A New Dimension to Account for Data Error and Volume Support

Jared L. Deutsch, Jeff B. Boisvert and Clayton V. Deutsch

Data used in geostatistical models can come from a wide variety of sources, each with differing measurement errors and sampling volume. Accounting for these differences is complicated. A new dimension is introduced to simplify accounting for these factors. The new dimension (measured in units of distance) modifies the scalar normalized distance and can be easily implemented in existing kriging or inverse distance estimation methods. The use of d *in a number of case studies results in improved estimates with lower mean squared errors and a higher covariance between the estimates and truth. A program* `derror` *is introduced to calculate the value of* d *and other programs, such as* `kt3d` *and* `invdist`, *are modified to use the new dimension.*

1. Introduction

There are a wide variety of data sources used in geostatistical modeling, each with different errors and volume support. This poses a problem for the geostatistician; more weight should be given to data with low error and a large support volume with less weight given to erroneous, low volume data. The two problems of data error and volume support are generally considered separately, although they typically exist together. For example, data collected from diamond drilling is generally precise, but comes from a very small volume while blasthole data is collected from a larger volume but often has a greater sampling error than drillholes.

A number of methods have been proposed for reconciling data with different support volumes into a geostatistical model. One of the most recent techniques by Liu and Journel (2008) does this by using direct sequential simulation or error simulation where the block data are taken to be linear averages of a series of point values. This method is suitable for integrating large and fine scale data; however, it does not account for differing data measurement errors. The problem of integrating seismic data with well data for petroleum applications has been studied extensively (Yao and Journel, 2000; Deutsch et al., 1996 to name a few). These consider the much larger scale of seismic data and differences in precision between well log data and seismic.

Existing methods for accounting for data types with differing errors and volume support are often difficult to implement and cumbersome so they see little use. To simplify accounting for data error variances and volume, a new dimension *d* (units of distance) is introduced to lower the estimating weights given to erroneous, low volume data while increasing the weights given to precise, high volume data. This additional dimension is implemented for kriging and inverse distance estimation. A number of synthetic case studies are presented to demonstrate that rewarding precise, high volume data results in a lower mean squared error and higher predictive capabilities.

First, the calculation of *d* and its integration into estimating methods by modifying the scalar normalized distance is discussed. A number of case studies are considered to show how *d* lowers the estimating weights for poor data and increases them for high quality data which results in better estimates. The problem of accounting for dependent errors is also considered and a solution proposed.

2. The New Dimension: *d*

The new dimension *d*, with units of distance, effectively moves poor data away from the location being estimated. Low quality data receives a high *d* value which increases the distance between this data and the location being estimated. The result is that a greater estimating weight is assigned to the other, higher quality, data while the low quality data receives lower weights. The effects of increased data error and volume support on *d* are shown schematically in Figure 1.

The change caused by the introduction of *d* can be seen in Figure 2. In this 2-dimensional example, *a* is a hard data point and *b* is a soft data point. The estimated value at □ is considered to be a hard data point, similar to *a*, therefore both points to have the same value of *d*: d_{base}. The error associated with *b* means that *b* is effectively moved further away from the unknown location by the addition of *d*.

3. Accounting for Data Error

Consider a random variable, *Z,* and estimation of $z^*_{\square}$ given *n* data points, z_i. The linear estimation of $z^*_{\square}$ is given by Equation 1 where λ_i are the weights assigned to the data values with mean m_i. The estimation variance, σ^2_E, can

then be calculated (Equation 2) where $C_{i,j}$ is the covariance between data points *i* and *j* and σ^2 is the global variance.

$$z^*_{\square} - m_{\square} = \sum_{i=1}^{n} \lambda_i \cdot [z_i - m_i] \tag{1}$$

$$\sigma^2_E = \sigma^2 - 2\sum_{i=1}^{n} \lambda_i C_{i,\square} + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C_{i,j} \tag{2}$$

The variogram for this estimation problem, 2γ(**h**), for lag **h** is defined by Equation 3 where **u** is the location of the data and related to the covariance under the decision of stationarity (further details in Deutsch and Journel, 1998) by Equation 4.

$$2\gamma(\mathrm{h}) = E\left\{Z(\mathrm{u}) - Z(\mathrm{u}+\mathrm{h})^2\right\} \tag{3}$$

$$\gamma(\mathrm{h}) = \sigma^2 - C(\mathrm{h}) \tag{4}$$

If there is only one data point which is separated from $z^*_{\square}$ by a scalar normalized distance, **h**, then the above equations simplify so that the estimation variance, σ^2_E, is the value of the semivariogram for this separation distance, γ(**h**). The problem of mapping sample error variance, σ^2_{error}, to distance units for the new dimension, *d*, is the reverse problem. To calculate *d* the variance is mapped to distance units with the variogram as shown in Figure 3. For this calculation the range of *d*, a_d is chosen to be the maximum of the variogram ranges for numerical stability. This procedure assumes that the variance can be directly mapped through the variogram as in the case of only one data point. Note that this mapping procedure requires that the variogram be monotonic in nature, i.e. hole effect models would not be acceptable. This makes sense intuitively; increased error variance should not decrease *d*.

3.1 Application to Kriging

As *d* has units of distance, the variances can be worked into kriging by modifying the calculation of the scalar normalized distance, **h** (Equation 5). This is a small change that was made in existing kriging software (`kt3d`).

$$\mathrm{h}_{ij} = \sqrt{\left(\frac{x_i - x_j}{a_x}\right)^2 + \left(\frac{y_i - y_j}{a_y}\right)^2 + \left(\frac{z_i - z_j}{a_z}\right)^2 + \left(\frac{d_i - d_j}{a_d}\right)^2} \tag{5}$$

The range of *d*, a_d is taken to be the maximum of the Euclidean ranges a_i for numerical stability but could be modified. For example, under conditions of very sparse data, the penalty imposed on erroneous data could be reduced by increasing the range a_d.

The use of *d* to account for error variance is an approximate method. A theoretically correct approach that considers each data point at $\mathbf{u}_i$ to have an associated random error $\sigma^2_{error,i}$, was also developed. To account for this error when estimating at $\mathbf{u}_i$ using kriging, the error variance terms are added on the diagonal of the kriging matrix (this assumes independence of the errors). This assumption is reasonable as the errors being accounted for with this method are errors associated with the data type, not a systematic bias. The modified simple kriging matrix is shown in Equation 6. The process is similar for ordinary kriging, with the added constraint of the weights summing to unity.

$$\begin{pmatrix} C_{1,1} + \sigma^2_{error,1} & \cdots & C_{n,1} \\ \vdots & \ddots & \vdots \\ C_{1,n} & \cdots & C_{n,n} + \sigma^2_{error,n} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} C_{1,\square} \\ \vdots \\ C_{n,\square} \end{pmatrix} \tag{6}$$

Both of these methods (*d* value and error variance addition) are implemented in modification of `kt3d`: `kt3dd` (see accompanying electronic files). This program can use either a provided set of *d* values (calculated using `derror`) or error variances. These approaches are compared in a number of small case studies (section 5).

3.2 Application to Inverse Distance Estimation

The simple inverse distance (ID) estimator is considered to be a linear estimator with weights calculated by Equation 7 where *c* is the ID additive constant (generally 1/10 – 1/4 of the average data spacing) and ω is the ID

power (generally between 0.5 and 2.5) and the scalar distance **d**, between the data and unknown location. Similar to ordinary kriging, the weights are scaled to sum to unity.

$$\lambda_i^* = \frac{1}{(\mathrm{d}+c)^{\omega}},\ \lambda_i = \frac{\lambda_i^*}{\sum_{i=1}^{n} \lambda_i^*} \tag{7}$$

The calculation of the scalar distance between the data and unknown locations can be modified to include the *d* value (Equation 8). This small change can easily be made in existing inverse distance estimation programs. This method is implemented in `invdist`, a GSLIB compatible 2D inverse distance estimation program as well as `kt3dd` (see Appendix).

$$\mathrm{d} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (d_i - d_j)^2} \tag{8}$$

It is worthwhile to note that using the *d* value with inverse distance estimation still requires a variogram model to calculate the appropriate *d* values. The principal reason why inverse distance estimation is still used is because of the simplicity and no requirement for the practitioner to model the spatial correlation with a variogram. If no variograms have been calculated then these would have to be calculated to account for data error or volume support. The authors believe that if the disparity in data error or volume support is large and warrants correction, then the calculation of the variograms for this purpose would be worthwhile.

3.3 Implementation

To calculate the *d* values, a program `derror` was written. This program performs the interpolation shown in Figure 3 using a provided variogram model along with error variance values. The supplied variogram should be calculated using high quality data. If the variogram model was constructed using low quality data, this could manifest in a lower range of correlation and higher nugget effect. If this model were used then error would effectively be accounted for twice resulting in a very smooth estimate.

When calculating the *d* values for erroneous data, the nugget effect is not considered; only the structured portion of the variogram model is used. This means that the resulting *d* values are for all error variances and do not have a discontinuity at the nugget effect variance. If the variogram model is supplied with a standardized sill of unity, the variogram can be rescaled by a user supplied global variance or a variance calculated from the data. The error values associated with the data can be provided as either a standard deviation or variance for the calculation. The parameters for `derror` are:

```
Line
1                     Parameters for Derror
2                     *********************
3      START OF PARAMETERS:
4      wells.out                 -file with data
5      0  1  2  0  4  5  6       -   columns for DH,X,Y,Z,var,error,volume support
6      derror.out                -file for derror output
7      1                         -scale variogram by variance for error calc (1=yes)
8      -1.0                      -   variance for scaling (if negative from data)
9      1                         -error is in std.dev.(1) or var.(0)
10     0    0.001    0.010       -calculate d with volume for invdist (1=yes)
11     1    0.1                 -nst, nugget effect
12     1    0.9  0.0   0.0   0.0 -it,cc,ang1,ang2,ang3
13              32.0  32.0  10.0 -a_hmax, a_hmin, a_vert
```

The parameter file for `derror` is in the same style as many of the GSLIB programs and derivatives. **Line 4** specifies the location of the Geo-EAS compatible data file (Deutsch and Journel, 1998) and the column numbers for relevant variables in **Line 5**. The output file is specified in **Line 6** which will contain the drillhole, X, Y, Z, variable, error value and calculated *d*-value for use in `invdistd` or `kt3dd` (Appendix). If the variogram model provided needs to be rescaled then this can be flagged in **Lines 7 and 8**. This can occur if the variogram was modeled with a sill of unity but the original data is being used in which case the sill could be moved to match the original data. **Line 9** specifies if the error is provided as a standard deviation or variance. If inverse distance estimation is being used accounting for sample volume (Section 4), then **Line 10** gives the volume of point data, such as exploration

core, and the volume of the largest scale primary data used, such as blasthole samples. The variogram (**Lines 11-13**) is specified in standard GSLIB format (Deutsch and Journel, 1998).

4. Accounting for Volume Support

A number of methods have been proposed for integrating data with differing volume supports (reviewed in Section 1). These methods were predominantly for integrating data with support sizes that differed by multiple orders of magnitudes (block size compared to core size). The method proposed here is for volumes that differ by only a few orders of magnitude, such as blasthole data and exploration core. In a geologic setting without significant fine scale variability, small changes in sample volume relative to the block estimating size (0.1m compared to 10s or 100s of m) does not significantly alter volume variance.

However, small changes in sample size are important if there is significant geological variability at the very fine scale (nugget effect region). If there is a nugget effect, then larger scale data become more important. If the variogram is calculated using point scale data then the nugget effect is scaled for larger data, *o,* compared to point scale data, • (Equation 9). Here, point scale data is not considered to be point scale in the strictest sense (Journel and Huijbregts, 1978), it is considered to be the scale of the smallest available data type.

$$C_{\circ} = \frac{|V_{\bullet}|}{|V_{\circ}|} \cdot C_{\bullet} \tag{9}$$

This reduction is equivalent to reducing the nugget effect for the covariance function (Figure 4). In this case, the volume of the larger scale data is 3 times the volume of point scale data meaning that the nugget effect for the large data is 1/3 of the smaller. This decrease in the nugget effect is incorporated along the diagonal of the kriging matrix. It can be seen that the variance reduction (Equation 10) for using larger scale data depends on the ratio of the two volumes and the nugget effect calculated from the point scale data.

$$\sigma_{\circ}^2 = C_{\bullet} \cdot \left(1 - \frac{|V_{\bullet}|}{|V_{\circ}|} \right) \tag{10}$$

Different approaches to account for this reduction in variance are taken for kriging and inverse distance estimation, so the two will be discussed separately.

4.1 Application to Kriging

For kriging, the scaled nugget effect (and accompanying reduced variance) can be used on the diagonal of the kriging matrix, similar to the addition of an independent random error (Equation 11). This approach will be referred to as the variance scaling approach.

$$\begin{pmatrix} C_{1,1} + (|V_{\bullet}|/|V_1| - 1) \cdot C_{\bullet} & \cdots & C_{n,1} \\ \vdots & \ddots & \vdots \\ C_{1,n} & \cdots & C_{n,n} + (|V_{\bullet}|/|V_n| - 1) \cdot C_{\bullet} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} C_{1,\square} \\ \vdots \\ C_{n,\square} \end{pmatrix} \tag{11}$$

Only the diagonal elements of the kriging matrix are changed because at the small scale, the covariance is decreased only by the reduced nugget effect as changes to the structured portion of the variogram are insignificant. As the sample sizes being considered are much smaller than the ranges of the structured portion of the variogram, the average variogram (gamma-bar) values would be unchanged and therefore the variance contribution of nested structures would be unchanged (Journel and Huijbregts, 1978; Oz et al., 2002). The correction for differing support volumes is implemented in `kt3dd`, which can read in a sample volume size along with the data values. The kriging matrices are then adjusted using Equation 11.

An alternative approach is to calculate a *d* value by mapping the reduction in variance (Equation 10) to the variogram using the same procedure as for mapping error values (Figure 3). If differing volume support and data error are both being considered the independently calculated *d* values are summed for the final *d* value.

4.2 Application to Inverse Distance Estimation

For inverse distance estimation, the variance scaling approach cannot be used; thus, the *d* value must be calculated by mapping the reduction in variance to the variogram with the same procedure as described for

kriging. Calculation of the new scalar distance ||**d**|| is identical to the calculation for *d* stemming from error variances.

This calculation is implemented in `derror`, with parameters specified on **Line 10** in the parameter file. As when interpolating on the variogram for error values, the nugget effect is removed for the interpolation because it is considered to be stemming strictly from geologic variability.

5. Case Study 1: Effect of Error on Estimation Weights

The primary goal of accounting for data with different errors is to reduce the estimation weights given to erroneous data (such as blasthole data) relative to the weights given to high quality data (such as exploration drillholes). Consider the small synthetic data set in Figure 5. This data set was generated using sequential Gaussian simulation on a 30x30 grid using an isotropic variogram of $\gamma = 0.1 + 0.9Sph_{a=32}$. The resulting normal variable is transformed to follow a lognormal distribution with a mean of 0.68 and standard deviation of 1.13. Three high quality exploration samples (bold circles on Figure 5) are taken at grid locations of (5.5,2.5), (6.5,12.5) and (22.5,17.5) near an unsampled location (11.5,9.5).

In addition to the 3 high quality samples, 18 blasthole samples are taken with a higher error value in the upper portion of the grid, shown as thinly outlined circles in Figure 5. A normally distributed random error (mean = 0, standard deviation = *r*) was added to each of the blasthole samples. An independent random error was considered to simplify comparison of the weights assigned by different approaches. Similarly, the estimation technique used was ordinary kriging to simplify comparison. Three kriging-based approaches were then used to estimate the value of the unknown location: kriging with no correction for sample error, kriging with error variance addition and kriging using the *d* value. The estimation weights assigned to the 6 closest data (labeled in Figure 5) were then plotted as a function of error variance normalized by the sample variance (σ^2_E/σ^2) for each of these.

Inverse distance estimation was also considered (1) with no correction and (2) using the *d* value approach. For the inverse distance estimation, *c*=0.2 and ω=2.3 were chosen based on exploratory cross-validation testing.

5.1 Kriging with No Correction for Error

The simplest approach is kriging with no correction for sample error. The resulting weights assigned to the 6 closest data are summarized below (Table 1). These weights do not change with error variance. As the unknown location is close to the high quality samples, they received the majority of the estimation weights (0.766 compared to 0.234). Kriging with no correction for error is a non-optimal approach; erroneous data is not penalized compared to high quality data which can lead to poorer quality estimates. The addition of error variance or use of the *d* value aims to improve these estimates.

Table 1: Weights for 6 data locations using no correction for error, error variance addition and *d* values, assuming 10% error.

Datum	Weight (no correction)	Weight (error variance addition)	Weight (*d* values)
1 (High Quality)	0.229	0.229	0.228
2 (High Quality)	0.347	0.358	0.359
3 (High Quality)	0.190	0.194	0.195
4 (Erroneous)	0.070	0.058	0.068
5 (Erroneous)	0.177	0.143	0.165
6 (Erroneous)	0.108	0.094	0.103

5.2 Kriging with Error Variance Addition

Accounting for sample error by adding error variance on the diagonal of the kriging matrix was discussed in Section 3. This approach was implemented for this small case study and the effects of increasing error variance on the sample weights calculated. As the erroneous sample error variance increased, weights assigned to the erroneous samples decreased. The constraint of the weights summing to one meant that the high quality sample weights were increased accordingly. However, the weight assigned to sample 1 did not vary because it was outside the range of correlation for the erroneous samples 4-6 (Table 1).

5.3 Kriging with Calculated d Values

When the sample error was accounted for using the *d* value and modifying the scalar normalized distance accordingly, the results were similar to those obtained through error variance addition. The relative weights assigned to the erroneous data decreased while those assigned to the high quality samples increased. As before, the weight assigned to sample 1 did not vary. The procedure for calculating *d* values maps the error variance through the variogram, so any change in the variogram would be reflected in the weight curves.

The sum of the weights applied to the high quality and erroneous samples for kriging with error variance addition and *d* value method as well as inverse distance with the *d* value are plotted in Figure 6. The change in weights was approximately the same up to an error variance of 0.1 at which point the two methods diverged. The weights assigned using the error variance addition followed a quadratic curve while the *d* value weights followed a cubic. As the error variance increased, both methods decreased the weight assigned to erroneous data points substantially. This effect was even more pronounced for the *d* value method for this variogram.

5.4 Inverse Distance with No Correction for Error

Considering inverse distance estimation with no correction, the following weights were calculated. The weights differ significantly from the kriging weights because no correction is made for covariance between the samples and no spatial structure is considered (Table 2). These weights do not change depending on the error chosen, so erroneous data is not penalized. The high quality samples received 46.6% of the weight compared to 53.4% for the erroneous samples.

Table 2: Weights using inverse distance estimation for a select number of data points, assuming 10% error.

Datum	Weight (no correction)	Weight (*d* values)
1 (High Quality)	0.105	0.109
2 (High Quality)	0.293	0.303
3 (High Quality)	0.068	0.070
4 (Erroneous)	0.112	0.108
5 (Erroneous)	0.142	0.134
6 (Erroneous)	0.087	0.085

5.5 Inverse Distance with Calculated d Values

When using the calculated *d* values to modify the scalar distance, the weights assigned to erroneous data points decreased, which increased the weights assigned to the high quality data. All of the changes in weights followed a cubic curve which showed little change in weights until an error variance of approximately 0.1 was reached. Compared to the changes in kriging weights with the *d* value, inverse distance estimation weights changed more rapidly (Figure 6). The effect of *d* is a function of the parameters *c* and ω chosen for the inverse distance estimation and the variogram used to interpolate *d*.

6. Case Study 2: Effect of Volume Support on Estimation Weights

Consider the case study used to study the effect of error on estimation weights (Figure 5). If the high quality samples are considered to come from sample volumes V_o and the erroneous samples are drawn from $V_\bullet$, then the effect of the volume correction methods developed earlier on the estimation weights can be plotted as a function of $V_o / V_\bullet$. For this case study, the same variogram was used with a 10% nugget effect ($\gamma = 0.1 + 0.9 Sph_{a=32}$).

As before, three kriging based approaches (no correction, variance scaling and *d* value) as well as two inverse distance approaches (no correction, *d* value) are considered and the resulting total weights for high quality and erroneous data are plotted in Figure 7. The kriging and inverse distance estimation weights with no correction for support volume are the same as before (Sections 5.1 and 5.4, respectively) so will not be detailed again.

6.1 Kriging with Variance Scaling

The first case considered was kriging with variance scaling. The change in estimating weights caused by using this approach was the largest (Figure 7). As the volume ratio increased, the covariance terms along the diagonal for the high volume data converged to 0.9 causing the weights to converge to a constant value.

6.2 Kriging with Calculated d Values

Using calculated *d* values, the changes in estimating weights were not as large as when the variance scaling method was used. The *d* values converged to a constant value (approximately 2.33) as volume increased causing the assigned weights to converge. The *d* value was low enough that the weights were not significantly affected, even at a large volume ratio.

6.3 Inverse Distance with Calculated d Values

The inverse distance estimation weights showed an increase similar to kriging with the *d* value. The resulting weights converged as the *d* value approached the constant value of approximately 2.33, equal to γ(0.2). As the calculated *d* values depend on the initial structured part of the variogram, if the variogram increases sharply at the origin, the *d* values will not increase significantly.

7. Case Study 3: Effect of Error on Cross Validation

The incorporation of the *d* value into estimation techniques has the desired effect on sample weights; erroneous data receives less weight than higher quality data. The resulting estimate is smoother because of the increased variance contributed from the data but also has a smaller mean squared error to the truth because it respects high quality data while reducing the effect of low quality data.

Consider the larger, synthetic data set in Figure 8. The variable of interest is Gaussian (mean = 0.04, variance = 1.08) generated using sequential Gaussian simulation. The same variogram was used, $\gamma = 0.1 + 0.9Sph_{a=32}$. A large number of erroneous samples were taken only in the upper portion while a smaller number of high quality samples were taken across the area. The erroneous samples were permuted in the same manner as the previous small case study. These erroneous samples with a chosen error variance of r^2 along with the high quality samples were used to estimate the grid. The grid was estimated using two approaches: normal kriging and kriging with the *d* value. The nugget effect was removed from the variogram when kriging with the *d* value to avoid over smoothing the estimate by accounting for data error twice.

The resulting estimates using the *d* value were smoother because of error filtering even with the removal of the nugget effect. It should be noted that data exactitude is not guaranteed for erroneous data as the sample error is accounted for using this approach; this is not a drawback, it is a result of considering the error distribution of erroneous samples. Samples without error are reproduced.

7.1 Effect on Cross Validation Mean Squared Error and Covariance

Estimates made using normal kriging and kriging with the *d* value with the *d* value were compared to the truth with cross validation. These scatterplots for an error variance of 10% (Figure 9) were generated using a single realization, exact values of the mean squared error and covariance fluctuate slightly depending on the random error assigned to the samples. Using the *d* value approach, the mean squared error is lower and the covariance is higher for the same error variance. There is also no evidence of a significant bias using either method. Using the *d* values, the variance of the estimate was decreased due to increased smoothing.

7.2 Relative and Independent Error Reconciliation

There are two primary models under consideration for sampling error; an independent, random error (Equation 12) or a dependent, random error (Equation 13) where *r* is the standard deviation of the sample error and *Y* is a standard normal random variable (mean = 0, standard deviation = 1). Ultimately, the decision of whether the error is independent or dependent must come from knowledge about the sampling procedure and variable of interest. Often, the magnitude of the sampling error depends on the magnitude of the true value making a relative sampling error more realistic.

$$z_{\text{samp}}(u_i) = z(u_i) + Y_i \cdot r \tag{12}$$

$$z_{\text{samp}}(u_i) = z(u_i) + Y_i \cdot r' \cdot z(u_i) \tag{13}$$

This presents a challenge in calculating *d* values as the true value is unknown (if it were known then there would be no need for *d*). To account for a dependent error when calculating the *d* value, the error must be dependent on the sample value, not the true value (Equation 14). Using this approximation, the error variance also changes (Equation 15).

$$z_{\text{samp}}(u_i) = z(u_i) + Y_i \cdot r' \cdot z_{\text{samp}}(u_i) \tag{14}$$

$$\sigma_{\text{E}}^2 = r \cdot z_{\text{samp}}(u_i) \tag{15}$$

This error variance is dependent on the sample value and can be used to calculate *d*. An alternative approach would be to ignore the dependent nature of the error and use an average value, similar to an independent random error. Ignoring the dependent nature gives data sampled using the same technique the same *d* value, while using the approximate error variance (Equation 12) gives different *d* values depending on the magnitude of the sample. Using different *d* values depending on the magnitude of the sample could lead to a bias in the estimate as high valued erroneous samples will receive less weight than their low valued counterparts. Using the large 128x128 case study considered for cross validation, both of these approaches (dependent and an average independent error) are implemented to determine the bias.

Using the same parameters as the case study in Section 6.1 with a 10% relative error, the grid was estimated using, 1) a *d* value approach considering an independent random error of 10% and 2) a *d* value approach considering a dependent random error calculated using Equation 14. The results were cross validated and the covariance and mean squared error calculated (Figure 10, Table 3).

Table 3: Covariances and mean squared errors for independent and random error assumptions.

Method	Covariance	Mean Squared Error
d Value with Independent Random Error	0.640	0.487
d Value with Dependent Random Error	0.542	0.507

Using independent random error method yielded a higher covariance and lower mean squared error, both desirable features in cross validation. While it is possible to account for dependent random errors using the *d* value, it is recommended that an average independent error be considered for these calculations to avoid introducing a bias into the estimates.

8. Case Study 4: Effect of Volume on Cross Validation

Consider estimating the large data set in Figure 8 (Section 7) with 3x3 blocks instead of the 1x1 blocks used. This data set was upscaled to 3x3 blocks using linear averages of the 1x1 blocks and the results plotted in Figure 11. The high quality samples (in this case high volume) sample the average values shown in Figure 11 while the poor samples are taken from the 1x1 blocks used in the prior case study. Cross validation was performed using the volume correction methods discussed in Section 4 with a volume ratio of 9:1 in the same manner as the previous cross validation case study using the upscaled grid. The cross validation results are plotted in Figure 12.

The covariance increased and mean squared error decreased when kriging was done with the *d* value and variance scaling methods compared to regular kriging. No type of kriging showed a significant bias in results and all showed similar variances. When using inverse distance estimation with the *d* value, the covariance and mean squared error decreased. Although the covariance decreased, the slope of regression was closer to 1, 0.898 compared to 0.883. This decrease in covariance likely has to do with the choice of estimation parameters and the realization used.

9. Conclusions

The authors have proposed a number of methods to reconcile data error and volume support differences. Here, some guidelines for their use are presented.

To reconcile data with different errors, the use of an average independent error is proposed rather than a dependent error (Section 7). This will avoid introducing a bias into the estimate. For inverse distance estimation, the *d* value approach is the only approach that can be used, but either the error variance addition or *d* value approaches can be used for kriging. The authors recommend that cross validation with both of these methods be performed. If one method significantly out performs the other, then that method should be used. Otherwise, for relatively small error differences (less than 30% of sample variance), the *d* value approach is recommended. For higher errors, the error variance addition option is recommended. This recommendation is made to avoid giving

too little weight to erroneous data. At greater than 30% error, the *d* value approach can devalue low quality data too much.

To account for different volume supports, a *d* value approach and variance scaling method have been proposed. For inverse distance, the *d* value approach must be used; however for kriging either method can be used. Both methods gave very similar results, so either method can be used.

All of these methods penalize low volume or erroneous data by giving these data points lower estimating weights. The new dimension, *d,* does this by moving the data point in a 4^{th} dimension. The case studies considered show that erroneous, low volume data should be given less weight and that the resulting estimates are more accurate. An aside also demonstrated that it is better to consider an independent average error than account for a dependent error as this can introduce a significant bias into the estimate.

There are a wide variety of data sources encountered in geostatistical estimation, each with differing support volumes and error variances. These differences should be accounted for in estimation. The proposed new dimension and variance scaling methods are simple methods for incorporating data error and volume support into kriging and inverse distance estimation techniques. A program `derror` was presented that will calculate the *d* values making this a simple change to existing software for estimation.

References

Boisvert, J.B., 2010, Geostatistics with Locally Varying Anisotropy, *Ph.D. Thesis*, University of Alberta, 175 pp.

Deutsch, C.V. and Journel, A.G., 1998, *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York, 2nd Ed., 369 pp.

Deutsch, C.V., Srinivasan, S. and Mo, Y., 1996, Geostatistical reservoir modeling accounting for precision and scale of seismic data, Proceedings - 1996 SPE Annual Technical conference and Exhibition, 9-19.

Journel, A.G. and Huijbregts, Ch.J., 1978, *Mining Geostatistics*, Academix Press, London, 1st Ed., 600 pp.

Kravchenko, A. and Bullock, D.G., 1999, Comparative study of interpolation methods for mapping soil properties, *Agronomy Journal* 91, 393-400.

Liu, Y. and Journel, A.G., 2009, A package for geostatistical integration of coarse and fine scale data, *Computers & Geosciences* 35, 527-547.

Lu, G.Y. and Wong, D.W., 2008, An adaptive inverse-distance weighting spatial interpolation technique, *Computers & Geosciences* 34, 1044-1055.

Machuca Mory, D.F., 2007, Inference of the Nugget Effect and Variogram Range with Sample Compositing, *Centre for Computational Geostatistics* 9, 123.

Matérn, B., 1960, Spatial variation, of lecture notes in statistics Second ed., Vol.36: Springer, New York, First edition published by Meddelanden fran Statens Skogsforskningsinstitut, band 49, no. 5, 1960, 151 pp.

Oz, B., Deutsch, C.V. and Frykman, P., 2002, A visualbasic program for histogram and variogram scaling, *Computers and Geosciences* 28, 21-31.

Pyrcz, M.J. and Deutsch, C.V., 2006, Semivariogram Models Based on Geometric Offsets, *Mathematical Geology* 38 (4), 475-488.

Yao, T. and Journel, A.G., 2000, Integrating seismic attribute maps and well logs for porosity modeling in a west Texas carbonate reservoir: addressing the scale and precision problem, *Journal of Petroleum Science and Engineering* 28, 65-79.

Figures

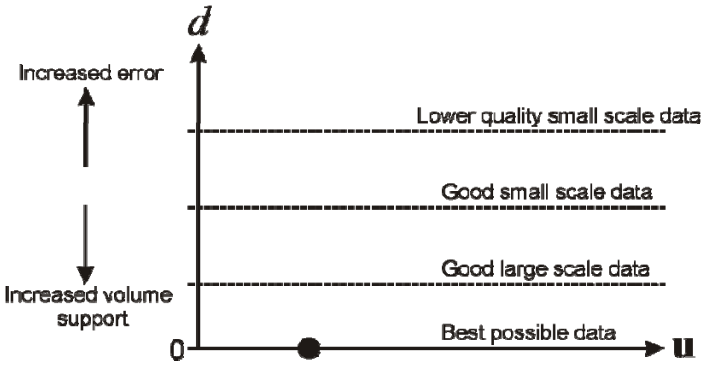

Figure 1: The effects of increased error and volume support on the new dimension.

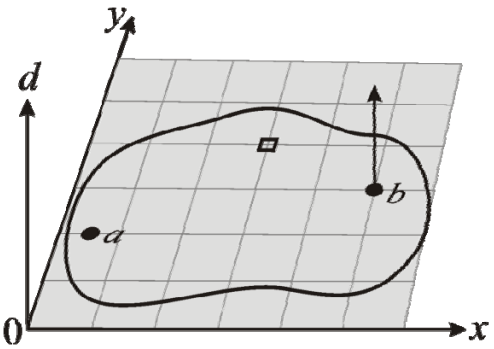

Figure 2: Estimating at ☐ given hard data point *a* and soft data point *b*. Here, *b* has more measurement error than *a* so is moved further away from ☐ by a distance *d*.

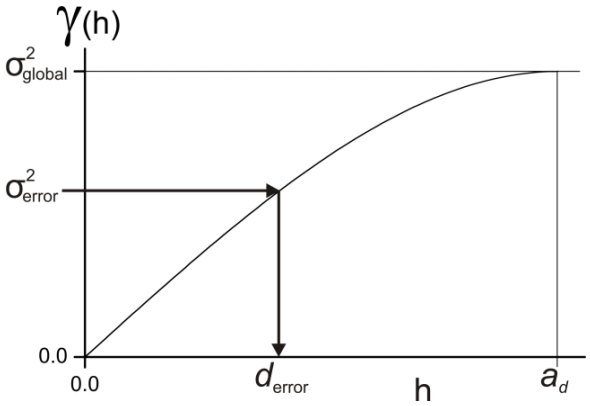

Figure 3: Mapping error variance to *d* using the variogram.

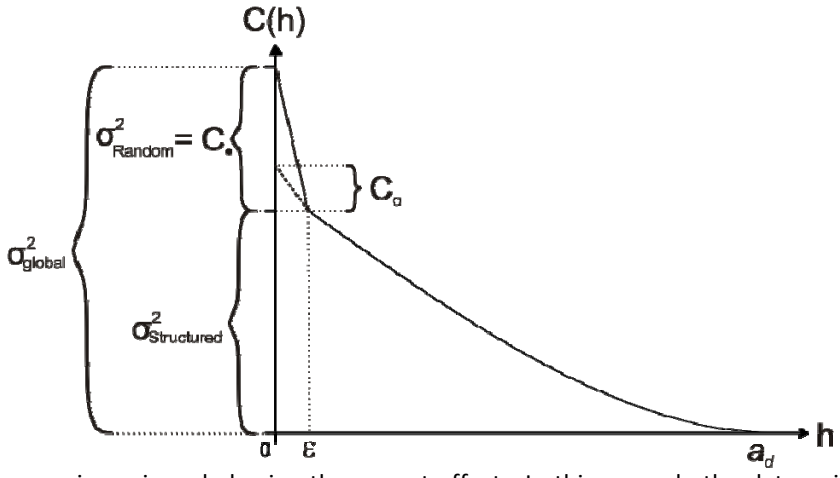

Figure 4: The random variance is scaled using the nugget effect. In this example the data point has a volume 3 times as large as point scale data which scales the nugget effect by 1/3.

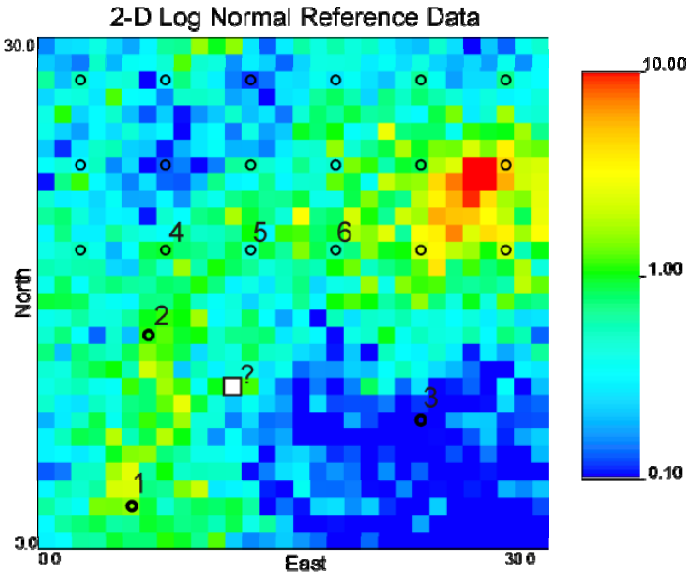

Figure 5: Fabricated data set to test weights (high quality samples with thick outline, erroneous samples with thin outline) and unknown location being estimated.

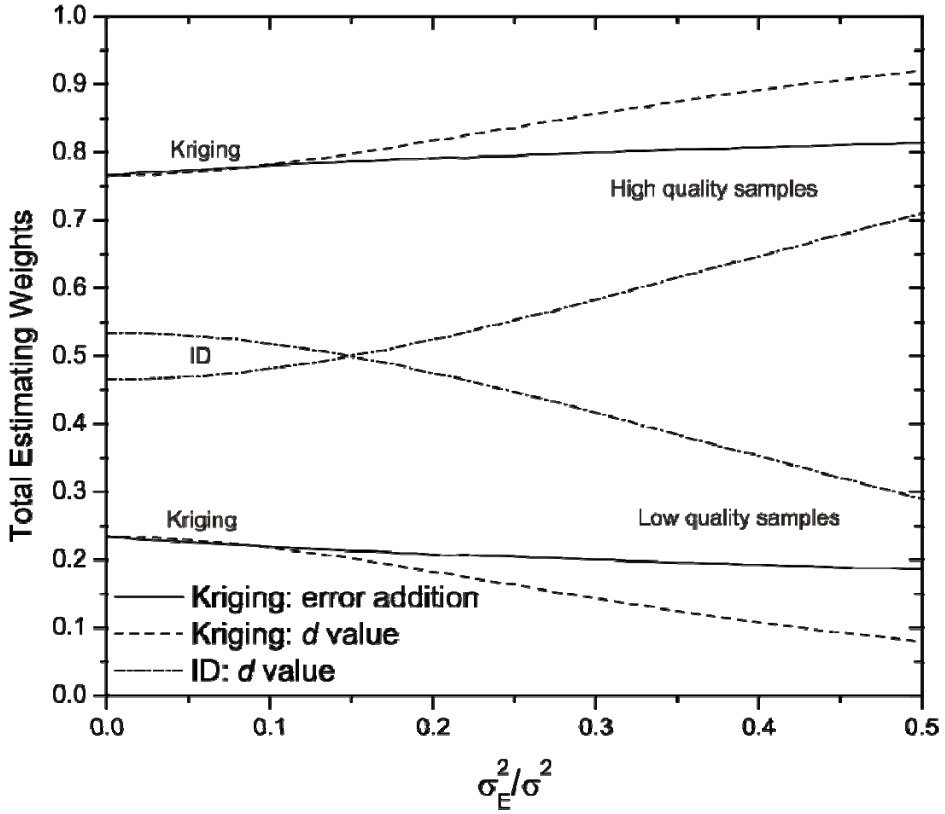

Figure 6: The total estimating weights assigned to the high quality samples and erroneous samples.

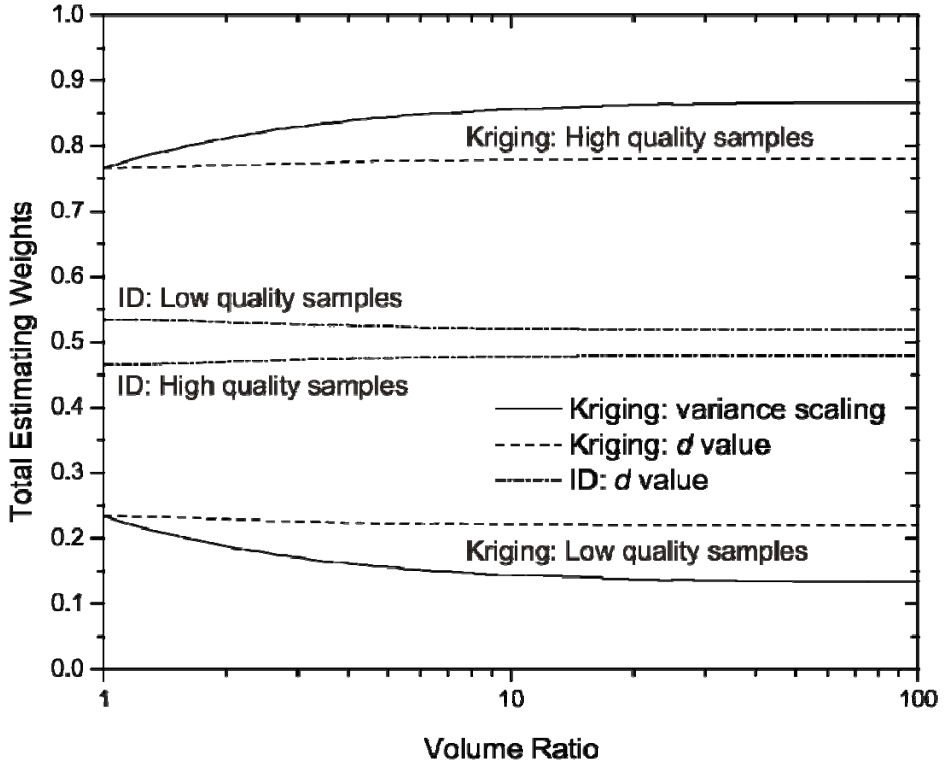

Figure 7: Total estimating weights assigned to the high quality samples and low quality samples. Note that total estimating weight change using the *d* value was significantly lower than the variance scaling method.

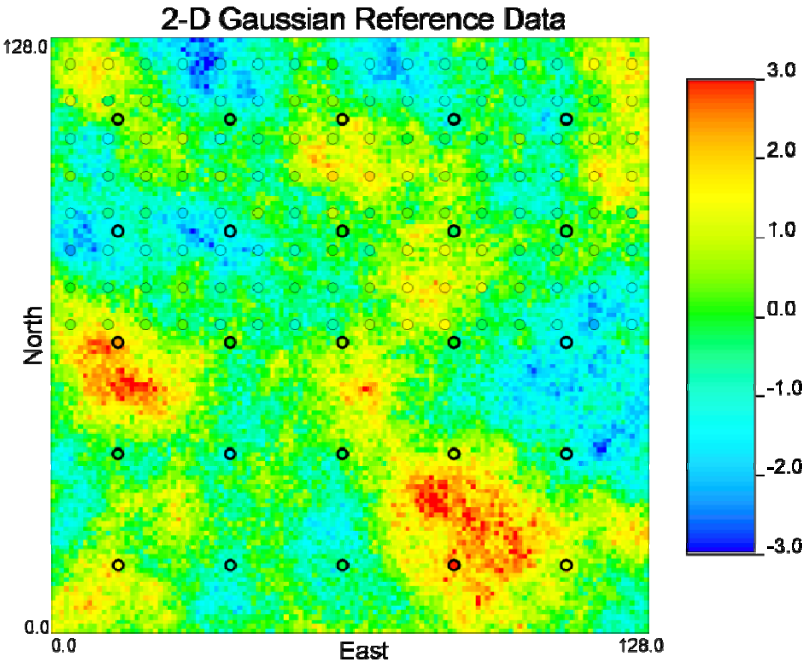

Figure 8: Simulated Gaussian reference data and locations of high quality samples (thick outline) and erroneous samples (thin outline) in the upper region.

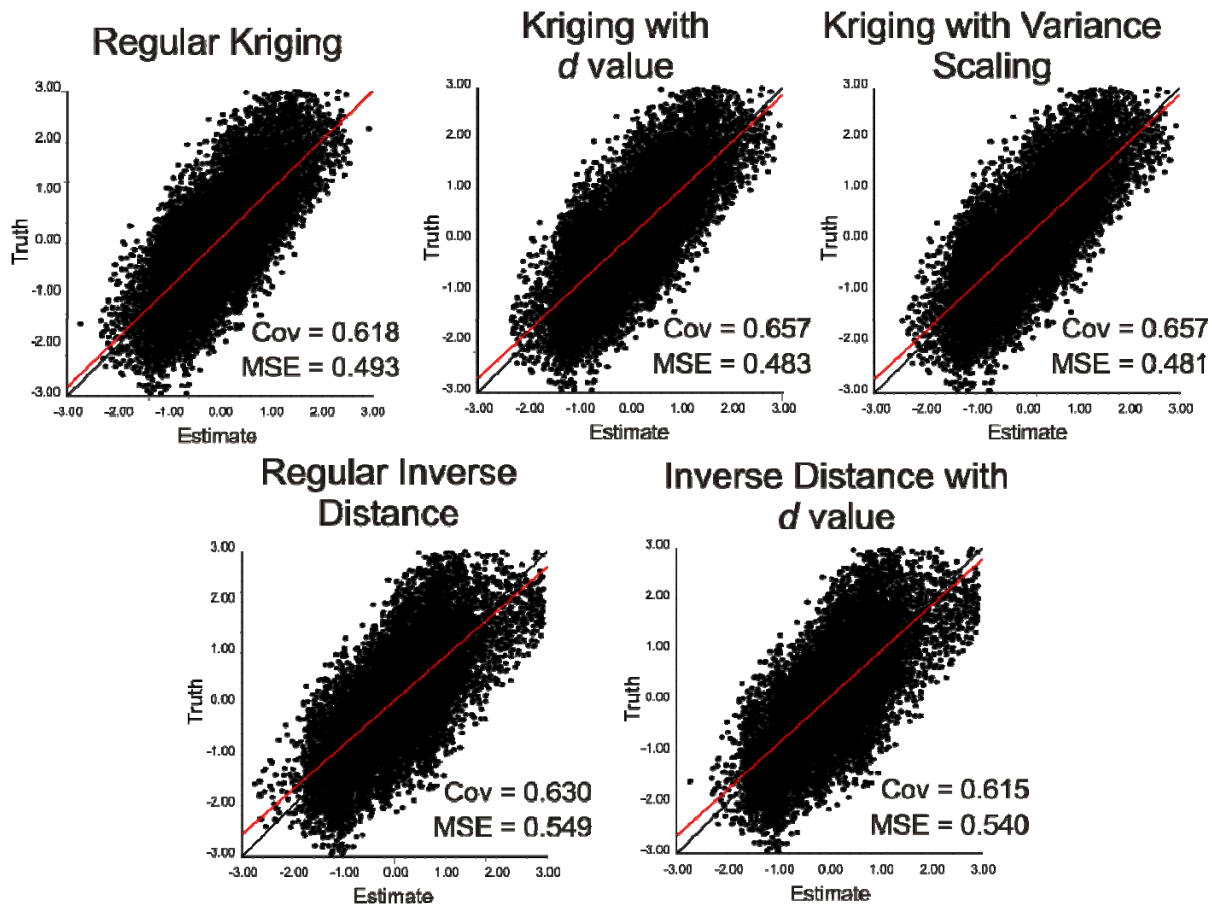

Figure 9: Cross validation scatterplots for 10% error variance. Note the changes in covariance and mean squared error using the *d* value and variance scaling.

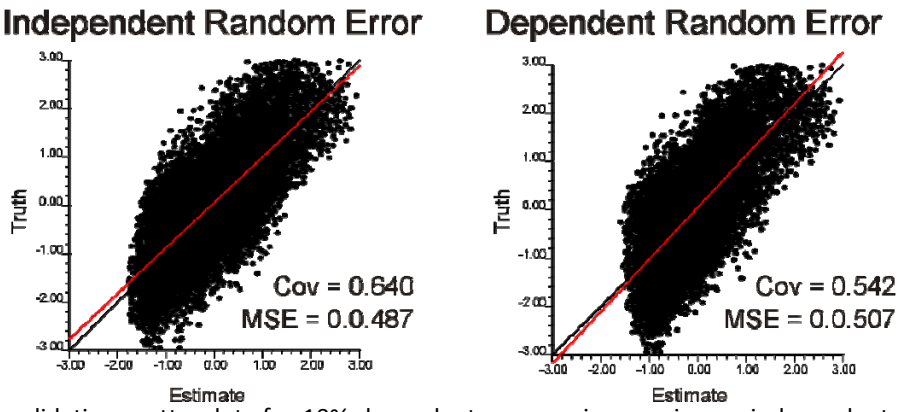

Figure 10: Cross validation scatterplots for 10% dependent error variance using an independent random error *d* value approach and dependent random error *d* value approach.

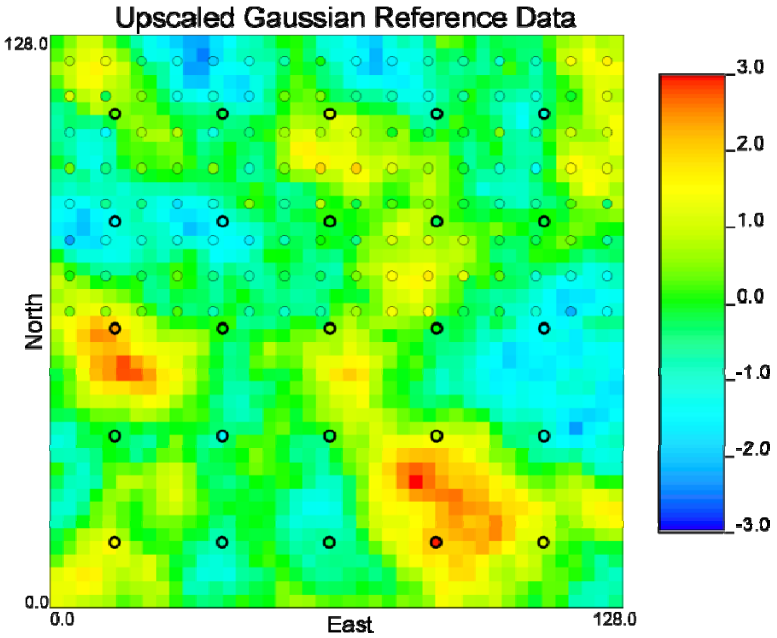

Figure 11: Upscaled Gaussian reference data used to cross validate volume support correction. Bold samples were taken from the upscaled reference data while thin samples taken from normal reference data (smaller volume).

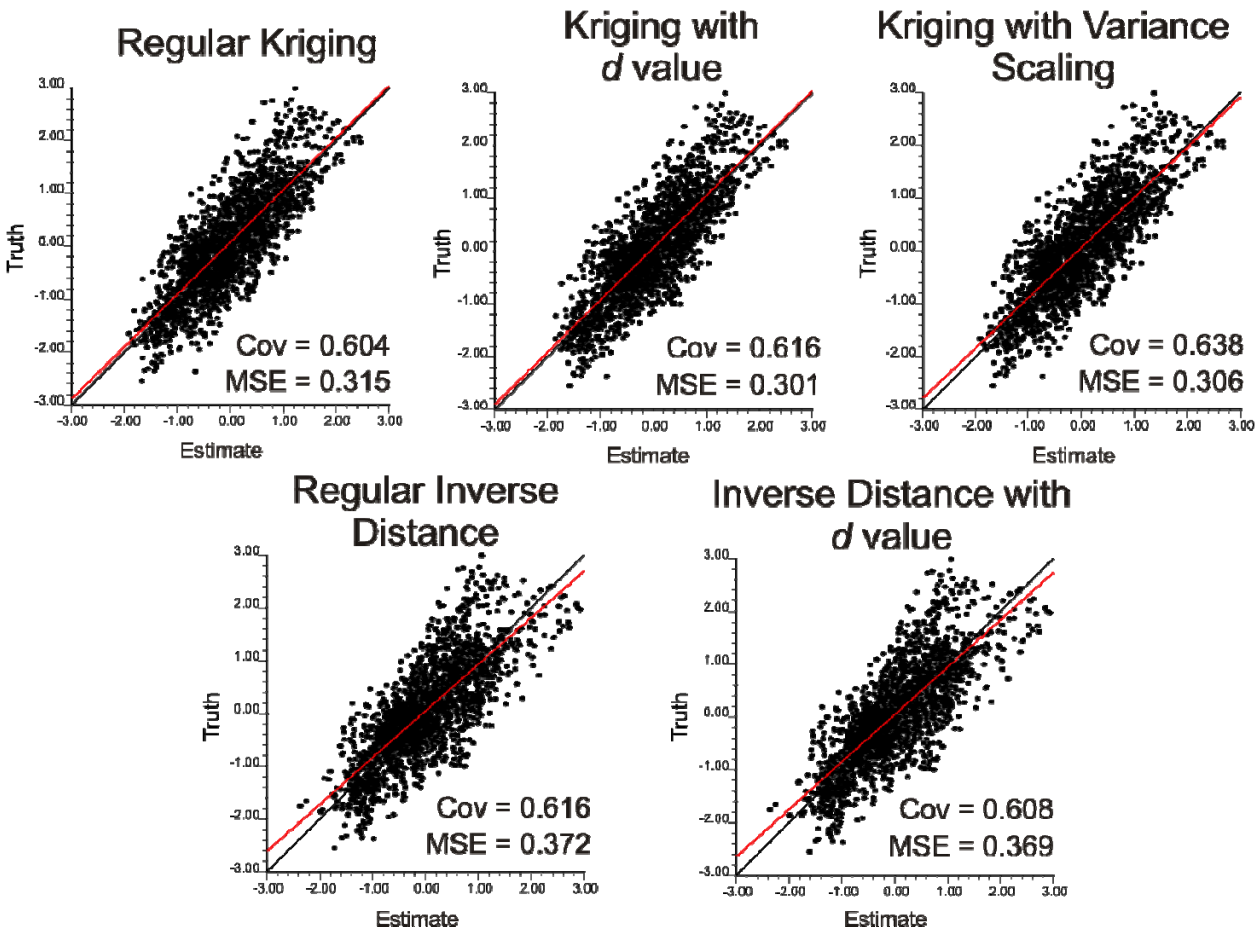

Figure 12: Cross validation scatterplots considering kriging and inverse distance variants with and without a volume correction.

Appendix

Examples, source code and parameter files for `invdist`, `kt3dd` and `derror` are included in the accompanying electronic files. This appendix details the modified parameter files for `invdist` and `kt3dd`.

The parameter file for `kt3dd` is almost identical to the parameter file for `kt3d`; a parameter file used with `kt3dd` can be used with `kt3d` with no adjustments. Only the modified parameters will be discussed here, a complete discussion of parameters for `kt3d` is available (Deutsch and Journel, 1998). **Line 5** specifies the drillhole ID, coordinates and variable columns. If using a *d* value or error addition option, then the column number is specified in `secvar`. This is then identified as either a variance (0), standard deviation (1) or *d* value (2). If using the volume scaling option, then the column for the data support volume is specified by `ivols` and the point volume (discussed in this paper) as `pointvols`.

This version of `kt3dd` has also been updated to perform inverse distance estimation using the search options available in kt3d. This is done by setting nst (**Line 26**) to -1. The nugget effect (**Line 26**) then corresponds to the inverse distance additive constant and cc(1) (**Line 27**) to the inverse distance power. Examples demonstrating this capability are included in the electronic files. This is particularly useful for performing cross validation to compare inverse distance estimation methods with kriging based methods.

```
Line
1                       Parameters for KT3D
2                       ******************
3     START OF PARAMETERS:
4     ../data/cluster.dat              -file with data
5     0  1  2  0  3  0   2   4   0.5   -   columns for
      DH,X,Y,Z,var,secvar,var(0)or stdev(1)or d(2),ivols,pointvols
6     -1.0e21   1.0e21                 -   trimming limits
7     0                                -option: 0=grid, 1=cross, 2=jackknife
8     xvk.dat                          -file with jackknife data
9     1   2   0    3    0              -   columns for X,Y,Z,vr and sec var
10    3                                -debugging level: 0,1,2,3
11    kt3d.dbg                         -file for debugging output
12    kt3d.out                         -file for kriged output
13    50   0.5    1.0                 -nx,xmn,xsiz
14    50   0.5    1.0                 -ny,ymn,ysiz
15    1    0.5    1.0                 -nz,zmn,zsiz
16    1    1      1                   -x,y and z block discretization
17    4    8                          -min, max data for kriging
18    0                                -max per octant (0-> not used)
19    20.0  20.0  20.0                 -maximum search radii
20     0.0   0.0   0.0                 -angles for search ellipsoid
21    0     2.302                      -0=SK,1=OK,2=non-st SK,3=exdrift
22    0 0 0 0 0 0 0 0 0                -drift: x,y,z,xx,yy,zz,xy,xz,zy
23    0                                -0, variable; 1, estimate trend
24    extdrift.dat                     -gridded file with drift/mean
25    4                                -  column number in gridded file
26    1    0.2                         -nst, nugget effect
27    1    0.8  0.0   0.0   0.0        -it,cc,ang1,ang2,ang3
28              10.0  10.0  10.0       -a_hmax, a_hmin, a_vert
```

Inverse distance estimation can also be carried out using the GSLIB compatible invdist. The data file in Geo-EAS format is specified in **Line 4**. Relevant columns for the coordinates, variable and *d* value (if desired) are given in **Line 5**. The trimming limits (**Line 6**), output gridded file (**Line 7**) and grid specifications (**Lines 8-9**) can be specified in standard GSLIB format. The maximum number of closest samples to consider can be set to (**Line 10**) and anisotropy in the *x* direction (**Line 11**). The inverse distance additive constant (**Line 12**) and exponent (**Line 13**) are the final parameters required.

```
Line
1                       Parameters for INVDIST
2                       **********************
3     START OF PARAMETERS:
4     wells.out                    -file with data
5     1   2   3   0                -   columns for X, Y, var, d-value
6     -1.0e21   1.0e21             -   trimming limits
7     invdist.out                  -file for gridded output
8     30   0.5    1.0              -nx,xmn,xsiz
9     30   0.5    1.0              -ny,ymn,ysiz
10    10                           -number of closest samples to keep
11    1.0                          -x anisotropy (>1 --> more continuous)
12    0.0                          -c (inverse distance additive constant)
13    1.0                          -omega (inverse distance exponent)
```