Conditional Standardization: A Multivariate Transformation for the Removal of Non-linear and Heteroscedastic Features

Ryan M. Barnett

Reproduction of complex multivariate features, such as non-linearity, heteroscedasticity and constraints, is a common goal when simulating multiple related geometallurgical variables. Traditional co-simulation modeling frameworks typically assume multivariate Gaussianity, whereby the relationships between variables are fully defined by the covariance matrix. Non-linear and heteroscedastic features are not captured by this statistic, and therefore will not be reproduced. In the case of independent simulation of related variables, linear transformation techniques such as principal component analysis (PCA) and minimum/maximum autocorrelation factors (MAF) may be used to decorrelate the variables prior to modeling, with back-transformation reinstating the original correlation. Unfortunately, as these techniques produce linear combinations of the original variables, complex relationships will be unaccounted for once again. Conditional Standardization is introduced as a simple and intuitive transformation for the removal of non-linear and heteroscedastic features. Given one or more conditioning variables and a variable is conditionally standardized by the subtraction of its conditional mean, and the division of its conditional standard deviation. The conditional mean and standard deviation functions may be obtained through binned partitioning of the distribution based on the probability class of the conditioning variables, or through continuous parametric regression. Distributions that approach linearity and homoscedasticity are produced, allowing for the more effective application of linear transformations or simulation methods.

Introduction

The multivariate Gaussian model is commonly implemented for geostatistical simulation of multiple related variables, due to the simplicity and mathematical tractability of the distribution. Unfortunately, geologic variables are rarely Gaussian in nature due to the existence of complex features such as non-linearity, heteroscedasticity and compositional or stoichiometric constraints. There are a number of transformation techniques that are available for the removal of these complex features, producing well behaved distributions that approach Gaussianity. As the dimensionality of the data to be modeled may render co-simulation frameworks impractical, there are additional transformations for the decorrelation of variables, allowing independent simulation to proceed without the need for cross-variograms.

The stepwise conditional transformation (Deutsch and Leuangthong 2003) decorrelates variables to produce a multivariate normal distribution, while accurately reintroducing complex features on the back-transformation. Although it has many powerful features, the data intensive nature of the stepwise transform and the fact that it does not address correlation beyond the zero lag distance may lead practitioners to other decorrelation methods such as principal component analysis (PCA) (Johnson and Wichern 1988) or minimum/maximum autocorrelation factors (MAF) (Swizter and Green 1984). These two linear transforms require less data, potentially allow for dimension reduction, and in the case of MAF, provide a more robust spatial decorrelation. Due to their linear nature, however, PCA and MAF do not capture complex multivariate features, and associated realizations may exhibit poor reproduction of complex features as a result.

Conditional standardization is proposed as a potential solution for these complex relationships, transforming non-linear and heteroscedastic data to approach linearity and homoscedasticity. In doing so, well behaved distributions that are more suitable for either co-simulation frameworks or linear decorrelation transformations are produced. The following paper will introduce this transformation, providing the simple theory, practical considerations, and a geometallurgic case study to demonstrate the technique. Parameters for the associated CCG programs are presented and discussed in the appendices.

Theory

Consider a bivariate distribution, consisting of two variables X and Z for n number of observations.

$$X_{1xn} = [x_1 \cdots x_n] \qquad Z_{1xn} = [z_1 \cdots z_n]$$

Suppose that the relationship between these two variables is non-linear and heteroscedastic in nature, such as the schematic bivariate distribution displayed in Figure 1. Observe that subtracting the Z values by a function which describes the mean of Z conditional to the value of X (red line in Figure 1), will yield residual values which have non-linearity effectively removed. Likewise, if the Z values are divided by a function which describes the standard deviation of Z conditional to the value of X, then a homoscedastic distribution will be produced. A bivariate conditional standardization is therefore given by Equation 1.

$$Z' = \frac{Z - E\{Z|_X\}}{\sqrt{Var\{Z|_X\}}} \tag{1}$$

The derivation of these conditional mean and standard deviation functions may be determined either parametrically through a form of regression or through a non-parametric partitioning of the conditioning variable in a manner similar to the stepwise conditional transform. This concept may also be extended to higher dimensions, where a variable is transformed conditional to two or more variables. The trivariate case is illustrated in Figure 2 and represented by Equation 2, where the transformed Z' variable is now conditional to the value of an additional Y variable. The conditional mean in Figure 2 is now represented by a plane, as opposed to a line in Figure 1. Non-linearity is seen to remain in the transformed distribution of Figure 2 because the bivariate relationship of the conditioning variables was not first addressed.

$$Z' = \frac{Z - E\{Z|_{X,Y}\}}{\sqrt{Var\{Z|_{X,Y}\}}} \tag{2}$$

The generalized form of the transformation for p number of conditioning variables takes on the form shown in Equation 3. The back-transformation of the data or simulated values is simply achieved by a rearranging of the forward transformation, producing Equation 4.

$$Z' = \frac{Z - E\{Z|_{X_1,\ldots,X_p}\}}{\sqrt{Var\{Z|_{X_1,\ldots,X_p}\}}} \tag{3}$$

$$Z = Z' \sqrt{Var\{Z|_{X_1,\ldots,X_p}\}} + E\{Z|_{X_1,\ldots,X_p}\} \tag{4}$$

Parametric vs. Non-Parametric Functions

The success of this transform is largely dependent on the calculation of conditional mean and standard deviation functions which accurately describe the non-linearity and heteroscedasticity of the distribution. A bivariate distribution is transformed in Figure 3 using several parametric functional forms and a discretized non-parametric function. While this figure applies the same parametric form to define both the mean and standard deviation functions for the sake of demonstration, the CCG program allows for independent consideration of each function.

As is the case in the case in Figure 3, the non-parametric approach is generally expected to produce superior results, since no assumptions of the functional form of the distribution must be made. Parametric application may still be considered as a viable option in cases where a low number of data, or high dimensionality makes the discretized non-parametric approach impractical.

Data Requirements of the Non-Parametric Approach

There is no strict rule regarding the number of classes that are required for partitioning the conditioning variable, or the number of data that are required in each bin for the subsequent calculation of mean and standard deviation. The fewer the classes, the more likely that complex features will remain within the partitioned bins following transformation. Conversely, increasing the number of classes may reduce the number of data in each bin leading to unstable calculation of the conditional statistics.

This is the same issue faced by the stepwise conditional transformation, where it was found that a general rule is to use between 10 and 20 classes for discretizing each conditional variable, with 10-20

data required as a minimum for the calculation of each bin (Leuangthong 2003). Based on observation this applies for conditional standardization as well, with between 10-20 data required as a minimum to calculate a stable mean and standard deviation for the partitioned bins. It then follows that between 10^n and 20^n data will be needed for the transformation, where n is the number of variables being considered.

Consequently, only in the case of very large datasets (>10,000 as an absolute minimum), will the discretized approach be applied beyond a trivariate system. Unfortunately, multivariate distributions exist for geologic datasets of less than 10,000 observations and greater than three variables. In these cases practitioners may choose between either using a parametric calculation of the conditional functions, or a 'nested' application of the non-parametric approach. To help ease this sensitivity, the CCG program is implemented to allow for a smoothing data search beyond the bin limits, as well as the enforcement of order relations.

Nested Application

Nested conditional standardization refers to using only one or two conditioning variables, to remove non-linearity and heteroscedasticity with the higher order conditioned variables. Removal of the selected complex relationships will oftentimes resolve the majority of the complexity between variables that are not directly transformed conditional to one another. This is not guaranteed, however, and careful decision making must take place regarding the conditioning variables for this nested application. Considerations may include: (1) Reproduction of the multivariate relationships which the primary resource variable holds with all secondary variables (resource variables becomes the first conditioning variable for all transformations), (2) Reproduction of a multivariate relationship between secondary variables where the ratio or correlation between them is of critical interest (one secondary variable must condition the other), and (3) Conditioning of variables that demonstrate dramatic bivariate complexity. In the case of very non-linear or heteroscedastic relationships between variables, it is unlikely that a well behaved distribution will be produced unless one variable directly conditions another. These considerations will often lead to difficult decision making, as not all of them may be satisfied.

Ordering

The ordering of variables for this transformation is not trivial, as the variogram of a conditioned variable will no longer reflect the spatial correlation of only the transformed variable. As was thoroughly investigated for the stepwise conditional transformation (Leuangthong 2003), the variogram of a conditioned variable will instead be a combination of the original variogram, the variograms of the conditioning variables, and the cross-variograms formed between them. As spatial structure of the conditioned variables will be altered, it is recommended that variables of greater spatial structure be chosen as the lower order conditioning variables. One can imagine that conditioning a very continuous variable with another that is largely composed of the nugget effect, could have very negative consequences on the reproduction of both variables' spatial structure. Practical considerations may weigh on this decision making, such as leaving spatial structure of the primary resource variable unaltered.

Case Study

A nickel laterite dataset is used to demonstrate this transformation, which is composed of 7740 homotopically sampled assays. A typical mining model for a nickel laterite deposit will require the Nickel (Ni) resource, as well as Iron (Fe), Silica Dioxide (SiO_2) and Magnesium Oxide (MgO), which all exert a critical influence on smelting extraction. Figure 4 displays the bivariate cross-plots between Ni, Fe and SiO_2, where heteroscedastic, non-linear and constraint features are clearly observed and highlighted. Optimization of plant design, stockpiling, and blend planning will require realistic reproduction of the univariate variabiilty for each variable, as well as the multivariate relationships between them.

Only a bivariate and trivariate application will be demonstrated, though the conditional standardization workflow in Figure 5 includes MgO for readers to understand how the transformation may be applied beyond three dimensions in a nested fashion. This workflow figure displays that an initial bivariate transformation of Fe conditional to Ni is executed to remove non-linearity and heteroscedasticity between the two variables. Next, a trivariate application will transform SiO_2 to remove

non-linearity and heteroscedasticity from its relationship with the Ni and previously condionally standardized Fe. Finally, nested conditioning of the MgO would take place using Nickel and conditionally standardized SiO2.

The non-parametric conditional standardization transform will be applied to a trivariate system of Ni, Fe and SiO_2, all of which have been normal score transformed (Figure 6). A normal score transform is not required prior to conditional standardization, but it is recommended if outlying values are present. The parametric conditional functions are calculated through least squares regression, which is very sensitive to outying values. Likewise the non-parametric, discretized approach may have instability along the margins of a distribution in the case sparesely populated outlier values. In either case, the normal score transform will have a mitigating effect on this outlier issue.

Conditional standardization was executed following the workflow in Figure 5, transforming the non-linear and heteroscedastic distribution that is seen in Figure 6, to the linear and homoscedastic distribution in Figure 7. Following this conditional standardization, the transformed Ni, Fe and SiO_2 were decorrelated using MAF to remove any remaining correlation beyond the zero lag distance, independently simulated using SGS (Deutsch and Journel 1998) and back-transformed. A second modeling workflow that does not include the conditional standardization step, but was identical in every other regard was also completed for comparison.

The bivariate cross-plots for the back-transformed realizations with and without the use of conditional standardization are displayed in Figures 8 and 9 respectively. The major non-linear and heteroscedastic features of the original distribution (Figure 4) are also displayed for comparison. Due to the number of data being considered in both the original and simulation cross-plots (where 1 in 150 data are displayed for the later), it may be difficult to visualize whether the appropriate density of data is being reproduced. Bivariate gaussian kernel plots were produced in Figure 10 to aid in this comparison, as they display the relative density of the original and simulated distributions. With all other factors being held constant between these two modeling workflows, a significant improvement is seen in the reproduction of the non-linear and heteroscedastic features when using conditional standardization.

Conclusions

Complex multivariate features may be of critical importance to both the understanding and realistic modeling of a geologic deposit. Failure to account for these complexities prior to the use of covariance based geostatistical modeling and transformation tools, will often result in poor reproduction of the original multivariate relationships.

Conditional standardization is a potential solution to this issue. After the removal of non-linear and heteroscedastic features, well behaved distributions are produced that more closely obey the Gaussian assumptions of traditional geostatistical techniques based on kriging or maximum entropy. Following simulation, the back-transformation then reinforces the original multivariate complexities. Conditional standardization has been successfully applied to a Nickel laterite case study, and the relevant programs are presented in the appendices to this paper.

References

Deutsch, CV, Journel, AG. 1998. GSLIB: Geostatistical Software Library and User's Guide: 2nd edition. New York: Oxford University Press.

Johnson, RA, Wichern, DW. 1988. Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall. p. 340 – 370.

Leuangthong, O. 2003. Stepwise Conditional Transformation for Multivariate Geostatistical Simulation, PhD Thesis, University of Alberta.

Leuangthong, O, Deutsch, CV. 2003. Stepwise Conditional Transformation for Simulation of Multiple Variables. Mathematical Geology. Vol.35, No.2: p155-172.

Rosenblatt, M. 1952. Remarks on a multivariate transformation. Annals of Mathematical Statistics. Vol.23, No.3: p.470-472.

Switzer, P, Green, A. 1984. Min/Max autocorrelation factors for multivariate spatial imaging. Stanford University: Department of Statistics, Technical Report No.6. 14p.

Figures

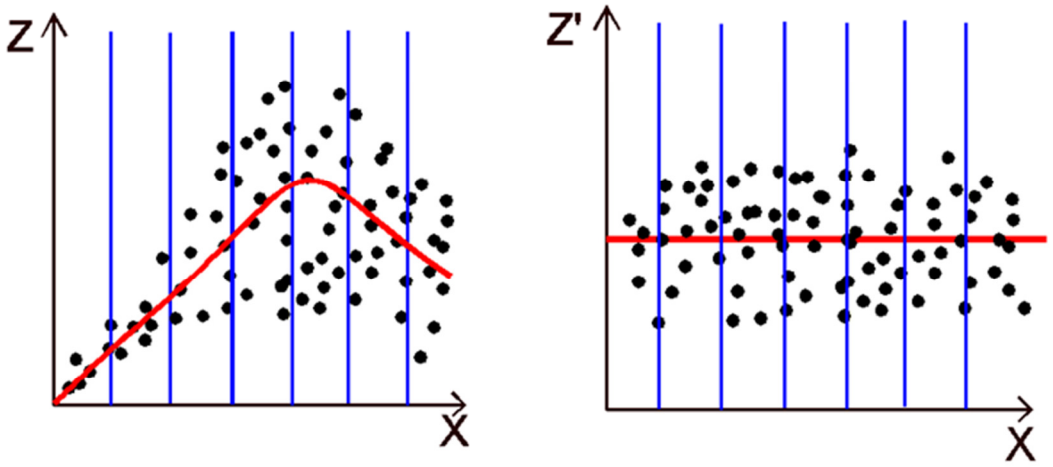

Figure 1: Schematic of a non-linear and heteroscedastic bivariate distribution that has been partitioned according to conditional probability classes of X (left). Subtraction of the conditional mean and division of the conditional standard deviation yields a linear and homoscedastic distribution (right).

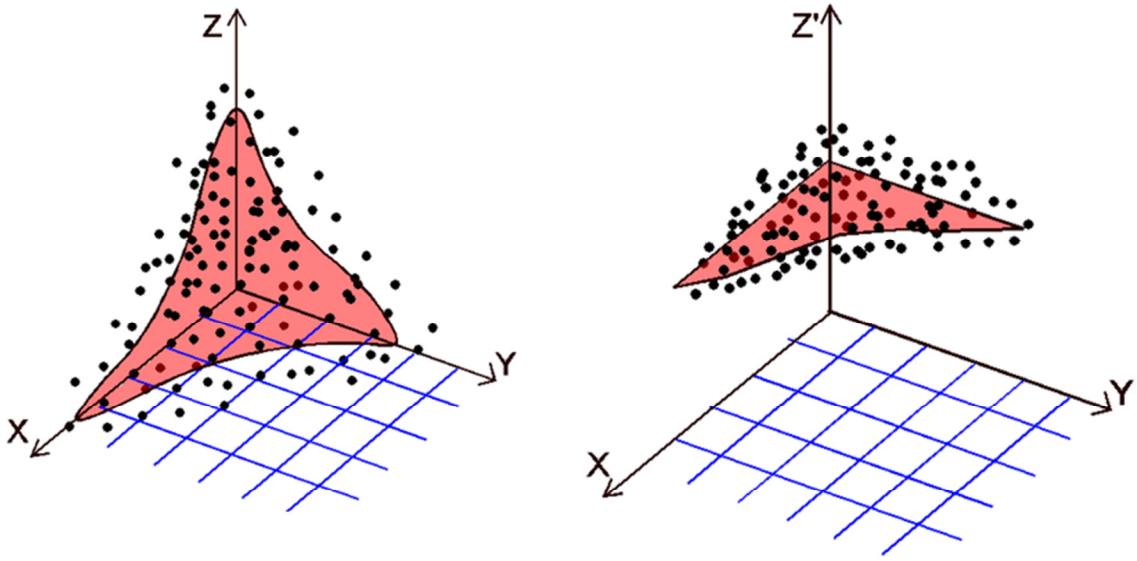

Figure 2: Schematic of a non-linear and heteroscedastic trivariate distribution that has been partitioned according to conditional probability classes of X and Y (left). Subtraction of the conditional mean and division of the conditional standard deviation yields a linear and homoscedastic distribution (right). Note that the bivariate non-linearity between X and Y has not been addressed prior to this transformation and therefore remains.

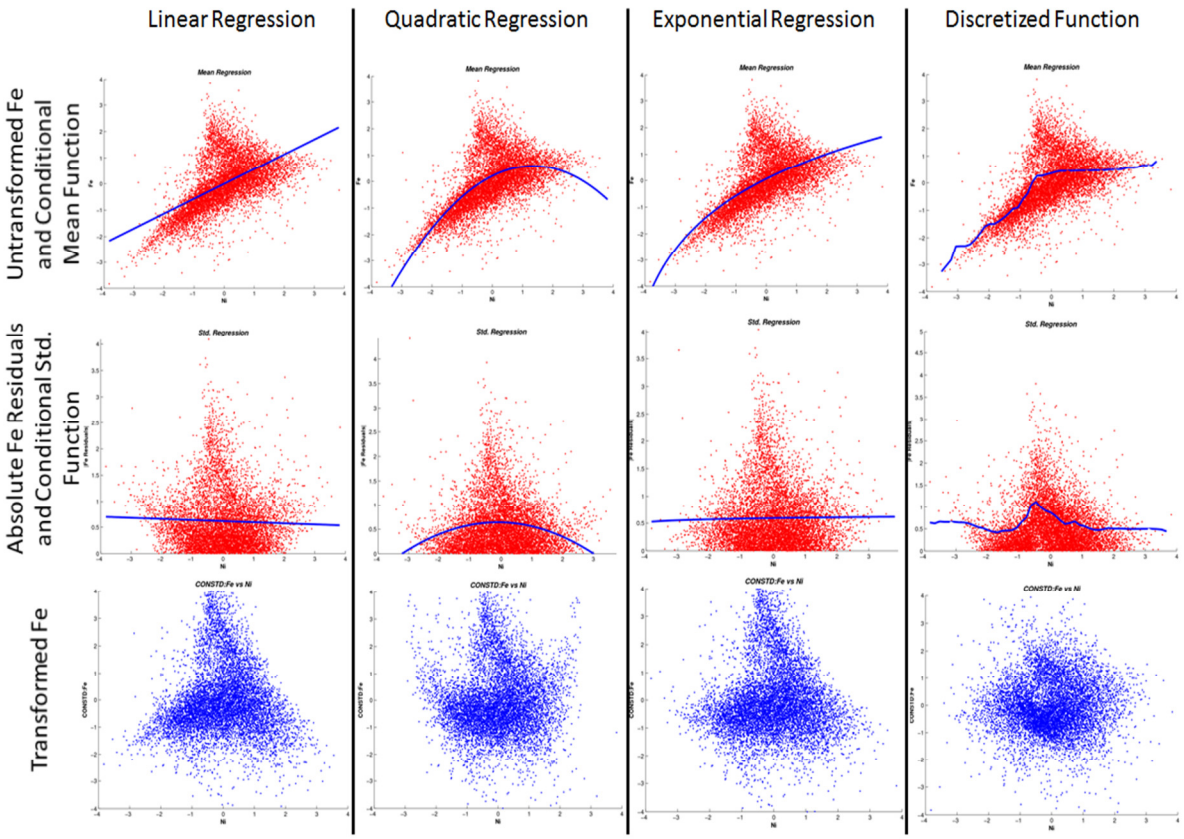

Figure 3: Demonstration of various calculations of the conditional mean function on the Ni and Fe cross-plot (top). The 'standard deviation' function is then calculated using the residuals of the preceding mean function (middle). The standard deviation function is calculated in an identical manner as the associated mean function in this demonstration, though this constraint is not present with the program. Resultant conditionally standardized distributions for each set of mean and standard deviation functions are shown at the bottom, with the discretized method (far right) exhibiting the only acceptable results in this case. This figure serves to reinforce that the parametric approach must only be applied where the distribution obeys the chosen functional form.

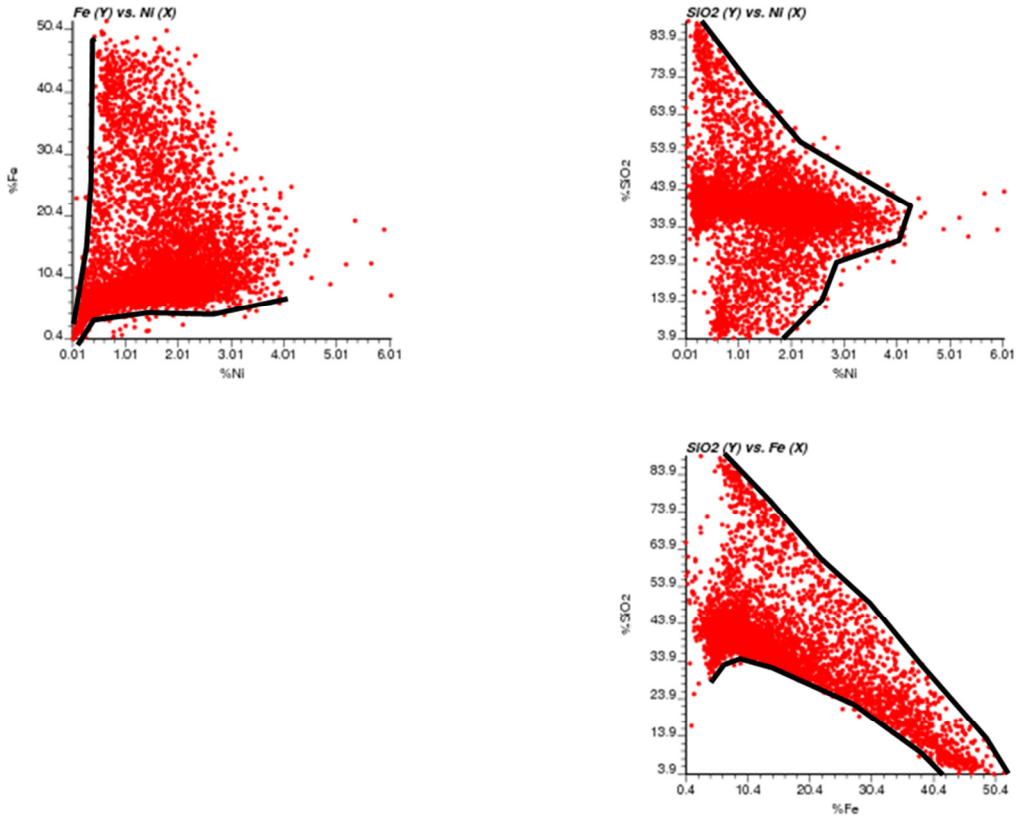

Figure 4: Cross-plots between the untransformed Nickel laterite variables: Ni, Fe and SiO_2. Major non-linear and constraint features are outlined.

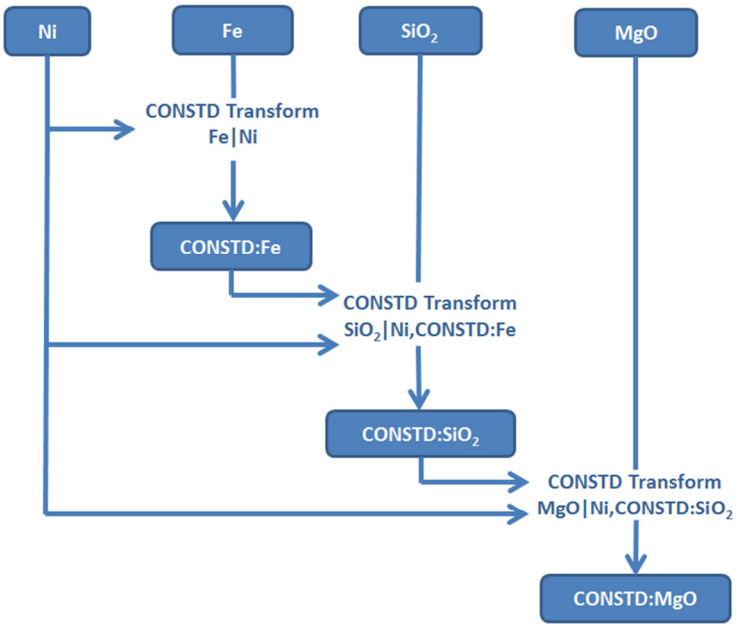

Figure 5: Sequential conditional standardization workflow for four variables of the Nickel laterite dataset

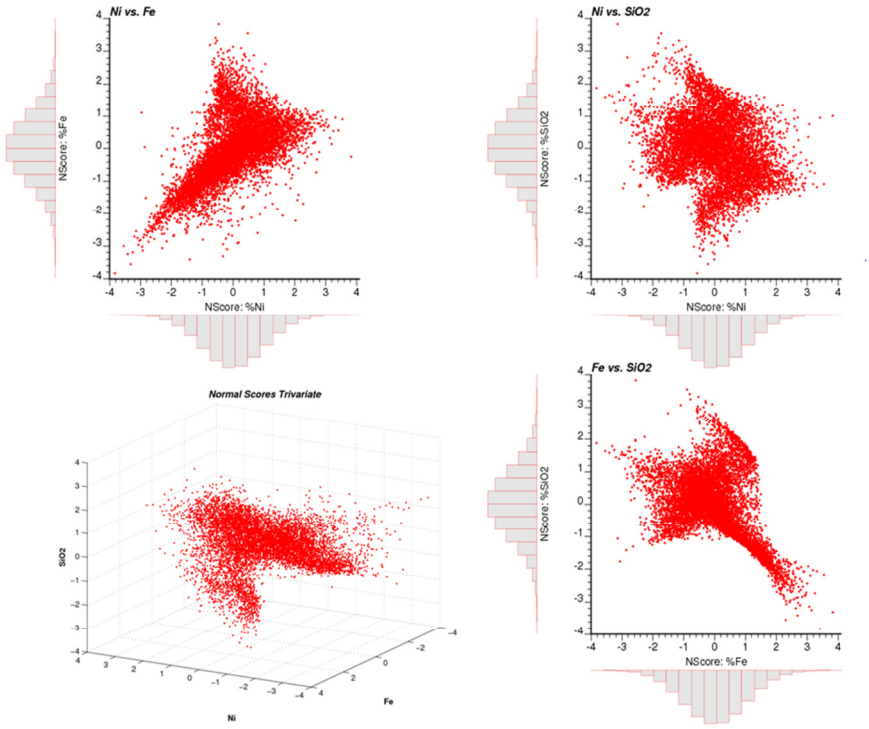

Figure 6: Trivariate (bottom left) and individual bivariate cross-plots between normal score transformed Ni, Fe and SiO_2. Marginal histograms display the univariate normality, but complex multivariate features clearly remain. This will be the 'pre-transform' distribution prior to conditional standardization.

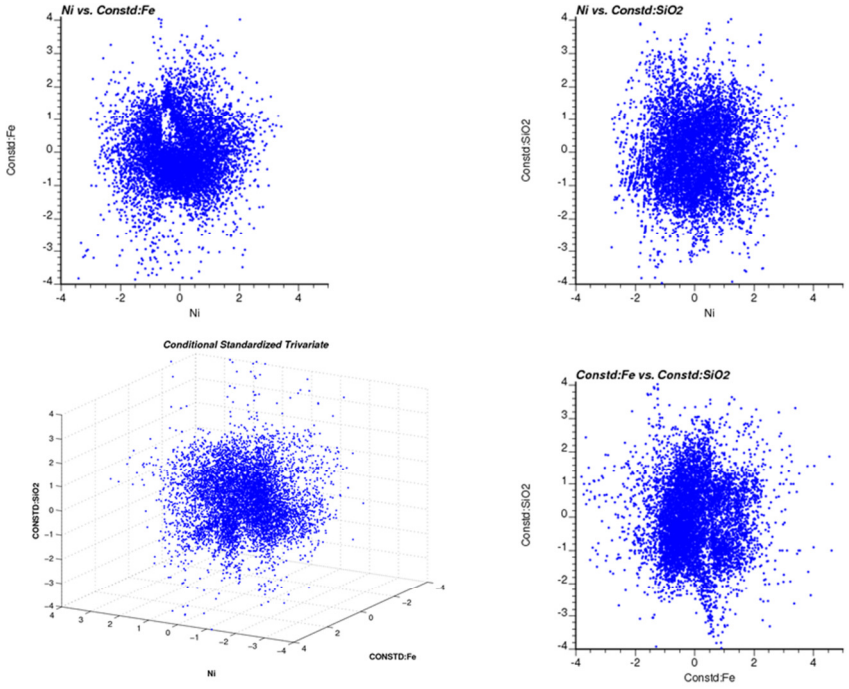

Figure 7: Trivariate (bottom left) and individual bivariate cross-plots between conditionally standardized Ni, Fe and SiO_2 following the workflow in Figure 5 with a non-parametric execution. The complex features from Figure 6 have been removed.

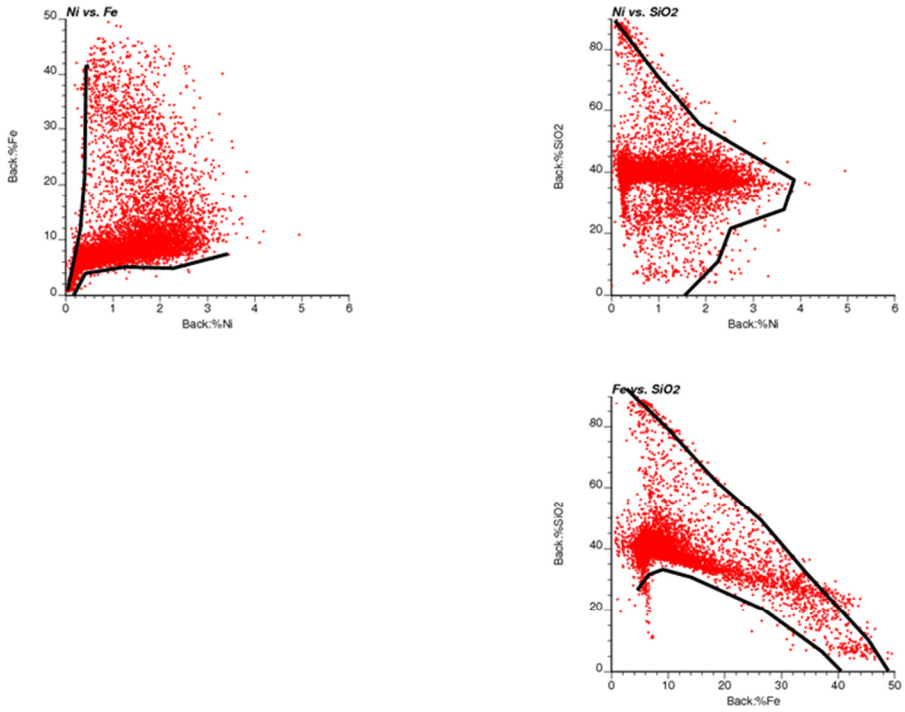

Figure 8: Cross-plots between the simulated (1 in 150 displayed) and back-transformed Nickel laterite variables: Ni, Fe and SiO_2 using conditional standardization and otherwise following an identical modeling workflow to the results shown in Figure 9.

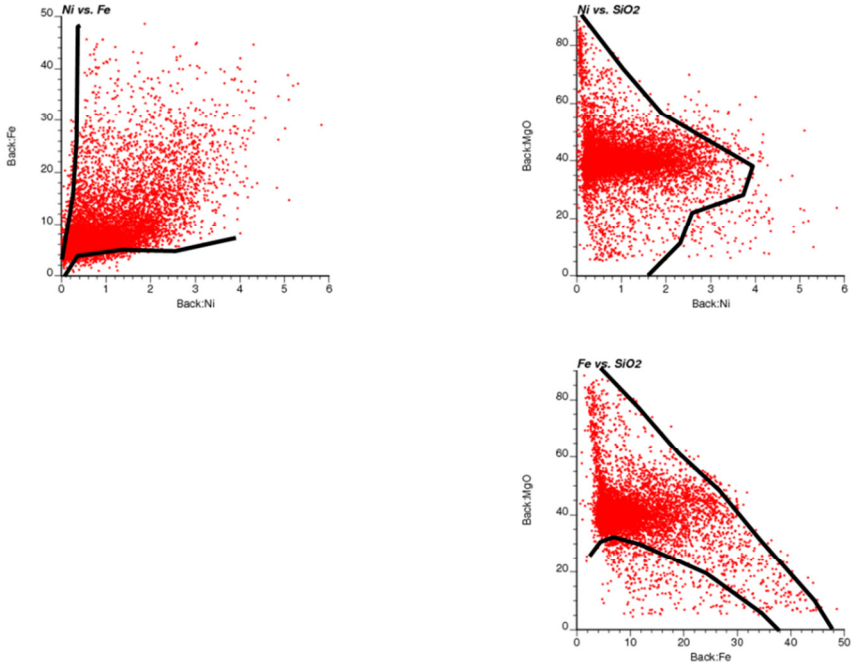

Figure 9: Cross-plots between the simulated (1 in 150 displayed) and back-transformed Nickel laterite variables: Ni, Fe and SiO_2, following a workflow that does not account for complex features.

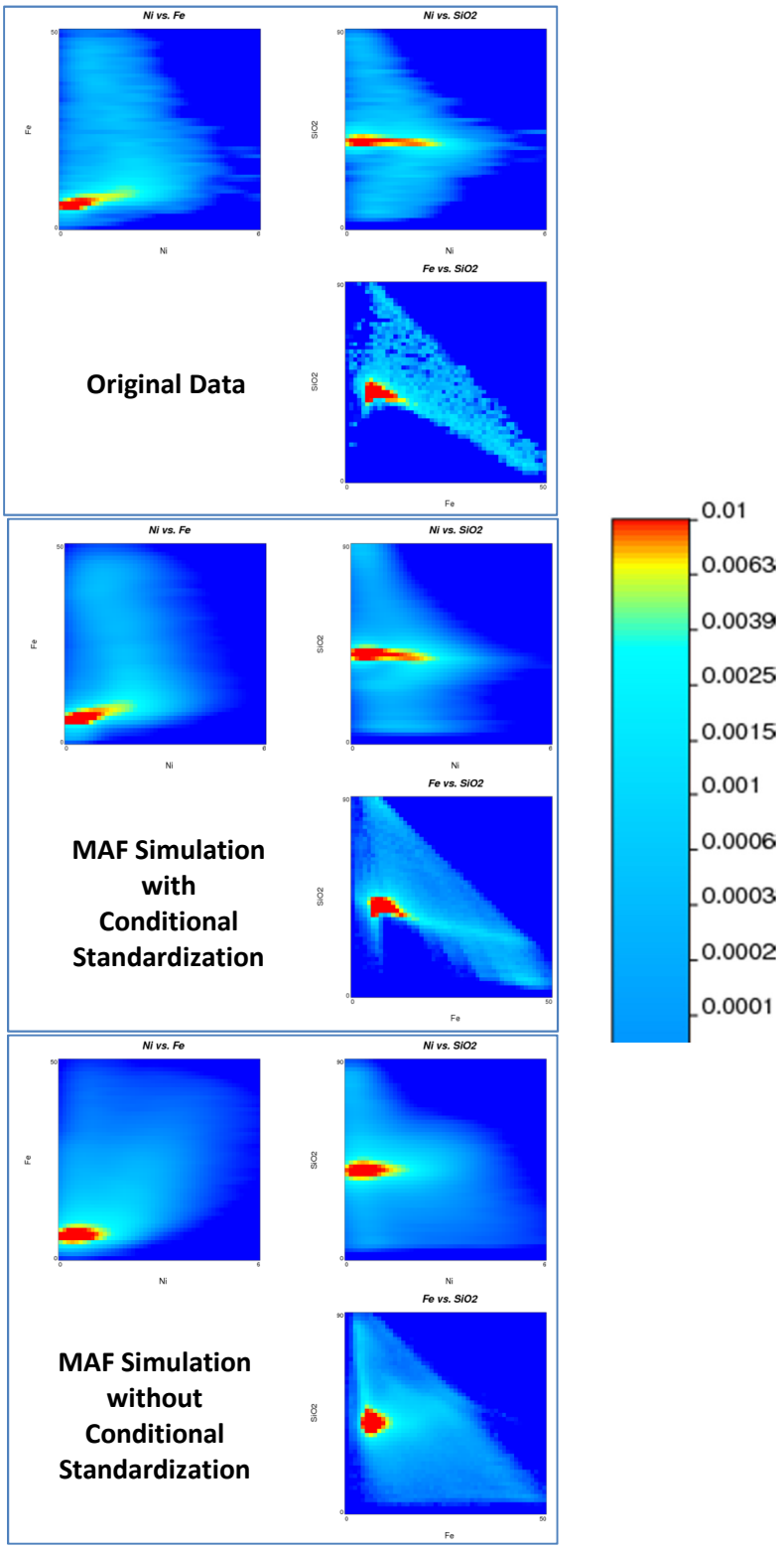

Figure 10: Bivariate Gaussian Kernel plots, displaying the relative distribution density for the original data (top), MAF with conditional standardization simulation (middle) and MAF without conditional standardization simulation (bottom). These plots correspond with the bivariate scatter plots in Figures 4, 8 and 9 respectively. The densities were calculated using Jeff Boisvert's biv_gauss_kernel program.

Appendix: Software

The first Conditional Standardization program, **constd**, is used to perform the forward transformation using a GSLIB (Deutsch and Journel 1998) format of implementation. The corresponding required parameters are shown in Figure 10 and are explained below:

- **datafl:** file with the input data to be transformed.
- **ixp,iyp,izp,iwp:** columns for the x, y, z and weight variables. Refer to Note1 in the parameter figure for additional considerations.
- **tmin, tmax:** trimming limits to filter out data.
- **xmin,xmax:** minimum and maximum value of the conditional x variable for allocating bins. Refer to Note2 in the parameter figure for additional considerations.
- **ymin,ymax:** minimum and maximum value of the conditional y variable for allocating bins.
- **ipar:** parametric (1) or discretized (0) calculation of the mean and standard deviation functions.
  If **ipar=1,** the following two lines apply:
- **xm_regress, ym_regress:** order of regression for the parametric calculation of the conditional mean. Refer to Note 3 in the parameter file for additional considerations. Note that if optimizing, all methods will be tested, with the functional form that produces the lowest mean squared error selected. When possible, it is advised to choose the functional form based on visual inspection of both the original distribution and resultant residual values, as mean squared error does not reveal issues regarding the bias of a function.
- **xstd_regress, ystd_regress:** order of regression for the parametric calculation of the conditional 'standard deviation'. Refer to Note 3 in the parameter file for additional considerations.
  If **ipar=0,** the following three lines apply:
- **nxdis, nydis:** number of discretizations for partitioning the x and y conditioning variables.
- **bxsize, bysize, nbmax:** multiplying factor of sample consideration limits in the x and y directions (refer to Note 4 in the parameter for additional details). Samples within each conditioning bin will be sorted by distance from the center, with samples above the **nbmax** threshold discarded. A large **bxsize/bysize**, a large number of **nxdis/nydis**, with a relatively low **nbmax** may therefore be used to improve stability of the function calculation in sparse regions of a distribution, while maintaining appropriate resolution in the dense regions.
- **iorder**: order relations for the x mean and x standard deviation functions (refer to Note 4 in the parameter file for specifications).
- **outfl:** file for output from the constd transform. This file contains the transformed z variable appended to the original data file.
- **outfltrn:** output file for the transformation table. Contains conditional bin limits and associated mean/standard deviation of the transformed variable.

```
                  Parameters for CONSTD
                  *********************

START OF PARAMETERS:
data.dat                        - file with data
4   3  2 0                      - columns for X, Y, Z and weight (Note1)
-1.0e21     1.0e21              - trimming limits
1 0                             - xmin,xmax(Note2)
1 0                             - ymin,ymax
1                               - method for trans.(1=parametric,0=discretize)
2 4                             - If 1, order of regression for x mean,y mean(Note3)
0 0                             - If 1, order of regression for x std.,y std.(Note3)
50 50                           - If 0, # x discretizations, # y discretizations
2 2 500                         - If 0,  xbin size, ybin size,maximum nsamples/bin.(Note4)
1 0                             - If 0, order rel. for x mean, x std(0=no,-1=decr.,1=incr.)
constd.out                      - output file for transformed values
constd.trn                      - output file for transformation table


**Note1: X and Z must always be specified as the conditional and transformed
         variables respectively. Y is only specified if there is a second
         conditional variable, and left as 0 otherwise
**Note2: xmax < xmin will automatically calculate the values from the trimmed data
         the same applies for the subsequent ymin/ymax line
**Note3: 0=optimize (lowest MSE),1=linear,2=quadratic,3=cubic,4=Exponential
**Note4: limits for each x conditional bin = (xmax-xmin) / (#x dis.) * (xbinsize)
         The same applies for y.
```

Figure 10: Parameter file for the forward constd program

The second Conditional Standardization program, **constd_b**, is used to perform the back-transformation. The corresponding required parameters are shown in 11 and are explained below:

- **datafl:** file with the input data to be transformed.
- **ixp,iyp,izp:** columns for the x ,y, and z variables. Refer to the Note in the parameter figure for additional considerations.
- **tmin, tmax:** trimming limits to filter out data.
- **trnfl:** file containing the transformation table from the forward constd program.
- **outfl:** file for output. This file contains the back-transformed variable appended to the original data file.

```
                  Parameters for CONSTD_B
                  ***********************

START OF PARAMETERS:
data.dat                          - file with data
4   0   3                         - cols for x,y,z variables (see Note)
80 500                            - trimming limits
constd.trn                        -file with constd forward transformation table
constd_b.out                      - output file for backtransformed values


**Note: x,y,z variables must align with the specification in the
        forward conditional standardization
```

Figure 11: Parameter file for the backward constd_b program