Unbiased Estimation through Non-Linear Transforms

John G. Manchuk and Clayton V. Deutsch

Most algorithms to deal with multivariate data are geared towards simulation and this is likely due to the challenge in evaluating the integral equations which describe the mean and variance through non-linear transforms applied to geological variables. Multivariate data are often put through a sequence of transformations to obtain a set of variables that can be approximated as Gaussian. Another desirable quality targeted by transformations is that the data are uncorrelated. This removes the need to model a complex linear model of coregionalization. Once transformed, simulation followed by back-transformation is applied; however, it may also be interesting to generate kriged maps of the data. To do this, estimates and estimation variance from kriging are used along with an assumption of multivariate Gaussianity to evaluate the integrals that define the mean and variance through the back-transformation process. The complexity of this problem warrants the use of methods such as Monte Carlo integration. This paper describes the integral equations involved and develops a program to achieve the back-transformation of kriged multivariate data through a variety of transformations.

1. Introduction

A variety of geostatistical modeling applications often involve data with multiple variables, potentially from multiple sources. In almost all cases the data are non-Gaussian which tends to complicate the statistical analysis and handling of the data throughout a study. A common approach to divert from non-Gaussian data is to apply a normal score transform; however, this only results in univariate Gaussian variables. The multivariate relationships with other variables may remain non-Gaussian. Moreover, the spatial distribution of each individual variable may not be Gaussian, that is, the normal score transform does not necessarily result in a variable from a Gaussian random field even though the marginal distribution is Gaussian. For example, the bivariate distribution of a variable with itself separated by a lag vector, $\mathbf{h}$, may not be bivariate Gaussian. We often make the assumption that each variable, after a normal score transform, is Gaussian as a whole.

Methods to deal with multivariate data are undergoing development. Some of the recent ones at the CCG that are found in this report include: paper 101 (Deutsch), 116 (Manchuk and Deutsch), 310 (Barnett and Deutsch), and Guidebook Series 13 (Barnett). Topics include multivariate distribution modeling and transformation, and general handling of and tools for multivariate geostatistics. Various transformations including the normal score transform can be used for multivariate data to accomplish certain tasks such as the removal of non-Gaussian features, linearization, to obtain a set of independent variables, and for dimension reduction. An issue with most transformations is they are non-linear. Sometimes, the link from transformed variables to their original state is convoluted and non-intuitive. For example, looking at a realization generated using sequential Gaussian simulation (SGS) of a variable that has gone through a stepwise conditional transform (Leuangthong and Deutsch, 2003) with other variables often does not reflect the original variable. Variables must be back-transformed to their original state.

A concern when involving forward and backward transformations in geostatistical modeling is bias (Chiles and Delfiner, 1999). One of the objectives with geostatistical modeling is to minimize bias, and there are several forms discussed in the background section of this paper. The focus of this paper is on the bias of the estimator. When back-transformation is considered in a Gaussian simulation context, simulated values (or vectors in the multivariate case) are back-transformed individually. Simulated values that have been drawn from a Gaussian distribution characterized by an unbiased estimator, or a nearly unbiased estimator, remain unbiased through the back-transformation process. However, this statement does not encompass systematic bias that is a result of incorrect parameterization or a systematic error in the data. Conversely, estimates made with an unbiased estimator are not necessarily unbiased through a back-transformation unless the transformation is linear.

The focus of this paper is on maintaining the unbiased property of estimates made using kriging through the back-transformation process. A CCG developed program for performing the inverse normal score transform of kriged estimates in an unbiased fashion is already available called POSTMG (Lyster and Deutsch, 2004). Other transformations sometimes used with multivariate data include: the additive logratio (ALR) transform for compositional data; the stepwise conditional transform (SCT); principal component analysis (PCA); minimum/maximum autocorrelation factors (MAF); and alternating conditional expectations (ACE). All are

described in the Guidebook by Barnett (2011). Back-transformation processes as applied to estimates and estimation variances through these transformations are described. A new version of POSTMG is described for applying a variety of back-transformations individually or in series.

2. Background

A nice property of the kriging estimator is that it is unbiased and when the underlying random variable is Gaussian, it is also conditionally unbiased. For an estimator to be unbiased, the following relation must hold (Equation 1), where $Z(\mathbf{u})$ is a random function and Z^* is an estimator.

$$E\{Z^*\} = Z \quad (1)$$

The bias of an estimator is expressed as Equation 2, which is the common form seen when deriving the simple kriging equations.

$$B(Z^*) = E\{Z^*\} - Z \quad (2)$$

When the random function, $Z(\mathbf{u})$, is Gaussian, simple kriging is also a conditionally unbiased estimator. This property is expressed as Equation 3, where $Z_0 = Z(\mathbf{u}_0)$.

$$E\{Z_0 \mid Z^*\} = Z^* \quad (3)$$

Variables encountered in Geostatistics are rarely Gaussian; however, kriging still provides unbiased estimates as long as the errors, ε, in Equation 4, are uncorrelated, $\mathrm{Cov}\{\varepsilon, \varepsilon\} = 0$, and homoscedastic, $\mathrm{Var}\{\varepsilon\} = \sigma^2$. Here, X is a non-random component of Z and is sometimes referred to as a drift or mean.

$$Z = X + \varepsilon \quad (4)$$

An example of a random variable that has heteroscedastic errors is a lognormal variable, where the variance depends on the data values. When kriging a lognormal variable, the estimation variance remains homoscedastic unless the proper transformation is applied (Chiles and Delfiner, 1999); of course one can make simple kriging estimates of $\log(Z)$, which is normally distributed.

In other cases where the distribution of a random function is not Gaussian or of another straightforward parametric form such as lognormal, a normal score transformation can be applied. This transformation is non-linear and estimates cannot be directly back-transformed without introducing bias. Letting any arbitrary transform be denoted by T, a bias results when the estimator of $Y = T(Z)$ does not coincide with the estimator of Z, that is, $Y^* \neq T(Z^*)$. Equality can be obtained through integration of the distribution with the knowledge that the estimate, Y^*, is the mean of a normal distribution with a variance equal to the estimation variance, σ_Y^2. Equation 5 defines the estimate, Z^*, computed from Y^* and σ_Y^2, where $f(y) \sim N(Y^*, \sigma_Y^2)$ is the probability density function of Y, and a and b are the minimum and maximum values of Z, which are usually constrained to some interval within $(-\infty, \infty)$, usually positive values between zero and some physically limited maximum.

$$Z^* = \int_{-\infty}^{\infty} T^{-1}(y) f(y) dy = \int_a^b z \cdot f\left(T(z)\right) dz \quad (5)$$

In a similar manner, the estimation variance associated with Z^* is calculated, Equation 6.

$$\sigma_Z^2 = \int_{-\infty}^{\infty} \left(T^{-1}(y) - Z^*\right)^2 f(y) dy = \int_a^b \left(z - Z^*\right)^2 \cdot f\left(T(z)\right) dz \quad (6)$$

Equations 5 and 6 assume the transformation does not result in a change in the integration areas defined by dz and dy. The existing CCG program, POSTMG, performs these calculations numerically using a series of quantiles in Equation 7, where $P(y_i)$ is the probability of y_i. The transformation function, T, is the normal score transform characterized by a transformation table contained in a text file output by the program NSCORE, or any of its available variants.

$$Z^* = \int_a^b z \cdot f\left(T(z)\right) dz = \frac{1}{N} \sum_{i=1}^{N} T^{-1}(y_i)$$
$$P(y_i) = \left(i - 1/2\right) / N \quad (7)$$

In the multivariate case, $\mathbf{Z}$ is a vector random function and back-transformation involves a vector of estimates, $\mathbf{Z}^*$ and an estimation covariance matrix, $\mathbf{\Sigma_Z}$. In cases where kriging of variables is done independently, the covariance matrix is diagonal and back-transformation is done independently. Using cokriging, the estimation

covariance matrix is not diagonal if variables have non-zero correlation or if cross-covariance exists. After applying a normal score transform to each of the variables and assuming the resulting distribution is multivariate Gaussian, the distribution function of $\mathbf{Z}$ is given by Equation 8, where $\mathbf{\Sigma_Y}$ is the estimation covariance matrix of the normal scores and $|\partial\mathbf{Y}/\partial\mathbf{Z}|$ is the Jacobian of the transformation.

$$f(\mathbf{Z}) = \frac{1}{(2\pi)^{m/2}\left|\mathbf{\Sigma_Y}\right|^{1/2}} \cdot \left|\frac{\partial \mathbf{Y}}{\partial \mathbf{Z}}\right| \cdot \exp\left[-\frac{1}{2}\left(T(\mathbf{Z}) - \mathbf{Y}^*\right)\mathbf{\Sigma_Y}^{-1}\left(T(\mathbf{Z}) - \mathbf{Y}^*\right)^T\right] \quad 8$$

From this, the estimate and estimation covariance matrix of $\mathbf{Z}$ is defined by Equation 9 and 10, where $\mathbb{R}^{\boldsymbol{Z}}$ is the real space where $\boldsymbol{Z}$ is defined.

$$\mathbf{Z}^* = \int_{\mathbf{Z}} \mathbf{Z} f(\mathbf{Z}) d\mathbf{Z} \quad 9$$

$$\mathbf{\Sigma_Z} = \int_{\mathbf{Z}} \left(\mathbf{Z} - \mathbf{Z}^*\right)\left(\mathbf{Z} - \mathbf{Z}^*\right)^T f(\mathbf{Z}) d\mathbf{Z} \quad 10$$

Evaluating these integral equations is more challenging than the univariate case because selecting a set of equally probable values is not straightforward and numerical evaluation of high dimensional integrals can suffer from the curse of dimensionality.

The normal score transform may be nested with other transformations where there are multiple variables. For example, PCA may have been applied first to remove correlation between co-located samples. In any case, PCA is a linear transformation defined by Equation 11, where $\mathbf{P}$ is an orthonormal matrix of eigenvalues derived form the covariance matrix of $\mathbf{Z}$, given in Equation 12, $e_k, k = 1, \ldots, m$ are the individual eigenvectors, and m is the number of variables. $\mathbf{D}$ is a diagonal matrix of eigenvalues.

$$\begin{aligned} \mathbf{Y} &= \mathbf{PZ} \\ \mathbf{P} &= \begin{bmatrix} e_1 & e_2 & \cdots & e_m \end{bmatrix}^T \end{aligned} \quad 11$$

$$Cov\{\mathbf{Z}\} = \mathbf{PDP}^T \quad 12$$

PCA defines the linear combination of $\mathbf{Z}$ that result in uncorrelated $\mathbf{Y}$. Considering a single variable, $Y_i = e_i^T\mathbf{Z}$, the mean, variance, and covariance with $Y_j, i \neq j$ are defined by Equation 13.

$$\begin{aligned} E\{Y_i\} &= e_i^T E\{\mathbf{Z}\} \\ Var\{Y_i\} &= e_i^T Var\{\mathbf{Z}\} e_i \\ Cov\{Y_i, Y_j\} &= e_i^T Var\{\mathbf{Z}\} e_j = 0, i \neq j \end{aligned} \quad 13$$

These relations hold regardless of the distribution of $\mathbf{Z}$, as long as the variance is finite. The back-transformation of kriging estimates and estimation variance for variables that have been transformed using PCA are straightforward since the transformation is linear. The vector of estimates, $\mathbf{Z}^*$, and estimation covariance matrix, $\mathbf{\Sigma_Z}$, of $\mathbf{Z}$ are computed using Equation 14, where $\mathbf{P}^T = \mathbf{P}^{-1}$ since $\mathbf{P}$ is orthonormal and $\mathbf{\Sigma_Y}$ is a diagonal matrix of estimation variances of $\mathbf{Y}$.

$$\begin{aligned} \mathbf{Z}^* &= \mathbf{P}^T\mathbf{Y}^* \\ \mathbf{\Sigma_Z} &= \mathbf{P}^T\mathbf{\Sigma_Y}\mathbf{P} \end{aligned} \quad 14$$

This assumes that the principal components are a stationary property of $\mathbf{Z}$, which results if second order stationarity is assumed for the random function. Another assumption is that the covariance between estimates is zero, which may not be the case if PCA does not eliminate the spatial correlation between variables. To assess more features of a random function transformed using PCA, other than the estimate and estimation variance, the shape of the distribution defined by $\mathbf{Z}^*$ and $\mathbf{\Sigma_Z}$ must be known. For example, determining the probability that $Z_i \leq z_i$ for one of the components of $\mathbf{Z}$ cannot be evaluated by determining the equivalent value of Y_i. If $\mathbf{Z}$ is multivariate Gaussian, then $\mathbf{Y} = \mathbf{PZ}$ is also multivariate Gaussian with a diagonal covariance matrix. In this case, the vector of estimates and the estimation covariance matrix define a normal distribution, $f(\mathbf{z}) \sim N(\mathbf{Z}^*, \mathbf{\Sigma_Z})$, and the probability that $Z_i \leq z_i$ can be assessed directly.

If $\mathbf{Z}$ is not multivariate Gaussian, then $f(\mathbf{z})$ cannot be inferred. Monte Carlo techniques can be used to infer $f(\mathbf{z})$ only in cases where the distribution characterized by $\mathbf{Y}^*$ and $\mathbf{\Sigma_Y}$ is known and it is possible to sample from it. Applying the normal score transform to $\mathbf{Z}$ will result in a distribution with Gaussian univariate marginals,

but not necessarily a multivariate Guassian distribution. Following this with the PCA transformation results in uncorrelated variables; however, because the distribution of the normal scores is not necessarily multivariate Gaussian, the linear combinations are not necessarily normally distributed (Johnson and Wichern, 2002). One might consider applying yet another normal score transform to the random function from PCA. Again, this does not necessarily result in a multivariate Gaussian random function. Moreover, inferring $f(\mathbf{z})$ may not be possible.

The PCA transform can be applied to transform kriging estimates and estimation variances when it is safe to assume the distribution of a multivariate data set is Gaussian after a normal score transform. This assumption can be relaxed significantly for simulation purposes. A similar transform to PCA is MAF (Switzer and Green, 1984), where two PCA transforms are used in succession to remove correlation at a lag of zero as well as at a lag, $\mathbf{h}$. For details, refer to the Guidebook by Barnett accompanying this year's report. Back-transformation concepts are identical to those for PCA since the transformation is also a linear combination defined by Equation 11 with the matrix $\mathbf{P}$ from MAF being different from that for PCA. The matrix is not orthonormal so $\mathbf{P}^T \neq \mathbf{P}^{-1}$. A potential advantage of MAF is that the assumption of stationarity of the transformation matrix is more acceptable, since MAF removes co-located and spatial correlation.

When data are compositional, a common class of transformations utilized is the logratio transforms. A few of these include the ALR, the centered logratio transform (CLR) (Aitchison, 1986), and isometric logratio transform (SLR) (Egozcue et al, 2003). The forward ALR transform is defined by Equation 15 and the inverse by Equation 16, where k is the variable chosen as the denominator. Note this transform results in $m-1$ variables.

$$X_{ij} = \log\left(Z_{ij} / Z_{ik}\right), \quad i = 1,...,n,\ j = 1,...,m,\ j \neq k \tag{15}$$

$$Z_{ij} = \frac{\exp\left(X_{ij}\right)}{\sum_{j=1,\, j\neq k}^{m} \exp(X_{ij}) + 1}, \quad i = 1,...,n,\ j = 1,...,m \tag{16}$$

As noted by Pawlowsky-Glahn and Olea (2004), there is no explicit form to back-transform estimates of log-ratio transformed random variables to estimates of the original variables; however, like with the normal score transform, integration can be used to approximate the solution. An approach using Gauss-Hermite quadrature is developed therein. The pertinent equations for the ALR transformation for back-transformation of the estimate and estimation variance are given by Equation 17 and 18.

$$\mathbf{Z}^* = \int_{\mathbb{R}^m} g_1(\mathbf{Y}) \exp(-\mathbf{Y}^T\mathbf{Y}) d\mathbf{Y} \tag{17}$$

$$\mathbf{\Sigma}_{\mathbf{Z}} = \int_{\mathbb{R}^m} g_2(\mathbf{Y}) \exp(-\mathbf{Y}^T\mathbf{Y}) d\mathbf{Y} \tag{18}$$

$\mathbf{Y}$ is defined by Equation 19, where R is lower triangular matrix from the Cholesky decomposition of the estimation covariance matrix of $\mathbf{X} = \text{alr}(\mathbf{Z})$ written as $\mathbf{\Sigma}_{\mathbf{X}} = R^T R$, and $\mathbf{X}^*$ is the estimate.

$$\mathbf{Y} = \frac{1}{\sqrt{2}}\left(R^{-1}\right)^T \left(\text{alr}(\mathbf{Z}) - \mathbf{X}^*\right) \tag{19}$$

The function $g_1(\mathbf{Y})$ in Equation 17 is given by Equation 20, and $g_2(\mathbf{Y})$ from Equation 18 is given by Equation 21, where agl is the additive generalized logistic or inverse ALR.

$$g_1(\mathbf{Y}) = \pi^{-\frac{m-1}{2}} \text{agl}\left(\sqrt{2}R^T\mathbf{Y} + \mathbf{X}^*\right) \tag{20}$$

$$g_2(\mathbf{Y}) = \pi^{-\frac{m-1}{2}} \left[\text{agl}\left(\sqrt{2}R^T\mathbf{Y} + \mathbf{X}^*\right) - \mathbf{Z}^*\right] \cdot \left[\text{agl}\left(\sqrt{2}R^T\mathbf{Y} + \mathbf{X}^*\right) - \mathbf{Z}^*\right]^T \tag{21}$$

Two other transformation that are considered in this work are conditional standardization (Barnett, 2011), and the SCT (Leuangthong and Deutsch, 2003). Both are very similar transformations that involve binning data in steps through each variable and either standardizing the data in each bin, or computing the normal score transform in each bin. Generalized equations for conditional standardization are given in Equation 18, where X_k are standardized random variables, $\mu_k(z)$ are conditional means for given values of X, and $\sigma_k(z)$ are conditional standard deviations.

$$X_1 = \frac{Z_1 - \mu_1}{\sigma_1}$$
$$X_2(z_1) = \frac{Z_2 - \mu_2(z_1)}{\sigma_2(z_1)} \qquad 22$$
$$\vdots$$
$$X_m(z_1, z_2, ..., z_{m-1}) = \frac{Z_m - \mu_m(z_1, ..., z_{m-1})}{\sigma_m(z_1, ..., z_{m-1})}$$

SCT is similar where the conditional standardization equation on the right hand side is replaced by the normal score transform. Both suffer from the curse of dimensionality and having enough data for a four or five variable application is rare. Continuous and likely non-parametric distribution models (Manchuk and Deutsch, 2011) are required in these cases; however, back-transformation of kriging estimates through such models is a point of future research.

The back-transformation of estimates for conditional standardization is straightforward since the conditional mean and standard deviation used in the transform are constants determined by the value of the estimate, $\mathbf{X}^*$. It is given by Equation 23, while the estimation variance is given by Equation 24.

$$Z_k^* = \sigma_k(z_1^*, ..., z_{k-1}^*) \cdot X_k^* + \mu_k(z_1^*, ..., z_{k-1}^*) \quad , k = 1, ..., m \qquad 23$$

$$\mathbf{\Sigma}_{Z,k} = \sigma_k^2(z_1, ..., z_{k-1}) \cdot \mathbf{\Sigma}_{X,k} \quad , k = 1, ..., m \qquad 24$$

This assumes the conditional mean and standard deviation are constant over the distribution defined by $\mathbf{X}^*$ and $\mathbf{\Sigma_X}$, which is likely not the case. For the moment, assume the conditional mean and standard deviations used in the transformation are continuous functions of $\mathbf{Z}$, rather than step functions defined by a set of bins. In this case, the estimate and estimation covariance matrix can be expressed as integrals similarly to the normal score transform. Using a change of variables defined by the conditional standardization transformation, the estimate is back-transformed according to Equation 25, where T is used again to denote a general transformation, in this case conditional standardization, and $|J|$ is the Jacobian of the transformation.

$$\mathbf{Z}^* = \int_{\mathbf{Z}} \mathbf{Z} f(\mathbf{Z}) d\mathbf{Z} = \int_{\mathbf{Z}} \mathbf{Z} \left| J_{T^{-1}(Z)} \right| f(T(\mathbf{Z})) d\mathbf{Z} \qquad 25$$

An expression for the estimation covariance matrix can be written similarly. Because the transformation is applied in a stepwise fashion, the Jacobian matrix is triangular and the determinant is the product of the diagonal elements given by Equation 26.

$$\left| J_{T^{-1}(Z)} \right| = \left| \frac{\partial z_1, ... z_m}{\partial x_1, ..., x_m} \right|^{-1} = \left(\prod_{k=1}^{m} \sigma_k(x_1, ..., x_{k-1}) \right)^{-1} \qquad 26$$

3. Methodology

A major challenge in back-transformation of estimates and estimation variance is when the problem is multivariate and when several transformations are performed in series to arrive at the set of variables that will be used. When multiple variables are involved, back-transformation involves multidimensional integrals that can become computationally demanding to solve. Back-transformation of a large grid of estimates quickly becomes unreasonable. Dealing with a series of transformations also increases the complexity because integration rules that are both numerically efficient and provide a good approximation to the estimate and estimation covariance through multiple back-transformations are not readily available, or have not been developed.

To use a multidimensional product integration rule such as Gauss-Hermite quadrature for back-transformation, all transformations would need to be expressed in a similar form as Equation 17 and 18, and we would have to assume that the final set of variables is multivariate Gaussian. This can be very challenging and intractable through such transformations as the SCT and conditional standardization, especially since the transformation function is usually non-parametric and discontinuous being defined by a series of bins. In this work, Monte Carlo (MC) integration (Asmussen and Glynn, 2007) is used to provide approximations to the estimate and estimation variance through the transformations used. Further research is needed for the

development of accurate integration rules and techniques for the diversity of transformations utilized in multivariate geostatistical modeling.

For high dimensional problems, MC integration can actually be more efficient than other numerical integration techniques, especially when the exact form of the function or shape of the domain is unknown. This is usually the case since we often do not know the exact shape of conditional distributions of uncertainty estimated form kriging unless they are Gaussian. MC integration is an approach that uses random sampling to evaluate an integral of the form of Equation 27, where $f(x)$ is a continuous function defined over domain D and N is the number of random samples used to estimate the integral.

$$F = \int_D f(x)dx \cong D \cdot \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{27}$$

Expressing the equations for a back-transformed estimate and estimation variance in terms of MC integration is shown in Equations 28 and 29, which are recognizable as the equations for sample mean and covariance.

$$\mathbf{Z}^* = \int_D \mathbf{Z} f(\mathbf{Z}) d\mathbf{Z} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_i \tag{28}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{Z}} &= \int_D \left(\mathbf{Z} - \mathbf{Z}^*\right)\left(\mathbf{Z} - \mathbf{Z}^*\right)^T f(\mathbf{Z}) d\mathbf{Z} \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{Z}_i - \mathbf{Z}^*\right)\left(\mathbf{Z}_i - \mathbf{Z}^*\right)^T \end{aligned} \tag{29}$$

Since $\mathbf{Z}$ cannot be sampled directly for MC integration, samples from the distribution that is assumed of a known form described by the estimate and estimation covariance matrix are used and back-transformed. In most cases, this is assumed a Gaussian distribution from which samples $\mathbf{Y}$ are drawn and used in Equations 30 and 31, where T^{-1} describes all back-transformations involved.

$$\mathbf{Z}^* = \frac{1}{N}\sum_{i=1}^{N} T^{-1}(\mathbf{Y}_i) \tag{30}$$

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \frac{1}{N}\sum_{i=1}^{N} \left(T^{-1}(\mathbf{Y}_i) - \mathbf{Z}^*\right)\left(T^{-1}(\mathbf{Y}_i) - \mathbf{Z}^*\right)^T \tag{31}$$

From many kriging applications, the estimation covariance matrix is rarely obtained, instead only diagonal elements or estimation variance of each variable are retained. In this case, the covariance structure is unknown and back-transformation of estimates cannot account for dependence of variables. Back-transformation is done for each variable independently and provides an approximation that is not necessarily accurate. This is not an issue when transformations that remove the dependency structure of multiple variables are used since estimation covariance is zero between different variables.

Generating samples,$\mathbf{Y}$, assuming the distribution is Normal involved first generating independent standard normal values. These are correlated and scaled based on the estimates and the Cholesky decomposition of the estimation covariance matrix, given by Equation 32, where $\mathbf{X}$ is a vector of random independent standard normal values, $\mathbf{Y}^*$ is the vector of estimates from kriging, and $\mathbf{L}$ is a lower triangular matrix from the Cholesky decomposition in Equation 33.

$$\mathbf{Y} = \mathbf{L}^T\mathbf{X} + \mathbf{Y}^* \tag{32}$$

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \mathbf{L}\mathbf{L}^T \tag{33}$$

4. Program and Parameters

A program called POSTMVMG (post-processing of multivariate, multi-Gaussian data) was written to back-transform estimates and estimation covariance matrices for several of the transformations using the MC integration approach. Various information required to perform the back-transformations, such as normal score transform tables and matrices from PCA, must be provided along with a file containing all the kriging results. Parameters are listed in Table 1.

Table 1: Parameters for POSTMVMG

Line	Parameters
1	3 -number of variables, N, to back-transform
2	kriging.out -file with kriged (normal score) mean and variances
3	1 2 3 - columns with normal score estimates (need N columns)
4	4 0 0 5 0 6 - columns with estimation variance-covariance
5	-998. 1E11 - trimming limits
6	2 -number of transforms applied to the data
7	1 2 -transformation flags in order they were applied
8	nscore1.trn - nscore transform file 1
9	0. 100. - minimum and maximum back-transformed values
10	nscore2.trn - nscore transform file 2
11	0. 100. - minimum and maximum back-transformed values
12	nscore3.trn - nscore transform file 3
13	0. 100. - minimum and maximum back-transformed values
14	0.580908 0.645403 -0.495984 - linear combination of 1 with 1 2 3
15	0.563624 0.120683 0.817168 - linear combination of 2 with 1 2 3
16	-0.587260 0.754248 0.293659 - linear combination of 3 with 1 2 3
17	-5. 5. -5. 5. -5. 5. - min and max back-transformed values
18	postmgmv.out -file for output
19	100 1 5412 -no. of points for Monte-Carlo integration, no. replicates, seed

Most parameters are explained adequately in the table. The kriging file on Line 2 must contain all estimates and estimation variances that will be involved in the back-transformation. It is not necessary to know the estimation covariance between different variables if this is not available. The program assumes they are zero in this case. When specifying the columns for estimation variance-covariance on Line 4, they must be put in the correct order. For a three variable problem, the columns would specify $\Sigma_{11}, \Sigma_{12}, \Sigma_{13}, \Sigma_{22}, \Sigma_{23}, \Sigma_{33}$, and they must all be present even if the cross-covariances are not available (columns are set to zero as in the table). On Line 7, as many transform flags as transforms were applied to the data are specified. The program currently applies the following back-transforms, listed by flag number: 1 – normal score transform; 2 – PCA; 3 – MAF; 4 – ALR. Conditional transforms including SCT and CST are planned for a future version.

If the normal score transform is used, the associated transformation table for each variable is supplied along with the minimum and maximum values the variable is limited to upon back-transformation. Lines 8 and 9 are repeated for as many variables as there were at the time the transformation was applied. For PCA or MAF, the transformation matrix is input as shown on Lines 14 – 16. The program computes the inverse transformation matrix. Minimum and maximum values for each variable are input on Line 17. For the ALR transform, the only required input is the variable used as the denominator when the transform was applied. The number of points for MC integration is specified on Line 19. The number of replicates defines the number of realizations to generate for each back-transformed value. Determining how many points and replicates to use is problem dependent and can be experimented with prior to application on large kriged models.

There is no limit to the number or order of transformations input to the program. For example, back-transformation can be done through an ALR transform followed by a normal score transform followed by PCA followed by a second normal score transform, if this was the order deemed best for the particular data set at hand. More transformation will result in increased program execution time. Depending on the complexity of the multivariate distribution described through the various transforms, more integration points may also be needed.

5. Examples

An example involving the nickel laterite data used throughout the guidebook by Barnett (2011) is provided. Three variables are used: nickel Ni, iron Fe, and silica SiO_2. Two transformation are applied, the normal score transform followed by PCA. Cross plots of the relationships of the original data and after each of the transformations are shown in Figure 1. Note that after each transformation, the data are clearly not multivariate Gaussian; however, the decision to move forward with kriging is made anyways under the assumption that the data are Gaussian in local regions. The transformation matrix (eigenvectors of the covariance matrix) resulting in the principle components is in Table 2. A principle component is calculated by multiplying the values of each variable by the values in the matrix, for example, $PC1 = 0.580908Ni + 0.645403Fe - 0.49598NiO_2$.

Mean vectors and covariance matrices of the original variables and after each transformation are provided in Table 3 and Table 3 respectively. The means are effectively zero after a normal score transform and

remain zero for the principle components. Variance of each variable is effectively one for each variable after a normal score transform and the covariances between variables are non-zero. After PCA, the variance of each variable is equal to the eigenvalues that result from PCA and the covariance between variables is zero. However, zero covariance is not an indication of independence. Highly irregular relationships between variables are still observed in Figure 1. Such complexity would warrant using the SCT or CST; however, these have not been incorporated into POSTMVMG at this time.

Table 2: Transformation matrix from PCA.

Combination	Ni	Fe	SiO_2
PC-1	0.580908	0.645403	-0.49598
PC-2	0.563624	0.120683	0.817168
PC-3	-0.58726	0.754248	0.293659

Table 3: Mean vectors after each transformation. V1, V2, V3 are used to indicate variables since they are not Ni, Fe, and SiO_2 after PCA.

Transform	V1	V2	V3
None	1.3185	12.5075	38.9652
NSCORE	0.0000	0.0000	-0.0002
PCA+NSCORE	0.0001	-0.0002	-0.0001

Table 4: Covariance matrices after each transformation.

Transform	Σ_{11}	Σ_{12}, Σ_{21}	Σ_{13}, Σ_{31}	Σ_{22}	Σ_{23}, Σ_{32}	Σ_{33}
None	0.8127	1.9957	-1.8367	93.3767	-59.5210	113.4578
NSCORE	1.0000	0.5646	-0.2415	1.0000	-0.4233	0.9998
PCA+NSCORE	1.8335	0.0000	0.0000	0.7707	0.0000	0.3955

The program KT3D from GSLIB (Deutsch and Journel, 1998) was used to kriging each of the principle component variables independently. The grid was 100 by 100 by 10 cells in x, y, and z respectively. An exponential variogram with zero nugget effect and a 10 to 1 horizontal to vertical anisotropy ratio was used for all variables. 40 nearest neighbors were used to estimate each grid cell. Resulting estimation and estimation variance maps are shown for slice $z = 4$ in Figure 2. This slice was chosen because it was one of the more interesting in terms of spatial structure and variation.

Before applying the back-transformation, an analysis to observe the effect on the number of points chosen for MC integration is done. Three cells were chosen from the kriging results with high, moderate, and low estimation variance respectively. These were back-transformed with the number of points ranging from 10 to 1000. The true value was estimated using 500,000 points. Convergence of the mean and variance using one replicate is shown in Figure 3. Covariance convergence is similar to variance and is not shown. All covariances tended to converge to zero with increasing points. Convergence properties are significantly improved by using more replicates as shown in Figure 4, where ten replicates were used. However, using ten times the number of points and one replicate gives similar results. In either case, a large number of points are required to obtain stable back-transformed values and one might consider using upwards of 1000 points, which is used in this example.

Results of back-transformation are shown in Figure 5. Maps are relatively smooth considering each value was back-transformed using a different set of random points for integration. Some noise is observed where the estimation variance is high, especially for SiO_2, which appears to be a lognormal variable. A comparison of the relationships between variables from back-transformed kriging estimates and of the original variables is shown in Figure 6. The overall shape is captured well. There are some interesting local phenomena in the kriging results, likely caused by discontinuities from using a local neighborhood in kriging or due to regions with low variance or variance instabilities. Using 1000 points and one replicate, the grid of 100000 cells took roughly one minute to process.

6. Conclusions and Future Work

The presented application makes it possible to construct models using kriging with multivariate data that has gone through a variety of transformations to obtain Gaussian data. Integration to obtain the back-transformed mean and covariance matrix is a challenging task; however, it is simplified using approaches such as Monte Carlo integration. More research is needed in this area including the addition of the stepwise conditional and conditional standardization transformations. An immediate step that can be taken to improve convergence is to adopt the class of quasi-Monte Carlo integration methods. They rely on low discrepancy sequences of random numbers instead of pseudo-random numbers that can result in significantly faster convergence. Another point of future research is to incorporate more complex boundaries other than a minimum and maximum value for each variable. Points that fall outside the boundary are ignored in the integration.

References

Aitchison, J., 1986. The statistical analysis of compositional data: Monographs on statistics and applied probability. Chapman & Hall Ltd., London, 416

Asmussen, S., Glynn, P.W., 2007, Stochastic simulation: algorithms and analysis. Springer Science + Business Media, 496

Barnett, R.M., 2011, Tools for multivariate geostatistical modeling. Centre for Computational Geostatistics, Guidebook Series 13

Barnett, R.M., Deutsch, C.V., 2011, Tools for Geometallurgical Modeling. Centre for Computational Geostatistics, Annual Report 13, Paper 310

Chiles, J.P., Delfiner, P., 1999, Geostatistics: modeling spatial uncertainty. John Wiley & Sons, 720

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3), pp. 279-300

Deutsch, C.V., 2011, A new multivariate transformation for complex distributions. Centre for Computational Geostatistics, Annual Report 13, Paper 101

Johnson, R.A., Wichern, D.W., 2002. Applied multivariate statistical analysis. Prentice Hall, Inc., 767

Leuangthong, O., and Deutsch, C.V. (2003) Stepwise conditional transformation for simulation of multiple variables, Mathematical Geology, Vol. 35, No. 2, pp. 155-173

Lyster, S.J., Deutsch, C.V., 2004. PostMG: A postprocessing program for multigaussian kriging output. Centre for Computational Geostatistics, Annual Report 6, Paper 405

Manchuk, J.G., Deutsch, C.V., 2011, A general program for data transformations and kernel density estimation. Centre for Computational Geostatistics, Annual Report 13, Paper 116

Pawlowsky-Glahn, V., Olea, R.A., 2004. Geostatistical analysis of compositional data. Oxford University Press, 304

Switzer, P., Green, A.A., 1984. Min/Max autocorrelation factors for multivariate spatial imaging. Stanford University, Technical Report 6, 14

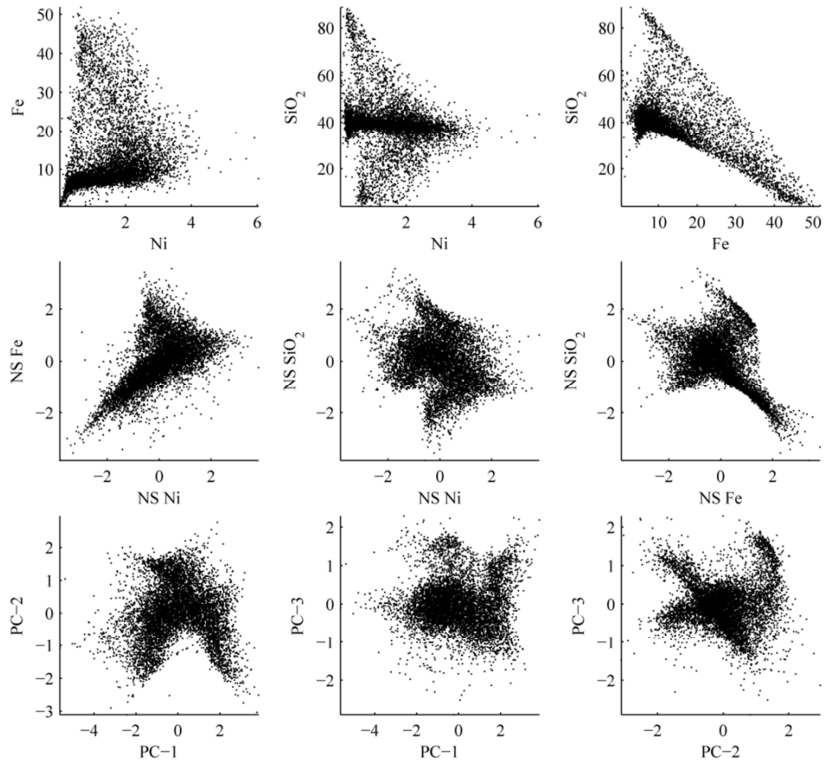

Figure 1: Relationships between original variables (top row), after normal score transform (middle row) and after normal scores + PCA (bottom row).

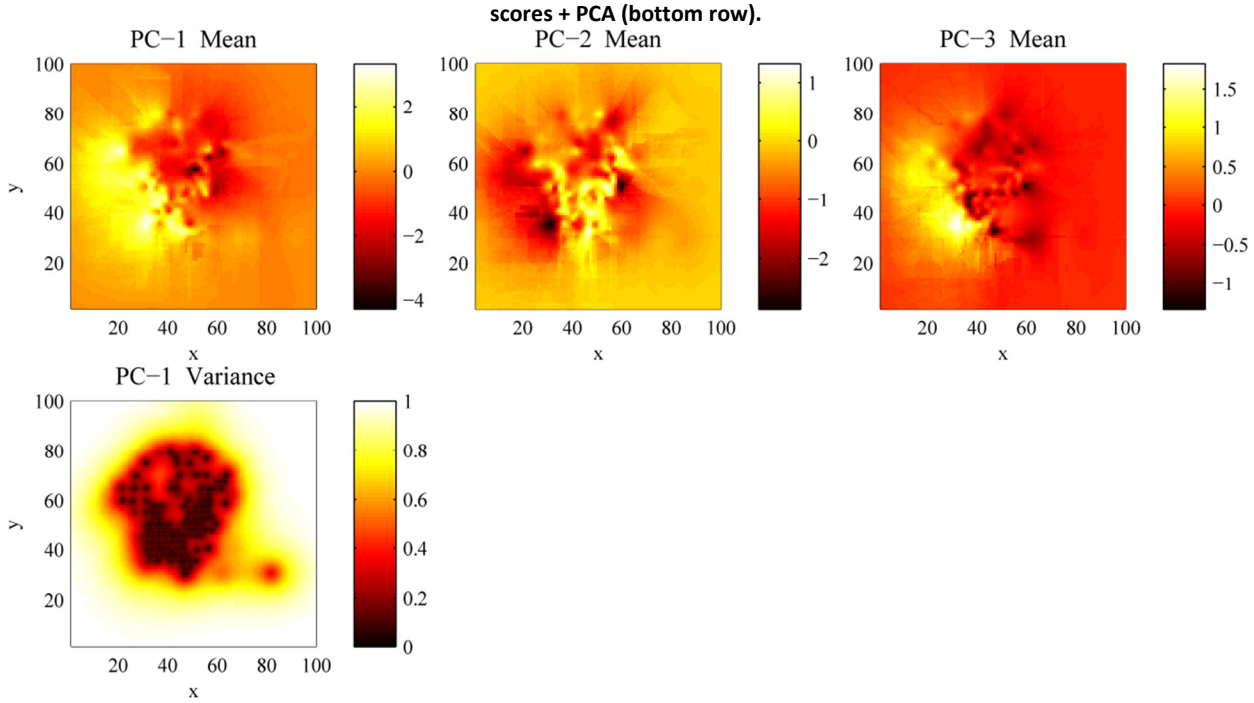

Figure 2: Kriging of principle components. The estimation variance of all three PC's was the same so only one is shown.

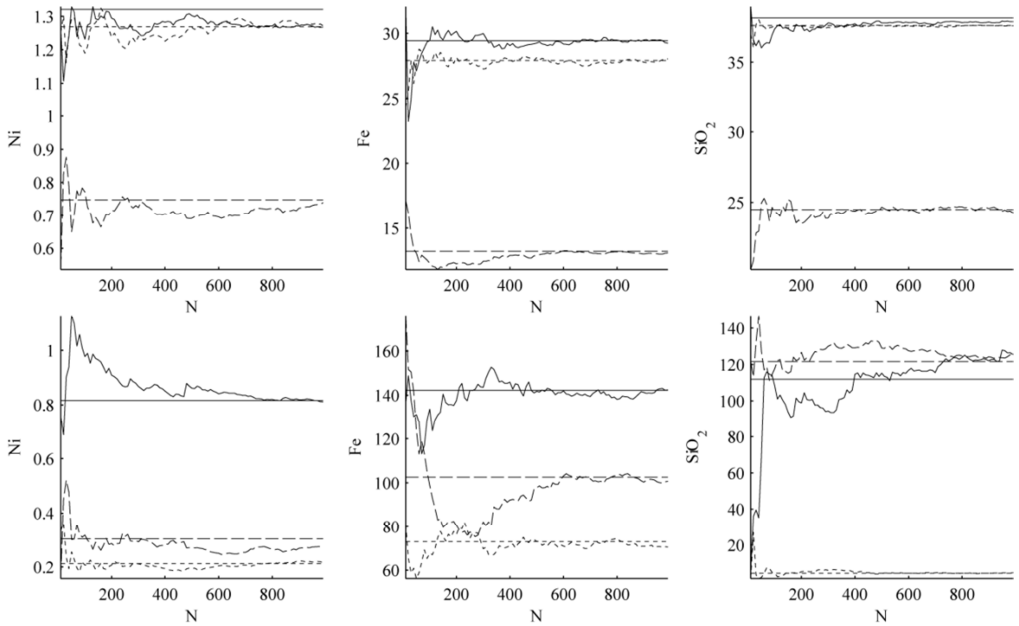

Figure 3: Convergence of back-transformed estimates and estimation variance with number of integration points, N. Line styles vary so mean and variance of each variable can be associated. Horizontal lines indicate estimates of the true values.

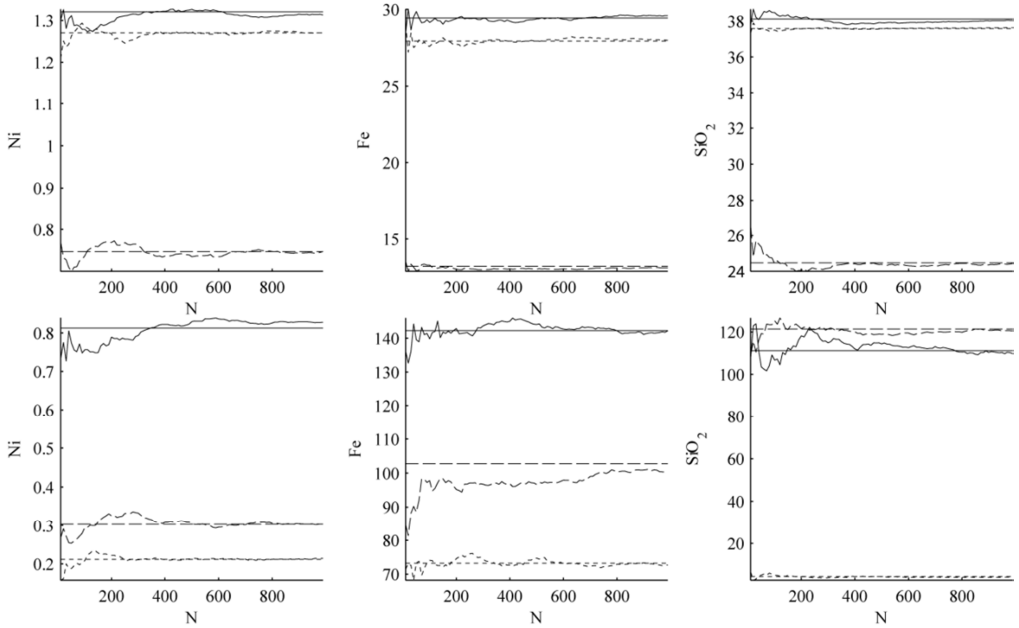

Figure 4: Convergence of back-transformed estimates and estimation variance with number of integration points, N and 10 replicates. Line styles vary so mean and variance of each variable can be associated. Horizontal lines indicate estimates of the true values.

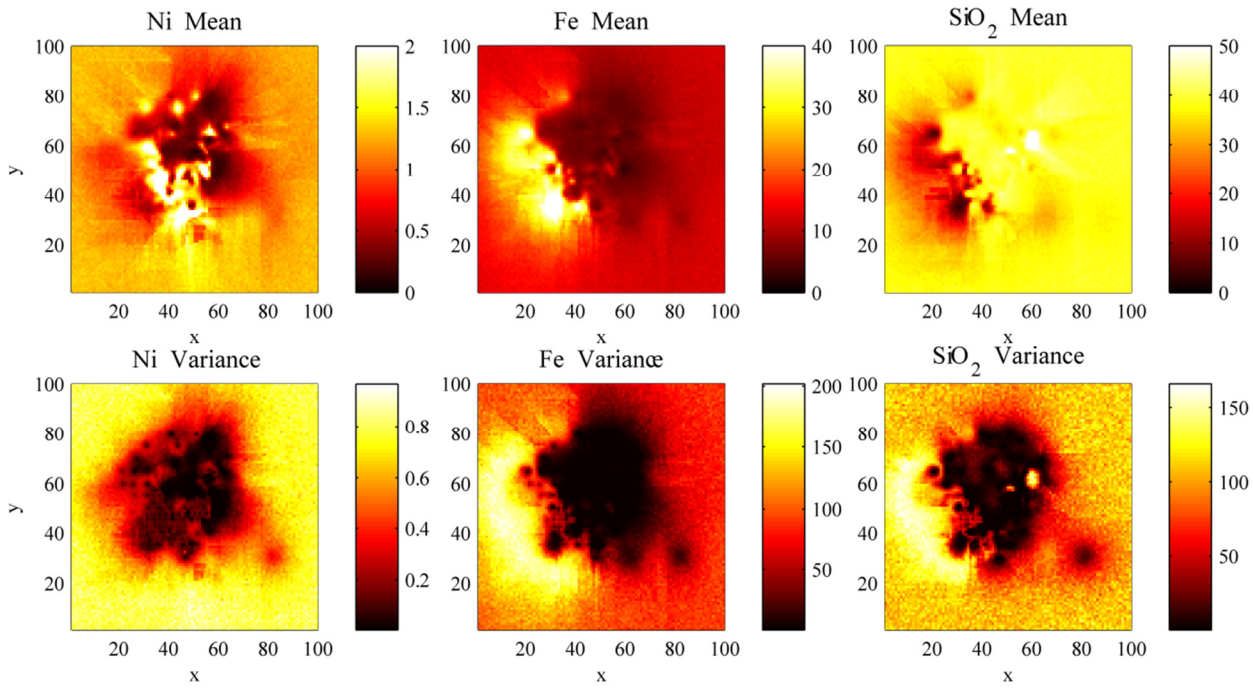

Figure 5: Back-transformed estimates of Ni, Fe, and SiO_2 through PCA and NSCORE.

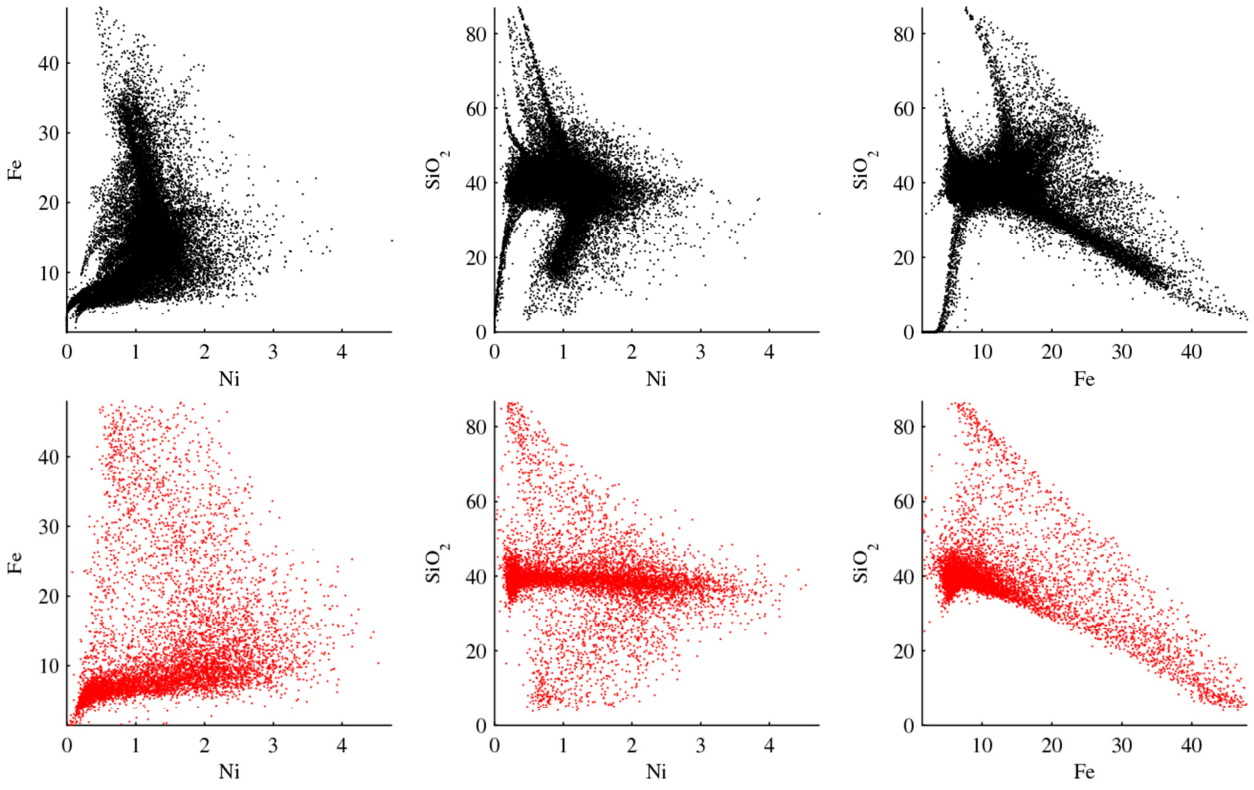

Figure 6: Cross plots showing relationships of back-transformed kriging results (top) and original data (bottom).