The Proportional Effect: What it is and how do we model it?

John Manchuk

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

One of the outstanding issues in the use of direct simulation is correctly handling the proportional effect. Current implementation of kriging in simulation algorithms assumes the variance is homoscedastic. In most instances this will not be the desired outcome when working with original units directly. A proportional effect must be modeled and incorporated to reproduce variables exhibiting heteroscedastic variance. The proportional effect can be understood from the bivariate form of a distribution. A relationship between the mean and corresponding conditional standard deviation can be derived for such distributions as the Gaussian and lognormal. The relationship between mean and conditional standard deviation is the proportional effect. For Gaussian data, regardless of the mean the conditional standard deviation is constant or homoscedastic. In the lognormal case, the standard deviation is proportionately related to the mean, or heteroscedastic. We would like to model this mean-standard deviation relationship for any distribution, even those with no analytically defined bivariate form. Moving window statistics do not give the proportional effect.

Introduction

Moving to the use of unstructured grids and incorporation of variables with different support volumes is continuously pushing the development of direct kriging and direct simulation algorithms. Implementation of direct algorithms avoids the non-linear normal score transform, which would have major implications when dealing with various block sizes and support volumes. However, the normal score transform implicitly accounts for other aspects of data in original units including heteroscedastic behavior. By working with data in original units and under a direct formalism, one must ensure any heteroscedastic behavior is incorporated correctly and reproduced in results.

Heteroscedasticity of data in original units can be best discussed through use of the lognormal distribution. A relationship between the mean and standard deviation exists for this distribution and is analytically known. This extreme in terms of the proportional effect can be contrasted with the Gaussian distribution that exhibits homoscedastic variance or no proportional effect at all. Actual methods do exist for determining if data exhibit a proportional effect; however, they cannot be incorporated into any direct algorithm as they are qualitative in nature. Two of these methods include moving window statistics and analysis of **h**-scatterplots from variogram calculation.

A third method for measuring the proportional effect can be adapted from an algorithm initially meant for building a lookup table of conditional distributions for use in a direct simulation algorithm, dssim-hr (Oz et al, 2002). This algorithm did reproduce the input histogram, as was its intention, but the resulting variance remained homoscedastic even for data with a proportional effect. By correcting some numerical issues with the lookup table construction algorithm

including cumulative distribution function (CDF) interpolation and extrapolation, it can be used to model the proportional effect for any set of data.

Background

In current literature, the proportional effect is derived from effects of non-stationary spatial data on the variogram and observed through the use of moving window statistics. An explanation of the proportional effect as differences in variability in high average valued areas to low average valued areas is given by Chiles and Delfiner (1999). They suggest that local variograms should be calculated and if all have similar structure and differ by a multiplicative factor, the variograms indicate a proportional effect.

Journel and Huijbregts (1978) derive the proportional effect by examining the variogram within two different neighborhoods from the same data set. This method is reiterated by Cressie (1991) as well. In this way, the variogram depends not only on a lag distance *h*, but also on the location x_0. Given two neighborhoods centered on **u** and **u'**, the dependence of the variogram on location can be removed by dividing out a function of the experimental mean at each location, $f[m^*(\mathbf{u})]$ and $f[m^*(\mathbf{u'})]$. The variograms at **u** and **u'** then differ by a proportional effect defined by the function of the experimental mean (Jounel and Huijbregts, 1978). This function is derived by calculating variograms from various different neighborhoods and examining differences. Considering a lag distance of zero, the variogram is an expression of the variance and the process is identical to moving window analysis.

The proportional effect is explained similarly by Isaaks and Srivastava (1989). Various possible mean profiles are shown for 1-dimensional data along with the actual data such that the variance can also be inferred. Regardless of the fitting approach, be it smoothing splines or some form of kernel, the comparison is equivalent to moving windows in one dimension. Moving window statistics were also applied across two dimensional data to compare the mean and standard deviation with reference to the correlation coefficient. A definition of the proportional effect is also given by Armstrong (1998) as a relationship between the variogram and the square of the local mean grade.

Proportional Effect for Lognormal Distribution

One of the inherent features of lognormally distributed data is the proportional effect and it is analytically defined. If the natural logarithm of a set of data is normally distributed, then it is lognormally distributed in its original units. Equations 1 to 3 describe this relationship.

$$\ln\left(Z \cong LogN(m,\sigma)\right) = X \cong N(\alpha,\beta)$$

$$\beta^2 = \ln\left(1+\frac{m^2}{\sigma^2}\right) \qquad \text{(1 to 3)}$$

$$\alpha = \ln(m) - \frac{\beta^2}{2}$$

where m and σ are the mean and standard deviation of Z, which is lognormally distributed, and α and β are the mean and standard deviation of X, the natural logarithm of Z. This relationship is very convenient for kriging of lognormal data. By kriging a location **u** with nearby X as conditioning we would acquire an estimate and estimation variance, which could be converted to

the correct lognormal mean and variance by rearranging Equations 2 and 3. Rearranging Equation 2 yields the proportional effect of lognormal data where σ is a function of *m* and β.

$$\sigma^2 = m^2\left(e^{\beta^2} - 1\right) \tag{4}$$

Unlike the relationship of the variogram to the proportional effect discussed in the previous section, Equation 4 relates the proportionate variance to the mean and a stationary variance. This implies that there is more to the variogram relationship developed in the literature:

$$\gamma(h,m) = f(m)\cdot\gamma(h) \tag{5}$$

where γ (*h*,*m*) is the variogram at lag *h* and mean *m*, *f*(*m*) is a function of the mean modeled by examining various local variograms, and γ(*h*) is a common or stationary variogram model. Equation 5 can be paralleled with the variogram structure for lognormally distributed data given by Equation 6, which is determined by replacing the variance with $\sigma^2 - \gamma(h)$ and dividing through by σ^2.

$$\gamma_Z(h) = 1 - \frac{m^2}{\sigma^2}\left[e^{\beta^2\cdot(1-\gamma_Y(h))} - 1\right] \tag{6}$$

Here, γ(*h*,*m*) is equivalent to $\gamma_z(h)$ and γ(*h*) to $\gamma_Y(h)$. As in Equation 5, $\gamma_Y(h)$ is the stationary variogram model because it is calculated over normal scored data. By making this comparison, a more general from of Equation 5 should be sought after. It could be derived by first modeling the proportional effect directly in the form of Equation 4.

$$\gamma(h,m) = f(m)\cdot g\left(\gamma_S(h)\right) \tag{7}$$

γ(*h*,*m*) is now related to a function of the mean as well as a function, *g(·)* of the stationary variogram $\gamma_S(h)$.

The Proportional Effect in a Qualitative Sense

Two methods for determining if the proportional effect exists in a data set will be discussed: (1) moving window statistics; and (2) analysis of **h**-scatterplots. Both provide an indication of any heteroscedastic behavior of data in original units, but neither can be utilized in a direct simulation algorithm for reasons to be identified in the following sections.

Moving Window Statistics

Practically any statistic of interest can be calculated from a region containing known data. We are interested in the mean and standard deviation from moving window statistics to extract any relationship as an indication of the proportional effect. The method of moving windows was applied to a few well known data sets to show any heteroscedasticity including Walker Lake (Isaaks and Srivastava, 1989) and the Spatial Interpolation Contest rainfall data (Dubois, 1998), see Figure 1. Moving windows were also applied to the lognormal equivalent of an unconditional Gaussian simulation for reference.

Results indeed show heteroscedastic behavior in the lognormal data. Visually inspecting the plot for SIC rainfall may indicate no proportional effect because of the variability, although the correlation of 0.626 may lead someone to believe a proportional effect does exist. In any case,

however, it is likely that there exists a set of data that are normally distributed and oriented in such a manner that moving windows shows a proportional effect. In parallel, there may be a configuration of lognormal data that moving windows indicates is homoscedastic. Moving window statistics are not a reliable measure of the proportional effect, but may give an indication of its existence for a set of data.

For the case where moving windows define a very obvious relationship between the mean and standard deviation, the problem of incorporating it into a kriging algorithm remains. The relationship is a global measure. There is no way to extend its use to calculate a proportionate variance for a kriged estimate.

h-Scatterplots

If we know the bivariate form of a distribution, we can directly access the proportional effect since the conditional mean and standard deviation are known. A numerical account of a distribution's bivariate form can be drawn from an **h**-scatterplot. For a particular plot, conditional mean and variance values can be calculated by binning the data at (**u**+**h**) using cutoffs along the data at **u**. This was carried out for an exhaustive Gaussian and lognormal data which were unconditionally generated, see Figure 2. Dashed lines on each plot indicate the actual analytical proportional effect.

An advantage over moving window statistics of this method is that the associated homoscedastic variance can be determined for a particular lag **h**. Using the lognormal case as an example, if we determine the normal equivalent of each z(**u**), z(**u**+**h**) pair and perform the same binning exercise, the resulting variance will be approximately homoscedastic and equal to β^2. Two notable disadvantages exist for this method: (1) we never have access to an exhaustive version of the variable of interest; and (2) actual data sets are either too small or irregularly spaced to get reliable results, especially for a specific lag distance. Using **h**-scatterplots from irregularly spaced variogram calculation would only degrade results as the lag distance is not a discrete value.

The Proportional Effect in a Quantitative Sense

Both moving window statistics and analysis of **h**-scatterplots cannot be used to effectively model the proportional effect. The result of moving windows is not quantitative and there are just not enough data to apply the **h**-scatterplot method. A method must be devised to extract a proportional effect model directly from the distribution of a set of data. Much the same way as conditional variances can be calculated analytically for Gaussian and lognormal distributions, we can numerically extract the approximate proportional effect relationship for any distribution.

Construction of a model for the proportional effect of any distribution is primarily based on the analytical relationship from the lognormal distribution. What is sought after is a relationship linking a mean in original units and homoscedastic variance with the proportionate variance as in Equation 4.

$$\sigma_Z^2 = f\left(m_Z, \sigma_Y^2\right) \tag{8}$$

where σ_Z^2 is the proportionate variance, m_Z is a mean in original units and σ_Y^2 is a stationary variance. The mean and stationary variance may come from kriging a set of data directly using a standardized variogram. Originally, it was thought that the proportional effect was only a

relationship between the mean and variance; however, Equation 8 makes sense because the stationary variance is analogous to a correlation for a bivariate normal distribution. Intuitively, the mean-variance relationship should depend on σ_Y^2.

To define Equation 8 for any distribution, a grid of mean and variance values are calculated from a set of conditional distributions, which are created using methods explained in the Appendix. The procedure is as follows:

1. For an adequate set of non-standard normal mean and variance values
    a. Build the conditional distribution in original units, see the Appendix.
    b. Interpolate a set of quantiles to calculate its mean, m_Z, and variance, σ_Z^2
2. Model the conditional variance surface with the non-standard normal variance and original-unit mean as independent variables.

Once the surface is fit, it can be used in a direct kriging or simulation framework to determine the heteroscedastic variance and aid in construction of conditional distributions. An additional fit that would be useful during estimation or simulation is the relationship between the non-standard mean m_Y and both m_Z and σ_Y^2. This surface could be used to determine the correct non-standard normal distribution required to build the associated condition distribution.

Example using Lognormal Data

As a preliminary test of the modeling procedure, a set of 50 data points that are analytically lognormal were created for the input distribution. By doing this, results can be compared to the exact analytical relationship. The input data have a mean and variance of 20 and 400 respectively. Minimum and maximum values for tail extrapolation were chosen at 0 and 250. Cardinal splines and Bezier curves achieved a very good fit of the distribution when compared to the actual lognormal distribution for these data, see Figure 3. There are some discrepancies in the tails; however, the method is not specific for lognormal distributions and in most cases we do not know the tails at all.

To see how well the analytical proportional effect of lognormal data was reproduced, 500 conditional distributions were modeled and 1000 quantiles were interpolated for mean and variance calculations. Proportional effect contours were plotted against the analytical equation, see Figure 4. Each contour is defined by a stationary variance and a set of 50 means. Note that mean-variance pairs that exist with less than 1 % probability were ignored. The plot shows that the numerically calculated contours compare well with analytical ones aside from two notable differences: (1) when the normal variance is extremely low (first three contours) and the mean is less than the global mean, there are numerical precision problems; and (2) when the mean is high and the variance is high, sample contours depart from analytical contours. This can be explained by using a fixed maximum for the upper tail. Variance will increase until conditional distributions reach near the maximum, at which point the region conditional distributions cover on the global CDF for given quantiles decreases along with the variance, see Figure 5.

Conclusion

The proportional effect can be numerically modeled for any distribution. It will be in the form of a function dependent on a mean and variance. Results are sensitive to methods used for modeling the global distribution of a set of data as well as corresponding conditional distributions, especially regarding assumptions for lower and upper tails. Improper lower or upper tail assumptions can lead to very different proportional effect models. Having a basis for modeling a proportional effect will allow extension of the theory and algorithm to application in direct kriging and simulation.

References

Armstrong M., *Basic Linear Geostatistics*, Springer-Verlag Berlin Heidelberg, 1998

Boyd S. and Vandenberghe L., *Convex Optimization*, Cambridge University Press, 2004

Chiles J.P. and Delfiner P., *Geostatistics Modeling Spatial Uncertainty*, John Wiley and Sons, Inc., 1999

Cressie N., *Statistics for Spatial Data*, John Wiley & Sons, Inc., 1991

Dubois, G., 1998. Spatial Interpolation Comparison 97: Foreword and Introduction: Journal of Geographic Information and Decision Analysis, v. 2, no. 2, pp. 1-11.

Hearn D. and Baker M.P., *Computer Graphics with OpenGL*, Pearson Prentice Hall, Third Edition, 2004

Isaaks E.H. and Srivastava R.M., *An Introduction to Applied Geostatistics*, Oxford University Press, 1989

Journel A.G. and Huijbregts CH.J., *Mining Geostatistics*, Academic Press Inc. 1978

Manchuk J., Leuangthong O., and Deutsch C.V., *A New Approach to Direct Sequential Simulation that Accounts for the Proportional Effect: Direct Lognormal Simulation*, Center for Computational Geostatistics, Report Six, 2004

Mortenson M.E., *Mathematics for Computer Graphics Applications*, Industrial Press Inc., Second Edition, 1999

Oz B., Deutsch C.V., Tran T.T., and Xie Y., *DSSIM-HR: A FORTRAN 90 Program for Direct Sequential Simulation with Histogram Reproduction*, Center for Computational Geostatistics, Report Four, 2002

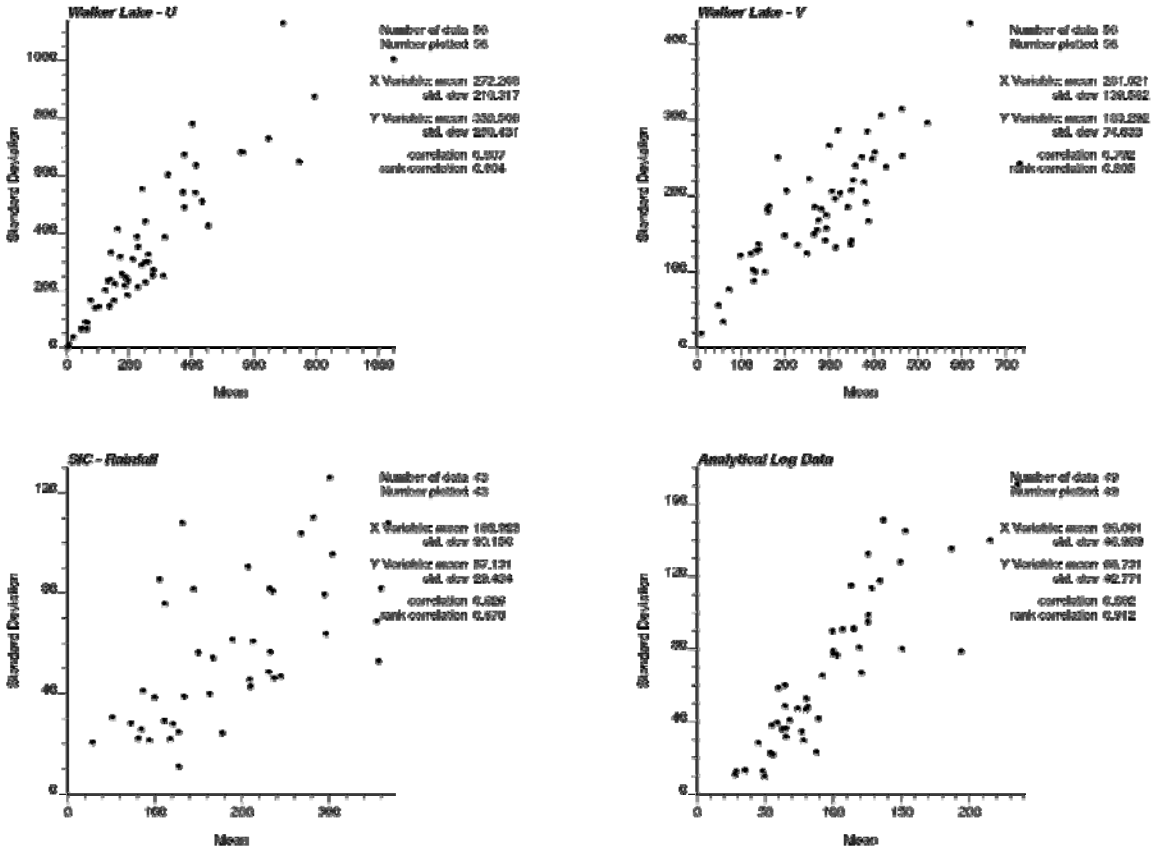

Figure 1: Moving window statistics for Walker Lake U and V variables, SIC rainfall data, and lognormally distributed data.

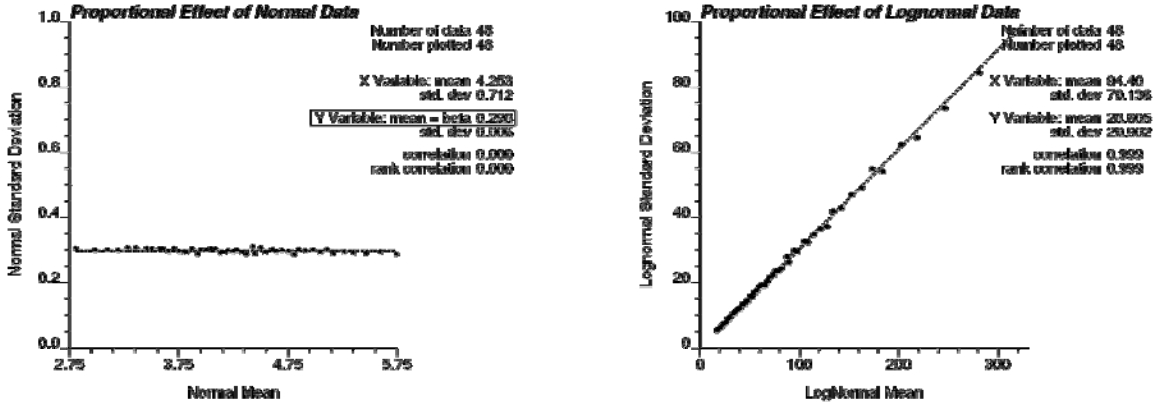

Figure 2: Analysis of the proportional effect of Gaussian and lognormal data using h-scatterplots. Dashed lines indicate the actual analytical proportional effect.

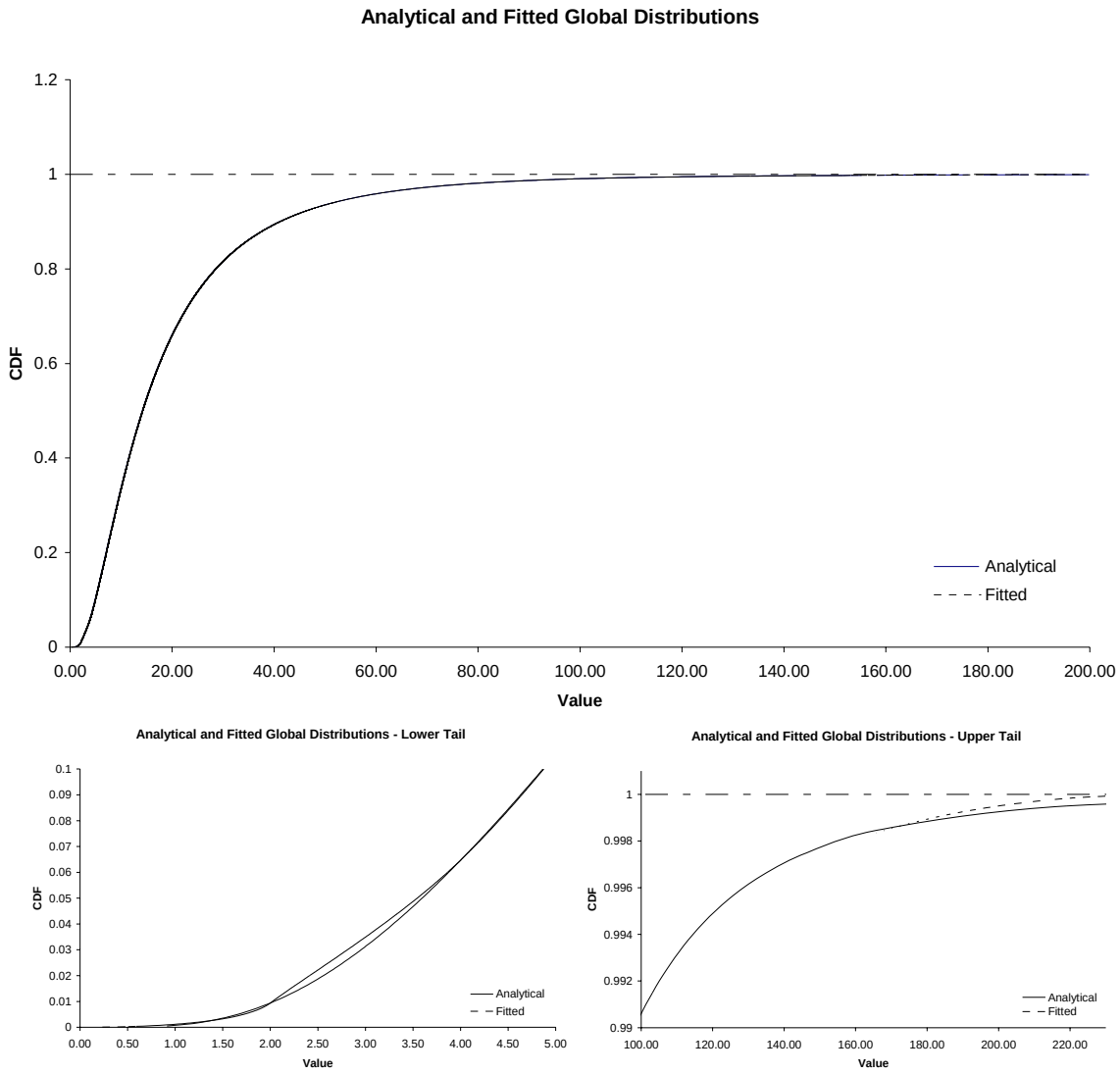

Figure 3: Analytical lognormal distribution and corresponding fit using cardinal splines and Bezier curves.

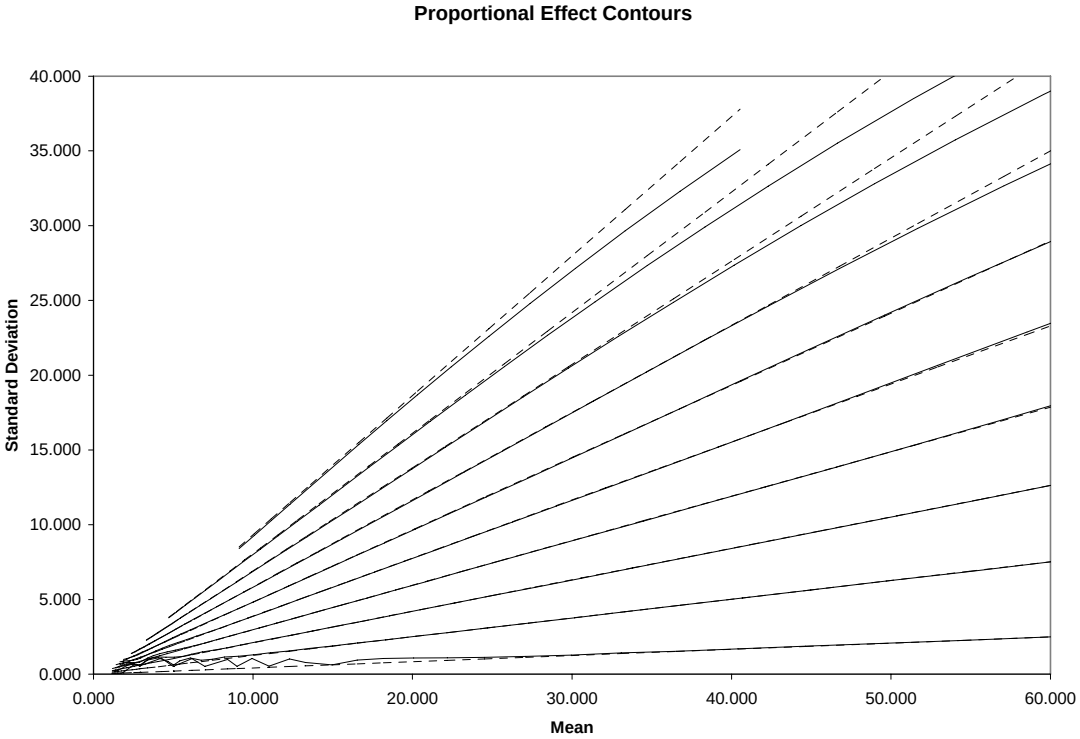

Figure 4: Proportional effect contours calculated numerically (solid lines) and analytically (dashed lines). Contour represents constant normal variances, which range from 0.05 (lowest line) to 0.95 (top line) in increments of 0.1.

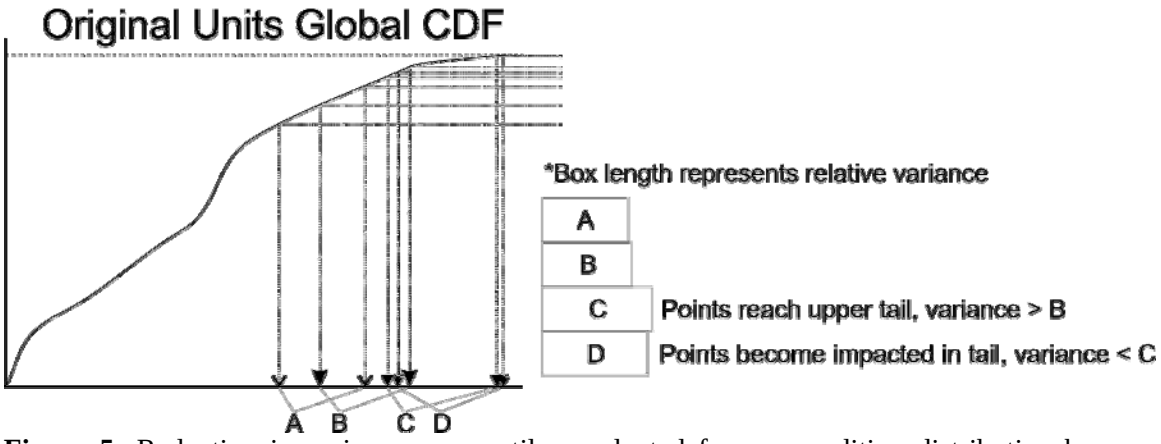

Figure 5: Reduction in variance as quantiles evaluated from a condition distribution become impacted in the upper tail of the global CDF.

Appendix A – Cumulative Distribution Function Modeling

An important aspect of modeling the proportional effect is modeling cumulative distribution functions since values will be interpolated and extrapolated from global and conditional distributions in original units. These distributions must be fit by some non-parametric means. A method already exists and was initially designed for a direct simulation algorithm to ensure histogram reproduction. Some deficiencies were noted with this method, but after making alterations it can be used as a component to modeling the proportional effect.

The CCDF modeling method as employed in dssim-hr implemented a data transformation from non-standard normal distributions to original units for a set of quantiles. Combining the quantiles, *q*, and resulting original unit values, *z*, from the transform provided a set of data describing a conditional original-units distribution, see Figure A-1. Some deficiencies of this method are a result of the interpolation and extrapolation scheme used to obtain the conditional distributions. All (*z*,*q*) pairs are acquired from interpolated or extrapolated points along the global distribution. Inaccurate results can be produced from this procedure since the relationship between quantiles for the conditional distributions and those obtained along the global CDF form a non-linear relationship. This relationship is in fact known:

$$q_{global} = G\left(G^{-1}(q)\cdot\beta+\alpha\right) \qquad \text{A-1}$$

Where q_{global} is the quantile along the global CDF, q is the quantile for the CCDF, and α and β are the mean and standard deviation of a non-standard normal distribution.

Implications of Equation A-1 become noticeable especially when a particular α and β cause poor characterization of either the lower or upper tails of the global original-units CDF. Poorly informed distribution tails is yet another problem to be discussed. Figure A-1 below shows the difference between values interpolated linearly along the global CDF using the above method to those interpolated along the CCDF built only from known data. This difference will cause problems regarding implementation in a simulation algorithm; however, it will not drastically affect results in modeling a proportional effect. Calculated mean and variance values of a CCDF will be similar with either interpolation scheme. To reduce error, interpolation should be carried out on the set formed by known original-units values and quantiles defined by the inverse of Equation A-1.

$$q = G\left(\frac{G^{-1}\left(q_{global}\right)-\alpha}{\beta}\right) \qquad \text{A-2}$$

When α is low or high, one of the tails of the CCDF will be poorly informed especially when combined with a low β. When α is high, the CCDF relies heavily on the extrapolation method used to model the upper tail of the global distribution. The reverse is true for the lower tail. Poor or incorrect choices for lower or upper tail shapes may drastically alter results regarding modeling a proportional effect.

A new scheme has been developed to aid in the modeling of conditional distributions under the same framework described above. Given a particular non-standard normal distribution, original-unit quantiles are transformed to their respective conditional values. Interpolation is then carried out using those points. However, if conditional quantiles when transformed to global exist in the

tails of the CDF, these points are to be extrapolated from the tails of the global distribution. For high values of α, the last known value from the set z may have a low conditional quantile relative to its global value. Modeling a tail under these circumstances would likely cause more error.

Interpolation and extrapolation methods for this new scheme have been updated as well. A class of splines referred to as cardinal splines (Hearn and Baker, 2004) can be used to model a cumulative distribution function. One requirement for any CDF is positive definiteness, which cardinal splines can work with. Extrapolation is accomplished using another class of curves called Bezier Curves (Mortenson, 1999). Cardinal splines and Bezier curves can be combined in a piece-wise fashion to define a CDF. One disadvantage to using these curves, however, is they are not invertible. Numerical optimization techniques such as line search algorithms are required to interpolate points (Boyd and Vandenberghe, 2004). If the same conditional distribution construction scheme used to create Figure A-2 is made using the spline fitting technique, results are more acceptable, see Figure A-3.

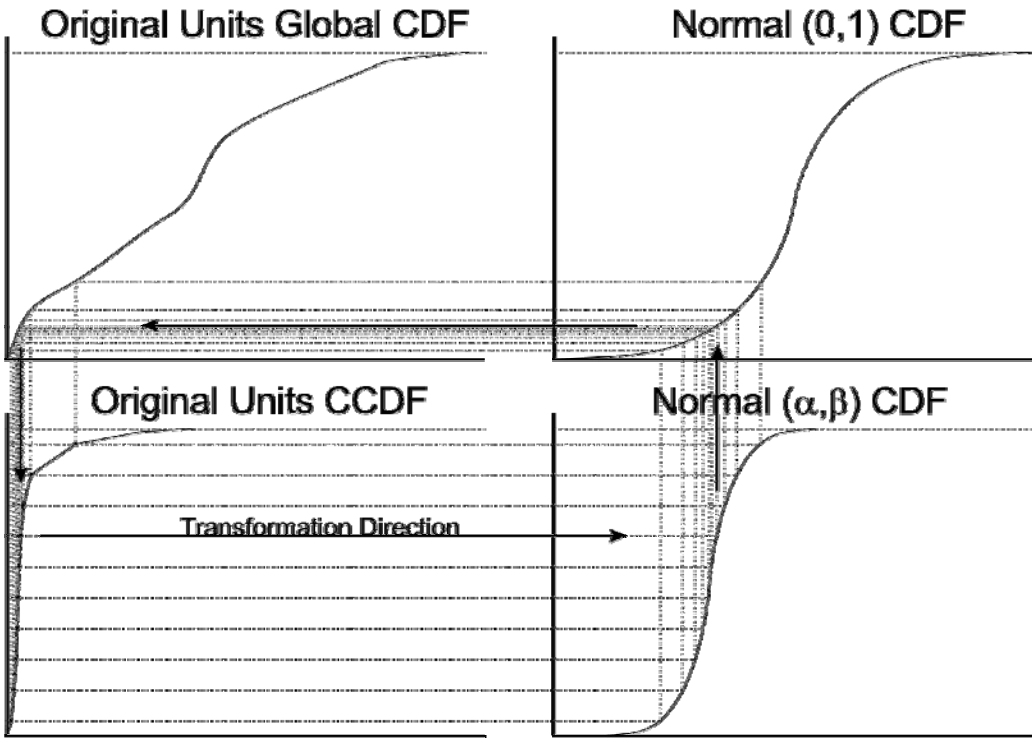

Figure A-1: Transformation method to build conditional distributions in original units.

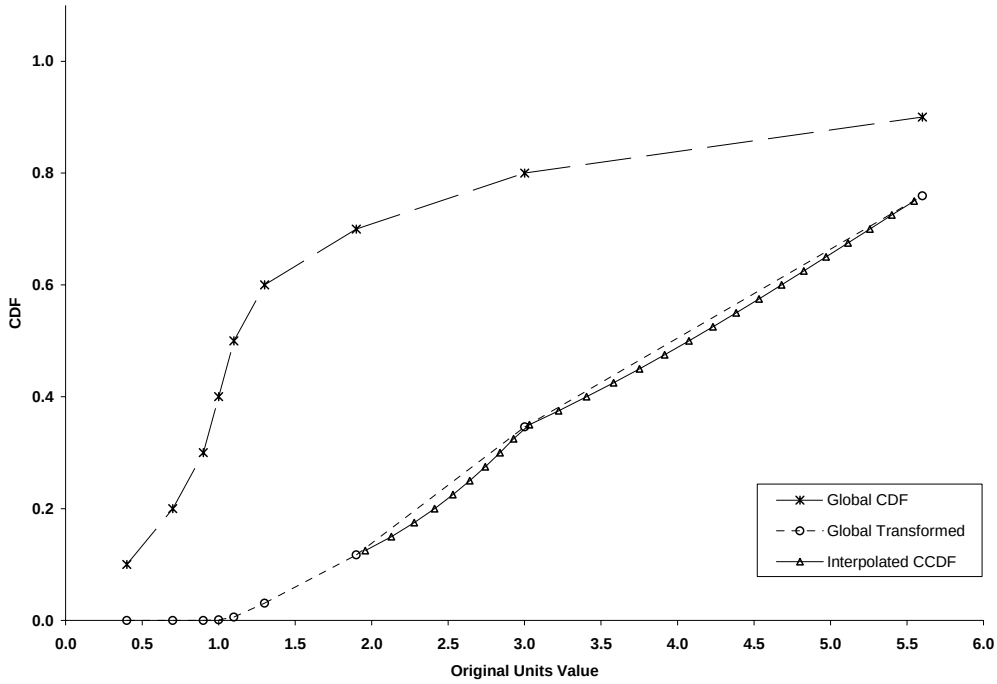

Figure A-2: Global and conditional distribution functions using linear interpolation.

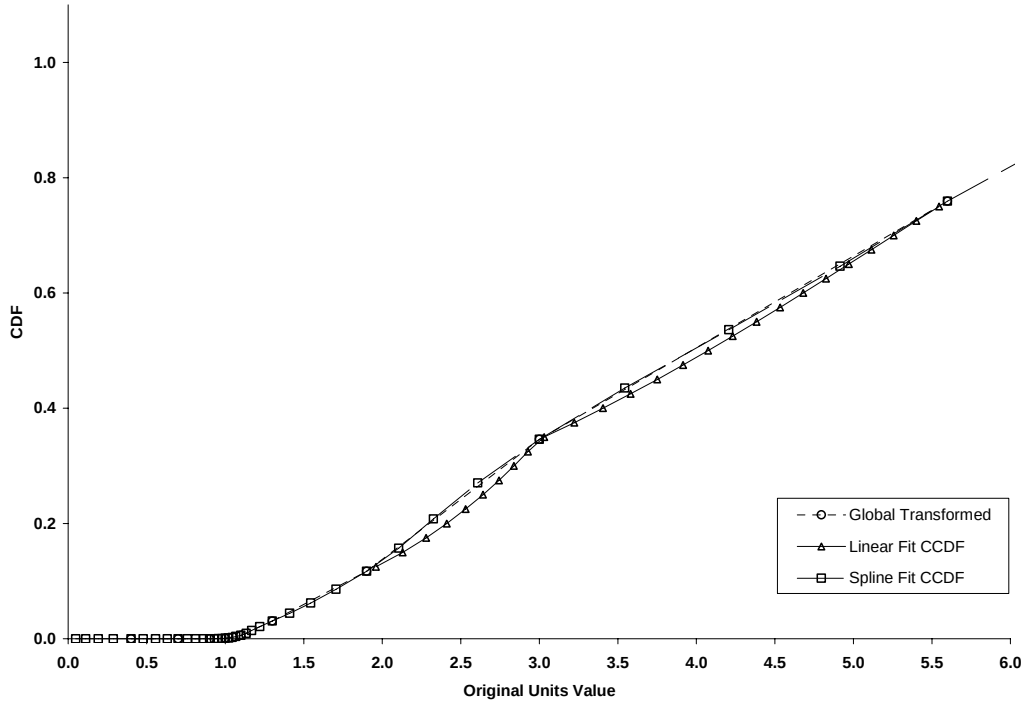

Figure A-3: Conditional distribution built using points linearly interpolated from the global distribution and points interpolated from the global using cardinal splines.