Stepwise Conditional Transformation in Estimation Mode

Clayton V. Deutsch

Centre for Computational Geostatistics (CCG)
Department of Civil and Environmental Engineering
University of Alberta

Stepwise Conditional Transformation was devised as a multivariate transformation technique to transform correlated variables to independent Gaussian distributed variables. In a geostatistical context, the stepwise transform is almost always applied in simulation mode. The estimates of stepwise transformed variables cannot be simply back transformed; the results are likely to be biased. There is often a need to provide single best estimates of multiple variables. Averaging many simulated values is tedious and sensitive to the number of realizations. This short note documents a back transform approach analogous to MultiGaussian (MG) kriging so that stepwise estimates can be directly back transformed in an unbiased manner.

Stepwise Transformation and Back Transformation

Consider K original variables z_k, $k=1,\ldots,K$. The transformation of the first variable is the normal score transformation $y_1=G^{-1}(F_{z1}(z_1))$. Subsequent transformations are made conditional to previous variables:

$$y_k = G^{-1}\left(F_{z_k|z_j,\, j=1,\ldots,k-1}(z_k)\right), k = 2,\ldots,K$$

The derivation of the conditional distributions becomes increasingly difficult as k increases because there is a decreasing number of data for to infer the conditional distributions. The resulting y variables are Gaussian distributed and uncorrelated, hence independent. This is a very nice property for simulation.

Simulated y-values can be back transformed using the appropriate conditional distributions. The first y-variable is the normal score back transformation $z_1= F_{z1}^{-1}(G(y_1))$. Subsequent back transformations are made with conditional distributions.

Bias in Estimation

Gaussian transformation and back transformation is at the heart of stepwise conditional transformation. The Z distribution is either the original Z_1 distribution or conditional distributions. The transformation and back transformation is shown on Figure 1; the equations are recalled below:

$$y = G^{-1}\left(F_Z\left(z\right)\right) \quad \text{and} \quad z = F_Z^{-1}\left(G\left(y\right)\right)$$

This quantile-to-quantile transformation procedure is well established. Conditional distributions in Gaussian space are fully defined by a Gaussian mean and variance or standard deviation. Original variable Z distributions are rarely Gaussian; they are often of non-negative variables and positively skewed. This means that conditional distributions in original units are also positively skewed, although the shape will change. Figure 2 shows a schematic example of the back transformation of a conditional distribution. The conditional distribution on the left represents a symmetric Gaussian distribution fully described by a Gaussiam y_{sk} and standard deviation σ_{sk}. The median and mean are the same in Gaussian units: $M_y=m_y=y_{sk}$. The median in Z units is a direct back transformation of the Gaussian median. The mean is *not* a direct backtransformation of the Gaussian mean. The back transform of the mean is, in fact, the median, which can be far from the mean if the distribution is highly skewed. The result would only be close if (1) the distribution of Z is symmetric, or (2) the variance of the conditional distribution approaches zero. The back transformation of the mean has been addressed by MG kriging (Verly, 1984).

The MG Approach to Back Transformation

Quantiles can be back transformed from Gaussian units to original Z units. We must back transform a large number of quantiles to calculate the mean in Z units. We denote this by a series of p^l, $l=1,..,L$ values:

$$z^l = F_Z^{-1}\left(G\left(\sigma_k G^{-1}(p^l)+y^*\right)\right),\ l=1,...,L$$

The quantiles must be unbiased, that is, they must fairly represent the range of Y/Z values. The p^l values could be regularly distributed between 0 and 1. Alternatively, the p^l values could be drawn by MCS. The advantage of regularly spaced quantiles is fast convergence to the mean. The advantage of MCS is an improved prediction of the variance – extreme values have a better possibility to be drawn. This author prefers regular spaced quantiles; 200 or more.

The PostMG program (Verly, 1984; Lyster and Deutsch, 2004) performs the back transformation with a user-defined number of quantiles. The back transformation of quantiles is shown on Figure 3 – the same principle as Figure 1. This principle will be extended to stepwise estimates.

PostProcessing Stepwise Estimates

Simulated stepwise values can be associated to quantiles and, therefore, directly back transformed. The back transformation is straightforward with the correct conditional distribution. Kriging of stepwise transforms does not, however, provide quantiles that can be directly back transformed. The result is a kriged mean and variance in stepwise/Gaussian units for each variable:

$$\left(y_{SK,k},\sigma_{SK,k}\right),\ k=1,...,K$$

The first variable ($y_{SK,1},\sigma_{SK,1}$) is exactly the result of MG kriging. The other variables relate to conditional distributions and not the normal score transform of an original variable. The stepwise variables are uncorrelated. A schematic illustration of a bivariate conditional distribution provided by kriging the stepwise transforms is shown on Figure 4 – the yellow ellipses. This distribution must be back transformed to provide a distribution in original Z units.

The K-variate distribution of stepwise variables must be transformed. We discretize a K-variate standard Gaussian distribution by a regular n_{dis} intervals:

$$y_i = G^{-1}\left(\frac{i-1/2}{n_{dis}}\right), i = 1, \ldots n_{dis}$$

The paired values do not receive equal weight. The weight of a set of $y_{k1}, y_{k2}, \ldots, y_{kK}$ is proportional to the product of the marginal distributions. The y values are independent; therefore the product is the correct weighting:

$$wt_k \approx \prod_{j=1}^{K} e^{\left(\frac{-y_{k,j}}{2}\right)^2}$$

These weights are standardized. Figure 5 shows a bivariate example. The gray scale shading is proportional to the weighting. The y values are non-standardized according to:

$$y_{k,i}^{ns} = y_{SK,k} + y_{k,i} \bullet \sigma_{SK,k}$$

These values are back transformed as if they were simulated stepwise values and weighted according to the standardized weights.

Computer Code

The stepwise conditional transformed values, the transformation table, and kriged values (plus kriging variances) are required. The transformed values and transformation table are generated by the `sctrans` program. `kt3d` can be used for the kriging. The back transformation program has been named `postSCT` (like the `postMG` or `postsim` programs). The parameters:

```
2                              -number of variables
kt3d01.out                     -file with variable: 1
1 2                            -  columns for mean and variance
kt3d02.out                     -file with variable: 2
1 2                            -  columns for mean and variance
sctrans.trn                    -trans table
20                             -discretization level
postSCT.out                    -file for output
```

The number of variables will normally be 2 or 3. A file containing the mean and variance of each stepwise/Gaussian variable is required. The values are assumed missing of the mean is outside of the interval -9 to 9 or the variance is outside of 0 to 2 (of course, the variance should always be less than or equal to 1). The mean and variance values would normally come from simple kriging. The trainsformation table from `sctrans` is required. The discretization level applies to each variable. Setting a value of 20 for a bivariate problem would mean a discretization of 400 points. The output file contains the mean and variance of each variable in original units. The program could easily be modified to extract additional statistics such as quantiles and probability intervals.

Small Example

Figure 6 shows an original data cross plot and the result of a small stepwise simulation. 50 classes were used in the stepwise transformation. The data statistics are reproduced very close. 100 data in a 2-D plane were extracted for testing. Figure 7 shows the results of kriging the stepwise transformed variables. The top figures are the mean and variance of the first variable. The bottom row is the second variable. The data locations are readily seen as the zero-kriging variance locations. Figure 8 shows the mean and variance output of postSCT. The mean values are now unbiased. The variance values are heteroscedastic, that is, the high mean values are more variable than the low values. This is due to the skewed nature of the original histograms. A cross plot of the estimated values is shown on Figure 9. The values are, of course, smoother. A discretization of 20x20 (400) points were used for the transformation.

Conclusions

Estimates are unquestionably important in resource estimation. This short note presents a methodology for kriging and back transformation of stepwise transformed variables while avoiding bias. An approach analogous to the MG approach to kriging back transformation is applied. Code is provided.

References

Deutsch, C.V., *"Order Relations Correlation and Tail Extrapolation for Stepwise Conditional Transformation,"* Paper 2005-109, Center for Computational Geostatistics, September 2005.

Leuangthong, O., *"Stepwise Conditional Transformation"*, Ph.D. Thesis, University of Alberta, 2003.

Leuangthong, O., and Deutsch, C.V., "*Transformation of Residuals to Avoid Artifacts in Geostatistical Modelling with a Trend*," Mathematical Geology, accepted September 2003.

Leuangthong, O., and Deutsch, C.V., *"Stepwise Conditional Transformation for Simulation of Multiple Variables,"* Mathematical Geology, Vol.35, No. 2, pp. 155-173.

Lyster, S. and Deutsch, C.V., *"PostMG: A Postprocessing Program for Multigaussian Kriging Output"*, Paper 2004-405, Center for Computational Geostatistics, September 2004.

Ortiz, J.M. and Clayton V. Deutsch. "*Uncertainty Upscaling*", In Centre for Computational Geostatistics, Volume 5, Edmonton, AB, 2003.

Verly, G., "*Multivariate Kriging*", Ph.D. Thesis, Stanford University, 1984.

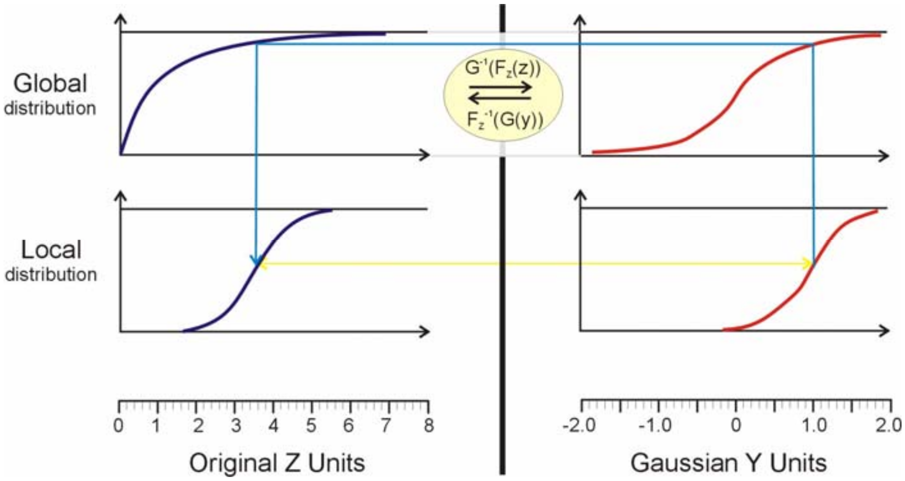

Figure 1: Schematic illustration of Gaussian transformation and back transformation. The equations in the central ellipse permit quantiles of the original data to be transformed to Gaussian values (and back transformed).

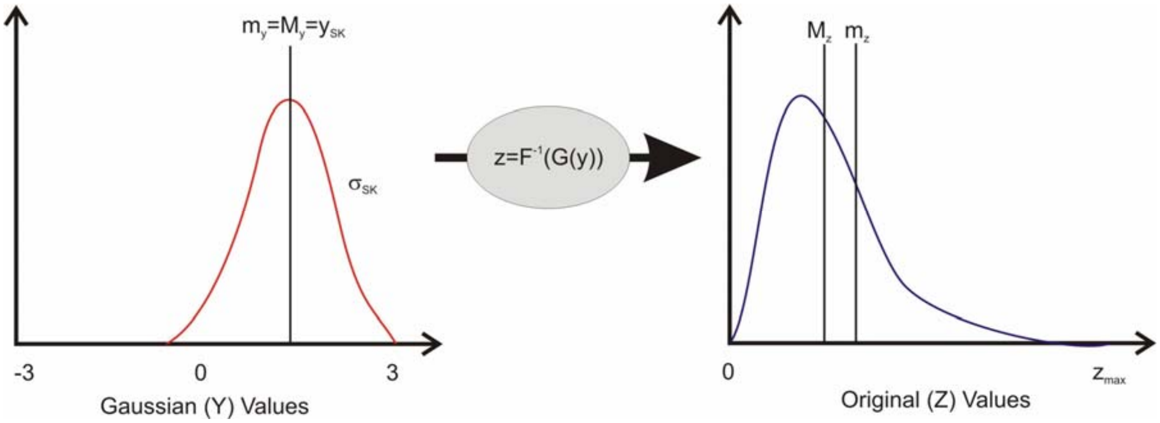

Figure 2: Schematic illustration of the Gaussian back transformation of a conditional Gaussian distribution. Note that the mean and median in back transformed space are not the same.

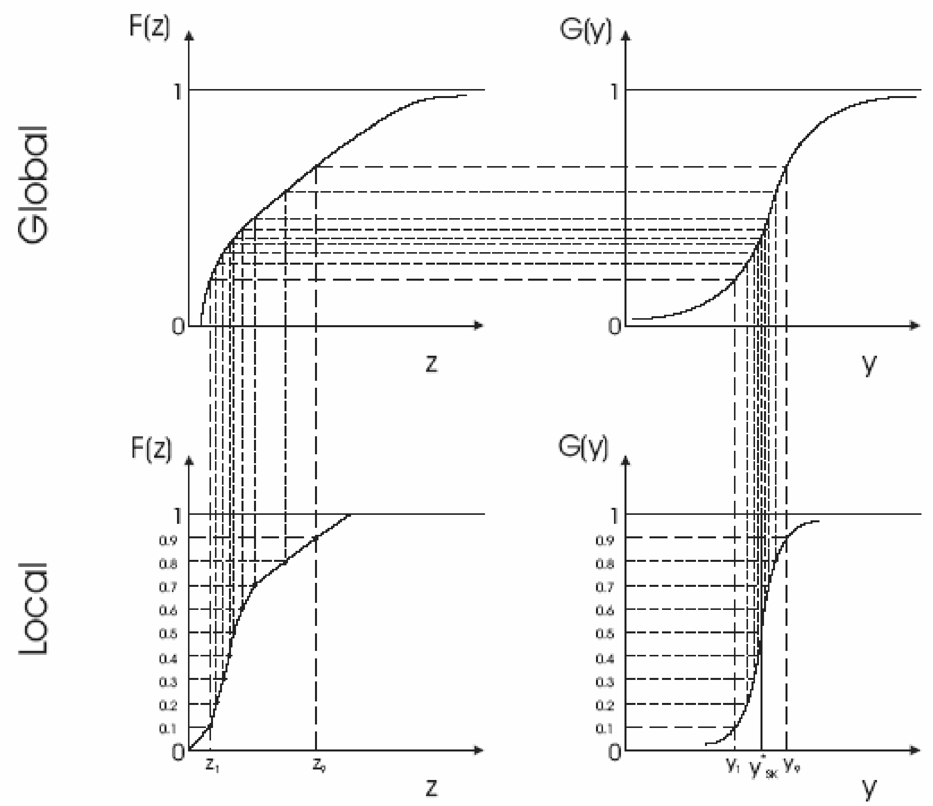

Figure 3: The transformation process used by `PostMG`. Working from the lower right distribution of the local multigaussian (G(y)) distribution, the values are transformed to the global G(y) distribution, then the global F(z) distribution of the original data, then finally the local F(z) values are found for the various quantiles. The local mean and variance are found from these quantiles (Ortiz and Deutsch, 2003).

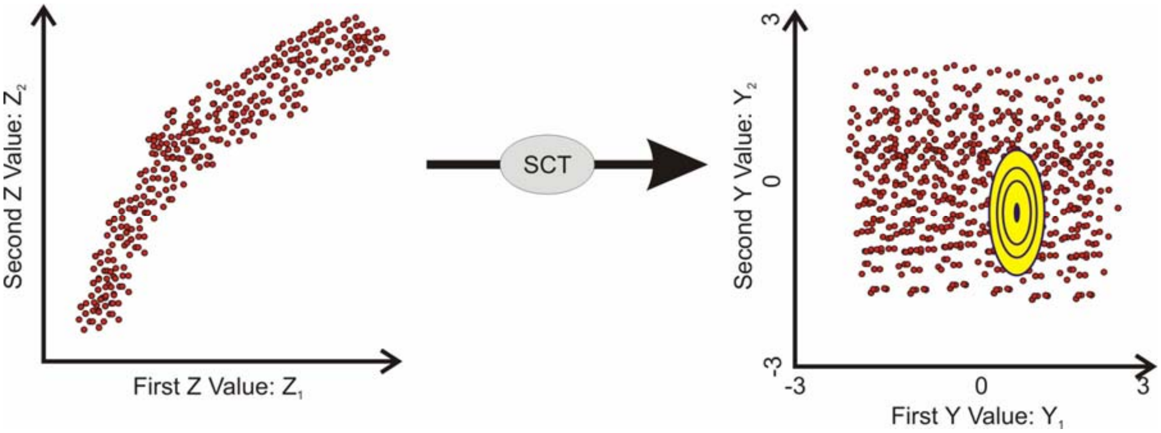

Figure 4: Schematic illustration of a bivariate conditional distribution provided by kriging the stepwise transforms (the yellow ellipses). This distribution must be back transformed to provide a distribution in original Z units.

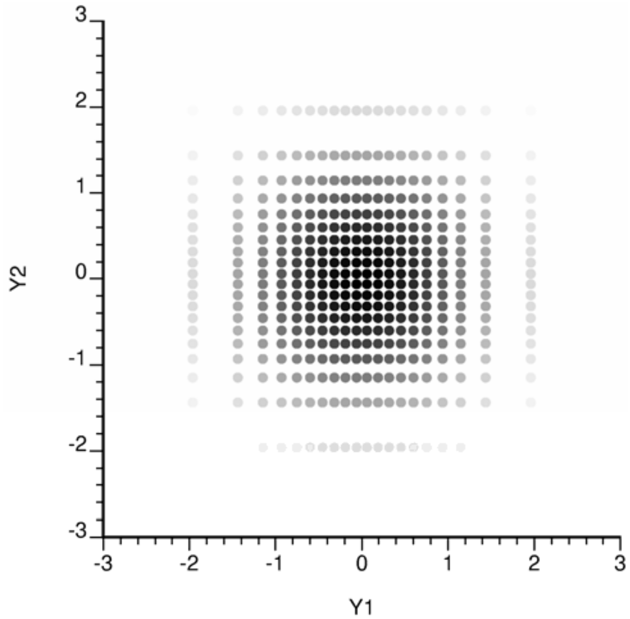

Figure 5: Discretization in standardized units. The color is the weight to the points, that is, the product of the marginal Gaussian frequencies.

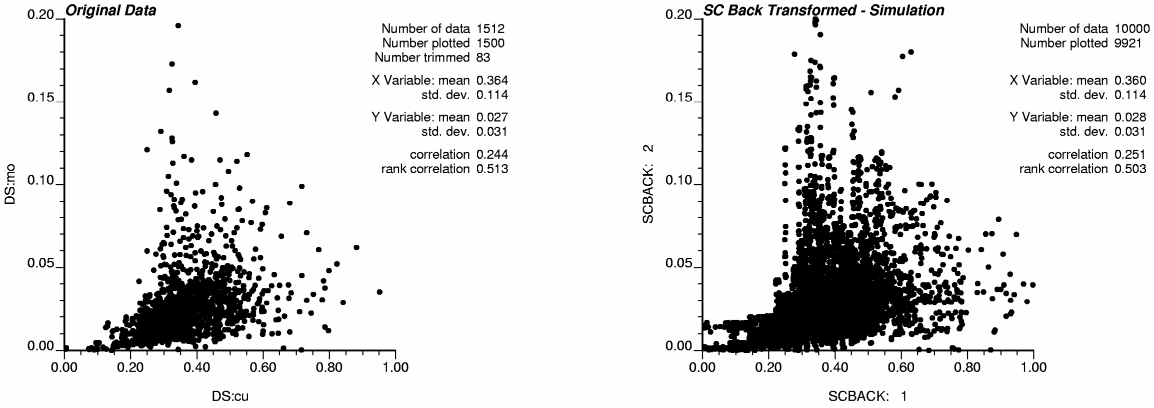

Figure 6: An original data cross plot (left) and the result after simulating and back transforming 10000 values. Note the close reproduction of the statistics.

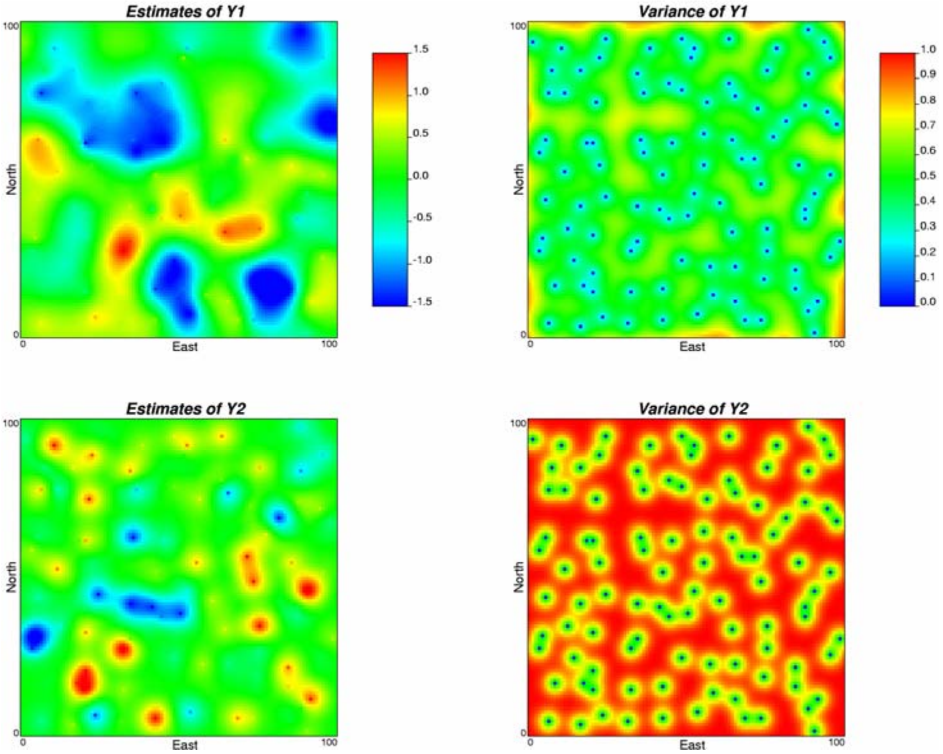

Figure 7: Kriged results using stepwise transformed variables. The top figures are the mean and variance of the first variable. The bottom row is the second variable. The data locations are readily seen as the zero-kriging variance locations.

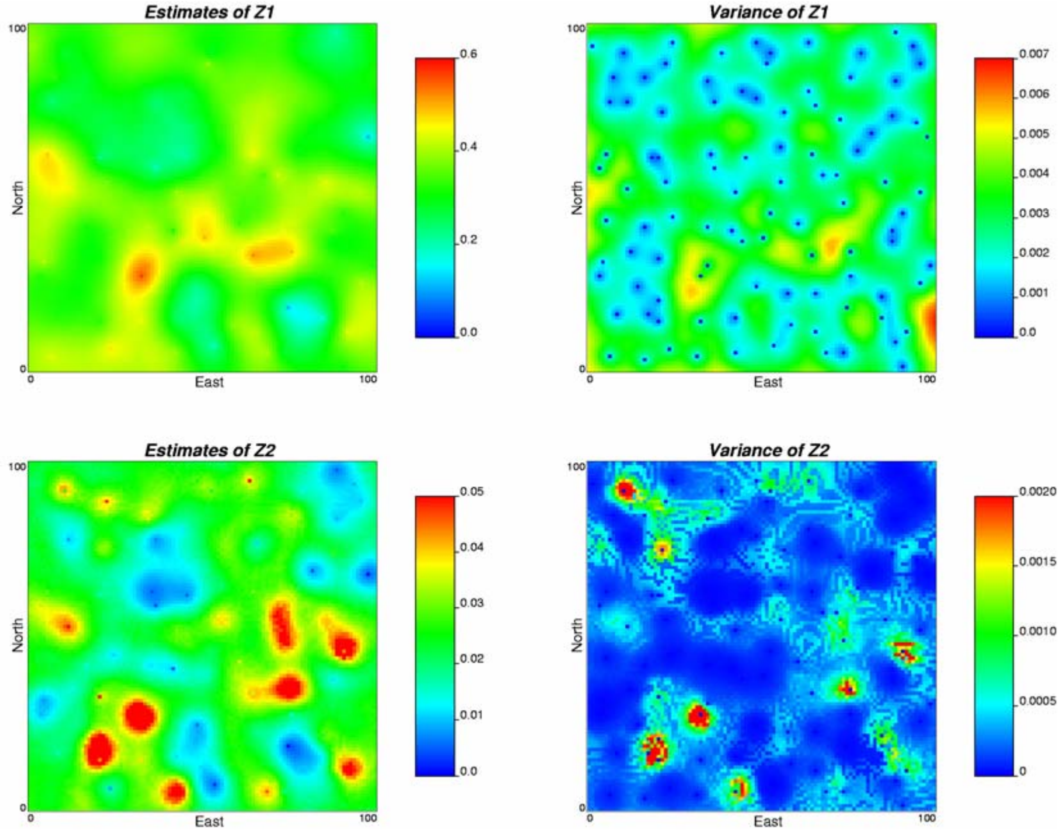

Figure 8: Results from postSCT: means and variances in original units. Note the strong proportional effect in the original variables. These estimates are like kriging – unbiased.

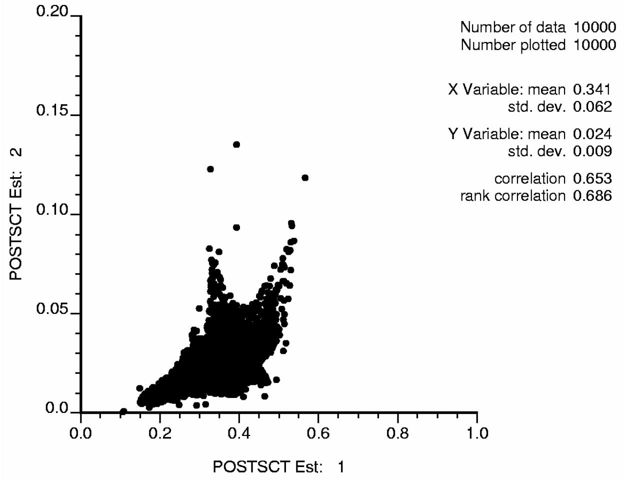

Figure 9: Scatterplot of mean values from postSCT: the results are smooth, but unbiased.