# The Promises and Pitfalls of Direct Simulation

Oy Leuangthong

Department of Civil & Environmental Engineering,
University of Alberta

**Abstract**

*The idea of direct simulation is to simulate in the space of the original data units, with minimal assumptions or transformations about the data distribution. A common approach to direct simulation is to proceed in a sequential fashion: direct sequential simulation (DSS). While the idea is not new, full development of the framework remains to be seen. The benefits of multiscale data integration, avoidance of the "Gaussian disease", and flexible distribution considerations are offset by problems with histogram reproduction, the pervasive influence of Gaussianity, and proportional effect reproduction. This paper examines the promises and pitfalls of direct simulation with some illustrative examples, and also discusses the future of DSS as a practical alternative for natural resource characterization. The future of DSS requires a procedure to account for the dependency between the local variance and mean.*

## Introduction

Over the last decade, direct simulation has been proposed as a viable alternative to the venerable Gaussian simulation approaches. The idea of direct simulation is to simulate in the space of the original data units, with minimal assumptions or transformations about the data distribution. Behind this key idea is the principle of simple kriging. Journel (1994) first showed that the covariance of simulated values reproduces the target covariance model *if* the simulated values are drawn from a distribution centred about the simple kriging (SK) mean and a variance given by the SK variance. Indeed, Bourgault (1997) showed this to be true for different distributional shapes including the uniform, dipole and of course, the Gaussian distribution. Caers (2000) also shows this for a uniform, double exponential, double exponential with a spike, and a "bootstrapped" distribution.

Covariance reproduction without relying on the Gaussian framework seeded the idea for direct simulation. The key premise for why direct simulation works lies in the orthogonality between the SK estimate, $Z^*(\mathbf{u})$, and the squared error which forms the basis for the SK error variance, $\sigma^2_{SK}(\mathbf{u})$. This can be thought of in terms of projections where the squared error, $[Z(u)-Z^*(u)]^2$, is orthogonal to the space of all finite linear combinations of the random variables (RV), $Z(\mathbf{u}_\alpha)$, $\alpha = 1,..., n$ (Journel and Huijbregts, 1978) (see Figure 1). The kriging estimate, $Z^*(\mathbf{u})$, lies in this space as it is a linear combination of the RVs, $Z(\mathbf{u}_\alpha)$, $\alpha=1,..., n$:

$$Z^*(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{u}_\alpha) \tag{1}$$

The squared error term, $[Z(\mathbf{u})-Z^*(\mathbf{u})]^2$, represents the distance to the unknown true value, $Z(\mathbf{u})$. Based on Projection Theory, there is a unique and exact solution that yields the linear coefficients, $\lambda_\alpha$, $\alpha=1, ..., n$, such that this distance is minimized (Journel and Huijbregts, 1978).

This solution is referred to as the projection of $Z(\mathbf{u})$ onto this space. The corollary to kriging lies in the fact that the weights, $\lambda_\alpha$, $\alpha=1, \ldots, n$, are determined such that the expected squared error, $E\{[Z(\mathbf{u})-Z^*(\mathbf{u})]^2\}$, is minimum. This visual interpretation of simple kriging can also be thought of as satisfying the Generalized Theorem of Pythagoras (Anton and Rorres, 1991).

Orthogonality of the kriged estimate and the squared error leads to an error variance that is independent of the data, commonly referred to as the homoscedascity of kriging. Under a Gaussian paradigm, this poses no problems; in fact, it would be exactly right. In reality, natural phenomena rarely possess characteristics similar to the Gaussian distribution. This is particularly evident upon examining the relationship between the local average and the local variability, which, contrary to the homoscedasticity inherent in kriging, often reveals the presence of a strong relationship between the two statistics. This relationship is referred to as heteroscedasticity, more specifically it is the proportional effect (Journel and Huijbregts, 1978). This poses the most significant problem for direct simulation.

This paper presents the promises and pitfalls of direct simulation with some illustrative examples. Five main areas of discussion are highlighted: (1) principle of simple kriging, (2) implementation of direct simulation, (3) multiscale data integration, (4) histogram reproduction, and (5) accounting for the proportional effect. Finally, the future of DSS is discussed.

## The Simple Kriging Principle

Reproduction of the covariance only requires that the conditional probability distributions have a mean and variance given by simple kriging (Journel, 1994). Journel proved this by showing firstly, the detailed simulation of a variable at location $\mathbf{u}$, then adding this simulated value to simulate the next location, $\mathbf{u}'$, and finally checking the covariance between these two simulated variables.

Consider a stationary random variable, $Z(\mathbf{u})$, with zero mean and unit variance. Firstly, construct a simulated value such that it can be decomposed as

$$Z_s(\mathbf{u}) = m(\mathbf{u}) + R_s(\mathbf{u})$$

where $m(\mathbf{u})$ is the expected value at location $\mathbf{u} \in$ domain, $\mathbb{A}$, and $R(\mathbf{u})$ is a random variable drawn from a distribution with zero mean and variance, $\sigma^2(\mathbf{u})$. The local mean is given by kriging mean (Equation 1), and the variance is given by the SK variance:

$$\sigma^2_{SK}(\mathbf{u}) = 1 - \sum_{\alpha=1}^{N} \lambda_\alpha C(\mathbf{u} - \mathbf{u}_\alpha) \tag{2}$$

where $C(\mathbf{u} - \mathbf{u}_\alpha)$ is the covariance between the location $\mathbf{u}$ and the data located at $\mathbf{u}_\alpha$, $\alpha=1, \ldots, n$, $\sigma^2_{SK}(\mathbf{u})$ is the simple kriging variance, and the weights, $\lambda_\alpha$, $\alpha=1, \ldots, n$ are obtained by solving the normal equations:

$$\sum_{\beta=1}^{n} \lambda_\alpha C(\mathbf{u}_\beta - \mathbf{u}_\alpha) = C(\mathbf{u} - \mathbf{u}_\alpha), \quad \alpha = 1, \ldots, n \tag{3}$$

This simulated value is added to the conditioning data set, and simulation is performed at the next location $\mathbf{u}' = \mathbf{u}_{n+1}$ with the following kriged mean and variance:

$$Z^*(\mathbf{u}') = \sum_{\alpha=1}^{n} \lambda_\alpha z(\mathbf{u}_\alpha) + \lambda_{n+1} Z_s(\mathbf{u}) \tag{4}$$

$$\sigma^2_{SK}(\mathbf{u}') = 1 - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{u}' - \mathbf{u}_\alpha) - \lambda_{n+1} C(\mathbf{u}' - \mathbf{u}) \tag{5}$$

Note that the weights $\lambda_\alpha$, $\alpha=1,\ldots,n+1$ are *not* the same as the weights $\lambda_\alpha$, $\alpha=1,\ldots,n$ obtained from solving the system in Equation 3. The simulated value is given as

$$Z_s(\mathbf{u}') = Z^*(\mathbf{u}') + R_s(\mathbf{u}')$$

The covariance between the two simulated variables is then examined:

$$\begin{aligned} C(\mathbf{u} - \mathbf{u}') &= E\{Z_S(\mathbf{u}) \cdot Z_S(\mathbf{u}')\} \\ &= E\{Z^*(\mathbf{u}) \cdot Z^*(\mathbf{u}')\} + E\{Z^*(\mathbf{u}) \cdot R_S(\mathbf{u}')\} + \\ &\qquad E\{Z^*(\mathbf{u}') \cdot R_S(\mathbf{u})\} + E\{R_S(\mathbf{u}) \cdot R_S(\mathbf{u}')\} \end{aligned} \tag{6}$$

where $E\{Z^*(\mathbf{u}) \cdot R_S(\mathbf{u}')\}$ and $E\{R_S(\mathbf{u}) \cdot R_S(\mathbf{u}')\}$ are zero since $Z^*(\mathbf{u})$ and $R_S(\mathbf{u}')$ are independent of each other and $R_S(\mathbf{u})$ and $R_S(\mathbf{u}')$ are also independent. The remaining portions of the right hand side are non zero since the kriged mean at the second location depends on the mean and randomly drawn value at the first location.

Expanding and simplifying the remaining two terms yields

$$E\{Z^*(\mathbf{u}') \cdot Z^*(\mathbf{u})\} = \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1}\left[1 - \sigma^2_{SK}(\mathbf{u})\right] \tag{7}$$

$$E\{Z^*(\mathbf{u}') \cdot R_S(\mathbf{u})\} = \lambda_{n+1} \sigma^2_{SK}(\mathbf{u}) \tag{8}$$

Equations 7 and 8 are substituted into Equation 6:

$$\begin{aligned} C(\mathbf{u} - \mathbf{u}') &= \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1}\left[1 - \sigma^2_{SK}(\mathbf{u})\right] + \lambda_{n+1}\sigma^2_{SK}(\mathbf{u}) \\ &= \sum_{\beta=1}^{n} \lambda_\beta C(\mathbf{u}_\beta - \mathbf{u}) + \lambda_{n+1} \\ &= C(\mathbf{u}' - \mathbf{u}) \end{aligned}$$

It is by this logic that Journel (1994) proved that so long as the conditional mean and variance are provided by simple kriging, covariance reproduction could be achieved without making any assumptions about the distributional shape. This is an exciting result as it opened the way for geostatisticians to consider simulation outside of the Gaussian framework without the inference effort required under the indicator paradigm.

**DSS Methodology**

A common approach to simulation is to proceed in a sequential fashion; thus, Direct Sequential Simulation (DSS) was coined (Xu and Journel, 1994). The sequential simulation framework is straightforward:

1. Select a random path visiting all locations.
2. At each location:
   a. Search for all nearby data of different types and/or scale and previously simulated nodes (e.g. $P$ data types with $n_p$ samples).
   b. Perform simple kriging to determine the parameters corresponding to the conditional distribution, $F(Z(\mathbf{u})\big|Z_p(\mathbf{u}_1),\ldots,Z_p(\mathbf{u}_{n_p}))$, $p=1, \ldots, P$.
   c. Draw a simulated value from this conditional distribution using Monte Carlo simulation. This simulated value is added to the conditioning data set.
3. Proceed to next node and repeat Step 2, until all locations are simulated.

The virtues of simplicity cannot be understated. The sequential algorithm was proposed by Johnson (1987), and is common in most geostatistical literature (Isaaks, 1990; Goovaerts, 1997; Deutsch and Journel, 1998; Chiles and Delfiner, 2002; Sinclair and Blackwell, 2000). There are other approaches for simulation, including the matrix approach (Davis, 1987) and turning bands (Journel and Huijbregts, 1978); however, the simplicity and efficiency of sequential simulation has made it the most popular approach in practice.

Indicator and Gaussian simulation have long been the "standard" geostatistical methods of choice in modern practice. Unlike sequential Gaussian simulation and sequential indicator simulation, the promise of DSS is that neither pre- nor post-processing steps are required. There is no need for data transformation, whether it is to a Gaussian or an indicator formalism. This sequential approach is common in mainstream numerical modelling, regardless of whether that modelling is performed under a parametric or non-parametric model.

**Multi-Scale Data Integration**

The current motivation for development of the direct simulation framework is the promise of integrating multiple data types from different sources and of different scales. Integrating data of different volume supports is neither new nor difficult in theory. Cokriging using average covariance/variograms permits consideration of multiscale data. In fact, the generalized cokriging equations are straightforward to obtain.

Consider $P$ stationary random variables, $Z_p$, $p=1,\ldots,P$ with mean $\mu_p$ defined on support $V_p$ centred at location $\mathbf{u}_{\alpha p}$, where $\alpha = 1,\ldots, n_p$ and $n_p$ is the number of available data of type $p$. It is not necessary that the volume supports $V_p$, $p=1,\ldots,P$ be constant.

$$Z(\mathbf{u}_{\alpha p}) = \frac{1}{V_p}\int_{V_p} Z_p(\mathbf{u}_{\alpha p})du$$

Without loss of generality, consider the residual of $Z_p$, $Y_p = Z_p - \mu_p$. Simple cokriging of the residual yields the following simple cokriging (SCK) variance:

$$\sigma^2_{SCK} = \overline{C}(V_i(\mathbf{u}), V_i(\mathbf{u})) - \sum_{p=1}^{P} \sum_{\alpha=1}^{n_p} \lambda_{\alpha p} \overline{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p}))$$

where

$$C(V_i, V_j) = \frac{1}{|V_i||V_j|} \int_{V_i} dy \int_{V_j} C(y - y') dy'$$

and the weights are determined by simultaneously solving the $\sum_{p=1}^{P} n_p$ equations that constitute the simple co-kriging system of equations:

$$\sum_{p'=1}^{P} \sum_{\beta=1}^{n_{p'}} \lambda_{\beta p'} \overline{C}(V_p(\mathbf{u}_{\alpha p}), V_{p'}(\mathbf{u}_{\beta p'})) = \overline{C}(V_i(\mathbf{u}), V_p(\mathbf{u}_{\alpha p})), \quad p = 1, \ldots, P \tag{9}$$

The resulting cokriging estimate and estimation variance correspond to the conditional expectation and variance of the RV $Y_p(\mathbf{u})$.

Greater efficiency can be achieved by simultaneously cokriging $M$ multiple data types, where $M \leq P$. This is simply achieved by changing the column vector of weights and right hand side covariance into an $M$ x $P$ matrix. An additional index is required to indicate the variable to be estimated. For this purpose, the $m$, $m=1, \ldots, M$, index is introduced.

$$Y^*_i(\mathbf{u}) = \sum_{p=1}^{P} \sum_{\alpha=1}^{n_p} \lambda^1_{\alpha p} Y_p(\mathbf{u}_{\alpha p})$$

$$\vdots$$

$$Y^*_M(\mathbf{u}) = \sum_{p=1}^{P} \sum_{\alpha=1}^{n_p} \lambda^M_{\alpha p} Y_p(\mathbf{u}_{\alpha p})$$

Solving for the weights of the resulting co-kriging system requires little additional effort since the large left hand side data to data covariance matrix (in Equation 9) only has to be inverted once. Matrix multiplication of the inverted covariance matrix with the additional *M-1* columns of the right hand side covariance will give the weights to estimate the other *M*-1 additional variables. In fact, most solvers can be modified to solve systems of simultaneous equations with multiple right hand sides without explicitly solving for an inverse. The only additional computation required in order to simultaneously estimate the collocated data types is the determination of the right hand side volume to volume covariance between the location to be estimated and the nearby data of *P* types.

While cokriging of one variable gives the conditional expectation and variance of the RV, simultaneous cokriging of multiple RVs gives the conditional mean vector and covariance matrix of the *M* RVs. Simulation using these distributional parameters must still be performed.

All this is fine in the context of estimation where cokriging can be performed in the space of the data; however, in the context of Gaussian simulation, which is the most common simulation

method in practice, using average statistics after a non-linear transformation and back transforming to original units does not work. Consider three numbers: 1, 2 and 10. The average of these three numbers is 4.33. Now consider an exponential transform, $e^x$ where $x$ is the data. This transform gives: 2.718, 7.389 and 22026.470, respectively. The average of the transformed values is 7345.524, which after back transformation yields 8.902. This is clearly not the same as the average in original space. Thus averaging in a non-linear space, such as Gaussian space, does not provide an appropriate method of accounting for multiscale data. This provides, yet, another impetus for pursuing DSS.

**Histogram Reproduction**

The topic of histogram reproduction is quite broad. It not only encompasses the obvious global distribution reproduction, but it also addresses the challenge of inferring the local distribution based on only two parameters. While this is sufficient information for a parametric model like the Gaussian model, it is often inadequate for more realistic non-parametric distributions.

The lack of a distributional assumption requirement is an obvious benefit for DSS. Natural phenomena rarely follow a parametric form such as the Gaussian distribution, and while quantile transformation permits a change from one distribution to any another, there is nothing that says we *should* transform the data to a parametric form. That data transformation is a widely accepted part of the modelling work flow speaks volumes about our strong and continued reliance on simple, yet restrictive mathematical models.

In fact, one could argue that the affect of data transformation on the true spatial distribution of the data may be undesired. Transformation to and back-transformation from Gaussian space yields some disturbing results when applied to skewed distributions. While statistical fluctuations are an inherent property of stochastic simulation, it is expected that these deviations should be reasonable and unbiased. For any one realization, minor fluctuations from a zero mean and unit variance are expected; however, when these values are back transformed to original units a slight shift in the mean in normal space may translate to a more significant shift of the mean in original units. Similarly, the combined fluctuation of the mean and variance in normal space may translate to more noticeable shifts in original space. This is particularly true for skewed distributions, which is the case for some real phenomena. Fixes to this particular problem have been proposed (Journel and Xu, 1994); yet this can be avoided altogether if we do not perform any data transformation prior to modelling – hence direct simulation.

Although Journel (1994) showed that covariance reproduction was achievable without any distributional assumptions, histogram reproduction remained a challenge. Most of the last decade has seen the majority of research focussed on this specific issue in DSS. Soares (2001) proposed to determine the local cumulative distribution function (cdf) by sampling from part of the global cdf. Caers (2000) suggested the use of a posterior correction of the histogram originally proposed by Journel and Xu (1994), in combination with an acceptance/rejection approach to determining the local cdf. Oz et. al. (2003) proposed the prior use of a Gaussian transform to determine a table of local distributions that could be accessed during DSS.

Despite the fact that DSS permits different shapes of the local distributions, the global distribution of simulated values tend to a symmetric, bell-shaped distribution characteristic of the Gaussian distribution (see Bourgault (1997) and Caers (2000)). This is a reflection of the pervasive influence of the Central Limit Theorem, sometimes referred to as the "Gaussian disease". Of the different approaches to infer the local distribution, only the approach proposed

by Oz et.al. (2003) is successful at reproducing the global distribution without need for a post-simulation histogram correction.

While histogram reproduction is key to the success of any simulation approach, this is not a significant obstacle in the widespread consumption of DSS. Actually, the work conducted in the past decade shows that there are any number of tricks and tools that can be employed to reproduce the histogram with varying degrees of desire. Although we intuitively understand that different distributions should exist to reflect different local regions, there is nothing in the prevailing DSS algorithms that will account for the proportional effect. The practicality and hence, viability, of DSS depends heavily on the promise of honouring the proportional effect.

**Proportional Effect**

By virtue of DSS' dependence on kriging, the resulting local variance is independent of the data values and the estimate, hence it is homoscedastic. In contrast, the variance of mineral grades or petrophysical properties found in a real deposit or reservoir often changes depending on the local mean – a property called heteroscedasticity. For example, it is common to find a low variance in a low valued area, and a correspondingly high variance in a high valued area. This heteroscedastic behavior is commonly referred to as the proportional effect (Journel and Huijbregts, 1978).

Consider the well-known Walker Lake data set and the lead pollution data from Dallas. A moving average approach was used with non-overlapping windows to determine the relationship between the local mean and variance. Figure 2 shows a very strong positive correlation for both data sets, in fact its relation appears quadratic, i.e.

$$\sigma^2(\mathbf{u}) = f\left(m(\mathbf{u})^2\right)$$

Note that this relationship is characteristic of real data (alternatively, it is sometimes shown as a linear relation between the standard deviation and the mean value), and it is more pronounced for a lognormal distribution (Armstrong (1998), Chilès and Delfiner (1999)). This relationship is neither new nor surprising. Journel and Huijbregts (1978), Isaaks and Srivastava (1991), Goovaerts (1997), and Chilès and Delfiner (1999) have all discussed the importance of the proportional effect in natural resource characterization. It is precisely in this aspect that direct simulation presents its biggest promise.

Yet there is a major flaw in the foundation of direct simulation. Its basis is founded in kriging, which yields a local variance that is data independent. As a result, it cannot produce models that will reproduce the heteroscedastic behaviour that would otherwise be found in real mineral deposits or reservoirs. Clearly, the flaw lies in the very fact that kriging is the engine behind the simulation. For it to fulfil its promise, direct simulation must be built on a method that yields dependent mean and variance.

**Future of DSS**

DSS is considered one of the future avenues for geostatistics. It is among the latest in a series of simulation approaches that have been introduced in the last two decades. Whether it will rank among the "standard" approaches remains to be seen, advances in particular areas will certainly be key to its popularity. DSS promises (1) the ability to integrate multiple scale data since no transformation of the data is required, (2) reduced reliance on the multiGaussian paradigm, (3) simplicity in methodology, and (4) flexibility to consider different local distribution shapes to account for multivariate non-stationarity.

These promises, however, are balanced by the pitfalls of DSS which include (1) the unavoidable influence of multiGaussianity due to the Central Limit Theorem, (2) problems in histogram reproduction which have led to ad hoc post-processing techniques, (3) the inability to account for spatial heteroscedasticity, specifically the proportional effect, and (4) flexibility in using different distribution shapes locally has not been shown to be practically advantageous or straightforward to implement.

A number of issues must still be resolved to show a real advantage to DSS. The practical significance of accounting for the proportional effect is enormous. Resolution of this issue will lend serious credibility to DSS in construction of realistic numerical models, for application in all natural resource sectors. A second area of research lies in inference of the multivariate distribution. Many authors have expended tremendous research energies into univariate distribution inference, yet the true multiscale data integration benefits of direct simulation will never be realized if the multivariate distribution cannot be properly inferred.

Although DSS was built on the principles of simple kriging, its future cannot remain anchored to simple kriging. It does not lie in the homoscedastic kriging variance, as real data show a very strong relationship exists between the variance and the data values. For it to be of practical significance and in fact, to prevent it from simply becoming an academic exercise, the underlying principle of DSS must permit a heteroscedastic variance that is data *dependent*. This is contrary to its simple kriging foundations.

**References**


Anton, H. and Rorres, C., *Elementary Linear Algebra: Applications Version*, 6th Edition, John Wiley & Sons, 1991.

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

Bourgault, G., Using Non-Gaussian Distributions in Geostatistical Simulations, *Mathematical Geology*, vol. 29, no. 3, 1997, p. 315-334.

Caers, J., Adding Local Accuracy to Direct Sequential Simulation, *Mathematical Geology*, vol. 32, no. 1, 2000, p. 815-850.

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

Davis, M.W., Production of Conditional Simulations via the LU Triangular Decomposition of the Covariance Matrix, *Mathematical Geology*, vol. 19, no. 2, 1987, p. 91-98.

Deutsch, C.V., *Geostatistical Reservoir Modeling*, Oxford University Press, 2002.

Deutsch, C.V. and Journel, A.G., *GSLIB: Geostatistical Software Library and User's Guide*, 2nd Edition, Oxford University Press, 1998.

Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.

Johnson, M.E., *Multivariate Statistical Simulation*, John Wiley & Sons, 1987.

Journel, A.G., Modeling Uncertainty: Some Conceptual Thoughts, *Geostatistics for the Next Century*, R. Dimitrakopoulos, ed., Kluwer Academic Publishers, 1994a.

Journel, A.G. and Xu, W., Posterior Identification of Histograms Conditional to Local Data, *Mathematical Geology*, vol. 22, no. 3, 1994b, p. 323-359.

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

Oz, B., Deutsch, C.V., Tran, T.T. and Xie, Y., DSSIM-HR: A Fortran 90 Program for Direct Sequential Simulation with Histogram Reproduction, *Computers & Geosciences*, vol. 29, 2003, p. 39-51.

Soares, A., Direct Sequential Simulation and Cosimulation, *Mathematical Geology*, vol. 33, no. 8, 2001, p. 911-926.

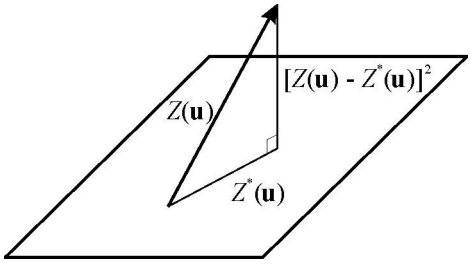


**Figure 1** Kriging in terms of projection theory (redrawn from Journel and Huijbregts, 1978; Anton and Rorres, 1991).

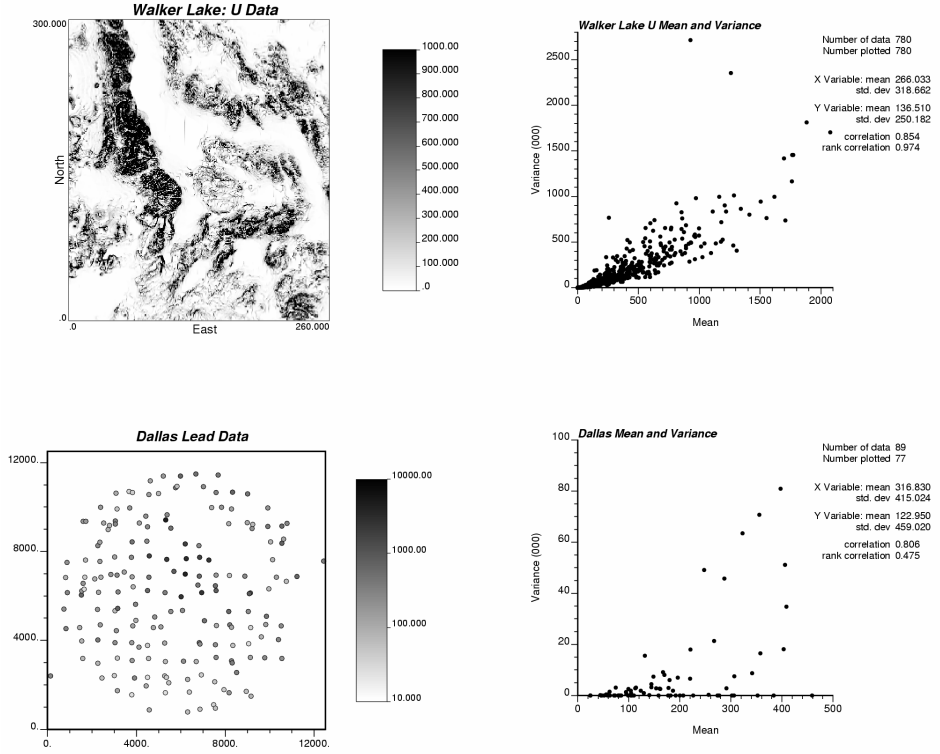


**Figure 2** Illustration of proportional effect for Walker Lake data (top), and the lead pollution data from Dall (bottom). Plan view of the data is shown on the left, and crossplots of local variance vs. local mean are shown on the right**.**