Theoretical Insight and Practical Implementation Details of Stepwise Conditional Transformation Technique

Oy Leuangthong (oy@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Clayton V. Deutsch (cdeutsch@ualberta.ca)
Department of Civil & Environmental Engineering, University of Alberta

Abstract

Gaussian simulation techniques are the most common and simple simulation approaches used in reservoir modeling. The use of Gaussian techniques requires that model variables be multivariate Gaussian; however, earth science phenomena are rarely Gaussian. Transformation techniques are applied to make the model variables Gaussian. The conventional technique is the normal score transform (also known as the 'graphical' or 'quantile' transformation). This technique generates univariate Gaussian distributions but does not enforce bivariate or higher order Gaussianity.

The stepwise conditional transformation technique introduced by M. Rosenblatt [7] promises to greatly simplify cosimulation of multiple variables. Several issues were identified for further research at the conclusion of the second CCG report[5]. This paper addresses the outstanding issues, including the effects of transformation ordering, independence at lag $\boldsymbol{h} > 0$*, influence of the cross covariance structure and scatterplot smoothing in presence of sparse data.*

Introduction

The increasing demand for realistic reservoir models has brought greater attention to the field of geostatistics. Gaussian techniques are commonly applied to create numerical models because of their simplicity. Implicit to this group of techniques is the requirement for multivariate Gaussianity; however, geologic data rarely conform to such well behaved distributions.

Application of conventional data transformation techniques can generate univariate Gaussian distributions, but do not ensure bivariate or multivariate Gaussianity. Instead, the multivariate distributions may show signs of non-linearity, mineralogical constraints, and heteroscedasticity (See Figure 1).

The stepwise conditional transformation technique is a powerful and robust technique for multivariate data transformation. The idea of applying this transform to geostatistical modeling was introduced in the previous CCG report [5]. This paper continues to explore the theoretical and practical aspects of applying Rosenblatt's stepwise conditional transformation in a geostatistical framework.

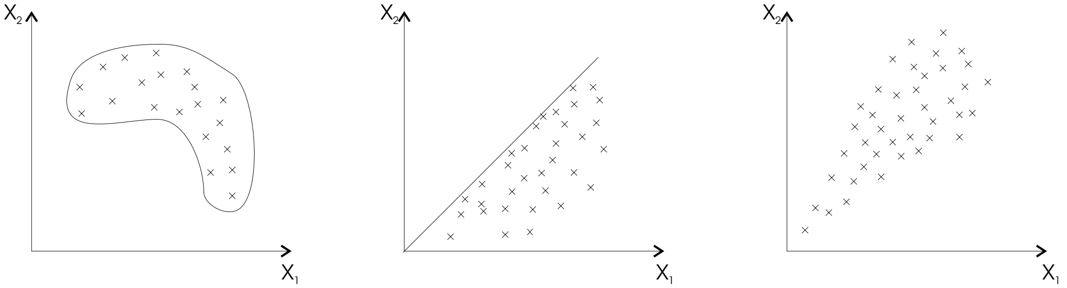

Figure 1: Examples of problematic bivariate distributions for Gaussian simulation: non-linear relations(left), potential mineralogical constraints(centre), and heteroscedasticity (right).

Recall of Stepwise Conditional Transformation

The stepwise-conditional technique is identical to the normal score transform in the univariate case. For bivariate problems, the normal transformation of the second variable is conditional to the probability class of the first variable. Correspondingly, for k-variate problems, the kth variable is conditionally transformed based on the (k-1) th variable, that is,

$$\begin{array}{ll} Y_1 & = G^{-1}[Prob(Z_1 \leq z_1)] \\ Y_{2|1} & = G^{-1}[Prob(Z_2 \leq z_2 \mid Y_1 = y_1)] \\ Y_{3|21} & = G^{-1}[Prob(Z_3 \leq z_3 \mid Y_2 = y_2, Y_1 = y_1)] \end{array}$$

Figure 2 shows the steps to accomplish this conditional transformation for a bivariate case. Once the data are separated into classes based on their conditional probabilities, each group of data is normal score transformed. Simulation is then performed on the normal score values and back transformation is performed in the reverse order of transformation. For example, Z_1 can be determined from Y_1 with the correct conditional distribution; Z_2 can be calculated from Z_1 and the simulated value of Y_2.

Conditional transformation of the data results in transformed secondary variables that are now artificial variables with little physical interpretation. It is a combination of both the primary and the secondary variable. Also, the multivariate spatial relationship of the original model variable is not transformed for $\mathbf{h} > 0$, that is, there is no modification of bivariate spatial distributions $Y(\mathbf{u})$ and $Y(\mathbf{u}+\mathbf{h})$, or trivariate distributions $Y(\mathbf{u})$, $Y(\mathbf{u}+\mathbf{h}_1)$ and $Y(\mathbf{u}+\mathbf{h}_2)$, etc....

The result of this transformation is independence of the transformed variables at $\mathbf{h} = 0$. Since each class of Y_2 data is independently transformed to a normal distribution, correlation between $Y_{2|1}$ and Y_1 is removed at $\mathbf{h} = 0$. Consequently, the simulation of multivariate problems does not require cosimulation due to the independence of the transformed variables. This is the primary motivation for transforming multiple variables in a step-wise conditional fashion.

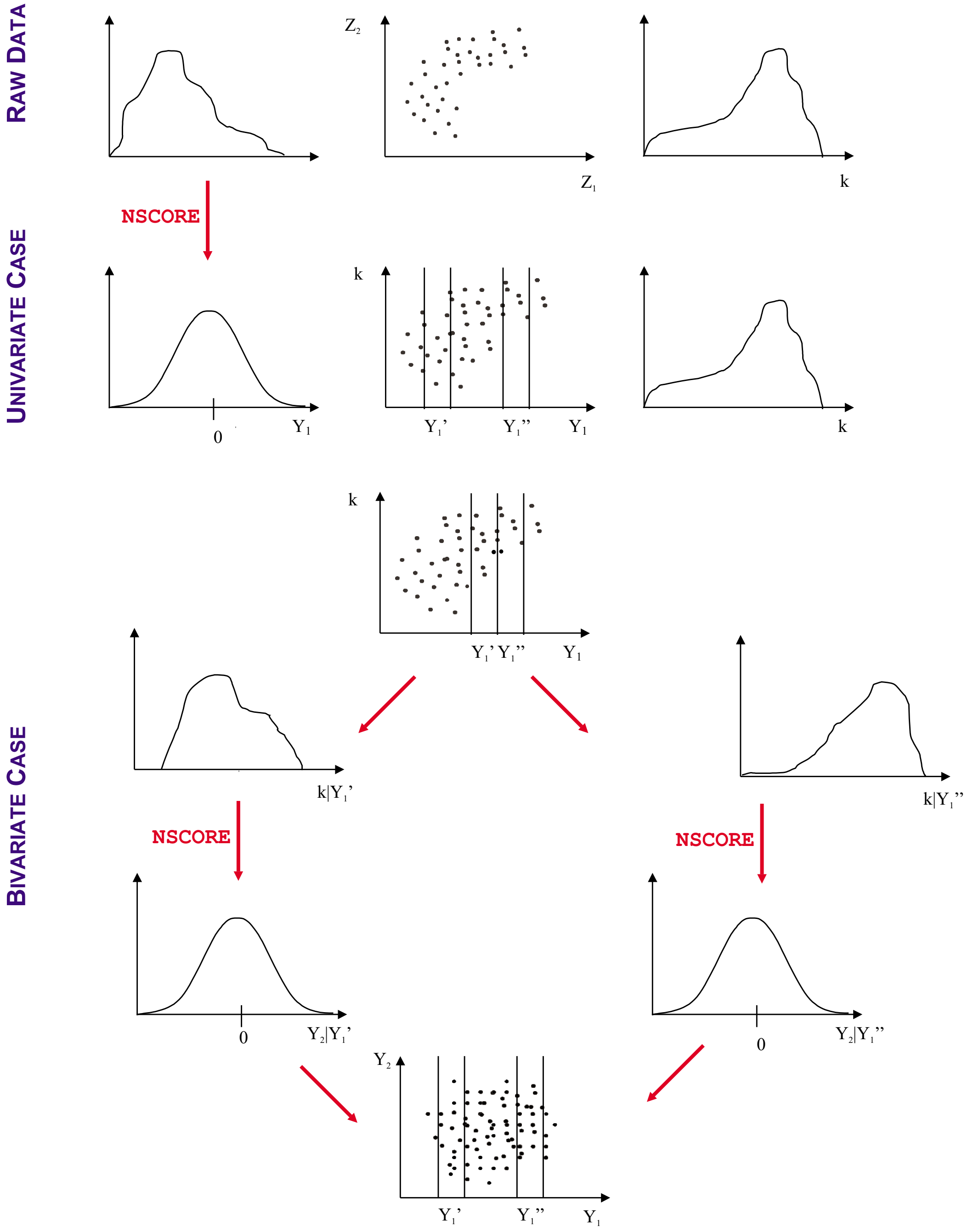

Figure 2: Processes involved in Stepwise Conditional Transformation, a) Distribution of data variables and the scatterplot of Z_2 and Z_1; b) Transform variable Z_1 into normal scored data, Y_1, and categorize Z_2 data conditional to probability classes of Y_1 ; c) Determine conditional distributions for each Z_2 class conditional to Y_1 data; and d) Normally transform each $Z_2|Y_1$ class independent from each other, and the scatterplot between one $Z_2|Y_i$ class and $Z_2|Y_{i+1}$ will show no correlation between the normally transformed classes.

Effect of Ordering

The first transformed variable is the primary variable on which all other variable transformations are based. Consider transforming two variables Z_1 and Z_2, transformation of the primary variable is identical to performing a normal score transform. Two possible scenarios exist for transformation: (1) choose Z_1 as primary variable and normal score transform to get Y_1, and then transform Z_2 to get $Y_{2|1}$; and (2) choose Z_2 as the primary variable to get Y_2, and then Z_1 is transformed to produce $Y_{1|2}$. In case (1), the simulation results for Y_1 would be identical to those obtained by conventional simulation using the normal scores of Z_1, and the same can be said for Y_2 in the second scenario.

Unlike the primary variable, simulation of the secondary variables does not produce the same results as conventional simulation. The secondary variable is a combination of the original variable and the normal scores of the primary variable.

For both ordering sequences, the semivariogram is calculated for both the primary and secondary variable. Sequential Gaussian simulation is independently performed for each transformed variable. Back transformation of the simulated values returns values to the original units. We determine the normal scores semivariogram of the simulated values. The resulting semivariogram of the primary variable is that obtained from the sequential Gaussian simulation. A comparison of the semivariograms for the same variable, when it is taken as (1) the primary variable, and (2) the secondary variable, will show the effect of ordering.

This methodology was applied to several petroleum-related examples. The first data set is the “two-well” data used in the GSLIB training course (see CCG Report 1). The second is East Texas core data where only vertical coordinates are available. For both data sets, porosity and log(permeability) are the two variables of interest, for which the effect of the transformation ordering sequence is examined. The first transformation order takes porosity as the primary variable, and the second transformation order takes log(permeability) as the primary variable.

Conditional simulation was performed with the “two-well” data. Figure 3 shows the comparison of the semivariograms for both ordering sequences using this data set. The semivariograms for porosity show that when porosity is chosen as the primary variable, the post- simulation semivariograms closely follow the input normal scores semivariogram - as it should. Conversely, the semivariograms corresponding to the scenario in which porosity is the secondary variable shows greater variability and a shorter range. Differences in the permeability variograms as a result of transformation ordering sequence are not so obvious; however, closer examination shows slight deviations of the secondary case from the primary case. In this instance, the secondary semivariograms for permeability have longer range.

Unconditional simulation was performed for the East Texas data. Figure 4 shows the comparison of the semivariograms for the East Texas data. Similar to the previous example, each scenario of ordering clearly shows departure of the secondary variable semivariograms from the direct semivariograms using the traditional normal scores. Unlike the Two Well example, the permeability semivariograms differ considerably after stepwise transformation. Further investigation showed that the stepwise transformation produced an artificial secondary variables with significantly different spatial correlation (higher nugget effect and longer range of correlation).

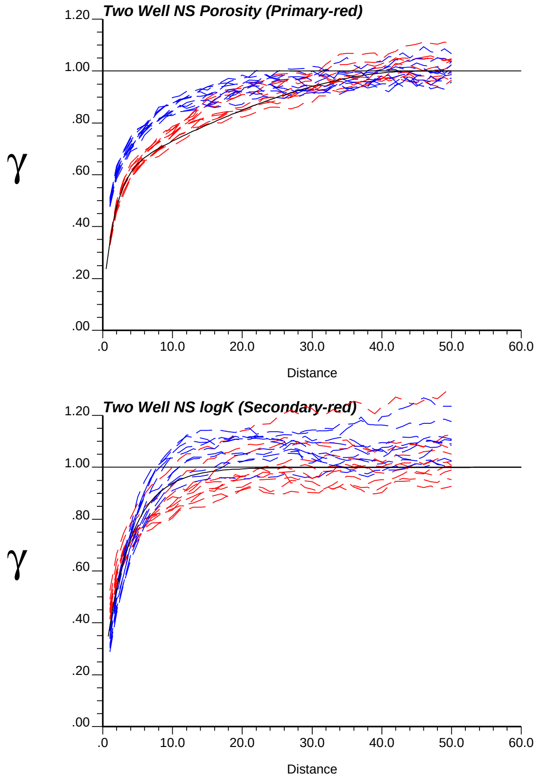

Figure 3: Effect of ordering using Two Well Data: normal scores variogram using simulated data for porosity (top) and log(permeability) (bottom). Variograms in red correspond to the simulation sequence in which porosity is the primary variable, blue variograms correspond to the simulations where log(permeability) is the primary variable. In both variograms, the blue variograms are higher.

The variogram departure of the conditionally transformed variable from that obtained using direct normal scores is attributed to the artificial nature of the transformed secondary data. The stepwise conditional transformation produces a variable that is a combination of the original primary and secondary data. As a result, the spatial variability of the new variable retains some of the spatial structure of both constituent variables. The exact nature of the contribution of the direct and cross covariance structures to the spatial variability of the new variable is explored in more detail in the following section.

Cross covariance structure

The main attraction of the stepwise conditional transformation is the resulting independence of the transformed variables at $\mathbf{h} = 0$. Consequently, the cumbersome modeling of the cross variogram in compliance with the linear model of coregionalization (LMC) can be avoided. A number of exercises were undertaken to gain a better understanding of the structure of the $Y_{2|1}$ variogram.

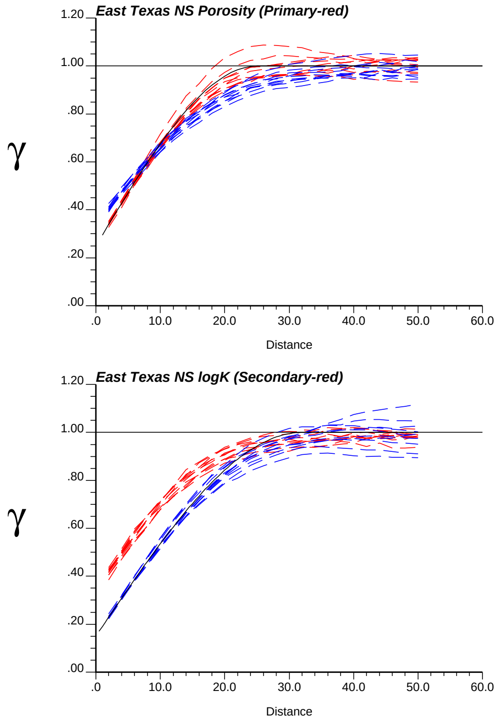

Figure 4: Effect of ordering using East Texas core data: normal scores variogram using simulated data for porosity (top) and log(permeability) (bottom). Variograms in red (i.e. the higher variograms) correspond to the simulation sequence in which porosity is the primary variable, blue variograms correspond to the simulations where log(permeability) is the primary variable.

Without loss of generality, consider two multi-Gaussian variables, Y_1 and Y_2, with the same direct isotropic variogram:

$$\gamma(\mathbf{h}) = 0.5Sph_{a=3}(\mathbf{h}) + 0.5Sph_{a=15}(\mathbf{h})$$

The correlation between Y_1 and Y_2 was chosen to be 0.70. Three different cross variograms were considered: short-range, intrinsic, and long-range. The “short-range” case gave a maximum variance contribution to the short range structure; while the “long- range” case gave the maximum variance contribution to the long range structure. The models are given below and illustrated in Figure 5.

$$\begin{array}{llcl} short-range: & \gamma(\mathbf{h}) & = & 0.50Sph_{a=3}(\mathbf{h}) + 0.20Sph_{a=15}(\mathbf{h}) \\ intrinsic: & \gamma(\mathbf{h}) & = & 0.35Sph_{a=3}(\mathbf{h}) + 0.35Sph_{a=15}(\mathbf{h}) \\ long-range: & \gamma(\mathbf{h}) & = & 0.20Sph_{a=3}(\mathbf{h}) + 0.50Sph_{a=15}(\mathbf{h}) \end{array}$$

For each case, stepwise conditional transformation was applied, direct and cross variograms were calculated and modeled, sequential Gaussian simulation was performed, simulated values were back transformed, and the resulting simulated direct and cross variograms

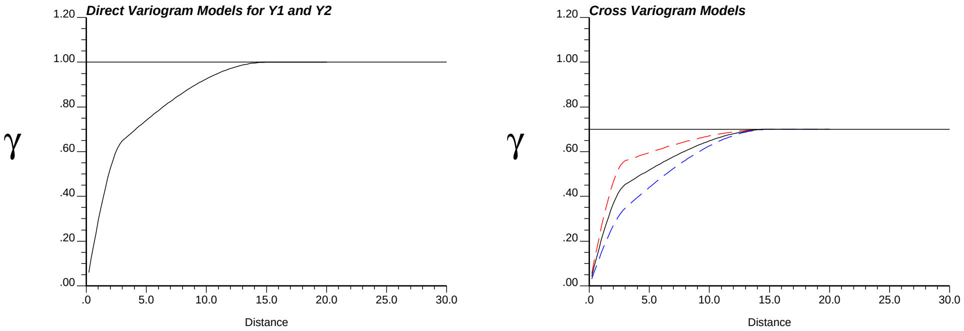

Figure 5: Direct variogram of Y_1 and Y_2 (left), and the three different cross variogram models (right) : short-range (red), intrinsic (black), and long-range (blue).

were examined. Figure 6 shows the direct variograms for $Y_{2|1}$ and the cross variogram of Y_1 and $Y_{2|1}$, following application of the stepwise transform.

In the case where greater contribution is given to the short-range structure, the cross variogram was slightly higher than zero over small lag distances and then returns to zero just past the short-range distance. Conversely, the long-range scenario showed that the cross variogram was negative over the short-range, and then increased to zero beyond the short-range distance. Unlike the two extreme cases, the intrinsic case showed independence of the transformed pairs, with no deviation from zero over all lags. These results confirm that the stepwise transformation implicitly assumes that the direct and cross variograms are intrinsic. Independence at $\mathbf{h} \geq 0$ is satisfied only for the intrinsic case.

Following simulation the values were back transformed and the cross variogram was checked for each scenario. Figure 7 shows the model cross variograms of the original variables and the cross variogram obtained after simulation of the conditionally transformed variables. The relative range of correlation is preserved, i.e. the short range model produces an average cross variogram with the shortest range of the three simulated scenarios. In all three cases, the range of correlation following simulation shows that the stepwise conditional transform reduces the overall range of correlation of the variables. As well, the variogram structure of the extreme cases (short- and long- range cross variograms) appear to be shifted towards the intrinsic model.

To gain a better understanding of the differences between the direct variogram of the transformed variable, $Y_{2|1}$, and the original Y_2 variable, a small analytical exercise was carried out. Two points separated by a distance $\mathbf{h}$ was considered (see Figure 8). At each point, Y_1 data is available and an estimate at $Y_2(\mathbf{u})$ is required using $Y_1(\mathbf{u})$ and $Y_1(\mathbf{u}{+}\mathbf{h})$. Both the original variables, Y_1 and Y_2, are homoscedastic, multi-Gaussian variables with univariate N(0,1) distribution.

To determine the covariance structure of the transformed secondary variable, $Y_{2|1}$, we first need to define its conditional distribution. The parameters that define the conditional distribution of $Y_{2|1}$ can be obtained by solving the kriging system of equations. The mean and variance of the conditional distribution are given by the kriged estimate and the error variance, respectively.

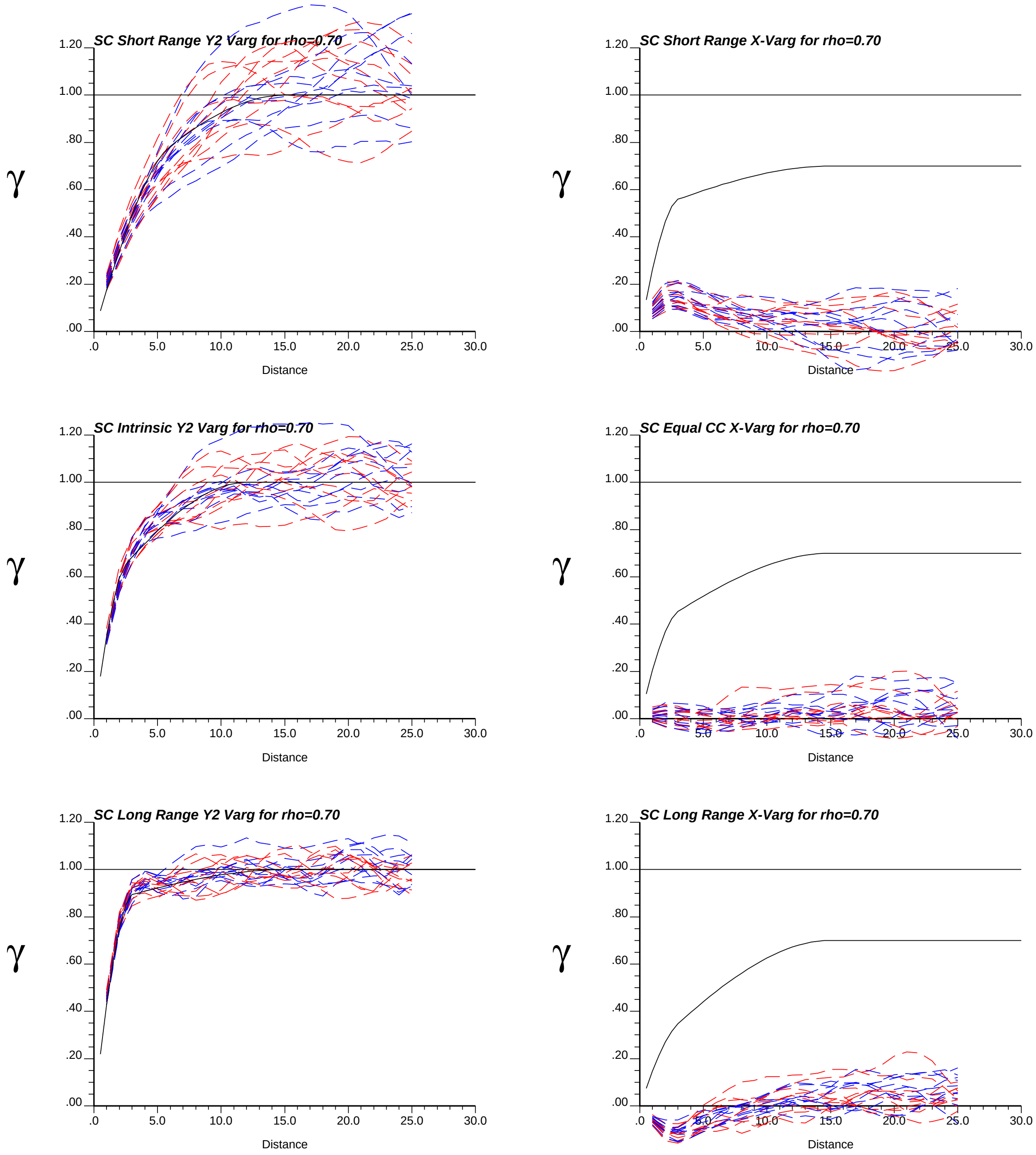

Figure 6: Direct variogram of $Y_{2|1}$ (left) and cross variogram of Y_1 and $Y_{2|1}$(right), after stepwise conditional transformation. The black line on the cross variograms represent the cross variogram model used to create the unconditioned simulation prior to transformation.

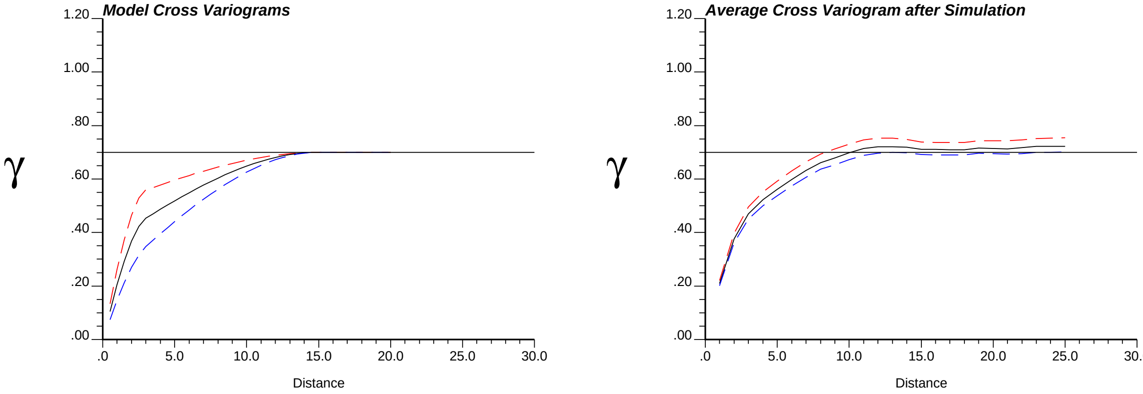

Figure 7: Input model of cross variogram of Y_1 and Y_2 (left), and the cross variogram obtained after simulating with stepwise transformed variables Y_1 and $Y_{2|1}$ (right). In both cases, the variograms follow the same color/line code: short-range (red, top dashed), intrinsic (black, middle), and long-range (blue, bottom dashed).

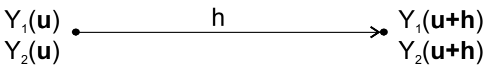

Figure 8: Schematic illustration of two points separated by a distance $\mathbf{h}$.

The simple kriging (SK) equations for this system are:

$$\sigma_E^2 \quad = \quad C_1(0) + \sum_{\alpha=1}^{2} \sum_{\beta=1}^{2} C(\mathbf{u}_\alpha - \mathbf{u}_\beta) - 2 \sum_{\alpha=1}^{2} C(\mathbf{u} - \mathbf{u}_\alpha)$$

Solving the kriging equations gives:

$$\sum_{\alpha=1}^{2} C(\mathbf{u}_\alpha - \mathbf{u}_\beta) \quad = \quad C(\mathbf{u} - \mathbf{u}_\alpha) \tag{1}$$

For two points, Equation 1 becomes:

$$\lambda C_1(0) + \lambda' C_1(\mathbf{h}) \quad = \quad C_{12}(0) \tag{2}$$
$$\lambda C_1(\mathbf{h}) + \lambda' C_1(0) \quad = \quad C_{12}(\mathbf{h}) \tag{3}$$

where $C_1(\mathbf{h})$ is the covariance for Y_1 at distance $\mathbf{h}$, and $C_{12}(\mathbf{h})$ is the cross covariance between Y_1 and Y_2. Note that $C_1(\mathbf{h}) = 0$ and $C_{12}(0) = \rho$. So equations 2 and 3 becomes:

$$\lambda + \lambda' C_1(\mathbf{h}) \quad = \quad \rho \tag{4}$$
$$\lambda C_1(\mathbf{h}) + \lambda' \quad = \quad C_{12}(\mathbf{h}) \tag{5}$$

Solving equations 4 and 5 yields the following SK weights:

$$\begin{aligned} \lambda &= \frac{\rho - C_{12}(\mathbf{h}) \cdot C_1(\mathbf{h})}{1 - C_1(\mathbf{h})^2} \qquad (6) \\ \lambda' &= \frac{\rho - C_{12}(\mathbf{h}) \cdot C_1(\mathbf{h})}{1 - C_1(\mathbf{h})^2} \qquad (7) \end{aligned}$$

Equations 6 and 7 shows the weights for the general case with two data points.

Intrinsic Coregionalization

For the special case of an intrinsic coregionalization, $C_{12}(\mathbf{h}) = \rho \cdot C_1(\mathbf{h})$ and the SK weights become:

$$\begin{aligned} \lambda &= \rho \\ \lambda' &= 0 \end{aligned}$$

The kriged estimate is:

$$\begin{aligned} Y_{2|1} &= \rho Y_1(\mathbf{u}) \\ \mu_{2|1} &= \rho Y_1(\mathbf{u}) \end{aligned}$$

The kriging variance becomes:

$$\sigma_E^2 = 1 - \rho^2$$

So the conditional distribution of $Y_{2|1}$ for the intrinsic case is $N(\rho Y_1(\mathbf{u}), 1 - \rho^2)$, which agrees with the conditional distribution obtained by applying Bayes postulate on conditional expectation.

Using the mean and variance of the conditional distribution, the covariance model of the transformed variable can be determined:

$$\begin{aligned} Y_{2|1}(\mathbf{u}) &= \frac{Y_2(\mathbf{u}) - \mu_{2|1}}{\sigma_{2|1}} \\ C_{2|1} &= E\{Y_{2|1}(\mathbf{u}) \cdot Y_{2|1}(\mathbf{u}+\mathbf{h})\} \\ C_{2|1} &= E\left\{\left(\frac{Y_2(\mathbf{u}) - \rho Y_1(\mathbf{u})}{\sqrt{1-\rho^2}}\right) \cdot \left(\frac{Y_2(\mathbf{u}+\mathbf{h}) - \rho Y_1(\mathbf{u}+\mathbf{h})}{\sqrt{1-\rho^2}}\right)\right\} \\ &= \frac{1}{1-\rho^2} E\left\{[Y_2(\mathbf{u}) - \rho Y_1(\mathbf{u})] \cdot [Y_2(\mathbf{u}+\mathbf{h}) - \rho Y_1(\mathbf{u}+\mathbf{h})]\right\} \\ &= \frac{1}{1-\rho^2} E\left\{Y_2(\mathbf{u})Y_2(\mathbf{u}+\mathbf{h}) - \rho Y_2(\mathbf{u})Y_1(\mathbf{u}+\mathbf{h}) - \rho Y_1(\mathbf{u})Y_2(\mathbf{u}+\mathbf{h}) + \rho^2 Y_1(\mathbf{u}) \cdot Y_1(\mathbf{u}+\mathbf{h})\right\} \end{aligned}$$

$$
\begin{aligned}
C_{2|1} &= \frac{1}{1-\rho^2}\Big\{C_2(\mathbf{h}) - 2\rho C_{12}(\mathbf{h}) + \rho^2 C_1(\mathbf{h})\Big\} \\
&= \frac{1}{1-\rho^2}\Big\{C_1(\mathbf{h}) - 2\rho(\rho C_1(\mathbf{h}) + \rho^2 C_1(\mathbf{h})\Big\} \\
&= \frac{1}{1-\rho^2}\Big\{C_1(\mathbf{h}) - \rho^2 C_1(\mathbf{h})\Big\} \\
&= \frac{C_1(\mathbf{h})(1-\rho^2)}{1-\rho^2} \\
&= C_1(\mathbf{h})
\end{aligned}
$$

Comparison of the above analytical result with the numerical results shown in Figure 6 for the intrinsic case shows that the numerical result deviates only slightly from the analytical solution. This deviation can be attributed to numerical artefacts, likely resulting from any or a combination of the following factors: search radius, search strategy, number of data and simulated data for simulation, and/or the actual algorithm used.

General Case with Two Points

Using the general SK weights given in equations 6 and 7, the mean and variance of the conditional distribution for the general case is:

$$
\begin{aligned}
\mu_{2|1} &= \lambda Y_1(\mathbf{u}) + \lambda' Y_1(\mathbf{u}{+}\mathbf{h}) \\
\mu_{2|1} &= \left(\frac{\rho - C_{12}(\mathbf{h})\cdot C_1(\mathbf{h})}{1-C_1^2(\mathbf{h})}\right) Y_1(\mathbf{u}) + \left(\frac{C_{12}(\mathbf{h}) - \rho\cdot C_1(\mathbf{h})}{1-C_1^2(\mathbf{h})}\right) Y_1(\mathbf{u}{+}\mathbf{h}) \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
\sigma_E^2 &= C_1(0) - \Big\{\lambda C_{12}(0) + \lambda' C_{12}(\mathbf{h})\Big\} \\
\sigma_E^2 &= 1 - \left\{\left(\frac{\rho - C_{12}(\mathbf{h})\cdot C_1(\mathbf{h})}{1-C_1^2(\mathbf{h})}\right) C_{12}(0) + \left(\frac{C_{12}(\mathbf{h}) - \rho\cdot C_1(\mathbf{h})}{1-C_1^2(\mathbf{h})}\right) C_{12}(\mathbf{h})\right\} \qquad (9)
\end{aligned}
$$

As before, the general covariance model is given as:

$$
\begin{aligned}
Y_{2|1}(\mathbf{u}) &= \frac{Y_2(\mathbf{u}) - \mu_{2|1}}{\sigma_{2|1}} \\
C_{2|1} &= E\{Y_{2|1}(\mathbf{u})\cdot Y_{2|1}(\mathbf{u}{+}\mathbf{h})\} \\
C_{2|1} &= E\left\{\left(\frac{Y_2(\mathbf{u}) - \mu_{2|1}}{\sigma_{2|1}}\right)\cdot\left(\frac{Y_2(\mathbf{u}{+}\mathbf{h}) - \mu_{2|1}}{\sigma_{2|1}}\right)\right\} \qquad (10)
\end{aligned}
$$

where $\mu_{2|1}$ and $\sigma_{2|1}$ are given in equations 8 and 9, respectively. It is clear that the covariance structure of the conditionally transformed variable (in equation 10) implicitly incorporates the cross-covariance structure of the original variables.

Independence at lag distances

The shape of the cross variogram structure after transformation was examined to investigate the issue of independence for distance lags greater than zero ($\mathbf{h} > 0$). The ideal case would be a cross variogram $\gamma(\mathbf{h}) = 0, \forall \mathbf{h}$, since independent simulation of the transformed variables, Y_1 and $Y_{2|1}$, assumes that the cross variogram $\gamma_{Y_1 \cdot Y_{2|1}}(\mathbf{h}) = 0, \forall \mathbf{h}$.

The same two data sets (Two Well and East Texas Core data) were used to study the effect of the transformation on the multivariate spatial distribution. For each data set, the cross variogram was calculated after the variables were transformed. At lag distances where the cross variogram deviated most from zero, the crossplots at that lag, $\mathbf{h}$, was generated. Figures 9 and 10 show the crossplot of the transformed variables at $\mathbf{h} = 0$, the resulting cross variogram and a crossplot taken at lag $\mathbf{h}$, corresponding to the maximum deviation of the cross variogram from $\gamma(\mathbf{h}) = 0$.

It is difficult to understand the numerical results from the crossplots. Distinguishing between class artifacts and actual structural trends is not obvious.

Transformation in Presence of Sparse Data

The first paper identified the lack of sufficient data as a major limitation to the practicality of this transformation technique. At the same time, it was suggested that the application of a smoothing technique should create more representative conditional distributions in the application of a stepwise transformation. This avenue of research was followed up by exploring the kernel density estimator.

Smoothing using a kernel density is characterized by the following probability distribution [8]:

$$\hat{f}(x) \;=\; \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{11}$$

where $K(\cdot)$ is a kernel function associated to some specified density function. Since we are primarily concerned with discretizing the bivariate distribution, the kernel density is chosen to be a non-standard bivariate Gaussian density distribution with specified correlation:

$$f_{xy} \;=\; \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot exp\left[\frac{-1}{2(1-\rho^2)} \cdot \left(\frac{(x-m_x)^2}{\sigma_x^2} - \frac{2\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right)\right]$$

where m_x and σ_x^2 is the mean and variance corresponding to the random variable X, m_y and σ_y^2 is the variance corresponding to the random variable Y, and ρ is the correlation coefficient between X and Y.

The general algorithm is to generate a bivariate density distribution centered about each data pair. The frequencies for each X and Y are then averaged to obtain density estimates for that particular pair. The result is a "cloud" of values centered about the data. Discretizing the bivariate distribution for the stepwise conditional transformation will then be accomplished using the smoothed bivariate distribution. The basic steps in

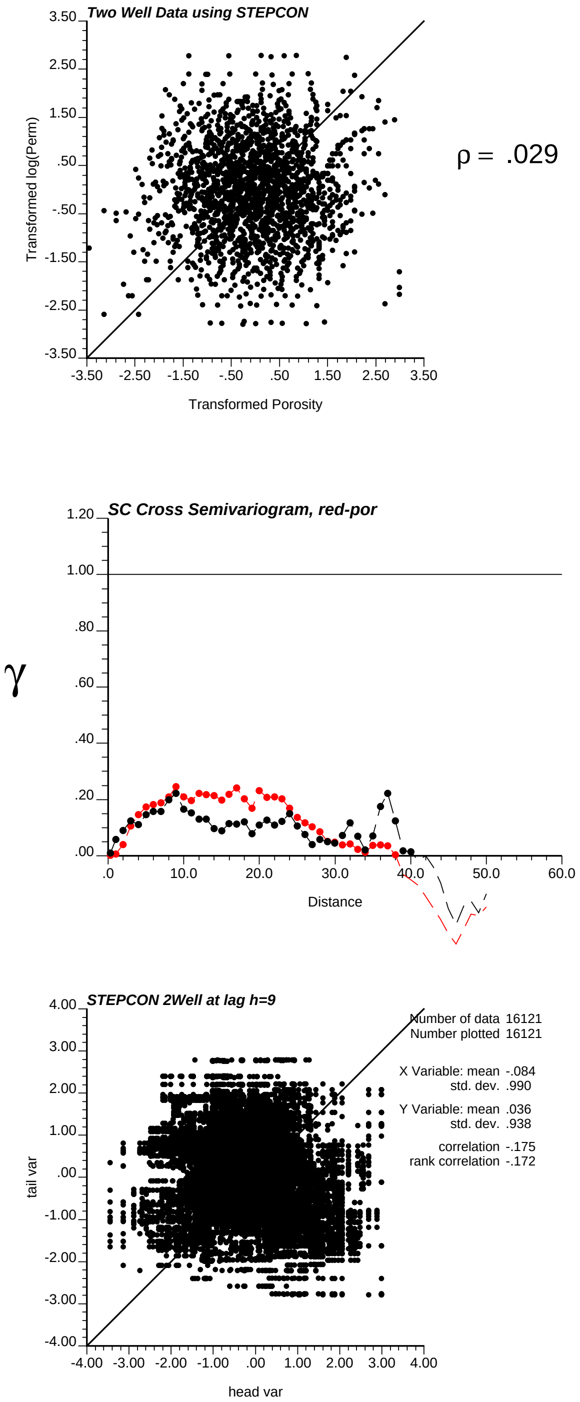

Figure 9: Two Well data: Cross plot of transformed data at lag **h**=0 (top) ,cross variogram of the transformed variables (centre), and of crossplot at lag **h**=9, corresponding to maximum value on cross-semivariogram (bottom).

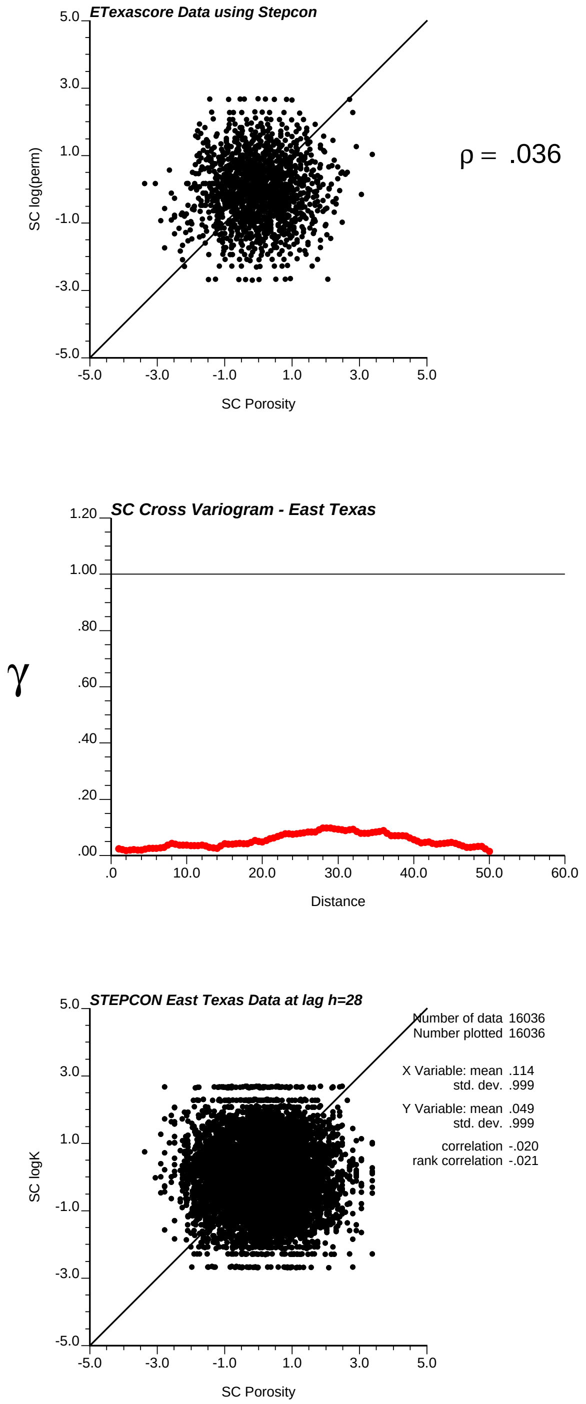

Figure 10: East Texas core data: Cross plot of transformed data at lag **h**=0 (top) ,cross variogram of the transformed variables (centre), and of crossplot at lag **h**=28, corresponding to maximum value on cross-semivariogram (bottom).

smoothing using a kernel estimator are (with user specified correlation coefficient, ρ, and variance for each variable, σ_1 and σ_2):

1. Using the scatterplot limits for both variables, discretize the scatterplot to create a regular grid of X and Y values.
2. Go to each data pair:
   - Set $m_x = x$ and $m_y = y$.
   - Visit each node in the new scatterplot grid and calculate the bivariate frequency using the non-standard Gaussian density function.
3. Visit each node in the grid again, and average all the calculated frequencies at each node.

Implementation

The program scatsmth_k was developed to implement this algorithm. Figure 11 shows the effect of variance and correlation coefficient on the smoothed bivariate distribution. Specification of a unit variance is considered high since the variance of normal scored data is 1.0 in Gaussian space. This translates to assigning each paired data the full variance of the sample distribution leading to an over-smoothing of the bivariate distribution. The specified correlation coefficient should correspond to the correlation between the normal scored data.

The programs stepcon and backstep were modified to perform the stepwise conditional transformation and the back transformation using the smoothed distribution, respectively. Parameters required for each program are given in the Appendix.

The data should first be transformed into normal scores using nscore. Using the normal score values of the multivariate data, the program scatsmth_k applied to smooth the bivariate distribution of the normal scores. Using the output bivariate transformation table from scatsmth_k, stepwise transformation can be performed for the smoothed distribution in stepcon. Independent simulation of the model variables can now proceed in Gaussian space. Back transformation of the simulated values is implemented by calling on the univariate and the bivariate transformation tables output from nscore and scatsmth_k, respectively.

This methodology was applied to several small petroleum-related data sets. The first data set is referred to as the Reservoir data, and consists of 164 paired samples of porosity and log(permeability). Figure 12 shows the results of the applying the above methodology to this data set. As expected, the cross plot of the smoothed bivariate distribution has a similar shape to that of the normal score data. The correlation coefficient after stepwise transformation using the smoothed distribution is not as low as that obtained using only the data. The conditional cumulative distribution function (ccdf) for each class of the primary data is more clearly defined as a result of the smoothing algorithm; however, this ccdf is not perfectly Gaussian. Although the algorithm and proposed methodology is sufficient for geostatistical simulation to proceed in the presence of sparse data, it is not meant to take the place of real data. A comparison of the cross plot of the original 164 data and the

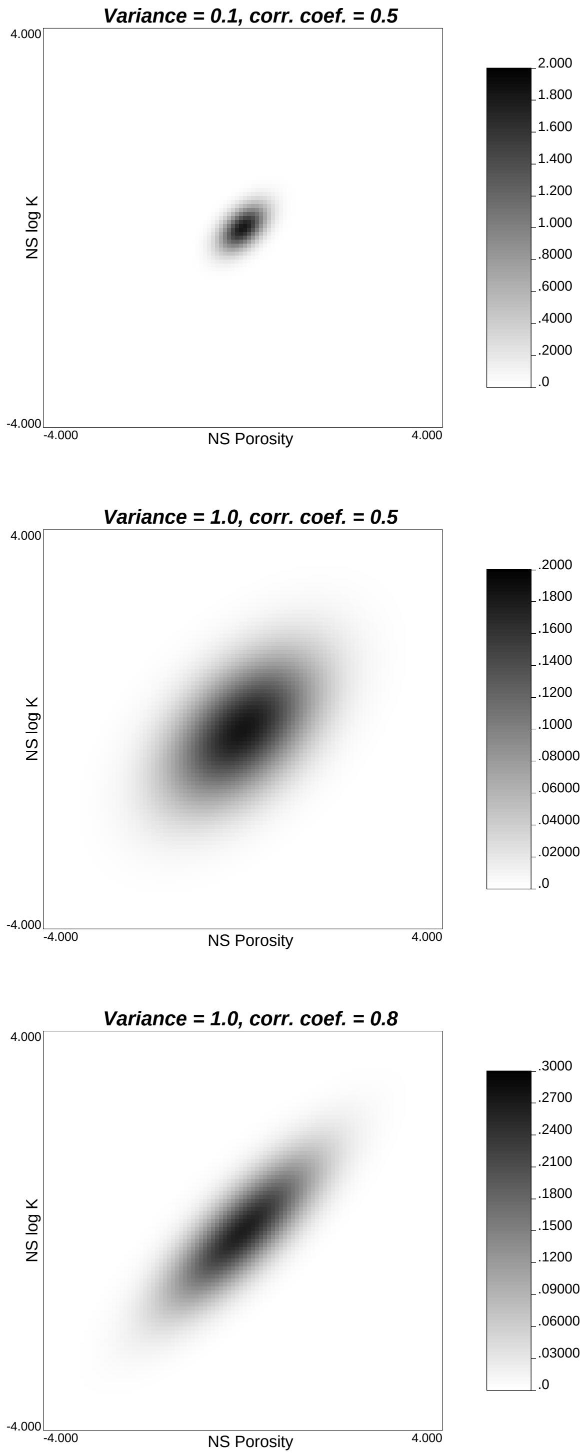

Figure 11: Effect of variance and correlation coefficient on the bivariate frequency distribution after smoothing using 1 data point. The top two figures show the effect of variance on the "spread" of the bivariate distribution centred at the data point, while the bottom two figures show the effect of the correlation coefficient on the skew angle of the bivariate distribution.

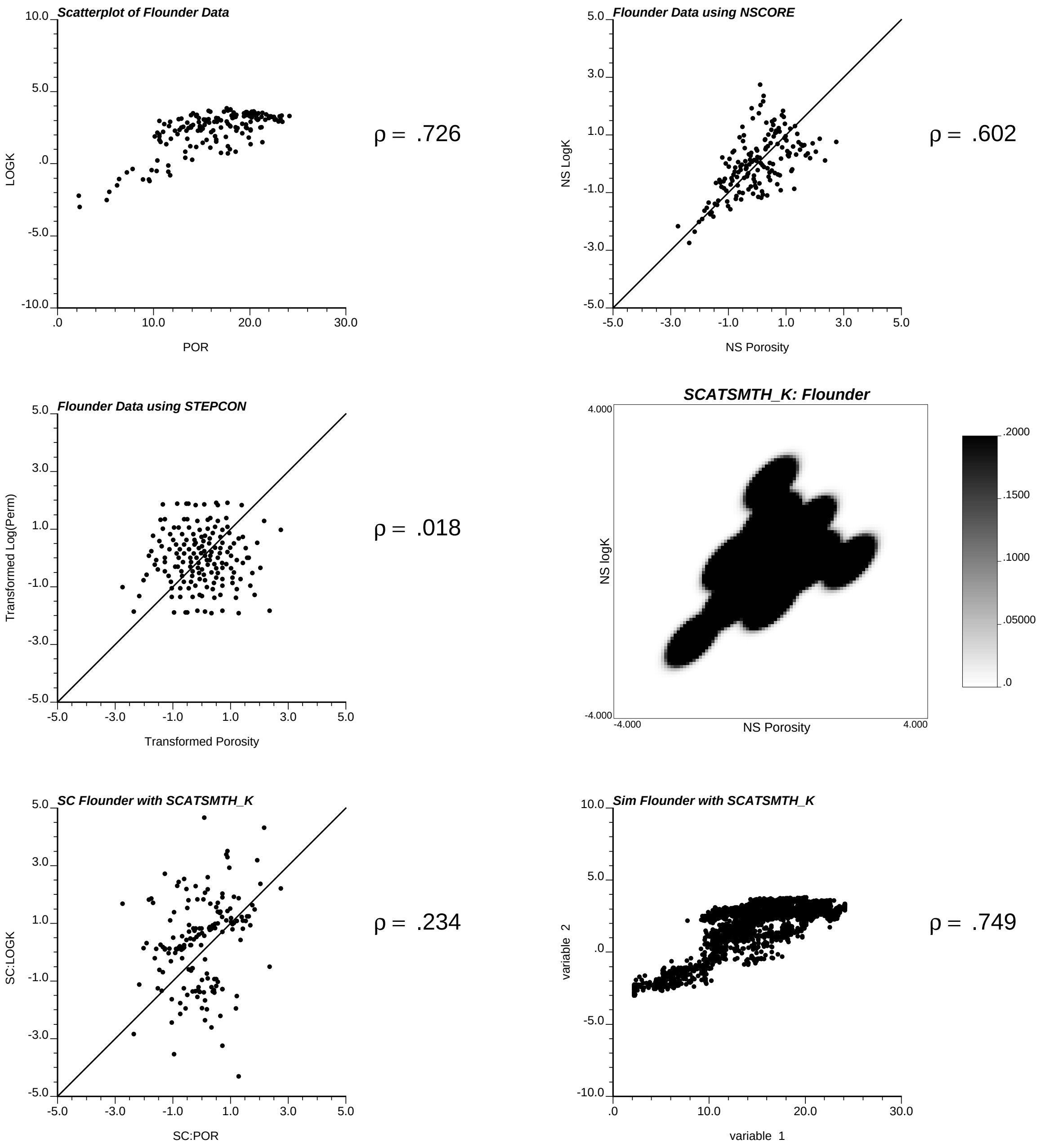

Figure 12: Reservoir data: Cross plot of the original data (top left), cross plot of the normal scored data (top right), cross plot of the stepwise conditionally transformed data using only the original 164 data values (middle left), smoothed bivariate distribution using kernel estimation (middle right), cross plot of stepwise conditionally transformed data using the smoothed distribution (bottom left), and a cross plot of the simulated values after back transformation (bottom right).

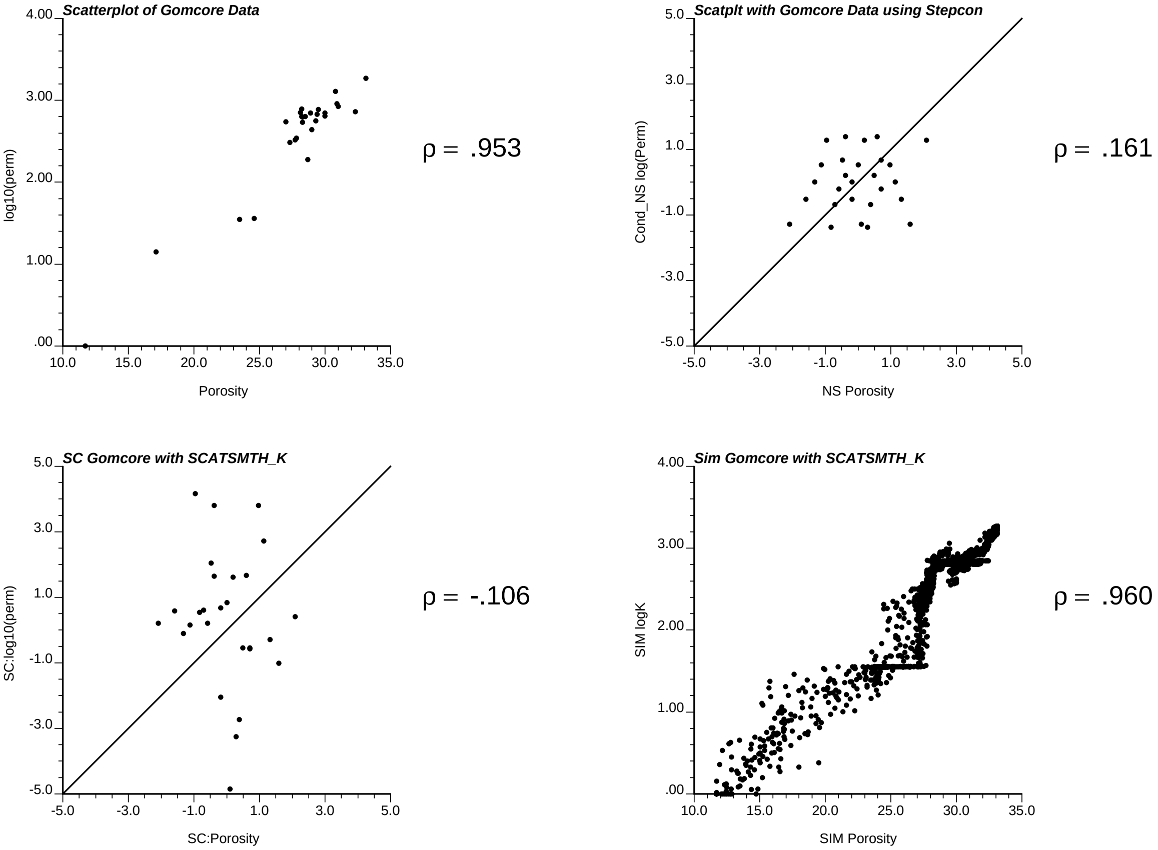

Figure 13: Gomcore data: Cross plot of the original data (top left), cross plot of the stepwise conditionally transformed data using only the original 27 data values (top right), cross plot of stepwise conditionally transformed data using the smoothed distribution (bottom left), and a cross plot of the simulated values after back transformation (bottom right).

simulated values after back transformation shows that the bivariate distribution of the data is reproduced with only a minor difference in correlation.

The second data set is known as the Gomcore data set. It consists of only 27 data pairs of porosity and log(permeability). Figure 13 shows several comparative cross plots. The two cross plots of the stepwise conditionally transformed variables resulting from (1) only the data, and (2) the smoothed distribution have similar correlation magnitudes, but with opposite signs. Simulation and back transformation of the transformed variables according to the smoothed distribution shows good reproduction of bivariate distribution with only a 0.007 difference in correlation.

Conclusion

The stepwise conditional transform presents great benefits in multivariate geostatistical simulation. Unlike the linear model of coregionalization, the model of regionalization invoked as a result of the transform is not easily defined analytically. Theoretical exercises showed that the covariance structure of the conditionally transformed secondary variables is a function of the cross covariance model between the original variables. The transformation results in non-physical secondary variables that simplify multivariate simulation without changing the underlying multivariate distribution between the original variables.

A bivariate distribution smoothing algorithm was presented for application in the presence of sparse data. With few data, a representative conditional distribution is difficult to infer. Smoothing of the bivariate distribution helps to identify a conditional distribution based on the available data, so that stepwise conditional transformation can be effectively applied. Simulation using the smoothed distributions reproduces the original bivariate distribution.

The theoretical and practical details addressed in this paper lend support to the advantages of applying this transformation method over conventional univariate techniques.

References

[1] T. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, 1958.

[2] C. V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, New York, 1992.

[3] P. Goovaerts. *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York, 1997.

[4] E. H. Isaaks. *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data.* PhD thesis, Stanford University, Stanford, CA, 1990.

[5] O. Leuangthong and C. Deutsch. Stepwise conditional transformation for simplified cosimulation of reservoir properties. In *Report Two*, University of Alberta, Edmonton, Alberta, CA, March 2000. Centre for Computational Geostatistics.

[6] G. R. Luster. *Raw Materials for Portland Cement: Applications of Conditional Simulation of Coregionalization.* PhD thesis, Stanford University, Stanford, CA, 1985.

[7] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[8] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York, 1992.

Appendix

Example parameter files `scatsmth_k.par`, `stepcon.par` and `backstep.par` are shown in Figures 14, 15 and 16, respectively. The parameters are self explanatory.

```
                   Parameters for SCATSMTH_K
                   *************************

START OF PARAMETERS:
../data/cluster.dat           - file with data
4   5   0                     - columns for X, Y, wt
-1.0e21   1.0e21              - trimming limits
-4.0    4.00                  - min and max
-4.0    4.00                  - Y min and max
scatsmth_k.out                - file for smoothed distribution output
scatsmth_k.trn                - file for transformation table
0.602                         - correlation coefficient
0.05   0.05                   - x and y variance for kernel density
```

Figure 14: Parameters for `scatsmth_k`.

```
                   Parameters for STEPCON
                   **********************

START OF PARAMETERS:
../data/cluster.dat           - file with data
3                             - number of variables to transform
13 14 15                      - columns for variable transformation
-1.0e21  1.0e21               - trimming limits
10                            - number of classes
1                             - smoothed distribution, yes=1,no=0
scatsmth_k.trn                - file for input transformation table
stepcon.out                   - file for output
stepcon.trn                   - file for output transformation table
```

Figure 15: Parameters for stepcon.

```
                Parameters for BACKSTEP
                ***********************

START OF PARAMETERS:
2                             - number of variables
file1.dat                     - data file number 1
file2.dat                     - data file number 2
-1.0e21  1.0e21               - trimming limits
10                            - number of classes
stepcon.trn                   - file with transformation table
scatsmth_kfl.trn              - file with input transformation table
nspor.trn                     - univariate transformation table for variable 1
nsper.trn                     - univariate transformation table for variable 2
backstep.out                  - file for output
```

Figure 16: Parameters for backstep.