - O_c|, \qquad c = 1, \ldots, C, \tag{7}$$

where $\overline{|\Delta O_c|}$ is the average change for component c, $O_c^{(m)}$ is the perturbed objective value, and O_c is the initial objective value.

The overall objective function may then be written as

$$O = \frac{1}{O^{(0)}} \cdot \sum_{c=1}^{C} \nu_c \cdot O_c. \tag{8}$$

The objective function O is normalized by its initial value, $O^{(0)}$, so that it starts at 1.0.

3. HISTOGRAMS

Consider the problem of assigning N probability values $p_i, i = 1, \ldots, N$, to evenly spaced z values between given minimum $z_{\min}$ and maximum $z_{\max}$. The equal spacing of the z_i values is

$$\Delta z = \frac{1}{N} \cdot (z_{\max} - z_{\min}), \tag{9}$$

with $p_i \geq 0$ for all $i = 1, \ldots, N, z_1 = z_{\min}$, and $z_N = z_{\max}$. The idea is to choose N large (100–500) so that the resulting distribution can be reliably used for subsequent stochastic/geostatistical simulation.

The final set of smoothed probabilities is established from an initial set of probabilities by successive perturbations. The perturbation mechanism consists of selecting at random two indices i and j such that $i \neq j$ and $i, j \in]1, N[$. The probability values at i and j are perturbed as follows:

$$p_i^{(\text{new})} = p_i + \Delta p$$

$$p_j^{(\text{new})} = p_j - \Delta p.$$

The incremental change Δp is calculated as $\Delta p = .005 \cdot U$, where .005 is a constant chosen to dampen the magnitude of the perturbation (found by trial and error) and U is a pseudorandom number $U \in [0, 1]$. Both $p_i^{(\text{new})}$ and $p_j^{(\text{new})}$ must be between 0 and 1; new pseudorandom numbers are drawn until this condition is met.

If the initial $p_i, i = 1, \ldots, N$, values are legitimate (i.e., $p_i \in [0, 1]$ for all i, and $\sum_{i=1}^{N} p_i = 1$), then any set of probabilities derived from multiple applications of this perturbation mechanism is also legitimate.

For simplicity, all perturbations that lower the global objective function (defined later as the weighted sum of component objective functions) are accepted and all perturbations that increase the objective function are rejected.

The following component objective functions have been considered:

1. For the mean,

$$O_m = [\bar{z} - m_z]^2, \tag{10}$$

where m_z is the target mean (from the data or specified by the user) and $\bar{z}$ is the average from the smoothed distribution

$$\bar{z} = \sum_{i=1}^{N} p_i \cdot z_i. \tag{11}$$

2. For the variance,

$$O_v = [s^2 - \sigma^2]^2, \tag{12}$$

where σ^2 is the target variance (from the data or specified by the user) and s^2 is the variance from the smoothed distribution

$$s^2 = \sum_{i=1}^{N} p_i \cdot z_i^2 - \bar{z}^2. \tag{13}$$

3. For a number of quantiles,

$$O_q = \sum_{i=1}^{n_q} [P_i - F(z_i)]^2, \tag{14}$$

where n_q is the number of imposed quantiles, P_i is the smoothed cdf value for threshold z_i, and $F(z_i)$ is the target cumulative probability (from the data). For example, the median porosity value of 12.5 would be specified as $F(z_1) = .5$ and $z_1 = 12.5$. The cumulative P-probability value associated with any z threshold can be calculated by summing the p_i values until the corresponding z_i value exceeds the threshold z; that is, first establish the index k threshold value

$$k = \text{ integer portion of } \frac{z - z_{\min}}{\Delta z},$$

then the p value is computed as

$$P = \sum_{i=1}^{k} p_i. \tag{15}$$

4. The smoothness of the set of probabilities $p_i, i = 1, \ldots, N$, can be measured by summing the squared difference between each p_i and a smooth $\hat{p}_i$ defined as an average of the values surrounding i; that is,

$$O_s = \sum_{i=1}^{N} [p_i - \hat{p}_i]^2, \tag{16}$$

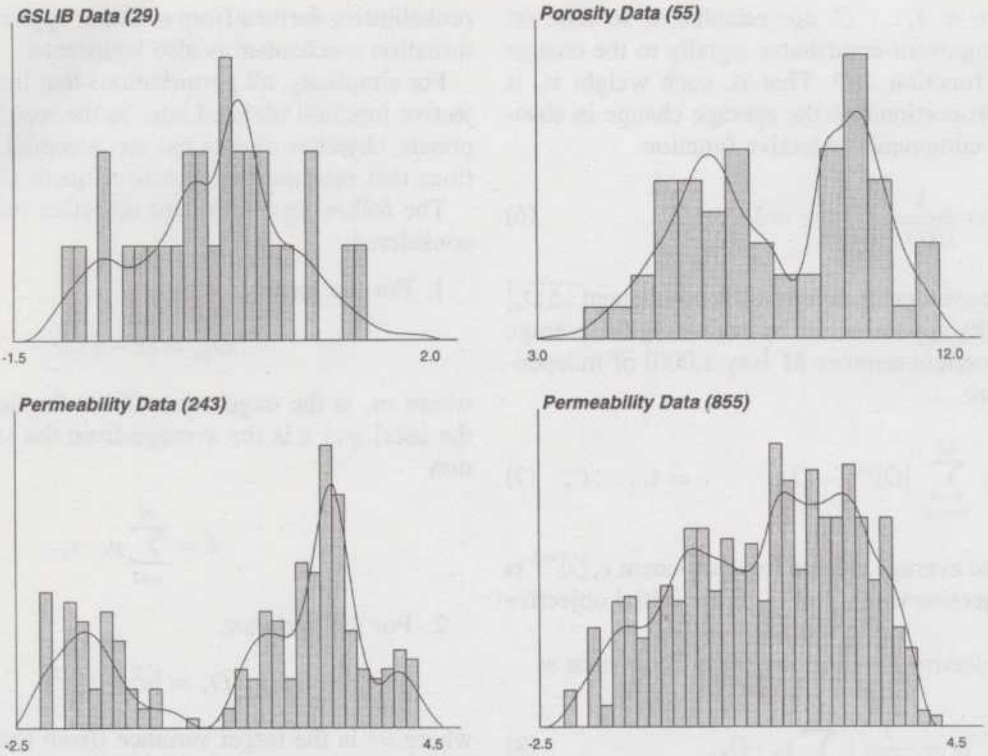

Figure 3. Four Examples of Histogram Smoothing: Clockwise From the Upper Left, 29 Data Taken From GSLIB (Deutsch and Journel 1992), 55 Porosity Data, 243 Permeability Data, and 855 Permeability Data. All four datasets are unrelated.

where the $p_i, i = 1, \ldots, N$, are the smoothed probability values and $\hat{p}_i, i = 1, \ldots, N$, are local averages of the p_i values

$$\hat{p}_i = \frac{1}{2 \cdot n_0} \sum_{k=-n_0, k \neq i}^{n_0} p_{i+k}, \qquad i = 1, \ldots, N, \quad (17)$$

where n_0 is the number of values in the smoothing window (say, 5–10), $p_i = 0$ for all $i \leq 1$, and $i \geq N$.

The global objective function is defined as the sum of the four components.

$$O = \nu_m \cdot O_m + \nu_v \cdot O_v + \nu_q \cdot O_q + \nu_s \cdot O_s, \quad (18)$$

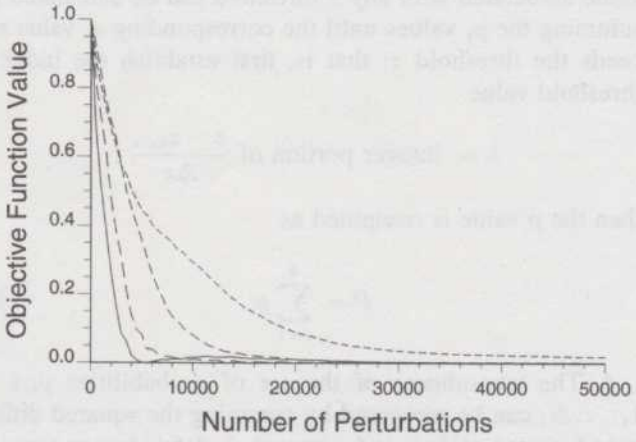

Figure 4. The Four Component Objective Functions (normalized to start at 1.0) Versus the Number of Perturbations. The plot only goes to 50,000 perturbations, although the program was allowed to go to 250,000 at which time the quantile component (not yet 0 at 50,000) was .0087: ——, mean; — — —, variance; – – –, smoothness; - - -, quantiles.

and the weights ν_m, ν_v, ν_q, and ν_s are computed such that each component has an equal contribution to the global objective function (as defined previously). Note that additional components could be added as necessary.

Examples

Figure 1 shows an initial set of randomly assigned probability values, two sets of intermediate probabilities, and the final *smoothed* probabilities. Two conventional smooth-

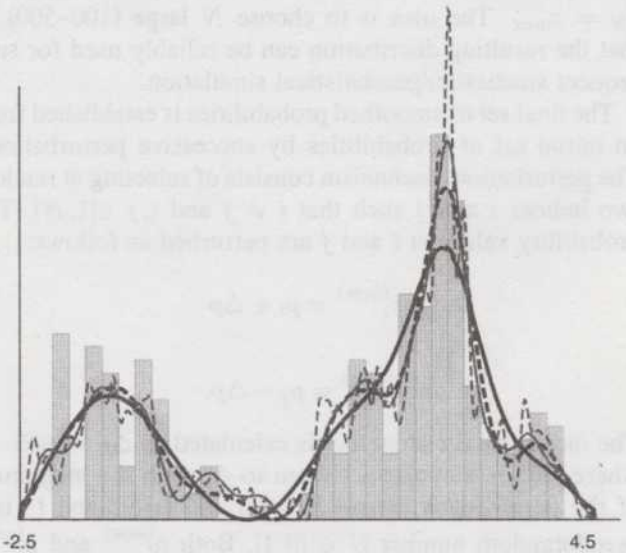

Figure 5. The Smoothed Probabilities for a Smoothing Half Window Varying From 1 to 20. In all cases the mean, variance, and 12 quantiles taken from the 243 original data are honored: – – –, half window = 1; ——, half window = 5; - - -, half window = 10, ——, half window = 20.

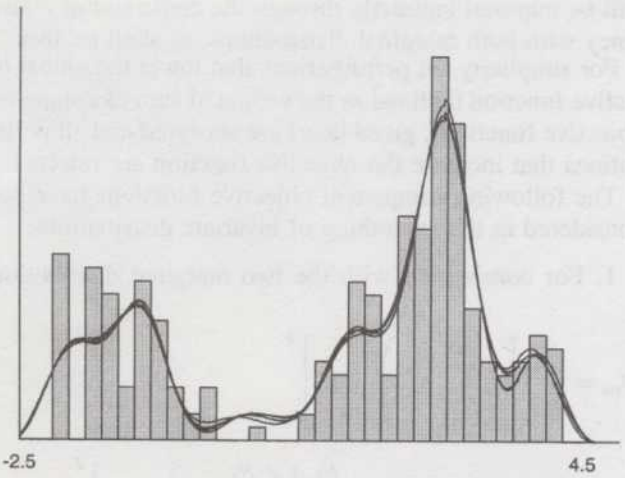

Figure 6. Five Sets of Smoothed Probabilities Obtained When Using Five Different Random Number Seeds. All other parameters were unchanged.

ing approaches have also been applied to the same data: The fit with two Gaussian distributions is shown in Figure 2, and the smoothed distribution using a Gaussian kernel function is also shown in Figure 2 (p. 268). By construction, the annealing-based solution honors the mean, variance, and quantiles better than either conventional approach. The importance of this depends on the application. Some considerations for selecting a method include the following questions: (1) What is the smoothed distribution going to be used for? (2) Are there physical reasons why the data should follow a specified parametric model? (3) How important is the skewness and fit in the tails of the distribution? The annealing-based approach appears to be a flexible approach that may be appropriate in many situations.

Figure 3 shows four additional sample histograms with their smoothed counterparts. In each case, the smoothed histogram contains 500 discretization points, the mean and variance were taken from the data, and a smoothing window (half width) of 10 was considered.

In all instances the final smoothed histogram honors all the input constraints. That is, the smoothing was considered finished when each component of the objective function dropped below .001.

Figure 4 shows the evolution of the component objective functions as the smoothing proceeds for the example of Figure 1. The components go to 0 at different rates. Figure 1 shows how the set of probabilities evolves as the smoothing proceeds.

Table 1. The mAD Statistic That Measures how Well the Smoothed Distribution Matches the Underlying Reference Distribution

Smoothing technique	*Gaussian*	*Lognormal*	*Exponential*
Annealing-based	.00188	.00919	.00236
Kernel	.00214	.00851	.00379
Fitted parametric model	.00146	.00408	.00118

NOTE: Results are shown for three reference-true distributions (Gaussian, lognormal, and exponential) and for three smoothing techniques (the annealing-based approach, a kernel approach, and a fitted parametric approach using the correct parametric model).

Figure 5 shows the smoothed distribution for smoothing half windows of 1, 5, 10, and 20. In general, there are no fixed guidelines about which is the best. An attempt at optimizing the smoothing parameters on the basis of cross-validation scores may be possible (Scott 1992). Significant details may be lost if the distribution is smoothed too much. On the other hand, artifacts of scarce data are preserved if the smoothing window is too small.

The distribution model obtained by simulated annealing is nonunique; a different random-number seed would lead to a different series of perturbations and ultimately to a different final set of probabilities. Figure 6 shows five smoothed distribution models obtained from five different random-number seeds. All five models appear very similar.

A small simulation study was carried out in an attempt to quantify the performance of the annealing-based smoothing methodology. The exercise consisted of smoothing distributions of 25 sample values drawn at random from reference probability distributions. The values were smoothed with the annealing-based approach presented previously, a Gaussian kernel approach, and then by fitting the parameters of the correct underlying parametric model. The mean integrated squared error optimal value of $(1.059n^{-1/5})^2 = .31$ for the kernel bandwidth was chosen.

Three reference distributions were considered: (1) the standard Gaussian distribution, (2) a lognormal distribution with a mean of 1.0 and variance of 4.0, and (3) an exponential distribution with a mean of 1.0. In each case, 100 randomly chosen samples of 25 values were considered.

To quantify the performance of the smoothing algorithms, the mean absolute deviation (mAD) for 100 probability values associated with evenly spaced z values from the .05 quantile and the .95 quantile were calculated. These mAD statistics are measures of difference between the smoothed distributions and the underlying true reference distributions. Table 1 shows the results. In all cases, knowing the correct parametric model yields the best result. The annealing-based smoothing method outperforms the kernel approach in two out of the three cases considered.

The documentation of a rigorous Monte Carlo study would make this article too long; however, the results of these three tests show that the simulated annealing smoothing method yields reasonable results.

4. SCATTERPLOTS

Consider the problem of finding a set of smooth bivariate probabilities $p_{i,j}, i = 1, \ldots, N_1, j = 1, \ldots, N_2$, applied to attribute z_1 (say, porosity) and attribute z_2 (say, permeability). There are many unknowns in the bivariate case; for example, if N_1 and N_2 are both 250, then the number of bivariate probabilities is $N_1 \cdot N_2 = 62{,}500$. Given a *smoothed* set of bivariate probabilities, the marginal univariate probabilities are retrieved by row and column sums; that is, the univariate distribution of z_1 is given by the set of probabilities

$$p_i = \sum_{j=1}^{N_2} p_{i,j}, \qquad i = 1, \ldots, N_1,$$

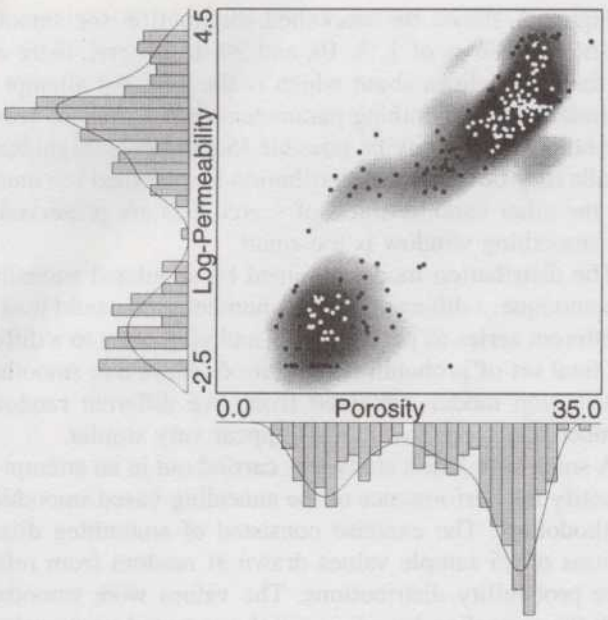

Figure 7. A Smoothed Scatterplot for 243 Porosity/Permeability Data.

and the z_2 distribution by

$$p_j = \sum_{i=1}^{N_1} p_{i,j}, \qquad j = 1, \ldots, N_2.$$

The approach taken here is to impose the already-smoothed univariate distributions on the smooth bivariate distribution model. The disadvantage of this approach is that more consistent distribution models could be obtained by establishing the bivariate distribution directly. The following advantages are considered more important:

1. The problem of computing smooth bivariate probabilities is easier once the marginal distributions are fixed. Otherwise, up to eight additional constraints, four per marginal as in Relation (18), would have to be added to the bivariate problem.
2. There is often more data informing one or both marginal distribution(s) than there are data pairs informing the scatterplot. Smoothing the marginals first allows this data to be easily accounted for.

The final set of smoothed bivariate probabilities is established from an initial set of probabilities by successive perturbations. The perturbation mechanism consists of selecting a bivariate index (i', j') and considering a perturbation of $p_{i',j'}^{(\text{new})} = p_{i',j'} + \Delta p$, with Δp chosen such that the candidate probability value $(p_{i',j'}^{(\text{new})})$ is positive: $\Delta p = .1(U - .5) \cdot p_{i',j'}$, where .1 is a constant chosen to dampen the magnitude of the perturbation (found by trial and error) and U is a pseudorandom number $U \in [0, 1]$.

The physical constraint on the sum of the bivariate probabilities

$$\sum_{i=1}^{i=N_1} \sum_{j=1}^{j=N_2} p_{i,j} = 1$$

will be imposed indirectly through the constraint of consistency with both marginal distributions, as shall be seen.

For simplicity, all perturbations that lower the global objective function (defined as the weighted sum of component objective functions, given later) are accepted and all perturbations that increase the objective function are rejected.

The following component objective functions have been considered in the smoothing of bivariate distributions:

1. For consistency with the two marginal distributions,

$$O_m = \sum_{i=1}^{N_1} \left[\left(\sum_{j=1}^{N_2} p_{i,j} \right) - p_i^* \right]^2 + \sum_{j=1}^{N_2} \left[\left(\sum_{i=1}^{N_1} p_{i,j} \right) - p_j^* \right]^2, \quad (19)$$

where p_i^* is the smoothed marginal probability of z_{1_i} and p_j^* is the smoothed marginal probability of z_{2_j}; both p_i^* and p_j^* are target values.

2. For the correlation coefficient,

$$O_c = [\rho - \rho^*]^2, \quad (20)$$

where ρ is the correlation coefficient from the bivariate probabilities and ρ^* is the target correlation coefficient (from the data or specified by the user).

3. For several bivariate quantiles

$$O_q = \sum_{i=1}^{n_q} [P_i - F(z_{1_i}, z_{2_i})]^2, \quad (21)$$

where n_q is the number of imposed quantiles and $F(z_{1_i}, z_{2_i})$ is the ith target cumulative probability. The cumulative P-probability value associated with any z_{1_i}, z_{2_i} pair can be calculated by summing the $p_{i,j}$ values for all z_1, z_2 threshold values less that the threshold values z_{1_i}, z_{2_i}.

4. The smoothness of the set of probabilities $p_{i,j}, i = 1, \ldots, N_1, j = 1, \ldots, N_2$, may be quantified by summing the squared difference between each $p_{i,j}$ and a smooth $\hat{p}_{i,j}$ defined as an average of the values surrounding i; that is,

$$O_s = \nu_s \cdot \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [p_{i,j} - \hat{p}_{i,j}]^2, \quad (22)$$

where the $p_{i,j}, i = 1, \ldots, N_1, j = 1, \ldots, N_2$, are the smoothed probability values and $\hat{p}_{i,j}, i = 1, \ldots, N_1, j = 1, \ldots, N_2$, are local averages of the $p_{i,j}$ values. These local averages are defined by values within an elliptical window. The major axis of the elliptical window is in the direction of greatest correlation and the anisotropy of the window is based on the anticipated degree of correlation; that is,

$$\text{angle from } Z_1 \text{ axis} = \tan^{-1}\left(\frac{\sum_{i=1}^{n} (z_{1_i} - \overline{z_1})^2}{\sum_{i=1}^{n} (z_{2_i} - \overline{z_2})^2} \right)$$

$$\text{anisotropy} = \sqrt{1 - \rho^{*2}},$$

where $z_{1_i}, z_{2_i}, i = 1, \ldots, n$, are the available data pairs, $\overline{z_1}$ is the corresponding average of z_1, $\overline{z_2}$ is the average of z_2,

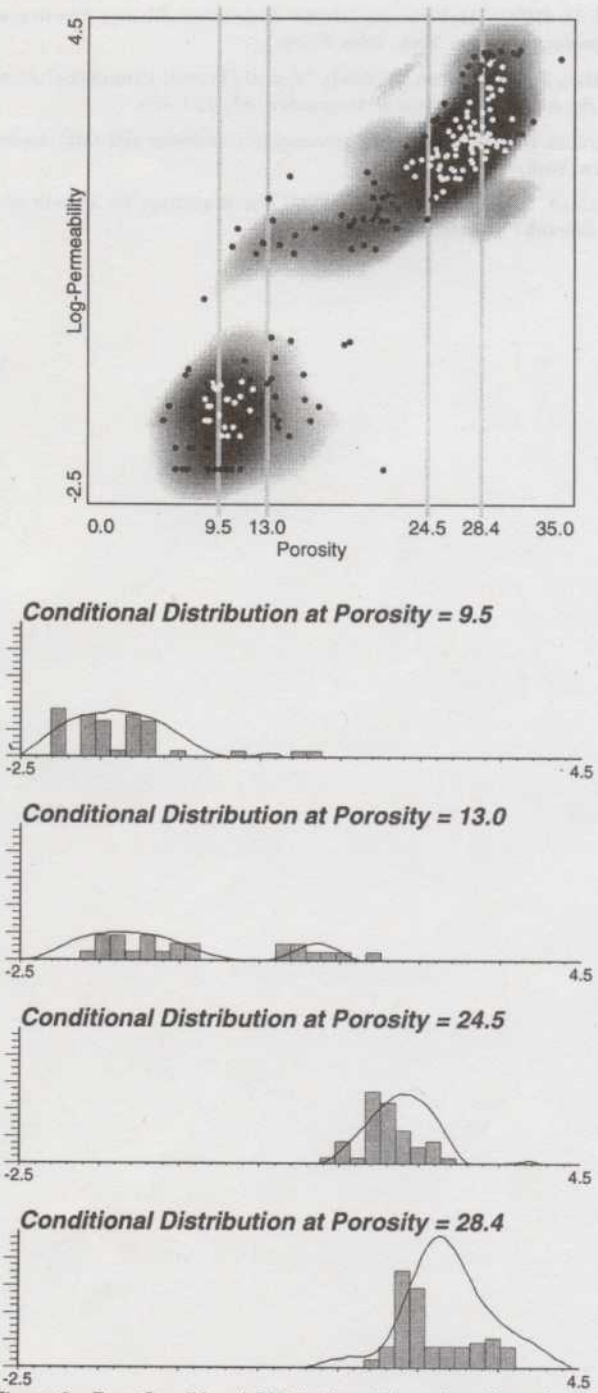

Figure 8. Four Conditional Distributions Through a Smoothed Scatterplot.

and ρ^* is the anticipated (target) correlation coefficient. The number of points in the smoothing ellipse is an input parameter.

Examples

Figure 7 shows a smoothed scatterplot for 243 porosity/permeability data pairs. The smoothed porosity and permeability distributions, shown below the porosity axis and to the left of the permeability axis, were calculated first. The bivariate distribution was then smoothed to honor these two marginal distributions. Figure 8 shows four conditional distributions or *slices* through the smoothed scatterplot. The final smoothed distribution is constrained to the marginal distributions, 11 by 11 bivariate quantiles, and a measure of smoothness. The computer time for this smoothing exercise was 303 seconds (about 5 minutes) on a Silicon Graphics Indy workstation.

5. REMARKS AND CONCLUSIONS

Accurate prediction of petroleum-reservoir performance requires representative numerical models of the spatial distributions of porosity ϕ and permeability K. The main idea behind geostatistical simulation techniques is to build numerical models that honor all of the available information. In addition to location-dependent well log and core data, geostatistical simulation techniques require global statistics such as the histogram of porosity, the histogram of permeability, and the scatterplot of porosity and permeability.

Typically, a geostatistical model may contain 1 to 100 million geological modeling cells. Were the data available, the histogram of 1–100 million ϕ or K values would not show the sawtoothlike spikes and gaps that appear in most sample histograms. For this reason, the sample histogram should be smoothed. An algorithm, based on simulated annealing, has been developed to smooth data histograms constrained to important summary statistics (data limits, mean, variance, specified quantiles, and measures of smoothness).

Bivariate histograms require many more classes to be defined than the univariate case. The smoothing algorithm developed for the univariate case has been extended to the bivariate case. Additional constraints include the linear correlation coefficient and consistency with both marginal distributions.

It is recommended that the histograms (possibly with declustering weights) of ϕ and K be smoothed prior to building stochastic models. Similarly, and more importantly, it is recommended that the scatterplot of porosity and permeability be smoothed prior to stochastic simulation.

ACKNOWLEDGMENTS

I thank A. G. Journel of Stanford University for his original suggestion of applying simulated annealing to the problem of modeling histograms. I also thank the management of Exxon Production Research Company for allowing publication of this article; this work was done while I was employed by Exxon Production Research Company.

[Received May 1994. Revised December 1995.]

REFERENCES

Deutsch, C. (1989), "DECLUS: A Fortran 77 Program for Determining Optimum Spatial Declustering Weights," *Computers & Geosciences*, 15, 325–332.

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

Johnson, N., and Kotz, S. (1970), *Continuous Univariate Distributions—1*, New York: John Wiley.

Jones, M. (1991), "On Correcting for Variance Inflation in Kernel Density Estimation," *Computational Statistics and Data Analysis*, 11, 3–15.

——— (1993), "Simple Boundary Correction for Kernel Density Estimation," *Statistical Computing*, 3, 135–146.

Journel, A. (1993), "Smoothing Sample Histograms Under Constraints," unpublished research notes, Stanford University, Dept. of Geological and Environmental Sciences.

Kirkpatrick, S., Gelatt, C., Jr., and Vecchi, M. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986), *Numerical Recipes*, New York: Cambridge University Press.

Scott, D. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley.

Sheather, S., and Marron, J. (1990), "Kernel Quantile Estimators," *Journal of the American Statistical Association*, 85, 410–416.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, New York, Chapman & Hall.

Silverman, B., and Young, G. (1987), "The Bootstrap: To Smooth or not to Smooth?" *Biometrika*, 74, 469–479.