

# **Geostatistical Modeling in Presence of Extreme Values**

by

Mahshid Babakhani

A thesis submitted to the Faculty of Graduate Studies and Research  
in partial fulfillment of the requirements for the degree of

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering

University of Alberta

©Mahshid Babakhani, 2014

# Abstract

Geostatistical modeling in presence of extreme values needs special attention. Certain extreme high values known as outliers require proper treatment or mineral resources may be overstated. A number of methodologies are proposed in this thesis to identify and manage outliers. The main goal of the outlier management strategies is to keep the outliers but control their influence on local block estimates, if required, by reducing their values.

The variogram is an essential geostatistical tool that is highly sensitive to outlier high values. There are alternatives to the experimental traditional variogram. The pairwise relative variogram is known to be a very robust alternative to the experimental traditional variogram in presence of outlier high values. There are two issues with the pairwise relative variogram; (1) an unknown sill, and (2) its convergence to the wrong variogram. Solutions to both of these problems are developed and documented in this thesis.

Another contribution of this thesis is a simulation-based approach for cutting outlier values. The volume of influence of an outlier is identified. The uncertainty in the mean of the volume of influence is established by simulation. Then, the outlier value is reduced until the estimated quantity of metal within the identified volume is reasonable relative to the distribution coming from simulation. The outlier management strategies presented in this thesis will provide practical assistance for geological modeling for engineering evaluations.

*To my parents for their endless love and support*

# Acknowledgments

I would like to express my sincere gratitude to Prof. Clayton V. Deutsch for his supervision, splendid guidance and encouragement during my master's research program. His support, encouragement and patience provided me with a great experience which is highly appreciated.

I am also grateful for the financial support from the Centre of Computational Geostatistics (CCG) and the University of Alberta.

I also would like to thank all CCG students, graduates, staff, and associates for their help. Special thanks to Brandon Wilde, Miguel Cuba, Ryan Burnett and Jared Deutsch for their kindly advice during my first couple of months in CCG group.

I also appreciate my family members including my mother, Azam; my father, Parviz; my sisters, Bitra and Banafsheh; and my brother Masoud, who have invaluable supported me throughout my life.

# Table of Contents

Chapter 1 – Introduction .....	1
1.1    General Problem Area .....	1
1.2    Problem Definition .....	2
1.3    Literature Review .....	3
1.3.1    Outlier Detection .....	3
1.3.2    Outlier Management for the Histogram .....	5
1.3.3    Outlier Management for the Variogram .....	8
1.3.4    Outlier Management for Block Modeling .....	9
1.4    Thesis Outline .....	10
Chapter 2 - Univariate Management .....	12
2.1    Univariate Statistical Approach .....	12
2.1.1    Cutting to Various Thresholds .....	12
2.1.2    Manual Method .....	17
2.1.3    Probability Plot Fitting .....	17
2.2    Spatial Outliers .....	21
2.2.1    Ranked Transformed of Data and Estimation .....	22
2.2.2    Sensitivity of Resource to Spatial Outliers .....	27
2.2.3    Effect of Outliers on Grades above Cutoff .....	28
2.3    Recommendation .....	30
2.3.1    Apply Univariate Analysis .....	30
2.3.2    Spatial Analysis .....	31
2.3.3    Devise a Combined Approach .....	31
Chapter 3 - Variogram Management .....	35
3.1    Traditional Variogram .....	35
3.2    Robust Alternatives to the Variogram .....	38
3.2.1    Correlogram .....	38
3.2.2    Normal Score Variogram .....	39
3.2.3    Indicator Variogram .....	39

3.2.4	Madogram and Rodogram .....	40
3.2.5	Pairwise Relative Variogram .....	40
3.3	Standard pairwise relative variogram.....	42
3.3.1	Sill of the Pairwise Relative Variogram .....	42
3.3.2	Standardization procedure.....	46
3.4	Correct Unbiased Estimation of Traditional Variogram .....	49
3.4.1	Relationship between the Correlation Coefficient of Normal and Lognormal Distribution .....	49
3.4.2	Unbiased Estimate of True Variogram .....	50
3.5	Recommendation.....	56
Chapter 4 - Estimation and Simulation.....		57
4.1	Volume of Influence.....	57
4.2	Move to Higher Dimension.....	59
4.3	Cutting Outlier.....	62
4.4	Recommendation.....	67
Chapter 5 – Conclusion.....		68
5.1	Contributions.....	68
5.2	Future Work .....	70
References.....		71
Appendix.....		73
A.1	Probability plot.....	73
A.1.1	Probability plot fitted model .....	73
A.2	Pairwise Relative variogram .....	75
A.2.1	Standard Pairwise relative variogram .....	75
A.2.2	Correct unbiased estimate of traditional variogram .....	77
A.3	Cutting outliers.....	77
A.3.1.	Identify the outliers and volume of influence .....	78
A.3.2.	Simulated uncertainty in local averages.....	80
A.3.3.	Managing outliers based on cutting level .....	82
A.3.4.	Check the results .....	84

# List of Tables

Table 1: Different outlier detections of the silver variable ..... 24

Table 2: Results of different methods of cuttings ..... 33

# List of Figures

Figure 1: Probability plot of Au grade (Deutsch, 2011) .....	4
Figure 2: Scatter plot of paired data with an outlier (Sinclair & Blackwell, 2002). No units are given, but variables $x$ and $y$ are grades measured on duplicate samples.....	5
Figure 3: Clustering of a data variable to produce an unbiased histogram (Sinclair & Blackwell, 2002). This figure is not scaled because it is a fraction of a larger domain as an example to illustrate the clustering method .....	6
Figure 4: Cumulative CV for specifying the breakpoint (Parker, 1991) .....	7
Figure 5: Metal at risk plot for goldcopper.dat (Deutsch, 2010) .....	8
Figure 6: Histogram of lognormal data.....	14
Figure 7: Cutting the values above 90% of the data to the value of 2.10. In plot (a) the means in section $A_1$ are reduced but the averages of $A_2$ decrease significantly. In plot (b) the histogram of the corrected mean is much thinner.....	15
Figure 8: Cutting the values above 95% of the data to the value of 3.56. In plot (a) the averages in both sections reduce properly and are more acceptable. In plot (b) the histogram of the corrected mean reduces well in the upper side of the population mean, but the deduction in the lower part is still high.....	16
Figure 9: Cutting the values above 99% of the data to the value of 12.14. In plot (a) the averages in both sections change slightly. In plot (b) the histogram of the distribution of the corrected mean is almost the same as the naive mean .....	16
Figure 10: Manually reduce the outliers of the 100 samples. In plot (a) the means in $A_1$ reduce more than the averages of $A_2$ . In plot (b) the histogram of the corrected mean is reduced properly .....	17
Figure 11: Probability distribution assumptions associated with the fitted non-parametric model (Deutsch, 2010).....	18
Figure 12: Probability plot of one sample of lognormal distribution with corrected outlier values for the arbitrary threshold (e.g., 90) .....	19

Figure 13: Probability plot of another sample with outlier high values that are less than lognormal fitted model.....	20
Figure 14: Reduce the outliers of the 2500 samples with lognormal fitted model. In plot (a) the means in $A_1$ reduce but the deduction of averages in section $A_2$ are minimized compared with other methods. In plot (b) the histogram of the corrected mean is reduced properly .....	21
Figure 15: Basic information of Ag value: (a) Histogram and, (b) Location map	22
Figure 16: (a) Location map of estimation Ag and, (b) the scatter plot of the true and estimated values .....	23
Figure 17: (a) Scatter plot of uniform rank transformed of true and cross estimation of Ag values and, (b) projection of data points on 135degree line .....	24
Figure 18: (a) Defining the range of CVE, (b) Scatter plot of silver values and CVE, (c) The location map of the value of CVE from the highest underestimation (blue) to the highest overestimation (red).....	26
Figure 19: Difference from the estimation average: (a) by reducing the data value of Ag and, (b) by removing each data value .....	28
Figure 20: The histogram of changing the average of estimation in absence of each data for the cutoff grade for silver .....	29
Figure 21: For the different cutoff value, the change in quantity (plot-a) and average of ore (plot-b) for the estimation value and four of the more important outliers.....	30
Figure 22: The variogram is calculated as the average of the distance squared of the points from the 45° line. These data are for illustration only and the units are arbitrary.....	36
Figure 23: Contour plot of a conventional traditional variogram which gives equal weight to the pairs with equal differences. The colored lines represent the contribution of traditional variogram for each pair of data.....	37
Figure 24: The weight scheme in the pairwise relative variogram (Wilde & Deutsch, 2006) .....	41

Figure 25: Two directional pairwise relative variograms for a standard lognormal distribution. The violet line is in the x direction. The black line is in the y direction .....	42
Figure 26: The relation between the CV and the sill of the pairwise relative variogram for lognormal distribution.....	43
Figure 27: Pairwise relative variogram for exhaustive lognormal data in figure (a) and two corrupted lognormal data with outlier high values in figures (b) and (c) in two directions. x is in violet and y is in black.....	44
Figure 28: The different value of the sill of the pairwise relative variogram for negatively skewed and positively skewed lognormal distribution.....	45
Figure 29: The light grey dots are the sill of the pairwise variogram for the lognormal distribution and the dark ones are for the uniform distribution.....	46
Figure 30: Figure (a) shows the pairwise relative variogram. The standard pairwise relative variogram is illustrated in figure (b).....	47
Figure 31: In figure (a) the pairwise relative variogram is illustrated in the black curve. In figure (b) the standard pairwise relative variogram is shown in a black curve with the traditional variogram shown in red .....	48
Figure 32: The scatter plots and histograms of normal and lognormal bivariate data .....	50
Figure 33: The relationship between two correlation coefficients based on their variograms.....	51
Figure 34: Portion sample of 1% of the data set. The traditional variogram of the lognormal data is shown as a red curve in all three figures. Figure (a) shows the traditional variogram (green curve), correlogram (violet curve) and normal score variogram (blue curve). Figure (b) shows the standard pairwise relative variogram (black curve) and figure (c) shows the correct estimation of the traditional variogram (black curve).....	54
Figure 35: The traditional variogram of the lognormal data is shown as a red curve in the two figures. Figure (a) shows the pairwise relative variogram in blue dash line and the standard pairwise relative variogram in black dash line for the exhaustive dataset. Figure (b) shows the pairwise relative variogram in blue dash line and the standard pairwise relative variogram in black dash line for the small sample of 1% of data.....	55

Figure 36: The location map of the Ag variable is shown in figure (a). The locations of the volume of influence of each outlier of the Ag variable are shown by different colors in figure (b).....	58
Figure 37: Reducing the effect of outlier value "a" by moving it further away by addition of d-value at " a " going to " a' " (redrawn from (Deutsch, Boisvert, & Deutsch, 2011)).....	59
Figure 38: Decreasing the local average by increasing the d value in two of the outliers of the Ag variable compare with uncertainty in the simulated local average. The more important spatial outlier (27.77 g/t) with <i>d</i> value of 60 m is shown in figure (a) and highest outlier (51.77 g/t) with <i>d</i> value of 63.5 m is shown in figure (b) .....	61
Figure 39: The outlier values of the Ag variable with local averages and the other values inside the volume of influence.....	63
Figure 40: The output file of the sgsim program contains the cutoff value for the median of uncertainty of the local average .....	64
Figure 41: Decreasing the local average by cutting more in outlier values compare with uncertainty in the simulated local average. The more important spatial outlier (27.77 g/t) should cut to 4.21 g/t is shown in figure (a) and the highest outlier (51.77 g/t) with should cut to 27.98 g/t to reach the cutting level of the median in average uncertainty is shown in figure (b).....	65
Figure 42: The check file to compare the OMS results with the simulated ones .	66
Figure 43: The comparison of Ag values and the Ag values with reduced outlier values .....	67
Figure 44: Parameter file of probplt_fit.....	74
Figure 45: Parameter file of lognormal_new .....	75
Figure 46: Parameter file of gam_pair .....	76
Figure 47: Parameter file of correct_pair.....	77
Figure 48: Parameter file of kt3d_new .....	79
Figure 49: Parameter file of sgsim_new .....	81
Figure 50: Parameter file of kt3d_new .....	83

Figure 51: Parameter file of kt3d\_new ..... 85

# List of Symbols

CV	Coefficient of variation
OMS	Outlier management strategy
CVE	Central variation error
MSE	Mean square error
$uc$	Under correction
$oc$	Over correction
$W_{uc}$	Weight of under correction
$W_{oc}$	Weight of over correction
$P_{uc}$	Proportion of under correction
$P_{oc}$	Proportion of over correction
CM	Combined measure
$\mathbf{u}$	A location in space
$Z(\mathbf{u})$	A random variable
$A$	An area or domain
$\mathbf{h}$	Lag separation vector
$m$	Mean
$\sigma^2$	Variance
$sd$	Standard deviation
$\hat{\gamma}(\mathbf{h})$	Calculated experimental variogram
$\hat{\gamma}_{corr}(\mathbf{h})$	Calculated correlogram
$\hat{C}(\mathbf{h})$	Covariance at lag separation vector of $\mathbf{h}$

$i(\mathbf{u}; z_c)$	Indicator transform value (0 or 1)
$\hat{v}(\mathbf{h})$	Calculated Madogram and Rodogram
$\hat{\gamma}_{PR}(\mathbf{h})$	Calculated pairwise relative variogram
$N(\mathbf{h})$	Number of pairs of data separated by a vector $\mathbf{h}$
$\hat{\gamma}_{PR,S}(\mathbf{h})$	Calculated standard pairwise relative variogram
$\rho(\mathbf{h})$	Correlation coefficient
$\hat{\gamma}_{PR,C}(\mathbf{h})$	Correct pairwise relative variogram as an estimate of true variogram
$\gamma_{ref}$	The traditional variogram of original data
$Z_{out}$	Outlier value
$\bar{\lambda}_{out}$	Average weight of outlier inside the volume of influence
$Z_{oth}$	Average of other data inside the volume of influence except outlier
$Z_V^*$	Average estimated values inside the volume of influence

# Chapter 1 – Introduction

## 1.1 General Problem Area

The mining industry is of great societal importance due to the quantity of resources needed to sustain and develop our quality of life. The mining industry extracts many different commodities including metals, coal and gold. A mineral resource is a portion of a deposit that may become of economic value. Mineral reserves are resources that are technically and economically feasible to extract. Determining the location, extent and grade in an ore body is an important part of resource estimation.

Resource estimation often proceeds by constructing block models that divide the deposit into a three dimensional array of blocks. The grade in each block is estimated using the nearby data. Block models help to understand the value of the mineral deposit and the best sequence of extraction. Global resource estimation is needed for financial planning in early years. In later years, during production, local estimation is required for specific goals such as mine planning, scheduling and grade control.

Geostatistics has evolved and proved its value since the late 1950's for estimation of grades. The integration of geology and geostatistics is routinely considered in resource estimation. A hierarchical framework is often applied for resource estimation. Some steps include the examination of the geology of the mineral deposit, careful data analysis, quantification of the spatial character of the grades in the deposit, then estimation and perhaps simulation of a block model. Simulation is used especially for uncertainty quantification.

The block model supports decision making; the drillhole data are used in the block model construction. There are concerns if the block model does not

accurately represent the deposit. The block model should be globally unbiased. Care must be taken not to overstate or understate the resource. Block models should minimize conditional bias to avoid overstating or understating the grade conditional to an estimate. Oversmoothing or undersmoothing will exacerbate conditional bias (Glacken & Snowden, 2001).

A recurrent problem in the estimation and simulation of a block model is extreme values or outliers. These values make the variogram noisy and difficult to interpret. They also introduce errors in the block model estimates or simulated values.

## **1.2 Problem Definition**

Most of the grades in mining deposits are characterized by highly skewed distributions which contain some extreme values. Extreme values are high grade measurements that will be used for resource estimation without change. The word outlier is used for high grade measurements that are treated somehow, for example, limited to a maximum grade value. An outlier can be defined as an observation that seems erratic when compared with the vast majority of the data values (Sinclair & Blackwell, 2002). If the outlier is a result of an error, it should be investigated and removed from the dataset. This study considers that the outliers are valid analysed samples; errors have been corrected in the dataset, but the unwarranted use of extreme values may introduce errors in the resource estimation.

Many of the variables in mining are grades that are mass fractions of some mineral species. Some of the data distributions have low grades outliers, but it is not very common (Rossi & Deutsch, 2013). The data distributions are mostly positively skewed since there cannot be negative grades. Extreme high grades are more common in precious metal deposits. These high grades have great economic value and must be treated carefully. Arbitrarily eliminating them to reduce their

influence could cause an underestimation of the value of the deposit. On the other hand, leaving outliers untreated may lead to overestimation particularly close to the high grade outlier samples.

Outliers influence the global distribution and preliminary resource estimates. Resource estimation requires statistical parameters for modeling. Outliers strongly affect the statistical parameters such as the mean, variance, correlation coefficient to secondary data and the variogram. The major issue is the estimation of high grade areas that may not exist.

Outliers have a strong impact on the variogram measure of spatial continuity. In a geostatistical study, the spatial continuity is an important aspect of local estimation (Isaaks & Srivastava, 1988). By improving the quality of the variogram, the estimation of local grades would be improved.

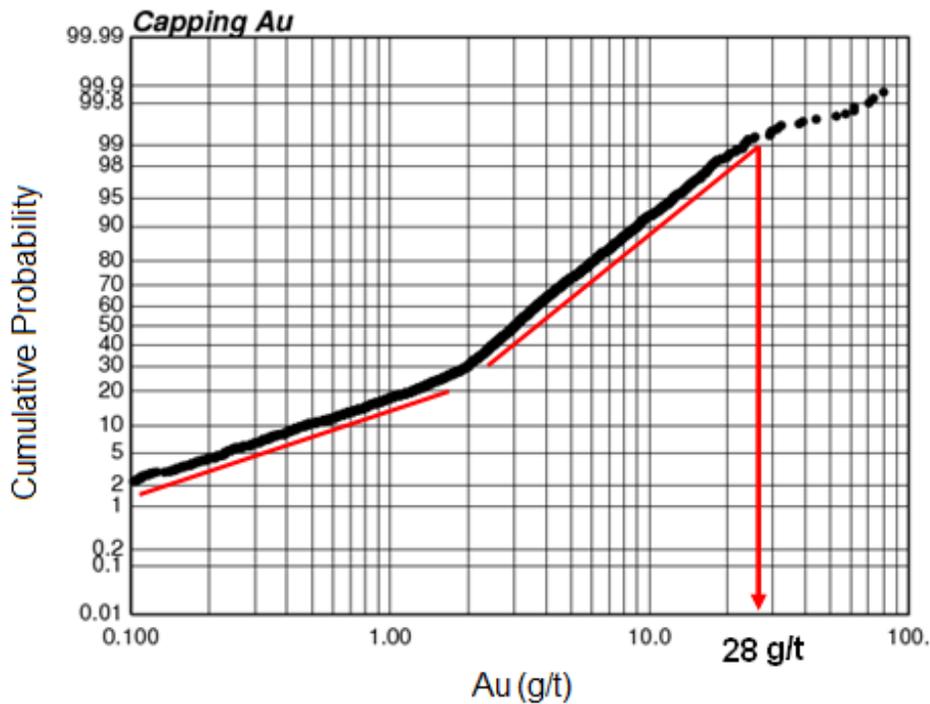
There is a need to manage the outliers through the resource estimation process. From the beginning of the procedure of resource estimation including collecting and interpreting the data to the end where reasonable decisions should be taken.

## **1.3 Literature Review**

### **1.3.1 Outlier Detection**

Outliers may be recognized by some statistical parameters such as the mean plus or minus two or three standard deviations (Sinclair & Blackwell, 2002). If data is outside of this interval, it might be considered an outlier. Outliers may be detected by graphical tools such as histograms, probability plots and scatter plots. Detecting extreme values with graphical and statistical methods does not necessarily indicate that the data is an outlier; these techniques illustrate the difference of these data with the vast majority of the dataset.

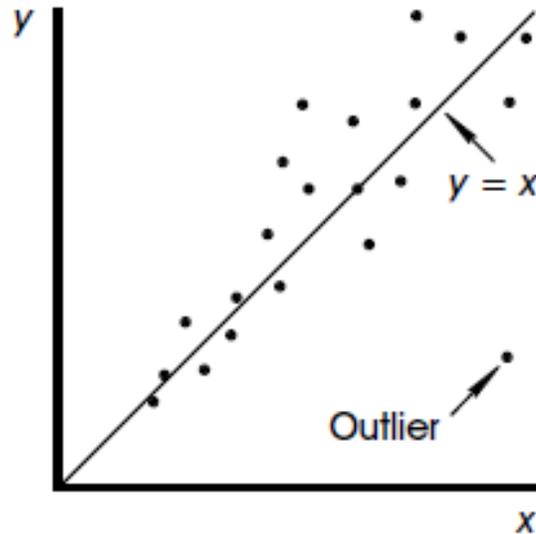
Histograms are an effective tool to show the grade characterizations. The shape of the distribution is observed. The lognormal cumulative frequency or probability plot is an appropriate tool to observe outliers. Probability plots are more useful than the histogram because each data point is shown. An example of a lognormal probability plot of gold grade is shown in Figure 1. The points at the high end of the distribution can be considered as extreme values. There is less than 1% of data above 28 g/t. Yet these values do not fit nicely with the rest of the high grades. These values may need to be controlled to reduce their influence on resource estimation.



**Figure 1: Probability plot of Au grade (Deutsch, 2011)**

Scatter plots can also be used in the data evaluation process to demonstrate the relationships between the grades and each of the geologic variables. Scatter plots

can be used for quick checking of outliers in both duplicate analytical data and correlated variables (Sinclair & Blackwell, 2002). In some cases, the data are expected to follow a well-defined relationship. If a data value is far from the 45° line, it can be considered as an outlier (Figure 2).



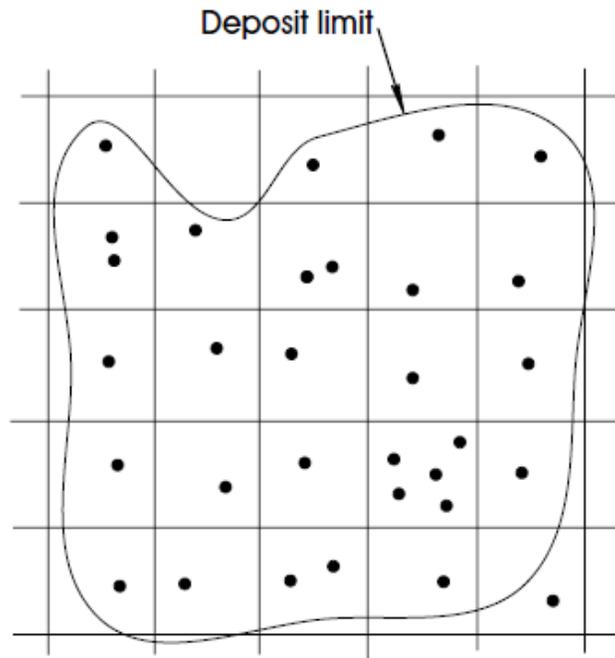
**Figure 2: Scatter plot of paired data with an outlier (Sinclair & Blackwell, 2002). No units are given, but variables x and y are grades measured on duplicate samples**

### 1.3.2 Outlier Management for the Histogram

Outliers can be left alone, cut or capped to a certain limit or treated specially in subsequent geostatistical modeling. There are often local customs within certain companies or mining regions; these local customs should be considered. Companies or geological provinces may have specific standards of capping such as 1.0 oz/ton or a fixed quantile of the distribution.

In much of resource estimation, the histogram is used in various methods to characterize the mineral deposit. The grades may be preferentially clustered, so

the influence of location bias in a histogram should be detected and removed. One method for correcting the location bias is the polygonal declustering method which weights the data values proportional to polygonal areas or volume influence (Isaaks & Srivastava, 1989; Sinclair & Blackwell, 2002). Another approach is cell declustering where each cell of a grid contains one or more data are equally weighted (Figure 3). This helps estimate an unbiased histogram.

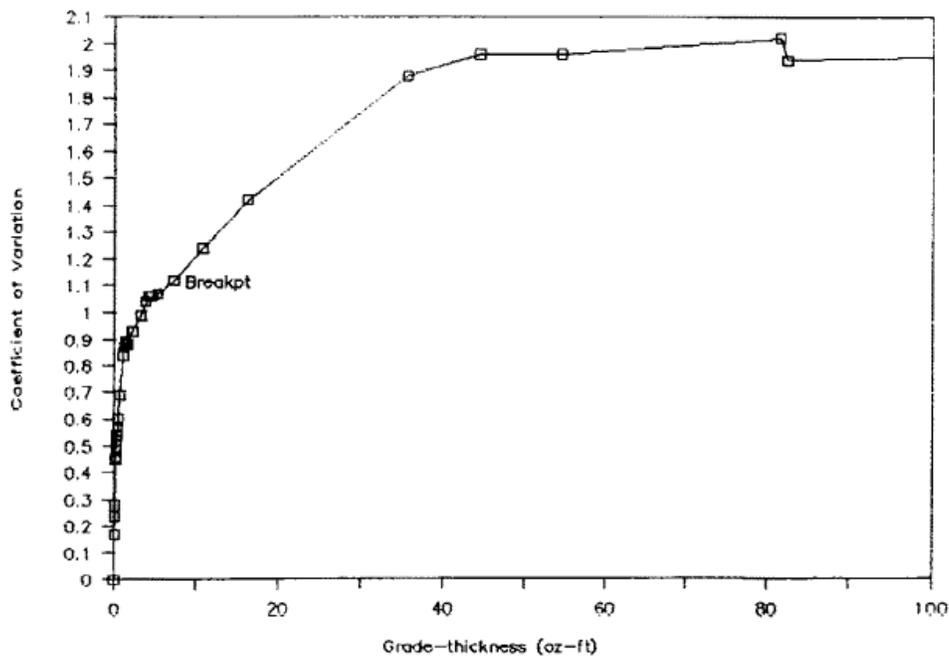


**Figure 3: Clustering of a data variable to produce an unbiased histogram (Sinclair & Blackwell, 2002). This figure is not scaled because it is a fraction of a larger domain as an example to illustrate the clustering method**

Some geostatistical methods have been developed to deal with outliers in highly skewed distributions. One such parametric method is disjunctive kriging (Matheron, 1976), which depends on the stationarity assumption and indicator. A nonparametric method is probability kriging, a technique that needs a separate variogram and kriging for each cutoff (Parker, 1991). The simplified indicator method to dealing with outliers is proposed (Journel & Arik, 1988), but an

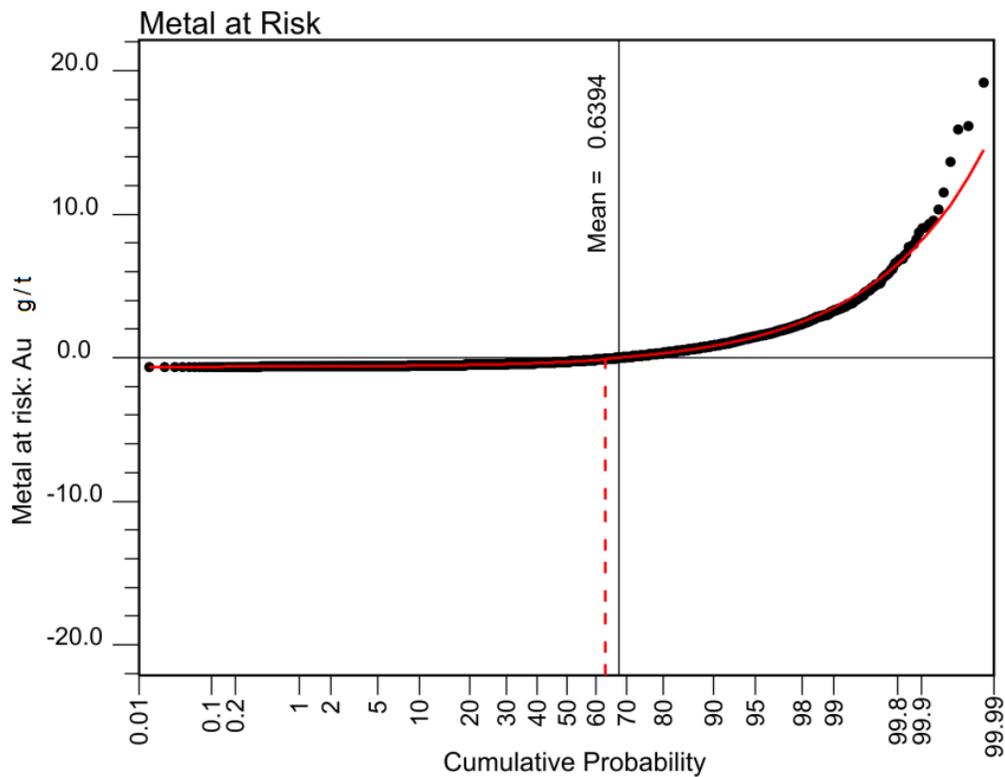
arithmetic average is used to model the mean of the last class, which may lead to an inaccurate estimate of the mean if the data considered in average calculation are outliers.

Parker proposed a method to merge the simplified indicator kriging and modeling of the class mean (Parker, 1991). This method uses the probability plot to create multiple subpopulations for recognizing the outliers by breaking the original distribution into two less skewed lower tail and strongly skewed upper tail. Then, a specific distribution is assigned to the upper tail of the grade distribution. The lognormal distribution could be one of the assigned distributions. The breakpoint for dividing the distribution into two is based on the cumulative CV (Figure 4). The point that the cumulative CV accelerates fast, the influence of outliers on upper tail increase and that point should be considered the breakpoint. These methods are all empirical and based on visual inspections.



**Figure 4: Cumulative CV for specifying the breakpoint (Parker, 1991)**

The metal at risk is another tool which is proposed to measure the potential effect of outliers (Deutsch, 2010). The metal at risk is defined as the difference between the data value and the mean. The metal at risk would be high for outlier high values in a lognormally distributed data. An illustration of metal at risk for the goldcopper.dat dataset is shown in Figure 5. The data points at the upper right of the figure indicate the high values which their effect decreased by reducing them to the solid line fitted model.



**Figure 5: Metal at risk plot for goldcopper.dat (Deutsch, 2010)**

### 1.3.3 Outlier Management for the Variogram

Variograms are an essential input in geostatistical calculations. The variogram is a measure of the spatial variability of the phenomena under study. The variogram is

a spatial correlation function that is used in kriging to estimate grades at the unsampled locations. The variogram model derived from a good experimental variogram has useful information of data. The variogram is also used to determine the effect of change of support in volume variance relations. A robust variogram is important in resource estimation.

The correlogram is defined as the correlation at points separated by distance  $h$ . The correlogram may be more stable than the variogram and more resistant to erratic data as it accounts for the local head and tails means and variances at different lag distances (Isaaks & Srivastava, 1989). The correlogram is commonly used when facing the high skewed data distribution in geostatistics.

Practice has shown that the noise is often best dealt with by the pairwise relative variogram. This relative variogram seems to be one of the most robust alternatives to the traditional variogram (David, 1988). The pairwise relative variogram is calculated using a method similar to the traditional variogram. The difference between them is a weight in the denominator, which serves to reduce the influence of outliers and clustered data.

#### **1.3.4 Outlier Management for Block Modeling**

Block models should minimize overstating or understating the grade. Oversmoothing where high grades blocks are underestimated and low grades blocks are overestimated is a result of conditional bias which can be solved by choosing the proper block size in kriging. Undersmoothing as another result of conditional bias happens where high grade blocks are overestimated and low grade blocks are underestimated which the solution could be change in grade relationship in either the block estimate or data assays (Glacken & Snowden, 2001).

One of the best unbiased block estimations is the kriging estimation, which is a distribution free estimator. In a skewed data set, the kriging estimation would not

be ideal. Robust kriging is presented to deal with the highly skewed data set (Cressie & Hawkins, 1980). The use of robust kriging in ore body modeling and resource estimation is discussed later by Costa (Costa, 2003). In robust kriging (RoK) the high values are picked by some choice of the threshold and edit the high values in RoK based on the kriging variance. This kriging variance illustrates the difference between the high value and the weighted median of local data.

The geostatistical model (e.g., block model) should be able to deal with outlier high values in the resource estimation process. A top-cut model is a model which is developed to deal with outliers and is supposed to reduce the conditional bias (Rivoirard, Demange, Freulon, & Lécureuil, 2012). This method is based on dividing the variables into truncated grades, weighted indicator above top-cut and residuals. The estimation in this procedure is based on the truncated grade and the indicator cokriging and a separate kriging for residuals, because the residuals are spatially independent.

#### **1.4 Thesis Outline**

In this thesis the goal is to find the best way to deal with outliers. The outlier values could cause problems in resource estimation.

This thesis consists of six chapters and an appendix. Chapter 2 is based on univariate statistical approaches to detect and manage the outliers. Spatial outliers are explained and detected. The probability plot which is one of the best tools to observe high values is used to reduce the value of outliers and manage them.

Chapter 3 describes the importance of a good variogram for interpretation and modeling. The experimental variogram is sensitive to high values. The pairwise relative variogram is used as an alternative to the traditional variogram in dealing with outlier high values. The two issues in the pairwise relative variogram, (1) an unknown sill, and (2) its convergence to the wrong variograms are discussed and resolved in this chapter.

Chapter 4 is proposed methodologies to reduce the effect of outliers in block estimation. The volume of influence of an outlier is defined. The value of an outlier is decreased to reduce its influence on the nearby grid blocks.

Chapter 5 presents a conclusion of this study and a discussion about conceivable adjustments to the proposed methods and future works.

The references are presented in Chapter 6. The FORTRAN programs and parameter files written and used for this study are shown in the appendix.

# Chapter 2 - Univariate Management

## 2.1 Univariate Statistical Approach

Data analysis using statistical tools is an essential step in every mineral resource estimate. Statistics provide mathematical tools that help in collecting, organizing and interpreting data. The probability distribution of one variable at a time is called a univariate distribution. Univariate statistical analyses such as the histogram, mean, standard deviation and probability plot are useful in summarizing the data and helping to understand the data quality. The univariate approach should be able to recognize the outlier high values in the process of data checking. The graphical tools such as the histogram and probability plot can be used in both identifying and managing the outlier high values.

Extreme values are valuable data as they define the richest areas, but they should be controlled to reduce their inordinate influence on modeling and estimation. The mean is a statistical measure that is very sensitive to extreme values and has a large effect on resource and reserve estimations. The proposed outlier management strategies such as cutting to a fixed threshold and using the probability plot to reduce the high values to a fitted population are possible ways to correct the sample mean and reach the population mean.

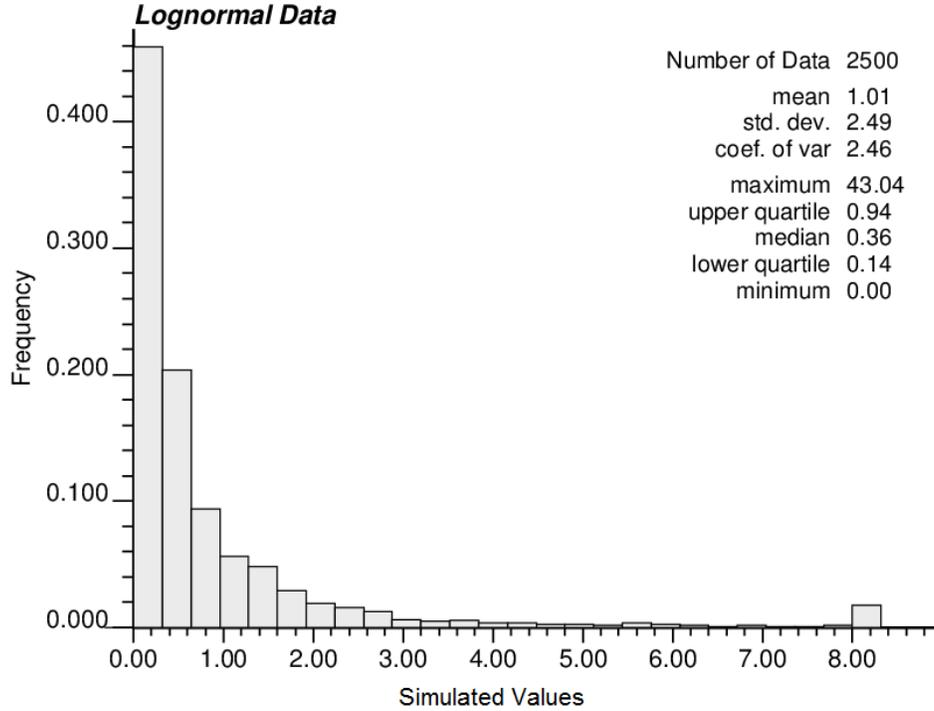
### 2.1.1 Cutting to Various Thresholds

Cutting or capping the outliers to a fixed threshold value is a common strategy when facing outliers. There is a possibility of some high estimation after capping depending on the choice of a capping threshold (Rossi & Deutsch, 2013). The proposed outlier management is based on the mean of the population and the mean of the samples drawn from the population. The hope is to identify samples

where the mean is higher than the population mean, and to figure out the characteristics of the high values that cause the higher mean. In real datasets, the distribution of data and the population mean are unknown. For that reason, a simulated lognormal dataset with a known mean is chosen to examine this method. Then, by using the spatial bootstrap sampling program (Deutsch, 2004), a fixed number of samples are taken from the dataset. The distribution of uncertainty to the mean is going to be Gaussian based on the central limit theorem. The goal is to manage the mean values above the population mean, but to leave the samples with mean values below the population mean.

The Gaussian and lognormal distribution are the two most important univariate distributions in a geostatistical study. The Gaussian distribution is important because it is used extensively in uncertainty characterization and simulation. It is fully characterized by its mean and standard deviation. The lognormal distribution is widely used in geostatistics because most of the mining and earth science variables are non-negative and highly skewed. The lognormal distribution is also characterized by its mean and standard deviation in either arithmetic or logarithmic units (Rossi & Deutsch, 2013). The univariate distribution that is used in here is lognormal distribution because the outlier high values exist more in lognormal distribution.

Consider a 2-D spatial domain with values from a lognormal distribution with a long tail and some very high values. The population mean is 1 and the standard deviation is 2.5. The histogram of this lognormal distribution is shown in Figure 6.



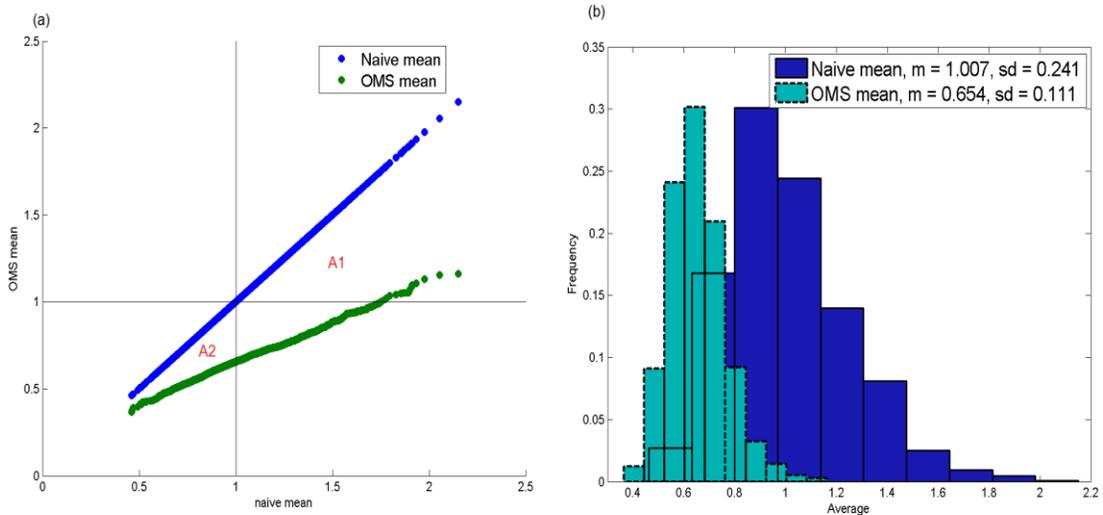
**Figure 6: Histogram of lognormal data**

Quantiles could be used to determine a cutting value based on the data distribution. For example sometimes the data are capped to the 0.9, 0.95 or 0.99 quantiles. In this lognormal distribution the three values of 2.10, 3.56 and 12.14 are the values for the P90, P95 and P99. Samples of 100 data are taken from this dataset for 2500 times. The distribution of uncertainty in the mean will be approximately Gaussian based on the central limit theorem. The mean of this distribution is close to the population mean. Figure 7, Figure 8, and Figure 9 shows the result of capping the data to the 90, 95 and 99 percentiles. In plot a's the naive true mean is plotted against the outlier management strategy mean. The proposed method attempts to reduce the risk of observing mean values that exceed the population mean values and to optimize (maximize) the difference between spaces  $A_1$  and  $A_2$ :

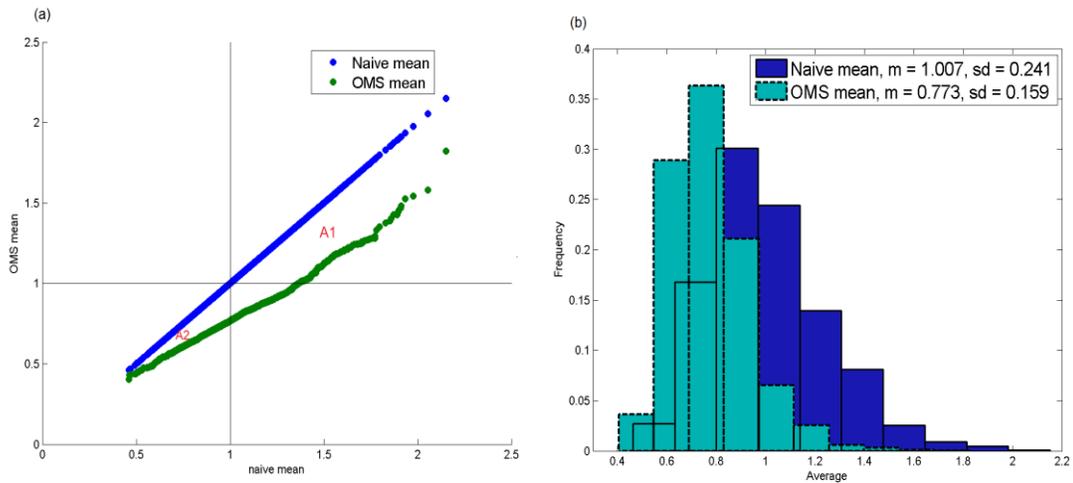
$$\text{Max} \quad \text{objective function} = A_1 - A_2 \quad (1)$$

Section  $A_1$  represents the space between the naive mean and the outlier management strategy corrected mean on the upper right quartile. The  $A_2$  illustrated this different space on the lower left quartile (e.g., Figure 7-a). The  $A_1$  section should reduce as much as possible because the means in this section correspond to those that are higher than the population mean. But the averages in section  $A_2$  should not reduce much because the averages in section  $A_2$  are already below population mean.

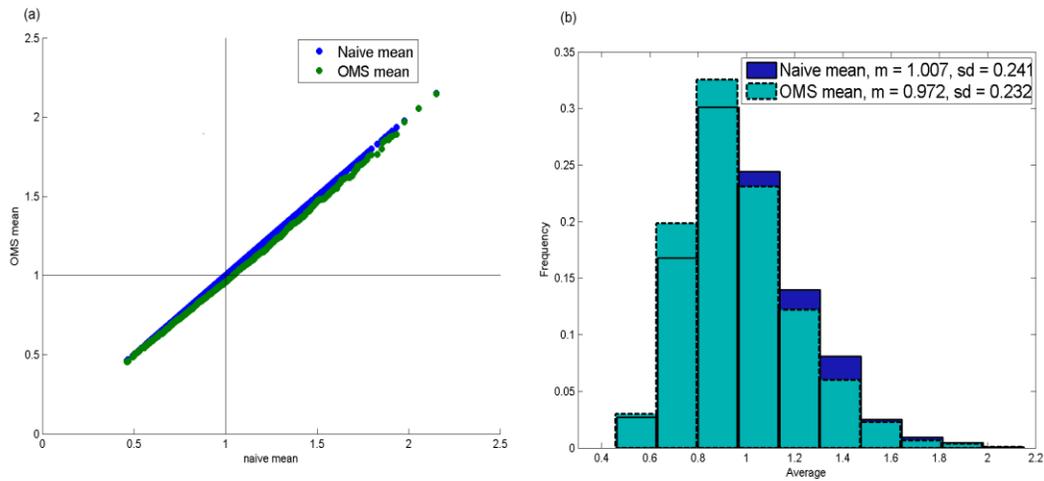
In plot b's the histogram of the distribution of the corrected mean is compared with the original data mean distribution.



**Figure 7: Cutting the values above 90% of the data to the value of 2.10. In plot (a) the means in section  $A_1$  are reduced but the averages of  $A_2$  decrease significantly. In plot (b) the histogram of the corrected mean is much thinner**



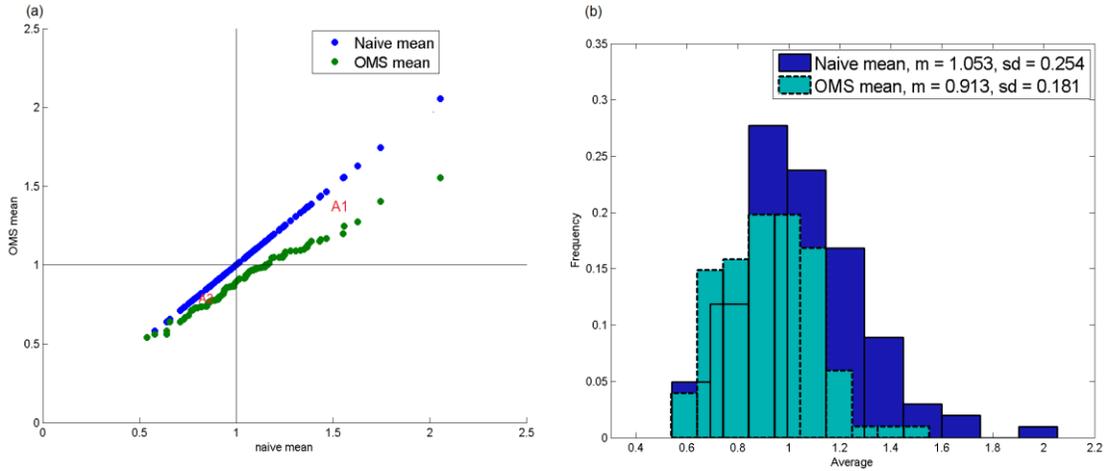
**Figure 8: Cutting the values above 95% of the data to the value of 3.56. In plot (a) the averages in both sections reduce properly and are more acceptable. In plot (b) the histogram of the corrected mean reduces well in the upper side of the population mean, but the deduction in the lower part is still high**



**Figure 9: Cutting the values above 99% of the data to the value of 12.14. In plot (a) the averages in both sections change slightly. In plot (b) the histogram of the distribution of the corrected mean is almost the same as the naive mean**

### 2.1.2 Manual Method

Cutting extreme value could be done manually by observing the probability plot of each sample and reducing the high values based on visual inspection. One hundred of the previous samples are taken randomly and the high values are reduced. The result is illustrated in Figure 10.

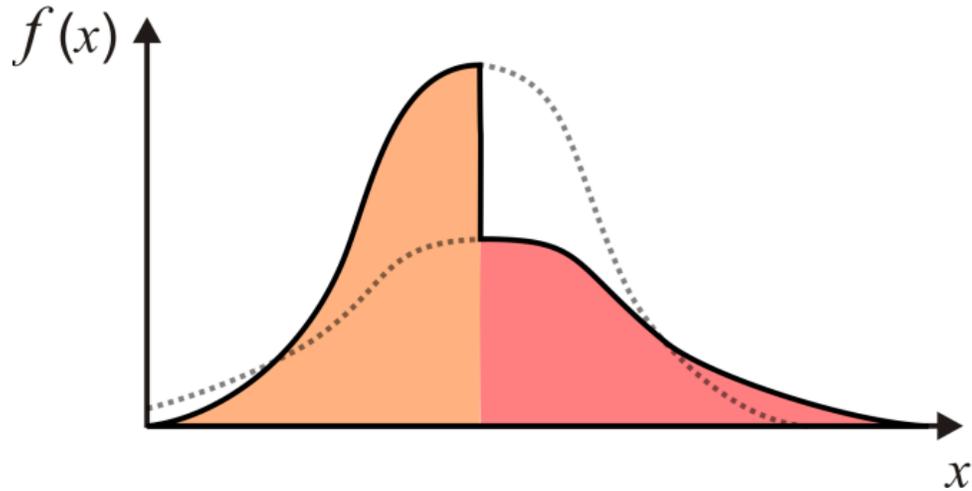


**Figure 10: Manually reduce the outliers of the 100 samples. In plot (a) the means in  $A_1$  reduce more than the averages of  $A_2$ . In plot (b) the histogram of the corrected mean is reduced properly**

### 2.1.3 Probability Plot Fitting

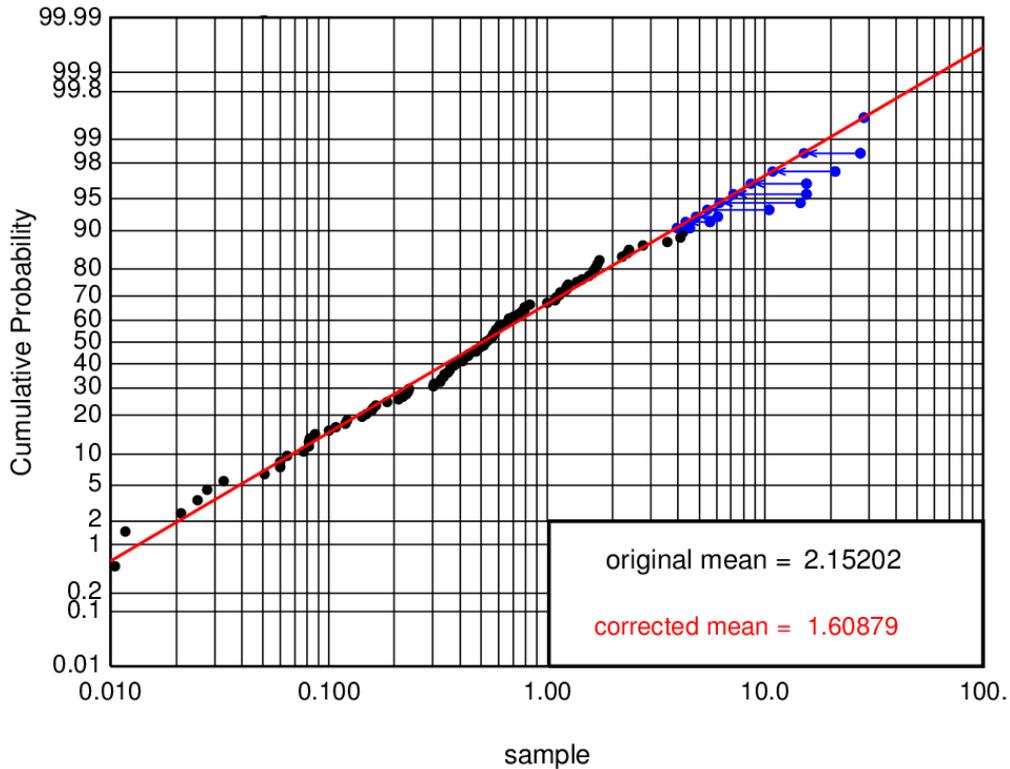
Extreme values could also be observed through probability plots. A lognormal probability plot with fitted lognormal distribution could be used to divide the data distribution into two lognormal distributions with an arbitrary threshold (Parker, 1991). A lognormal fitted model permits the extreme values to be fitted back to a straight line on the lognormal probability plot. The fitted nonparametric model used in this study assumes that the population is separated based on one

probability threshold. The assumed probability distribution functions in a fitted nonparametric model that separates two distributions (see Figure 11).



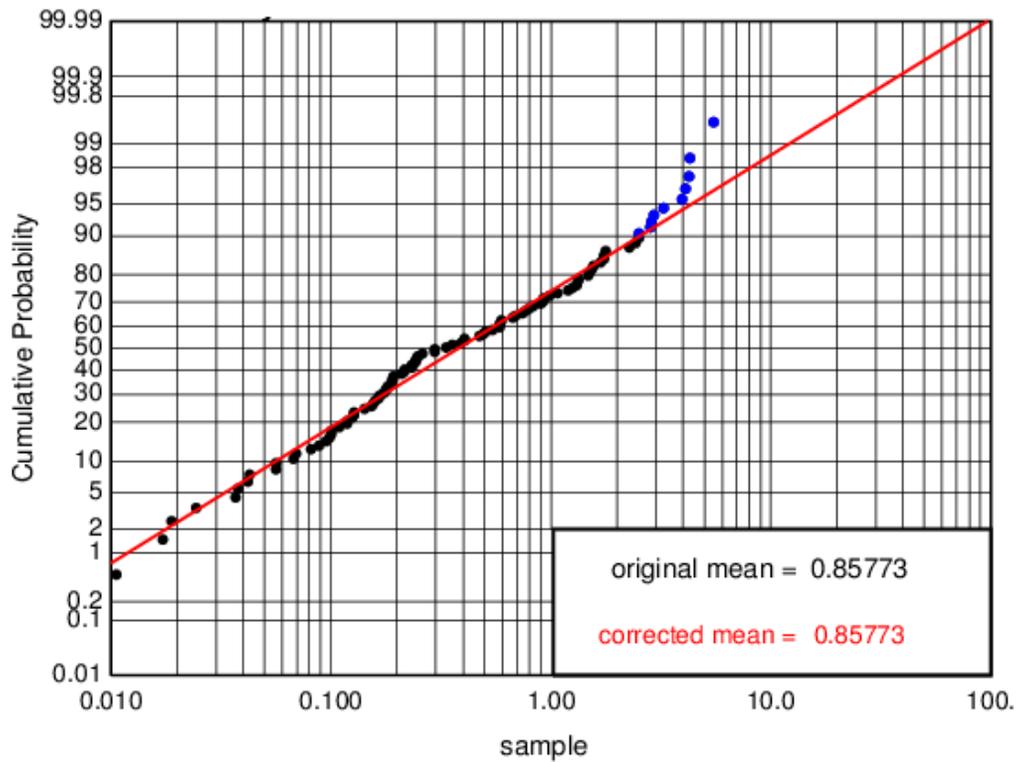
**Figure 11: Probability distribution assumptions associated with the fitted non-parametric model (Deutsch, 2010)**

One of the samples from the previous example of lognormal distribution with the highest mean is picked to show the procedure of identifying the outliers on the upper lognormal fit to the data values below an arbitrary threshold. After reducing the outlier values, the mean of the samples is computed and shown (see Figure 12)



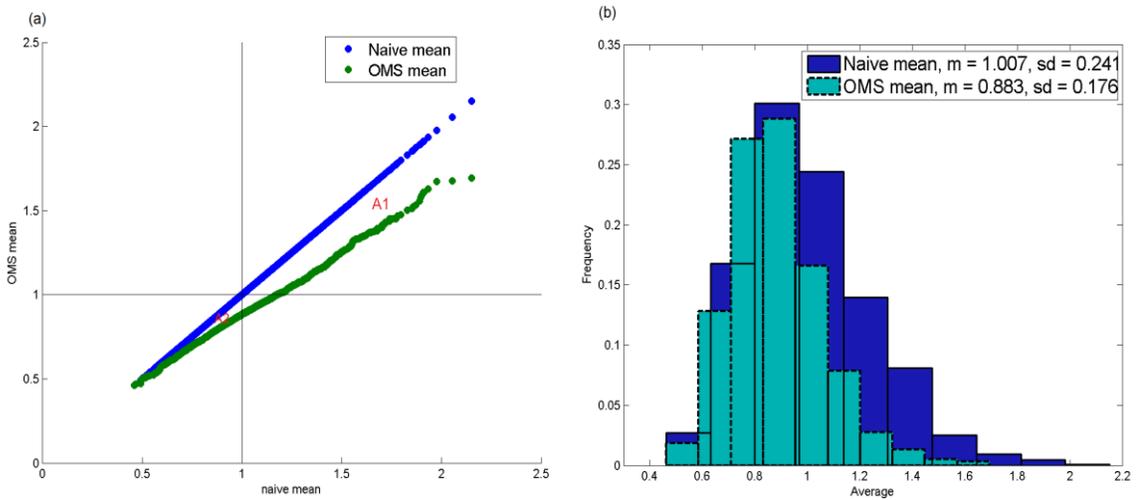
**Figure 12: Probability plot of one sample of lognormal distribution with corrected outlier values for the arbitrary threshold (e.g., 90)**

If the outlier high values are less than the red fitted lognormal they will be left alone because increasing the values might introduce a bias not based on data (see Figure 13). In some cases these high values do not cause a significant increase in the average of the samples, but sometimes they cause high averages that are actually less than the fitted model. Therefore, leaving them untreated would cause some under correction in the final results.



**Figure 13: Probability plot of another sample with outlier high values that are less than lognormal fitted model**

The result of correcting the means of all samples using the fitted lognormal probability is shown in Figure 14.

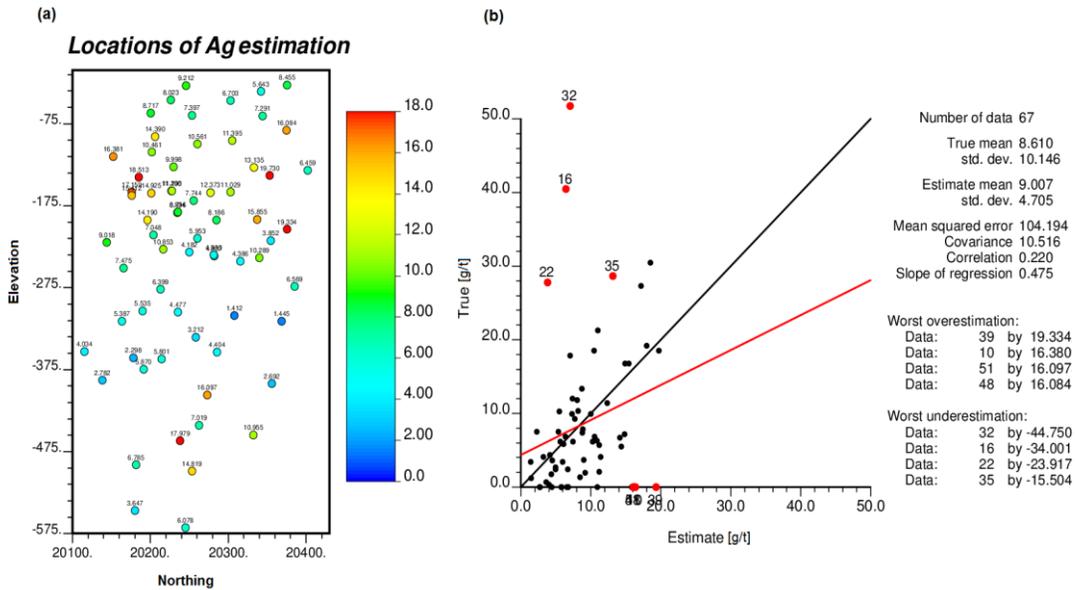


**Figure 14: Reduce the outliers of the 2500 samples with lognormal fitted model. In plot (a) the means in  $A_1$  reduce but the deduction of averages in section  $A_2$  are minimized compared with other methods. In plot (b) the histogram of the corrected mean is reduced properly**

## 2.2 Spatial Outliers

The distributions of data are often highly skewed. In Geostatistics a major concern is related to spatial outliers. These spatial outliers are data values that are unusually high with respect to their nearby data. An extreme high observation among other high values and few low values often has less of an impact on block model estimations. High values among very low values are considered as spatial outliers and may need to be controlled. The proposed detection technique based on the uniform rank transform of data and its estimation identify the spatial outliers.

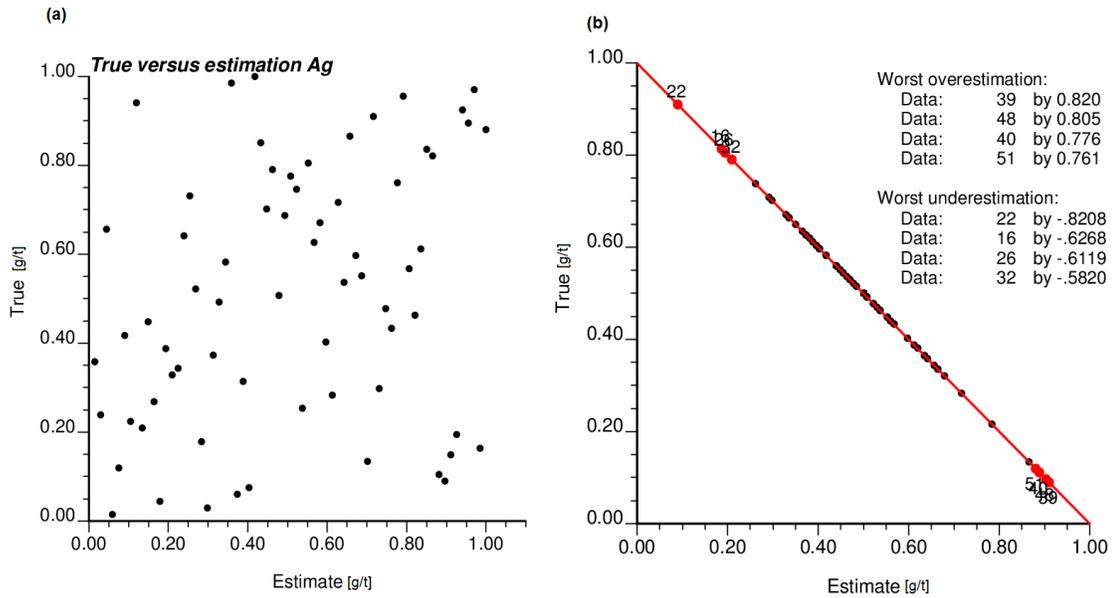




**Figure 16: (a) Location map of estimation Ag and, (b) the scatter plot of the true and estimated values**

On the scatter plot of the true versus estimation of the silver variable (Figure 16-b), the biggest value is 51.77 g Ag/t (data number 32), which is considered the worst underestimation at its location. This data is surrounded by both high and low values. The other data (data number 22) with the value of 27.77 g Ag/t is surrounded by two zero values and the other nearby data are low as well. The underestimation at the location of data 22 is higher than the underestimation at location of data 32.

In the proposed method for detecting the outliers, the data values and their estimations are rank transformed to become uniform between zero and one. These ranked true values and estimates are plotted and the projection of the data points on the 135 degree line is calculated (see Figure 17). The outliers detected by this method are different than simply considering the univariate distribution.



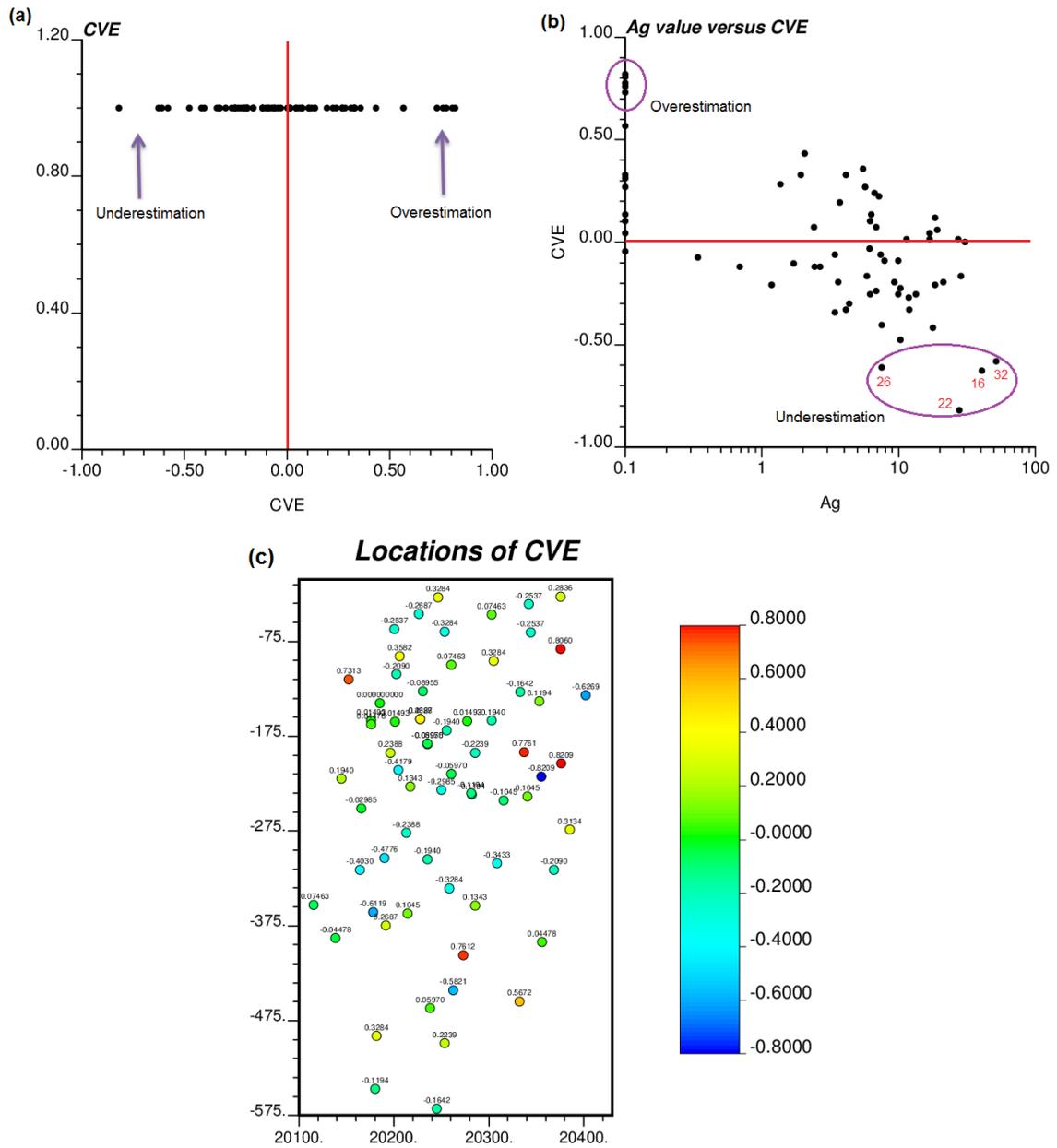
**Figure 17: (a) Scatter plot of uniform rank transformed of true and cross estimation of Ag values and, (b) projection of data points on 135degree line**

As shown in Table 1, data number 22 with the lower value compared to data 16 or data 32 causes a higher underestimation on its location and overestimation on nearby data (data number 39 and 40).

**Table 1: Different outlier detections of the silver variable**

Previous works				Current works			
Worst underestimation		Worst overestimation		Worst underestimation		Worst overestimation	
Data number	Data value	Data number	Data value	Data number	Data value	Data number	Data value
Data32	51.77 g/t	Data39	0 g/t	Data22	27.77 g/t	Data39	0 g/t
Data16	40.46 g/t	Data10	0 g/t	Data16	40.46 g/t	Data48	0 g/t
Data22	27.77 g/t	Data51	0 g/t	Data26	7.54 g/t	Data40	0 g/t
Data35	28.64 g/t	Data48	0 g/t	Data32	51.77 g/t	Data51	0 g/t

The 135 degree line in this paper is defined as the central variation error (CVE). The range of this line is between -1 for underestimation (high value outliers) and 1 for overestimation (low value outliers). The first two plots on Figure 18 show the CVE line and its behavior against the silver value. In other words, the negative part is defined by the estimations that are lower than their actual value and the positive section is related to the estimated values that are higher than their original data value. The values of CVE based on their location are shown in Figure 18-c. The range of the location map is changed from blue for the highest underestimation to red for highest overestimation.

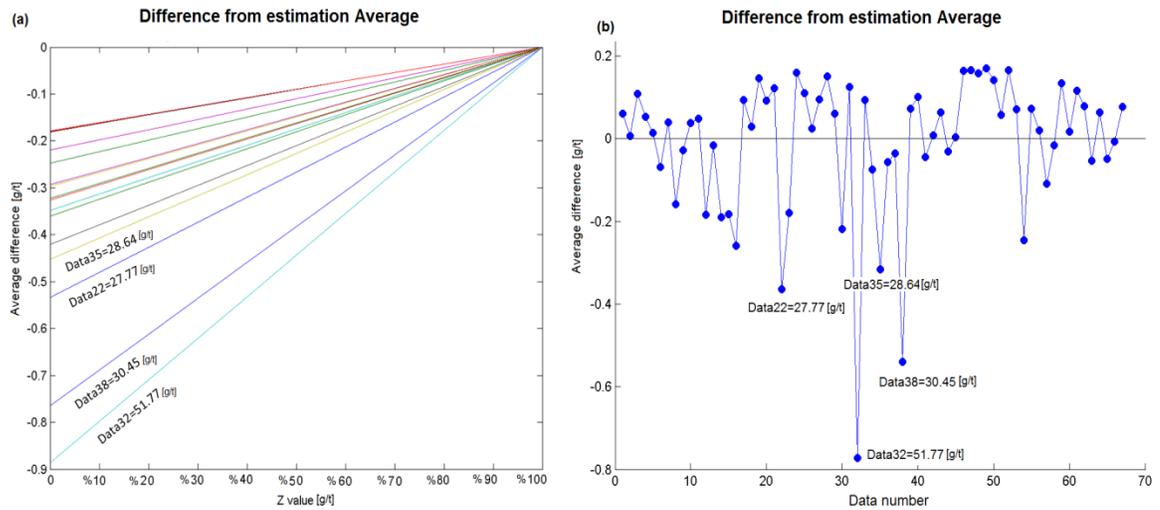


**Figure 18: (a) Defining the range of CVE, (b) Scatter plot of silver values and CVE, (c) The location map of the value of CVE from the highest underestimation (blue) to the highest overestimation (red)**

### 2.2.2 Sensitivity of Resource to Spatial Outliers

In a geostatistical study, the high value outliers are of greatest concern. In studying the outliers, it is preferable to define a threshold value. The values investigated are the resources such as global average, average above cutoff, fraction above cutoff and quantity above cutoff. Thresholds can be considered as a test of sensitivity of extreme values to these local and global resource estimates.

Values above a certain cutoff grade are of the greatest interest. To study this issue, consider the silver variable introduced above. The upper half of the data is selected. Then, by reducing the value of each data one at a time, the estimation resources will decrease (see Figure 19-a). Another approach is to eliminate each data value. To study the global average in this way, each time the estimation process is done without one of the data and the average of estimation values is computed (see Figure 19-b). As shown in Figure 19, data number 32 has the biggest value and the greatest effect on reducing the global average in both methods. Its influence is even more than that of data number 22, which might be considered as the more important spatial outlier. One explanation for this effect could be the higher underestimation of the value of data 22, which reduces its effect on the global mean with no cutoff threshold.

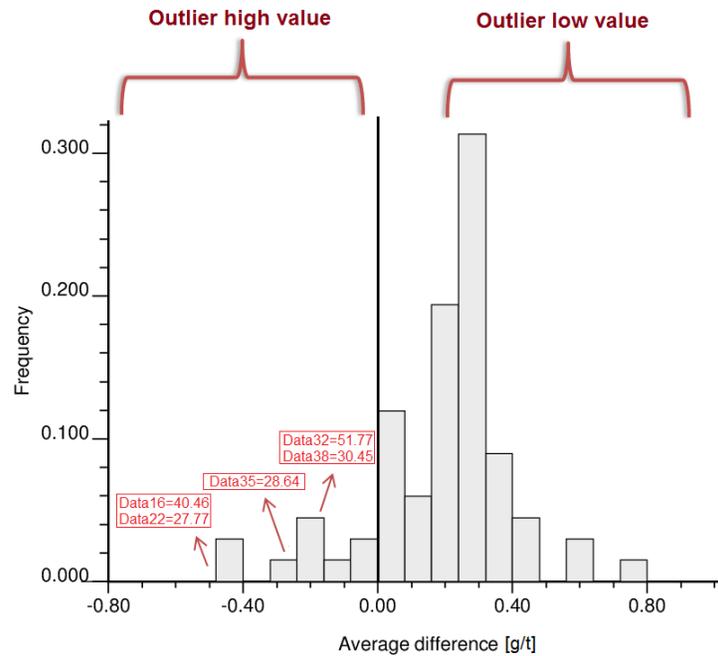


**Figure 19: Difference from the estimation average: (a) by reducing the data value of Ag and, (b) by removing each data value**

### 2.2.3 Effect of Outliers on Grades above Cutoff

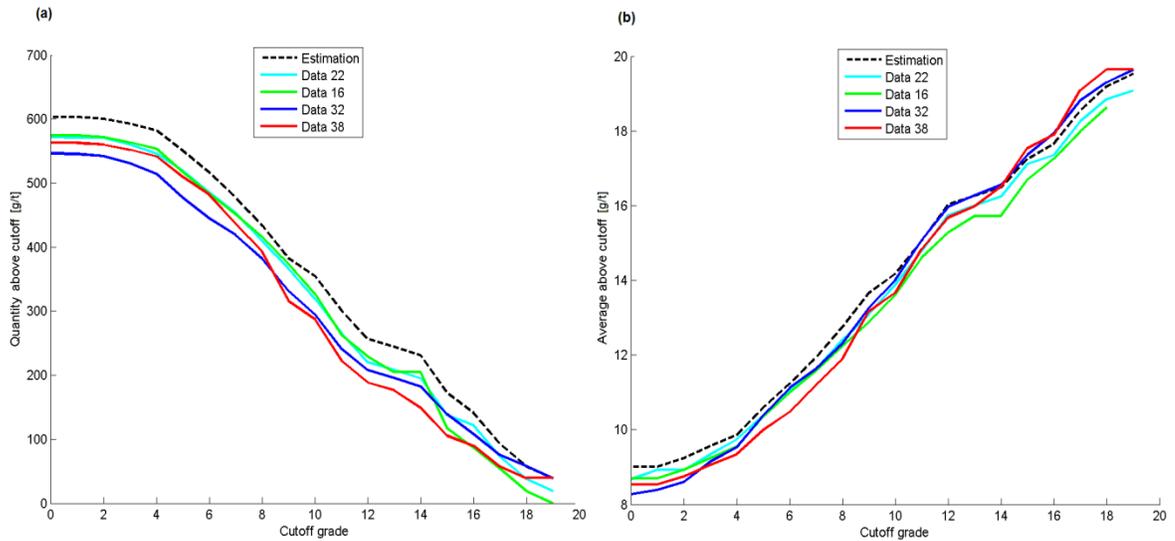
By defining the cutoff grade, the fraction of data above the cutoff, the average of ore grade and the quantity of resource above the cutoff are the factors that should be considered when faced with outliers. To see the impact of outlier high value on the average and quantity above cutoff, the silver variable is once again examined.

The silver values above this cutoff value are selected. The average of removing each data in turn is calculated and compared with the estimated average above the cutoff. The histogram of these changes in average is shown in Figure 20. The two outliers (data numbers 16 & 22) identified above have the largest impact on the average above cutoff.



**Figure 20: The histogram of changing the average of estimation in absence of each data for the cutoff grade for silver**

With a cutoff value of zero, all of the data values are treated as ore. The tonnage is equal to the entire area tonnage and the average ore grade will be the global mean (Isaaks & Srivastava, 1989) . As the cutoff increases, more values will be rejected as waste and the quantity of ore drops. Figure 21-a illustrates the decrease in quantity of silver value by increasing the cutoff grade. With a low cutoff, the highest outlier has a greater effect on the quantity of metal above the cutoff. Figure 21-b shows almost the same behavior by outliers. The two most important outliers (data numbers 16 and 22) have the greatest impact on the average above cutoff.



**Figure 21: For the different cutoff value, the change in quantity (plot-a) and average of ore (plot-b) for the estimation value and four of the more important outliers**

## 2.3 Recommendation

### 2.3.1 Apply Univariate Analysis

In univariate analysis, the statistics of the data is treated statistically. The high values above some thresholds are chosen to be cut. In this method the cutting threshold can be chosen arbitrarily. Cutting all the high values above some fixed value to that value is generally accepted among geostatisticians. Most of the real world mining data are highly skewed. Choosing the 95<sup>th</sup> or 90<sup>th</sup> percentiles can cause a huge deduction in the estimated resource. In this situation the outlier high values are over corrected. On the other hand, the 99<sup>th</sup> percentile may be too small and may not change the outlier high values enough. There should be some measure to decide the threshold that has the best results.

The other univariate analysis based on the lognormal fitted model has a different work flow other than just cutting to a fixed threshold; the amount of cutting

depends on the lognormal probability plot. This method seems to produce the best result with respect to not over correcting. Applying the fitted lognormal probability is simple and just needs a proper cut off threshold to reduce the outliers above that.

### **2.3.2 Spatial Analysis**

In spatial analysis extreme values could be examined spatially. Not all the high values are considered as outliers unless they have a significant influence on their nearby data. Spatial analysis concentrates more on outlier detection. If the interest is to find the spatial outlier then manage them, this method could be useful enough.

### **2.3.3 Devise a Combined Approach**

Univariate analysis suggests different threshold values to cut the outliers such as different quantiles, manual cutting and deduction based on the lognormal probability fitted model. In all of these cutting methods, there might be some under or over correction which means managing outliers inappropriately. Over correction is a problem that needs attention because it is the most common error introduced by over-zealous outlier management.

Small samples from a large simulated 2D data set are chosen to examine the results of cutting to each quantiles, the manual method and the probability plot that are shown in Table 2. After cutting the outliers, the mean of each sample is computed and the mean square error (MSE) of these averages from the mean of the samples before correction is computed.

$$MSE = \{(m^* - m)^2\} \quad (2)$$

Where  $m^*$  is the corrected mean after cutting and  $m$  is the original mean of the sample.

As mentioned above, avoiding under correction is more important. To avoid overstating the resource, weight can be introduced for computing a combined measure of both under correction and over correction.

$$\text{Under correction weight: } W_{uc} = \frac{2}{3} \quad (3)$$

$$\text{Over correction weight: } W_{oc} = \frac{1}{3} \quad (4)$$

The proportion of each error is computed and defined as a proportion, as shown in Table 2. Another factor to consider in using the combined measure is the difference of the true mean of the samples and the mean of under correction or over correction.

$$M_{uc} = \text{mean of under correction} - \text{true mean} \quad (5)$$

$$M_{oc} = \text{true mean} - \text{mean of under correction} \quad (6)$$

The combined measure (CM) in Equation 7 is computed using the parameters introduced through Equations 3 to 6.

$$\text{CM} = [W_{uc} * P_{uc} * M_{uc}] + [W_{oc} * P_{oc} * M_{oc}] \quad (7)$$

A higher MSE indicates a greater difference between the true mean and the outlier management strategy's mean. The highest MSE belongs to cutting to the 90<sup>th</sup> percentile method. This result is likely due to more aggressive cutting of high values. The manual method has the lowest MSE. The manual method is more subjective and would require more effort and time with real datasets. The other low MSE belongs to the fitted lognormal probability plot because the deduction in high values in this method is more reasonable.

**Table 2: Results of different methods of cuttings**

	MSE	Under Correction			Over Correction			Combined Measure (%)
		Proportion (%)	Mean of UC	$M_{uc}$	Proportion (%)	Mean of OC	$M_{oc}$	
Cut to 0.99	0.055	38.9	1.203	0.189	61.1	0.825	0.189	8.7
Cut to 0.95	0.083	8.2	1.130	0.116	91.8	0.741	0.272	8.9
Cut to 0.90	0.141	0.7	1.066	0.053	99.3	0.651	0.363	12.0
Manual	0.042	30.0	1.121	0.107	70.0	0.825	0.188	6.5
Fit to population	0.048	22.1	1.138	0.124	77.9	0.812	0.202	7.1

The under and over correction are the two main reasons that make it important to compare the true mean and the corrected mean and compute their differences. In all of these outlier management strategies the proportion of over stating is higher than that of under stating. That is why the combined measure is used to state which is the better strategy. The lower measure indicates that the outlier management strategy is more acceptable with respect to the amount of over and under correction. The manual method like the MSE parameter, has the lowest measure, which is 6.5%. In the other methods, the lognormal probability fitted model has the second lowest measure, a combined measure of 7.1%.

As a result, the outlier management strategy based on the lognormal fitted model has the best outcomes of MSE and combined measure compared to other cutting methods. Therefore, it is a better outlier management strategy.

# Chapter 3 - Variogram Management

The variogram is an important input parameter in conventional geostatistics. The variogram model specifies the spatial variability of the regionalized variable. Almost all variables in every stationarity subset that will be used in modeling require a variogram (Deutsch, 2011). The variogram is a spatial function that is used in kriging. The model derived from a good experimental variogram contains useful information regarding the spatial distribution of data. The variogram is also used to determine the effect of change of support in volume variance relations.

## 3.1 Traditional Variogram

Consider a stationarity random function  $\{Z(\mathbf{u}); \mathbf{u} \in A\}$  with stationarity mean  $m$  and variance  $\sigma^2$ . The variogram is the conventional measure of spatial variability. The variogram is defined as the expected squared difference between two data separated by a distance vector  $\mathbf{h}$ :

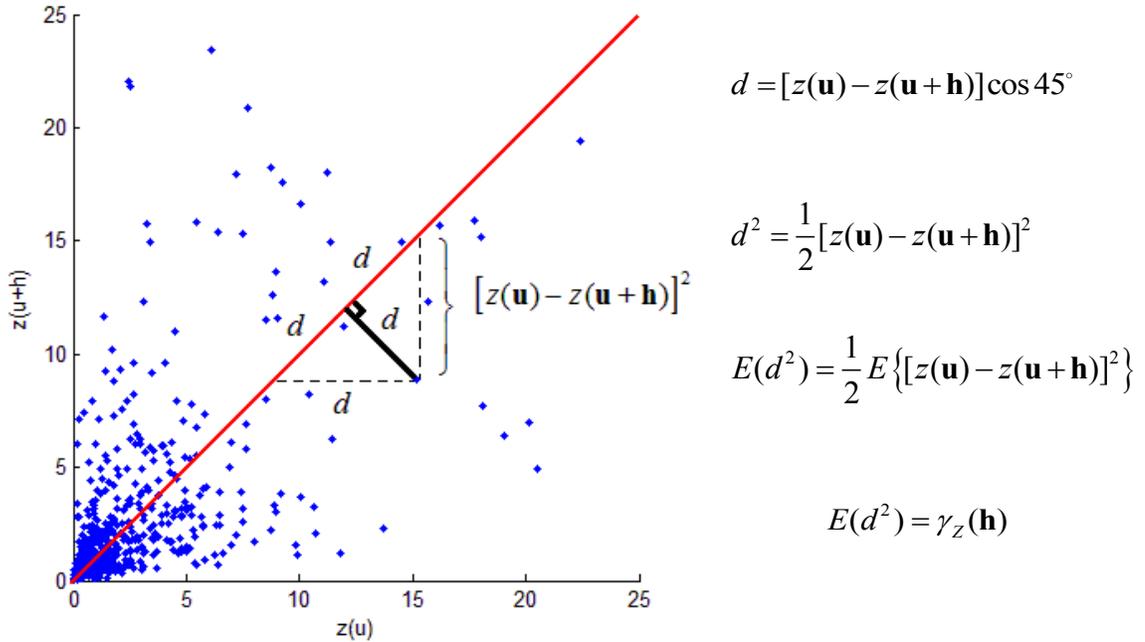
$$2\gamma(\mathbf{h}) = Var[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})] = E\left\{[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]^2\right\} \quad (8)$$

The variogram is sensitive to sparse data and extreme values. The traditional experimental variogram is estimated by calculating the squared differences of all pairs of data separated approximately by a lag vector  $\mathbf{h}$ :

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})]^2 \quad (9)$$

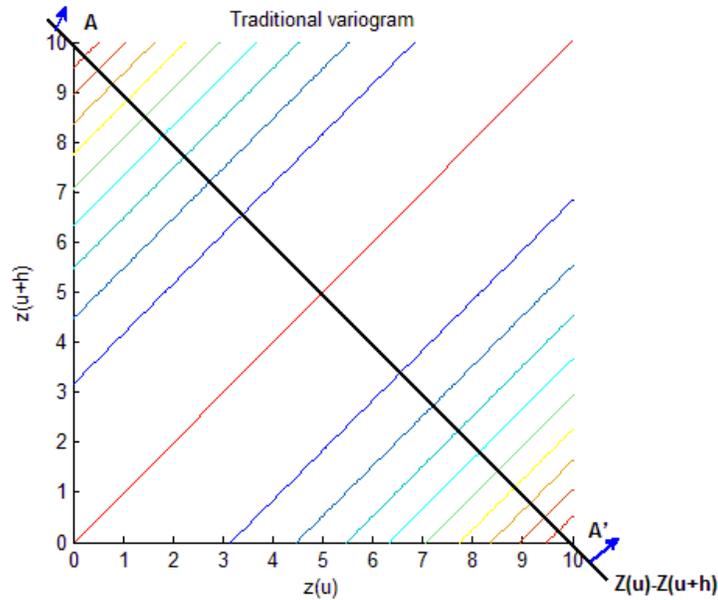
where  $N(\mathbf{h})$  is the number of data pairs separated approximately by a vector  $\mathbf{h}$ . For a given  $\mathbf{h}$ , the  $z(\mathbf{u})$  and  $z(\mathbf{u} + \mathbf{h})$  data pairs can be plotted on a cross plot. The moment of inertia of an h-scatterplot is interpreted as the variogram (Journel, 1983).

The experimental variogram is the average of the square of the distances between the point on the plot and the 45° line. See Figure 22.



**Figure 22: The variogram is calculated as the average of the distance squared of the points from the 45° line. These data are for illustration only and the units are arbitrary**

The variogram gives equal weight to all the data pairs. This causes problems when high values get paired with low values because the squared difference is very high. Figure 23 shows the quadratic weighting as the pairs depart from the 45° line.



**Figure 23: Contour plot of a conventional traditional variogram which gives equal weight to the pairs with equal differences. The colored lines represent the contribution of traditional variogram for each pair of data**

This weighting scheme of the traditional variogram makes it sensitive to extreme values. This sensitivity is increased with sparse data and the proportional effect, that is, the data available at short lag distances may be more variable than at large lag distances. A shortage of data makes variogram interpretation and modeling more difficult (Cressie & Hawkins, 1980; Genton, 1998a). In presence of extreme high values an alternative may need to be considered. Some alternatives to the traditional variogram include the correlogram, normal score variogram, indicator variogram, pairwise relative variogram, madogram and rodogram.

## 3.2 Robust Alternatives to the Variogram

### 3.2.1 Correlogram

The correlogram is related to the correlation of the points separated by distance  $\mathbf{h}$ . It is calculated by subtracting the value of the correlation from 1.0. The definition of the correlogram is as follows (Isaaks & Srivastava, 1988; Srivastava & Parker, 1989)

$$\hat{\gamma}_{corr}(\mathbf{h}) = 1 - \frac{\hat{C}(\mathbf{h})}{\hat{\sigma}(\mathbf{u})\hat{\sigma}(\mathbf{u} + \mathbf{h})} \quad (10)$$

Where

$$\hat{C}(\mathbf{h}) = \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i)z(\mathbf{u}_i + \mathbf{h}) - \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i) \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i + \mathbf{h}) \quad (11)$$

$$\hat{\sigma}(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i)z(\mathbf{u}_i) - \left[ \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i) \right]^2 \quad (12)$$

$$\hat{\sigma}(\mathbf{u} + \mathbf{h}) = \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i + \mathbf{h})z(\mathbf{u}_i + \mathbf{h}) - \left[ \frac{1}{n} \sum_{i=1}^n z(\mathbf{u}_i + \mathbf{h}) \right]^2 \quad (13)$$

The correlogram may be more stable than the variogram and more resistant to extreme values as it accounts for the local means and variances at different lag distances (Isaaks & Srivastava, 1989; Machuca-Mory & Deutsch, 2007). The correlogram is commonly used with skewed data. The convergence of the correlogram is close to the traditional variogram but not exactly the same because the stationary variance is used in the variogram instead of the local variances assumed in the correlogram (Wilde & Deutsch, 2006).

### 3.2.2 Normal Score Variogram

The normal score variogram is calculated in the same way as the experimental variogram but the original variable ( $Z$ ) is transformed to a standard normal distribution ( $Y$ ):

$$y = G^{-1}(F(z)) \quad (14)$$

Where  $G$  is a Gaussian or normal CDF and  $F$  is the CDF of the original data. This transformation makes the variogram more stable with respect to outliers and clustered data because the Gaussian transformation brings all data within a narrow range and removes the proportional effect to make the variogram better behaved.

The major issue with the normal score variogram is its bias with respect to the original variable variogram. This variogram converges to something different from the reference variogram. However; some methods have been proposed to back transform the normal score variogram to the true underlying variogram (Chiles & Delfiner, 1999; Wilde & Deutsch, 2006). The normal score variogram can still be noisy and sensitive to sparse data.

### 3.2.3 Indicator Variogram

The indicator variogram is another alternative to the traditional variogram in presence of extreme values. Data transformation to indicators removes the sensitivity of data to outliers because there are only 0's and 1's after the indicator transformation and these indicator values are used instead of the data values themselves (Gringarten & Deutsch, 2001). An indicator variogram is calculated in the same way as the traditional variogram calculation, by using the  $i$  rather than the  $Z$  values (Carr, Bailey, & Deng, 1985):

$$i(\mathbf{u}; z_c) = \begin{cases} 1, & \text{if } z(\mathbf{u}) \leq z_c \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

Where  $z_c$  is a chosen threshold value that could be, for example, the median. For extreme cutoffs, the indicator variograms may become unstable (Journel, 1983).

### 3.2.4 Madogram and Rodogram

The madogram is calculated using the absolute difference between the paired  $z(\mathbf{u})$  and  $z(\mathbf{u} + \mathbf{h})$  data instead of the squared difference that is used to calculate the traditional variogram (Cressie N. , 1993; Deutsch & Journel, 1998). This feature makes the madogram more stable in presence of high values, because it is not magnifying the differences by squaring them. The madogram is defined as:

$$\hat{v}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} |z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})| \quad (16)$$

One can go further and calculate the square root of the absolute difference used in the madogram (Deutsch & Journel, 1998; Lloyd, Berbeloglu, Curran, & Atkinson, 2004). This defines the rodogram:

$$\hat{v}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} |z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})|^{\frac{1}{2}} \quad (17)$$

The madogram and rodogram may be useful tools in identifying large scale structures. The madogram takes just a partial spatial dependence structure for bivariate vectors such as  $z(\mathbf{u})$  and  $z(\mathbf{u} + \mathbf{h})$ . The madogram cannot capture and characterize the full bivariate structure (Cooley, Naveau, & Poncet, 2006). These alternatives do not converge to the correct variogram.

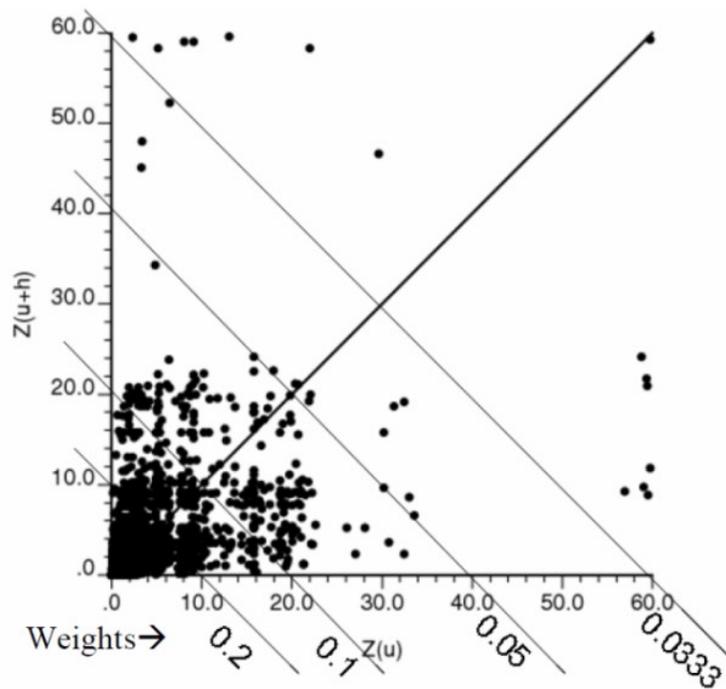
### 3.2.5 Pairwise Relative Variogram

Practice has shown that the pairwise relative variogram is very stable. The pairwise relative variogram is calculated using a method similar to the traditional

variogram. There is a weight in the denominator that serves to reduce the influence of outliers and clustered data. The definition of the pairwise relative variogram estimator for strictly positive variables is:

$$\hat{\gamma}_{PR}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \frac{[z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})]^2}{\left[\frac{z(\mathbf{u}_i) + z(\mathbf{u}_i + \mathbf{h})}{2}\right]^2} \quad (18)$$

The different weighting scheme in the pairwise relative variogram reduces the effect of outliers by giving them less weight, as shown in Figure 24.



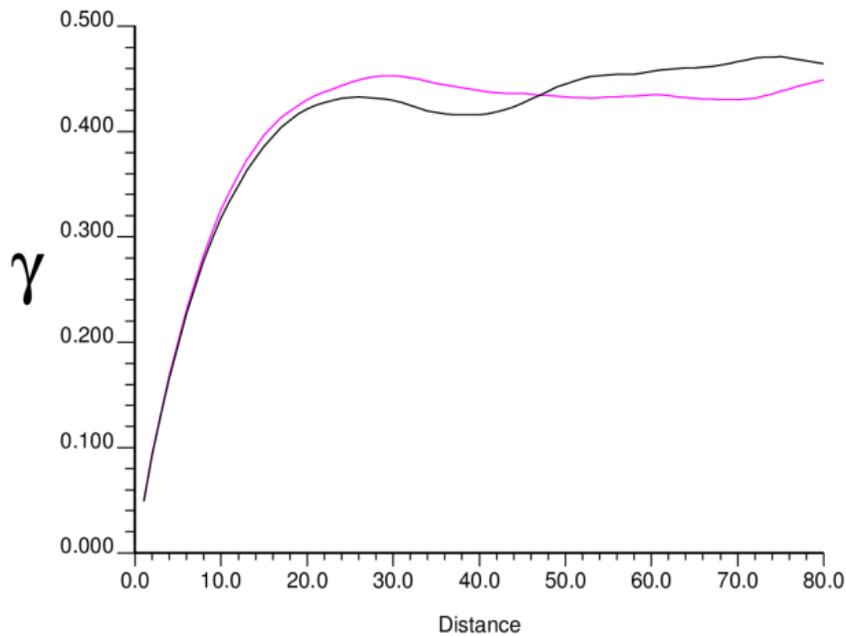
**Figure 24: The weight scheme in the pairwise relative variogram (Wilde & Deutsch, 2006)**

There are two main issues with the pairwise relative variogram: 1) the sill of this variogram is unknown, which makes it difficult to have a precise interpretation, 2) the standard pairwise relative variogram does not converge to the traditional variogram of the original data.

### 3.3 Standard pairwise relative variogram

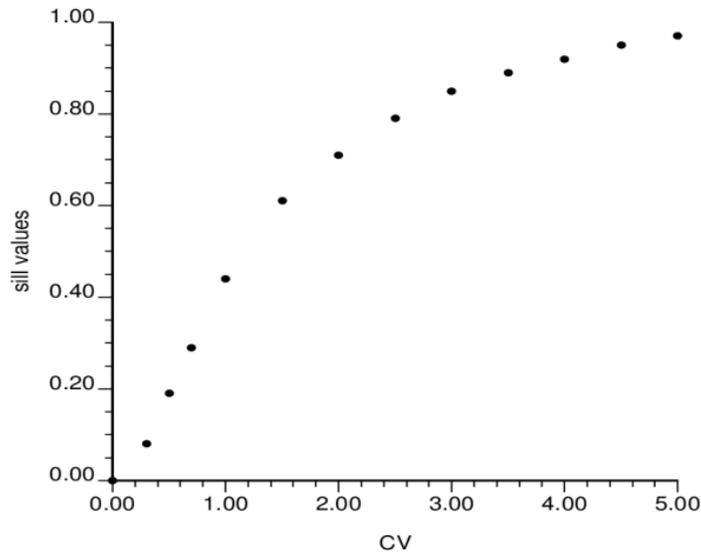
#### 3.3.1 Sill of the Pairwise Relative Variogram

The mean, variance and the shape of the distribution are the three important factors on which the sill of the pairwise relative variogram depends. To see this dependence more, in a lognormal distribution, with different values for the mean and variance, the sill of the pairwise relative variogram is computed. An interesting result is the dependence of the sill of the pairwise variogram on the ratio of the square root of the variance to mean not the mean or variance individually. This ratio is defined as the coefficient of variation (CV). The pairwise relative variogram of an exhaustive standard lognormal distribution with the CV of one is shown in two horizontal direction in Figure 25. The approximate value for the sill of the pairwise relative variogram is 0.44.



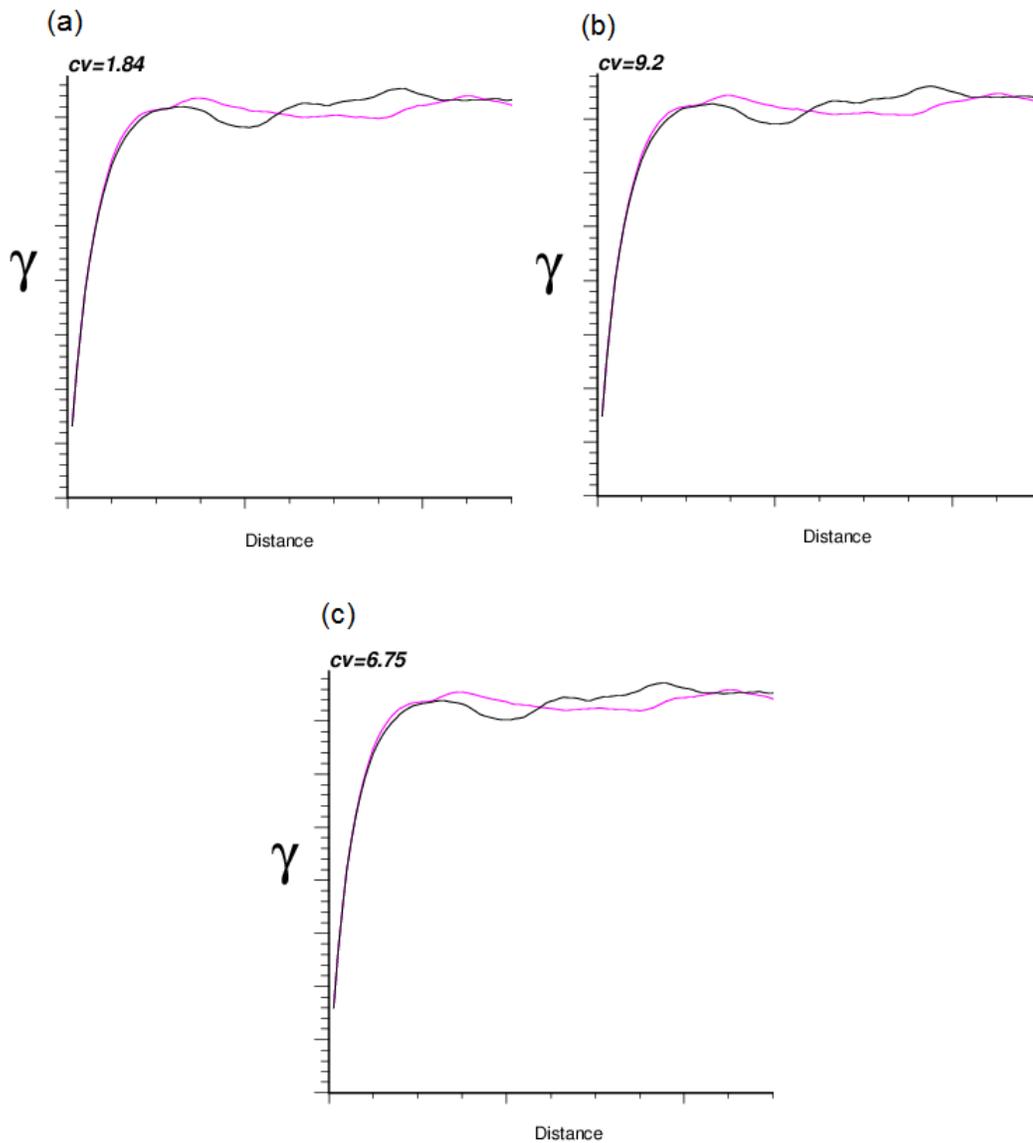
**Figure 25: Two directional pairwise relative variograms for a standard lognormal distribution. The violet line is in the x direction. The black line is in the y direction**

The sill of the traditional variogram is equal to the variance. The sill of the pairwise relative variogram in lognormal data has a connection with the coefficient of variation (CV). By increasing the CV, the value of the sill also increases (Figure 26).



**Figure 26: The relation between the CV and the sill of the pairwise relative variogram for lognormal distribution**

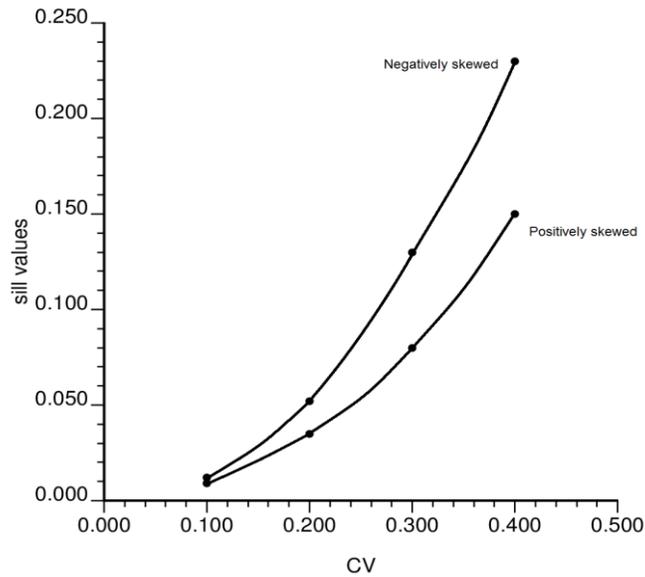
Outlier high values have little effect on the sill of the pairwise relative variogram. The corrupted lognormal distribution is used to show the low influence of outliers on the value of the sill (see Figure 27). Once the top 1% of the data is replaced by a high value (e.g., 1000) and the CV and pairwise relative variogram are calculated and the sill is observed (Figure 27-b). The same procedure is done for 2% of the data (Figure 27-c). The CV is changed by considering a different ratio of original data replaced by outliers. In corrupted data with high values by changing in CV the sill of the pairwise relative variogram remains the same because of the low influence of outliers on the behavior of the pairwise relative variogram.



**Figure 27: Pairwise relative variogram for exhaustive lognormal data in figure (a) and two corrupted lognormal data with outlier high values in figures (b) and (c) in two directions. x is in violet and y is in black**

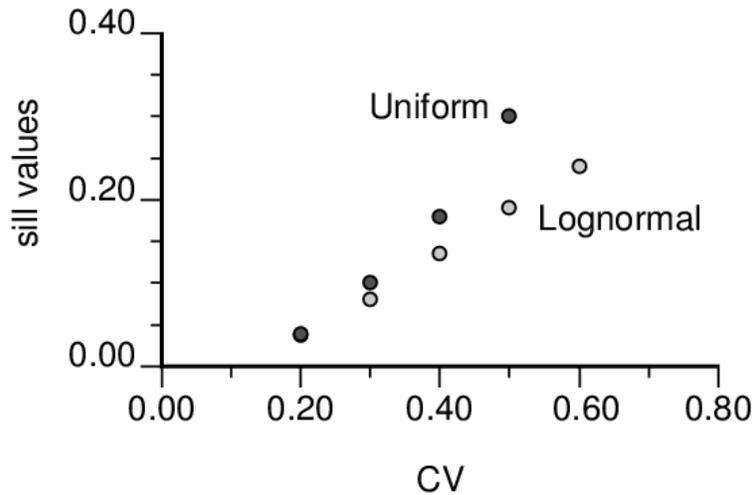
The other dependence of the sill of the pairwise relative variogram is the shape of the distribution. The lognormal distribution is used to determine the sill of the pairwise relative variogram. The difference of the value of the sill is examined for

positively and negatively skewed lognormal distribution. By increasing the value of CV, the sill value for the pairwise relative variogram is shown for negatively and positively skewed distribution in Figure 28. Outlier high values have more effect than outlier low values because the sill in the negatively skewed distribution increases more.



**Figure 28: The different value of the sill of the pairwise relative variogram for negatively skewed and positively skewed lognormal distribution**

Changing the type of the distribution is useful to demonstrate the dependence of the sill of the pairwise relative variogram on the shape of the distribution. The uniform distribution with a shape that is completely different from that of the lognormal distribution is used to show the relation between the CV and the sill of the pairwise relative variogram (see Figure 29). For a small value of the CV, the sill of the pairwise relative variogram for uniform distribution shows behavior similar to that of the lognormal distribution. But as the CV increases, there is more difference between the sills of the two distributions.



**Figure 29: The light grey dots are the sill of the pairwise variogram for the lognormal distribution and the dark ones are for the uniform distribution**

### 3.3.2 Standardization procedure

An important goal in a geostatistical study is to figure out a stable and standard variogram for an appropriate interpretation. The pairwise relative variogram is a very stable variogram estimator. The important thing is to find a way to standardize the sill of the pairwise relative variogram. Determining a general relationship for the sill given data of arbitrary shape, mean, variance and outliers would be intractable. There is another option that is simpler and more reliable. The data are transformed into a standard lognormal distribution with mean and variance of one. The sill of the pairwise variogram for this standard lognormal distribution is 0.44, so the pairwise relative variogram is standardized by dividing it by the constant 0.44:

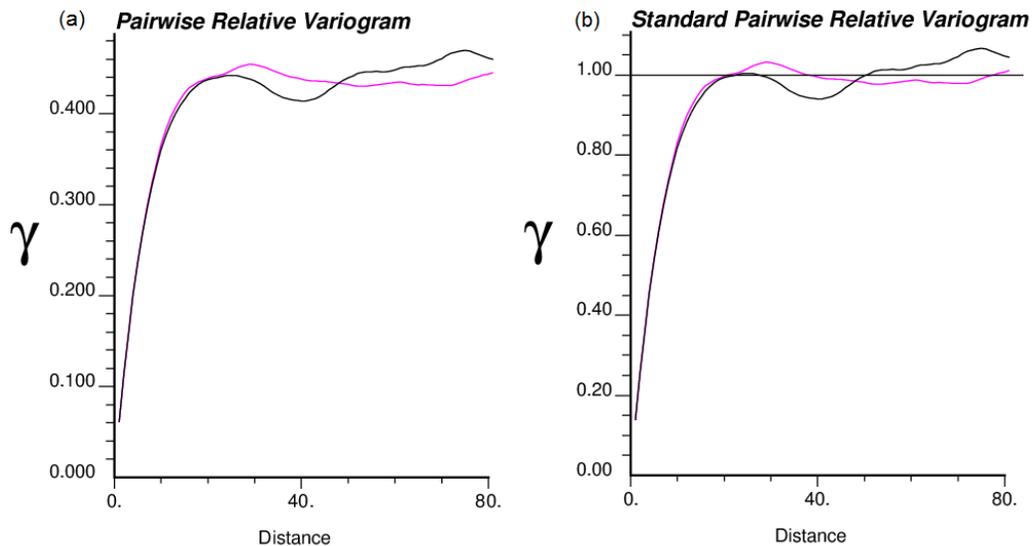
$$\hat{\gamma}_{PR,S}(\mathbf{h}) = \frac{\hat{\gamma}_{PR}(\mathbf{h})}{0.44} \quad (19)$$

The progress of calculating the standard pairwise relative variogram is simple. The lognormal algorithm is implemented in the FORTRAN program

lognorm\_new, which was developed from the lognorm program (Deutsch & Journel, 1998). The parameter file of this program is shown in the Appendix.

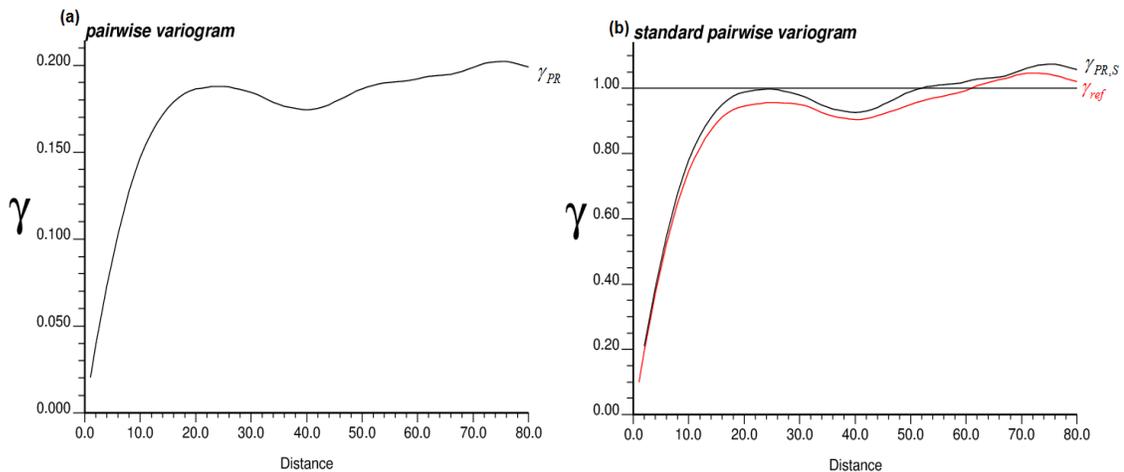
The standardization part of the pairwise relative variogram is implemented in the FORTRAN program gam\_pair, which was developed from the gam program (Deutsch & Journel, 1998). In the original program, the number 6 is defined for calculating the pairwise relative variograms. In the new program (-6) is the number for calculating the standardized pairwise relative variogram. Therefore, to write the progress in a script, first the executable nscore.exe transforms the data into a normal score. Then the executable lognorm\_new.exe transforms those data into the lognormal distribution with an arbitrary mean and standard deviation that are chosen to be one. Finally, the executable gam\_pair.exe calculates the standard pairwise relative variogram.

As an example, the previous exhaustive lognormal data is standardized in the standardization procedure and shown in Figure 30.



**Figure 30: Figure (a) shows the pairwise relative variogram. The standard pairwise relative variogram is illustrated in figure (b)**

Another method for estimating the sill of the pairwise relative variogram consists of drawing independent paired values drawn from the data distribution. This approach does not require the data to be lognormal or to have a CV of 1; it simply draws the original data values untransformed. The FORTRAN program PRSill (Deutsch, 2014) is used to draw paired values and compute the average pairwise relative variogram based on the independent samples. For testing, a set of lognormal data with a CV of 0.5 is chosen. The pairwise relative variogram of data is shown in Figure 31-a. The estimated sill based on the new approach is approximately 0.19. The standard pairwise relative variogram along with the traditional variogram of the data are shown in Figure 31-b. This approach is simple, but does not perform the correction to the specified coefficient of variation, nor does this approach handle zero grades.



**Figure 31: In figure (a) the pairwise relative variogram is illustrated in the black curve. In figure (b) the standard pairwise relative variogram is shown in a black curve with the traditional variogram shown in red**

### 3.4 Correct Unbiased Estimation of Traditional Variogram

#### 3.4.1 Relationship between the Correlation Coefficient of Normal and Lognormal Distribution

In the procedure of standardizing the pairwise relative variogram, the normal and lognormal transformation is applied on the data under study. There is an analytical relationship between the normal and lognormal distribution that permits transformation between these two distributions. Consider the  $Y(\mathbf{u}), Y(\mathbf{u}+\mathbf{h})$  as bivariate normal data with the mean of  $\alpha$ , standard deviation of  $\beta^2$  and correlation coefficient of  $\rho_Y(\mathbf{h})$ . The lognormal transformation shown in Equation 20 is used to transform the bivariate normal data  $Y(\mathbf{u}), Y(\mathbf{u}+\mathbf{h})$  to the bivariate lognormal data  $Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})$  with a mean of  $m_z$ , and variance of  $\sigma_z^2$ .

$$Z = e^{Y \cdot \beta + \alpha} \quad (20)$$

The mean and the standard deviation of the lognormal distribution are as follows:

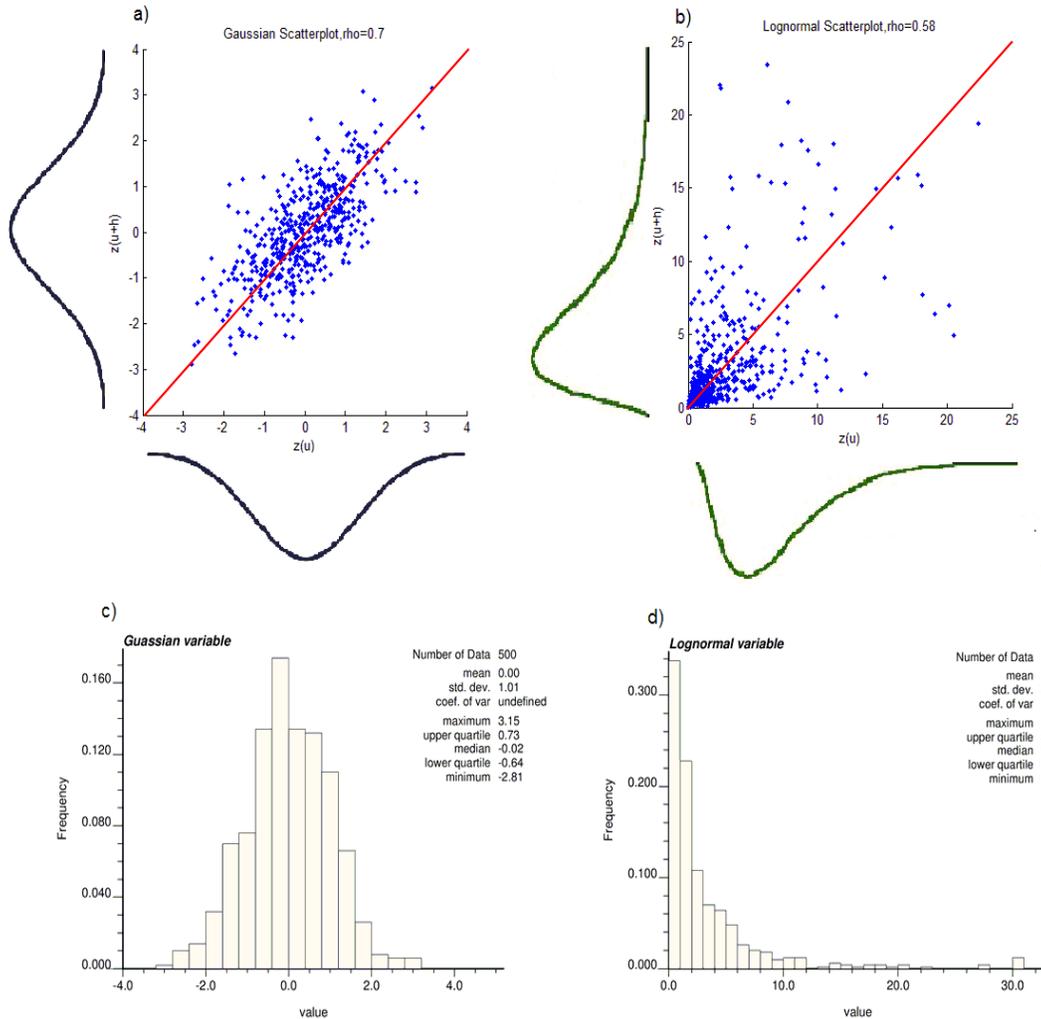
$$m_z = e^{\alpha + \frac{\beta^2}{2}}, \quad \sigma_z^2 = m_z^2 \left[ e^{\beta^2} - 1 \right] \quad (21)$$

The correlation coefficient of lognormal data  $Z$  can be clarified as a function of the mean and variance of the  $Z$  values and the correlation coefficient of normal data  $Y$  (Chiles & Delfiner, 1999):

$$\rho_z(\mathbf{h}) = \frac{\sigma_z^2}{m_z^2} \left[ e^{\beta^2 \rho_Y(\mathbf{h})} - 1 \right] \quad (22)$$

An unconditional Gaussian dataset of 500 data points is transformed to a lognormal distribution with a mean of 4 and standard deviation of 8 (CV=2). The histogram of two variables that follow a standard normal distribution and lognormal distribution, respectively, and the scatter plot of a standard bivariate normal distribution and bivariate lognormal distribution are shown in Figure 32.

The theoretical correlation coefficient is the same as that calculated from Equation 22.



**Figure 32: The scatter plots and histograms of normal and lognormal bivariate data**

### 3.4.2 Unbiased Estimate of True Variogram

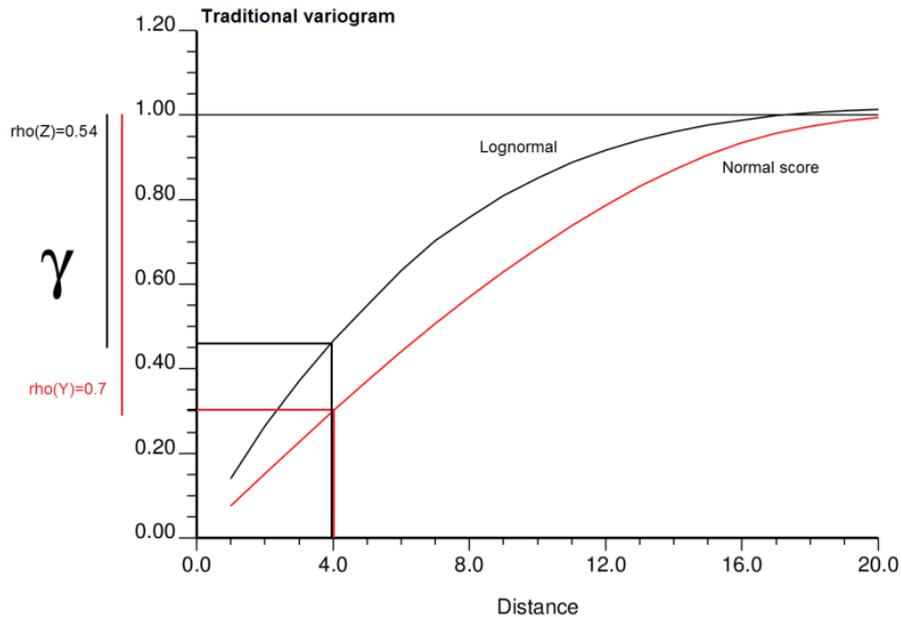
The relationship between the correlation coefficient of normal and lognormal data (Equation 22) is the key point in the process of calculating the correct variogram

of original data with the arbitrary CV based on the standard pairwise relative variogram.

If the correlation coefficient of normal data is known, the correlation coefficient of lognormal data can be derived from Equation 22. This correlation coefficient is comparable with the ones computed from the standard variogram. Figure 33 illustrates the traditional variogram of normal data and corresponding lognormal data with the CV=2. The correlation coefficients of these two distributions can be derived from their variogram values because in the standard form of the variograms:

$$\rho(\mathbf{h}) = 1 - \gamma(\mathbf{h}) \quad (23)$$

The correlation coefficient of lognormal data derived from the standard variogram and Equation 22 are is equal to the correlation coefficient coming from the data itself.



**Figure 33: The relationship between two correlation coefficients based on their variograms**

In the standardization procedure of the pairwise relative variogram, the data are transformed to a standard lognormal distribution. The correlation coefficient of normal data corresponding to this standard lognormal data is derived from the following equation, which is another form of Equation 22:

$$\rho_Y(\mathbf{h}) = \frac{\ln\left(\rho_Z(\mathbf{h})\frac{\sigma_z^2}{m_z^2} + 1\right)}{\beta^2} \quad (24)$$

Where the ratio of  $\frac{\sigma_z^2}{m_z^2}$  is equal to one because the CV of standard lognormal distribution is one. The parameter  $\beta^2$  and  $\rho_z(\mathbf{h})$  are as follows:

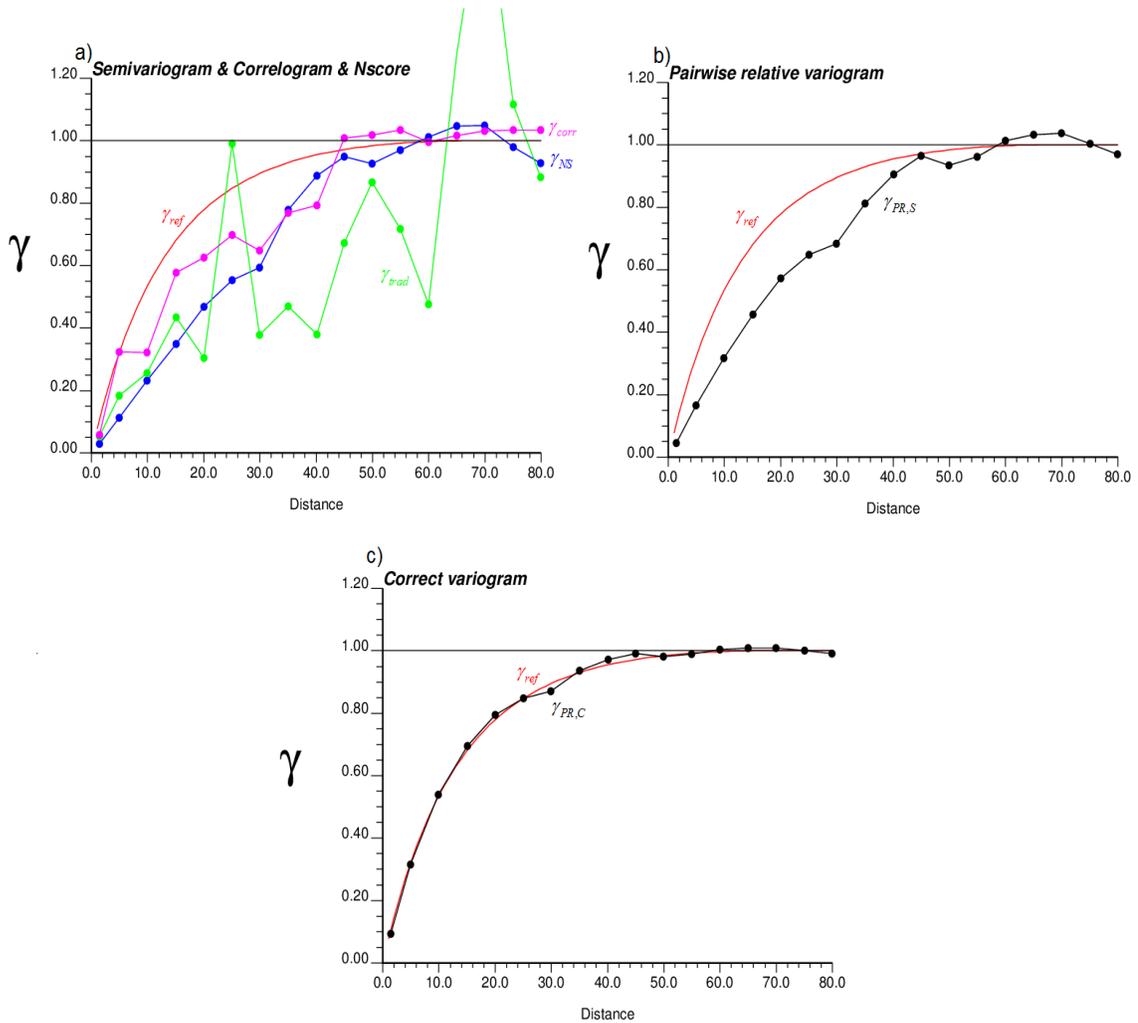
$$\beta^2 = \ln(1 + CV^2) = \ln(2) \quad , \quad \rho_z(\mathbf{h}) = 1 - \gamma_{PR,S}(\mathbf{h}) \quad (25)$$

By replacing the correlation coefficient of normal data, the standard pairwise relative variogram and the CV of the original data in Equation 22, the correct pairwise relative variogram as the unbiased estimate of the variogram of original data is derived:

$$\gamma_{PR,C}(\mathbf{h}) = 1 - \rho_Z(\mathbf{h}) = 1 - \frac{1}{CV^2} \left[ \exp\left(\frac{\ln(CV^2 + 1)\ln(2 - \gamma_{PR,S}(\mathbf{h}))}{\ln(2)}\right) - 1 \right] \quad (26)$$

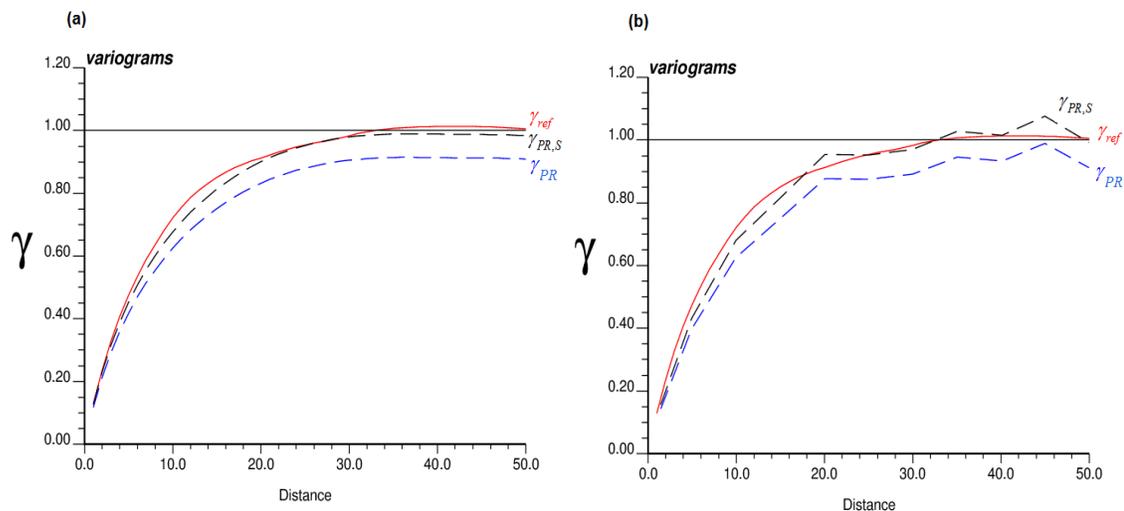
A program is implemented in FORTRAN code `correct_pair` which takes the standard pairwise relative variogram and the CV of the original data as an input and gives the correct unbiased estimation of the traditional variogram of true data.

An exhaustive data set of closely spaced values is used to show the convergence of the correct pairwise variogram to the true variogram. Consider a large grid  $1024 \times 1024$  (1,048,576 data points), 2-D normal score data that transfer to the lognormal distribution with the  $CV=4$ . The traditional variogram of this lognormal data is not noisy and it is shown as red curve in Figure 34. A random subset of 1% of the data is picked and sampled with replacement. This is a very small sample of the whole dataset with a highly skewed distribution which contains outlier high values. The traditional variogram of this small sample is very erratic and has no structure. The correlogram and the normal score variogram deal much better with the outlier (Figure 34-a). However, the pairwise relative variogram is the most robust variogram amongst them (Figure 34-b). The problem with the standard pairwise relative variogram in figure (b) is the incorrect convergence. Figure 34-c shows the correct estimation of the true underlying variogram.



**Figure 34: Portion sample of 1% of the data set. The traditional variogram of the lognormal data is shown as a red curve in all three figures. Figure (a) shows the traditional variogram (green curve), correlogram (violet curve) and normal score variogram (blue curve). Figure (b) shows the standard pairwise relative variogram (black curve) and figure (c) shows the correct estimation of the traditional variogram (black curve)**

The second method of standardizing the pairwise relative variogram based on the averaging the independent paired data values does not have the convergence issue. This standard pairwise relative variogram has the shape of the data because there is no data transformation or change in the CV of the original data. This method might not be as accurate as the previous method but it is simpler to apply. Another large exhaustive lognormal dataset with a CV of four is chosen. The traditional variogram along with the pairwise relative variogram and standard pairwise relative variogram of the whole data set is shown in Figure 35-a. A sample of 1% of the data set is selected that is a fairly small sample from a large dataset. The pairwise relative variogram and the standard pairwise relative variogram of this sample with the traditional variogram of the original data are shown in Figure 35-b.



**Figure 35: The traditional variogram of the lognormal data is shown as a red curve in the two figures. Figure (a) shows the pairwise relative variogram in blue dash line and the standard pairwise relative variogram in black dash line for the exhaustive dataset. Figure (b) shows the pairwise relative variogram in blue dash line and the standard pairwise relative variogram in black dash line for the small sample of 1% of data**

### 3.5 Recommendation

The standardized pairwise relative variogram is a robust alternative to the traditional variogram or the correlogram in presence of outliers. In the pairwise relative variogram, the unknown sill and the convergence of the pairwise relative variogram are the two major issues. The solution for these two problems is discussed in this thesis.

The proposed technique for calculating the standard pairwise relative variogram is simple, as it does not require the data to follow any particular distribution. The contribution of this thesis is the reliance on the transformation of the data into a standard lognormal distribution and the calculation of the standard pairwise relative variogram. This technique might have some shortcomings: 1) Mixing the population is one of the pitfalls in this approach. It means that in a population with two domains and two different distributions, this technique considers only one population and transforms all the data into a lognormal distribution. 2) Another pitfall is the tolerance in the calculation of the pairwise relative variogram. However, in case of not really good traditional variogram of original data because of outliers, the standard pairwise relative variogram could be the best option to study the spatial variability and other interpretations coming from the variogram calculation.

The transformations in the standardization procedure are useful for computing an unbiased estimate of the true underlying variogram of original data based on the analytical relationship between the correlation coefficient of normal and lognormal data. As some of the basic geostatistical calculations such as kriging and modeling need the variogram in its original units, it is highly recommended to use this unbiased estimate if the standard pairwise relative variogram is used for variogram modeling.

# Chapter 4 - Estimation and Simulation

Improving block grade estimates in presence of outliers is an important subject. Outliers can cause overestimation on nearby blocks. An important related problem is to know the locations that are being influenced by the outliers. The volume of influence changes based on the local data configuration and data values. The outliers for this chapter have been identified by a variety of graphical tools such as a histogram, probability plot and scatter plot. Chapter 2 presented more details on outlier identification.

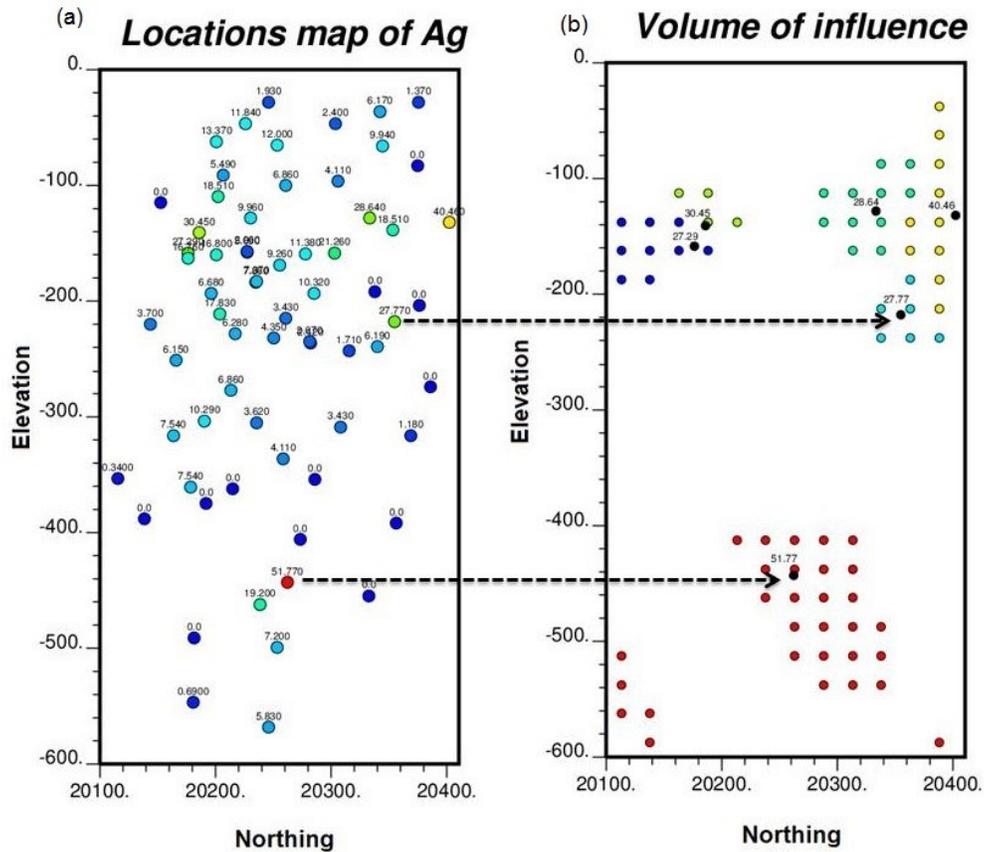
Two outlier management strategies will be discussed. The first method is based on moving the outlier further away to a higher dimension (Deutsch, Boisvert, & Deutsch, 2011) in order to manage the contribution of the outlier on the nearby blocks. The second method (also called the cutting or capping method) reduces the value of the outliers to manage the contribution of the outliers. The management is based on understanding what the local averages should be considering from the simulated local average. By managing the outlier values based on the uncertainty on the simulated local averages, the conditional bias will be decreased accordingly. Before explaining the outlier management strategies, the understanding of the volume of influence will be presented.

## **4.1 Volume of Influence**

The volume of influence of each outlier can be determined based on the data spacing, range of the variogram, search radius and number of data considered in the kriging estimation. The values above a fixed arbitrary threshold could be considered as outliers. As described in Chapter 2, more detailed local criteria could be considered, but this is a simple approach. In the kriging estimation program documented in the Appendix, all the grid blocks that receive a weight more than an arbitrary minimum value will contribute to the volume of influence

of an outlier. It is possible that the volume of influence of two outliers overlap, but that would mean the outliers are less consequential, that is, there are multiple high values supporting local estimation. The overlap happens where the location of the two outliers is close enough that some grid blocks are being influenced by both of the outliers. The simplest way to fix the overlapping is to allocate each block to the outlier that has the most influence on the block, that is, the outlier that receives the most weight.

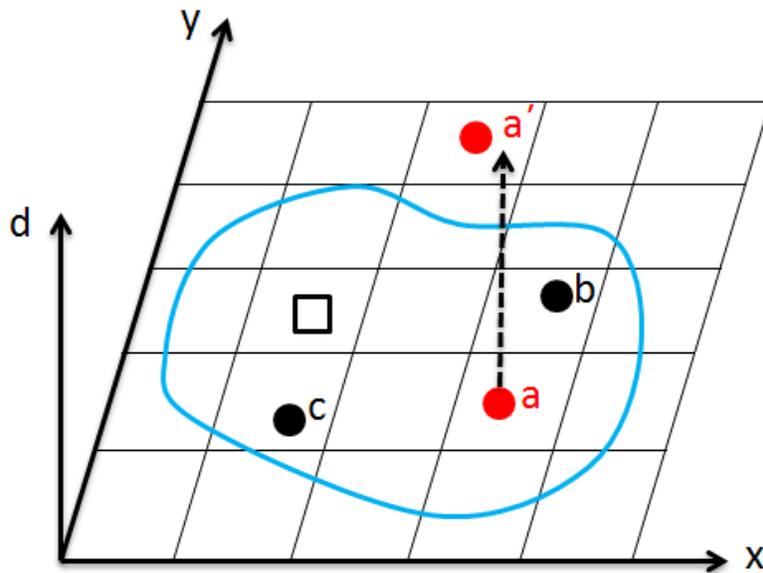
For illustration, consider a 2D mining dataset red.dat with a silver grade, Ag, that has some high value outliers. The location map illustrated in Figure 36 shows the locations of six outliers and the volume of influence of each outlier.



**Figure 36: The location map of the Ag variable is shown in figure (a). The locations of the volume of influence of each outlier of the Ag variable are shown by different colors in figure (b)**

## 4.2 Move to Higher Dimension

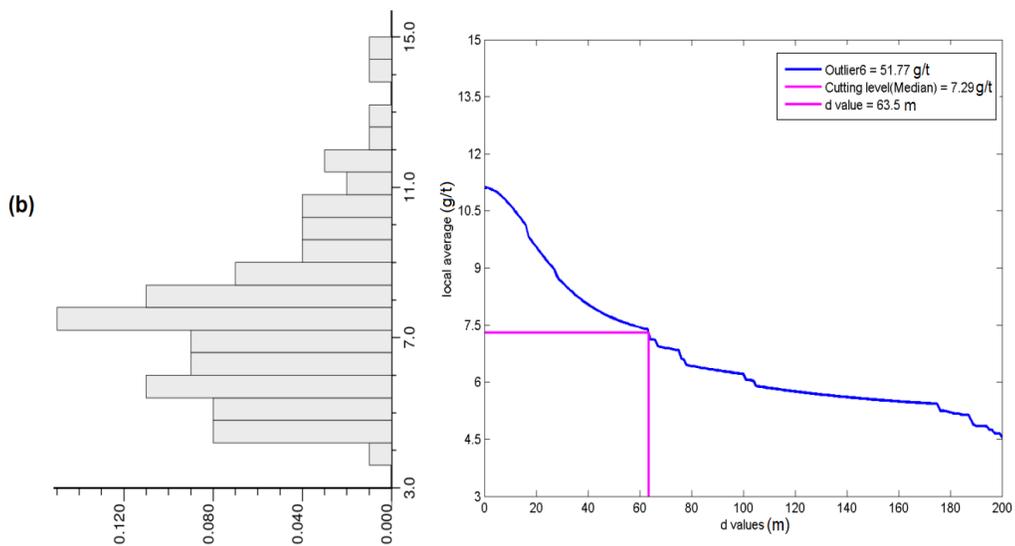
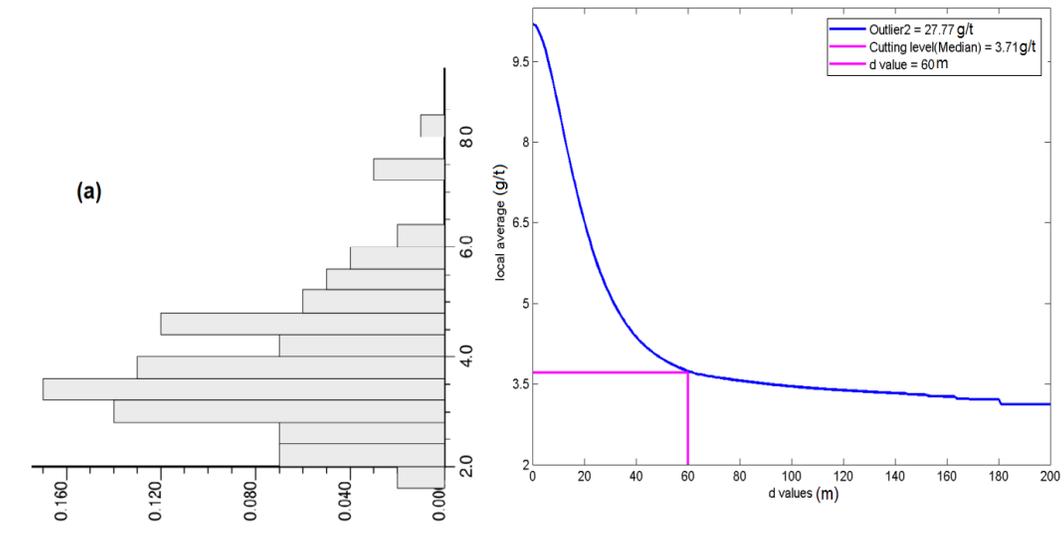
The main concern about outliers in resource estimation is that the outliers may be assigned too much weight within a region where the continuity of the outlier does not support the high estimated grades. The idea is to use an additional dimension  $d$  to reduce the influence of outliers on the grid blocks within the volume of influence in resource estimation. The workflow of the extra dimension  $d$ , which is in distance units, is to move the outlier value further away from the location being estimated (Deutsch, Boisvert, & Deutsch, 2011). This method is illustrated by a simple schematic 2D example in Figure 37. The goal is to estimate at the unsampled location, denoted as  $\square$ , with 3 data samples such that data  $a$  is considered to be an outlier. Inside the blue curve illustrates the volume of influence of the outlier.



**Figure 37: Reducing the effect of outlier value "a" by moving it further away by addition of d-value at " a " going to " a' " (redrawn from (Deutsch, Boisvert, & Deutsch, 2011))**

The average grade of the grid blocks inside the volume of influence is the main concern. Simulation could be used to determine the uncertainty in the average grade within the volume of influence. These values are more trustworthy than kriging because simulation is quite robust with respect to outlier high values and does not assign them high local continuity. The value for  $d$  for each outlier could be determined by calibrating the results of kriging to the local average of multiple realizations from simulation. Alternatively, a specific quantile (e.g., median) could be chosen to calibrate the distance value.

The silver variable  $Ag$  in the `red.dat` dataset is used to show the effect of moving outliers to the higher level of  $d$  on the local averages. For picking the cutting level, 100 realizations are simulated and the averages inside the volume of influence of each outlier are computed. The uncertainty in the mean tends to be normal because of the central limit theorem. In Figure 38, two of the outliers are analysed by increasing the  $d$  value; the local average inside the volume of influence decreases. The cutting level is defined based on the median of the uncertainty in the simulated averages, and the corresponding  $d$  value is considered for each outlier.



**Figure 38: Decreasing the local average by increasing the d value in two of the outliers of the Ag variable compare with uncertainty in the simulated local average. The more important spatial outlier (27.77 g/t) with d value of 60 m is shown in figure (a) and highest outlier (51.77 g/t) with d value of 63.5 m is shown in figure (b)**

The outlier value 27.77 g/t in Chapter 2 is known as the most effective spatial outlier regardless of its lower value compared to other outliers. The highest simulated average inside the volume of influence of this outlier is 8 g/t; however, the local average of estimated blocks being influenced by this outlier is about 10 g/t. With a good degree of confidence this outlier should move away with the  $d$  value of 60 m.

### 4.3 Cutting Outlier

Moving the data away from the grid block locations by a higher dimension is an interesting approach, but it is not fully implemented in commercial software and is non-intuitive to many practitioners. Another interesting reduction strategy relates to simply cutting to a reduced level. The same methodology could be used to calibrate the cutting level. This outlier management strategy is similar to the  $d$  value method, but instead of moving the outlier further away to reduce its effect on the nearby data, the value of each outlier is decreased to some cutoff level based on the simulated local average in the vicinity of each outlier.

This process of cutting can be done by two modified GSLIB programs `kt3d_new` and `sgsim_new` that are explained in detail in the Appendix. The first step is to run the kriging that identifies the volume of influence of each outlier. The goal is to observe the deduction in the local average for each outlier by decreasing the value of that outlier until the proper value is reached. Based on the kriging system of the equation shown in Equation 27, there is no need to do the kriging for each deduction in outlier value.

$$Z_V^* = (1 - \bar{\lambda}_{out})Z_{oth} + \bar{\lambda}_{out}Z_{out} \quad (27)$$

Where  $\bar{\lambda}_{out}$  is the average weight of the outlier in the volume of influence,  $Z_{out}$  is the outlier value before any treatment, and  $Z_V^*$  is the average estimated value

coming from the kriging program. The only unknown variable is  $Z_{oth}$ , which indicates all the other data values inside the volume of influence except the outlier value. After computing the  $Z_{oth}$  based on the other known parameter, the deduction in the average of estimated value can be computed simply by changing the value of the outlier in Equation 27 without running the kriging program again. The kriging output file related to these values for Ag outlier values is shown in Figure 39.

Estimation Averages for each outlier volume				
5				
Outlier ID	Outlier value	local average	outlier weight average	other values
1	27.290	10.370	0.166	6.996
2	27.770	10.203	0.276	3.516
3	28.640	14.791	0.244	10.330
4	30.450	16.633	0.315	10.281
5	40.460	12.999	0.277	2.478
6	51.770	11.110	0.161	3.322

**Figure 39: The outlier values of the Ag variable with local averages and the other values inside the volume of influence**

The second step is to run the simulation program to compute the uncertainty in local averages for each outlier based on the volume of influence resulted in last step. For this data set, 100 realizations are simulated providing 100 local averages for each outlier. An optional cutoff level can be chosen in this step to cut the outlier value to a level which the user can decide about that. The uncertainty in local average for each outlier tends to be normal so the value of the average and the median are close together. The `sgsim` output file in Figure 40 shows the cutoff level (median = 0.5) values for each outlier.

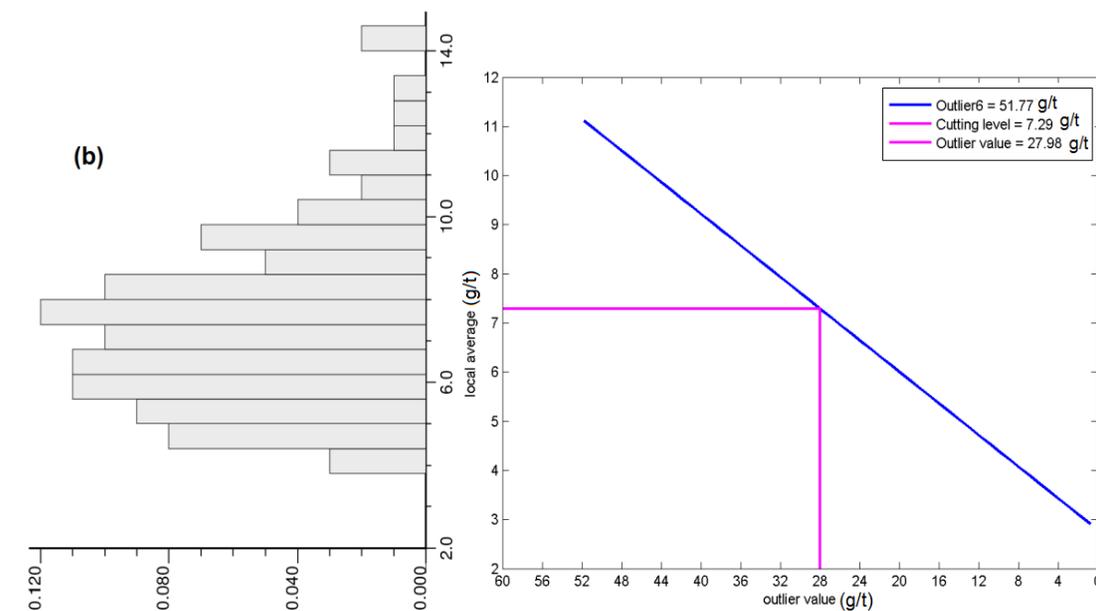
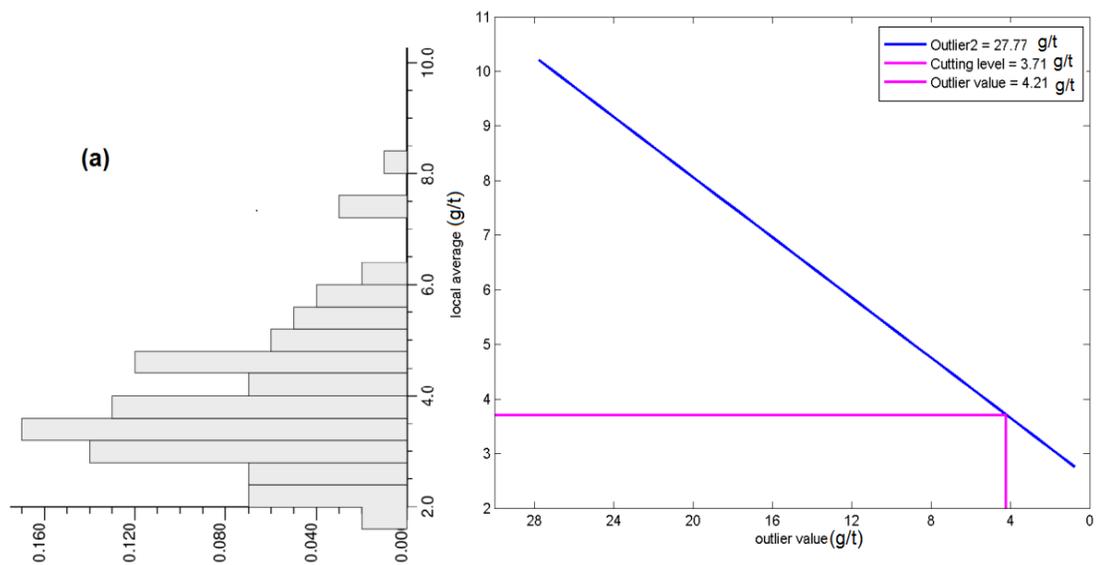
```

The cutting level and values
4      0.50
Outlier ID
Outlier value
local average
cutting value
1      27.29000      6.99711      6.32179
2      27.77000      3.91833      3.70798
3      28.64000      11.89186     11.56513
4      30.45000      14.63218     14.28099
5      40.46000      6.76617      6.63644
6      51.77000      7.53113      7.28606

```

**Figure 40: The output file of the *sgsim* program contains the cutoff value for the median of uncertainty of the local average**

The last step involves cutting the outlier value enough to reach the simulated cutting level value. The outlier management value can simply compute in Equation 27. The reduction in the local average by decreasing the outlier value for two outliers of the Ag variable is shown in Figure 41. The outlier value 27.77 g/t, which showed to be an important spatial outlier, should cut to 4.21 g/t to reach its cutoff level illustrated in Figure 41-a. The other outlier with the value of 51.77 g/t should cut to 27.98 g/t to reach to its level point (Figure 41-b).



**Figure 41: Decreasing the local average by cutting more in outlier values compare with uncertainty in the simulated local average. The more important spatial outlier (27.77 g/t) should cut to 4.21 g/t is shown in figure (a) and the highest outlier (51.77 g/t) with should cut to 27.98 g/t to reach the cutting level of the median in average uncertainty is shown in figure (b)**

For checking the result of the outlier management strategy, the kriging could be run again with the managed outlier values. Then, the average inside of the volume of influence of each outlier would be computed again. The output file of checking the outlier management cutting strategy is shown for each of the outlier managed values and their corresponding local averages (see Figure 42). The difference between the OMS local averages and the simulated one are close enough to accept the managed outlier values. The slight difference is due to the slight overlapping of the volume of influences.

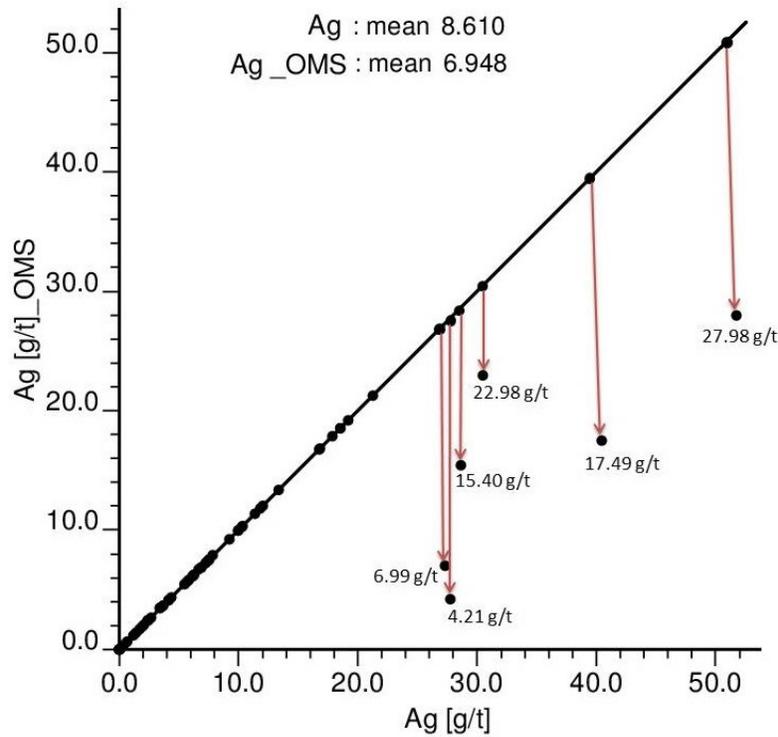
```

outlier management values and local average
6
outlier ID
outlier value
outlier managed value
original local average
simulated local average mean
OMS local average
1.0    27.2900    6.9961    10.3696    6.9971    6.7843
2.0    27.7700    4.2117    10.2027    3.9183    3.3307
3.0    28.6400    15.3993    14.7908    11.8919   11.4798
4.0    30.4500    22.9824    16.6329    14.6322   13.9112
5.0    40.4600    17.4906    12.9987    6.7662    7.0961
6.0    51.7700    27.9803    11.1101    7.5311    7.2861

```

**Figure 42: The check file to compare the OMS results with the simulated ones**

The variable Ag and the outlier management result of the Ag variable are compared in Figure 43. The six high values are reduced based on this methodology and the average of the Ag values is decreased respectfully.



**Figure 43: The comparison of Ag values and the Ag values with reduced outlier values**

#### 4.4 Recommendation

There are some considerations when applying this methodology. One of the important subjects that should be considered carefully is the way that the volume of influence is defined. The needed parameters should be chosen with caution in the kriging program. If the volume of influence defined is too small, the program tends to cut the outlier more. On the other hand, if the volume of influence is too large the outliers may not be treated properly.

The second focus is the cutting level based on the simulated local average. Choosing the median is fairly safe but a more conservative way may be to choose the lower quantile (25%).

# Chapter 5 – Conclusion

Outliers are one of the most important concerns in resource estimation. The identification outliers is sensitive to many factors such as the magnitude of the grade, the location of the outlier, the other data values in the area under study and how the outliers affect the estimation of nearby blocks.

The identified outliers must be managed carefully; simply removing them from the database will likely cause an underestimation. Leaving the outliers untreated will likely cause local overestimation. The outlier management strategy should reduce the value of the outliers without inducing an unwanted bias. In order to decrease an outlier, the volume of influence of each outlier must be defined.

## 5.1 Contributions

This thesis presents a method to identify spatial outliers that affect their nearby data by causing overestimation. An extreme high observation surrounded by other high values has a less significant impact on the estimation than an isolated high value among low values. A detection technique was developed based on the uniform rank transform of the data and its cross validation estimation.

This thesis also provides methodologies to manage outliers. One of the outlier management methods is based on one of the relative types of variogram estimation which reduces the effect of outliers to obtain a more robust estimator of spatial structure. Two of these methods are based on cutting the outlier values to get more robust spatial statistics. Another method decreases the effect of outliers by moving them further away from the unsampled location in a higher dimension.

The first contribution of this thesis is that it identified outliers by dividing the dataset and fitting two lognormal probability models. The modified FORTRAN code `probplt_fit` is programed to fit lognormal probability models to the

lower and upper portions of the dataset. If the second distribution contains any outlier values, it reduces them to the first fitted model. A small study showed that the uncertainty in the average becomes more accurate if there is a better estimation of the actual average of real data.

The second contribution is an understanding of how to deal with outliers in variogram calculations to get a more robust spatial model. The pairwise relative variogram appears as a robust variogram estimator. There were two issues with the pairwise relative variogram that were addressed. The unknown sill of a pairwise relative variogram was established, which is important for variogram interpretation and modeling. The second issue of convergence to the correct theoretical variogram was also fixed. The unbiased estimate of the traditional variogram driven from the pairwise relative variogram is more robust and can be used in modeling, kriging and other geostatistical calculations that require the variogram of original data.

The third contribution of this thesis is related to a second cutting method for the block estimation. Identifying the high values that cause overestimation on nearby data is the beginning of the procedure. Then, the grid blocks being influenced by these values are identified and referred to as the volume of influence for the outlier. The methodology amounts to reducing the outlier values based on the uncertainty in the local average within the volume of influence. This methodology helps to identify the high values that have more effect on the surrounded blocks.

The last contribution is to use the volume of influence of each outlier for an alternative management approach. The idea is to move the outliers further away in a higher dimension instead of cutting them to reduce their effect on nearby grid blocks. The goal of this methodology is the same as the cutting the outliers method.

## 5.2 Future Work

In the outlier detection research, the distribution of data is assumed to be lognormal. One aspect of future work would be to examine other distribution shapes.

There were some methods for detecting spatial outliers that were not used in the outlier management strategies developed in later chapters. An area of future work could focus on the estimation methodology to cycle through all the data to help in identifying the outliers.

Another area of future work is to come up with a theory for the standard pairwise relative variogram. This measure has not been theoretically justified despite the fact that examples show the result to be acceptable.

Dealing with zero grades in the pairwise relative variogram has been dealt with by the transforming to a lognormal distribution; however despiking the constant value is important. Random despiking may make the variogram look worse. This could be investigated more thoroughly.

The volume of influence identified in the outlier management strategies for estimation could be investigated more carefully. The interaction between outliers has not been dealt with completely. If a number of high values are close together that have overlapping in their volume of influence, it is possible that they are not outliers. Therefore, no treatment is required.

In defining the volume of influence the weight that the area of influence received is decided by the user. An area of future work would be to help establish the best value that should be chosen for this weight.

Finally, more case studies should be investigated to refine the proposed methodologies.

# References

- Carr, J. R., Bailey, R. E., & Deng, E. D. (1985). Use of indicator variograms for an enhanced spatial analysis. *Mathematical Geology*, 17(8), 797-811.
- Chiles, J. P., & Delfiner, P. (1999). *Modelling Spatial Uncertainty*. Geostatistics, Wiley series in probability and statistics.
- Cooley, D., Naveau, P., & Poncet, P. (2006). Variograms for spatial max-stable random fields. *Dependence in probability and statistics*, 373-390.
- Costa, J. F. (2003). Reducing the impact of outliers in ore reserves estimation. *Mathematical Geology*, 35(3), 323-345.
- Cressie, N. (1993). *Statistics for Spatial Data*, 2nd edn. Wiley, Ontario.
- Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Mathematical Geology*, 12(2), 115-125.
- David, M. (1988). Handbook of applied advanced geostatistical ore reserve estimation. *Developments in Geomathematics* 6.
- Deutsch, C. V. (2004). *A statistical resampling program for correlated data: spatial\_bootstrap*. Edmonton: CCG Report 6.
- Deutsch, C. V. (2011). *Guide to Best Practice in Geostatistics, Version 1.0*. Edmonton, Alberta, Canada: Centre for Computational Geostatistics, University of Alberta.
- Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical Software Library and User's Guide*. Applied Geostatistics Series. New York: Oxford University press, New York.
- Deutsch, J. L. (2010). *Fitting Probability Plot to Identify Multiple Population and Outliers*. CCG Annual Report 12.
- Deutsch, J. L., Boisvert, J. B., & Deutsch, C. V. (2011). New dimension to account for sample error and volume support in resource estimation. *Transactions of the society of mining, metallurgy, and exploration*, 598-605.
- Genton, M. G. (1998a). Highly robust variogram estimation. *Mathematical Geology*, 30(2), 213-221.
- Glacken, I. M., & Snowden, D. V. (2001). Mineral Resource and Ore Reserve Estimation. *The AusIMM Guide to Good Practice, The Australasian Institute of Mining and Metallurgy: Melbourne. P*, 189-198.
- Gringarten, E., & Deutsch, C. V. (2001). Teacher's Aide Variogram Interpretation and Modeling. *Mathematical Geology*, 33(4), 507-534.

- Isaaks, E. H., & Srivastava, R. H. (1989). *An introduction to applied geostatistics*. New York: Oxford University Press.
- Isaaks, E. H., & Srivastava, R. M. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical Geology*, 20(4), 313-341.
- Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology*, 15(3), 445-468.
- Journel, A. G., & Arik, A. (1988). Dealing with outlier high grade data in precious metals deposits. *Proceedings of Computer Application in the Mineral Industry—Canada: Balkema, Rotterdam*, 161-171.
- Lloyd, C. D., Berbeloglu, S., Curran, P. J., & Atkinson, P. M. (2004). A comparison of texture measures for the per-field classification of Mediterranean land cover. *International Journal of Remote Sensing*, 25(19), 3943-3965.
- Machuca-Mory, D. f., & Deutsch, C. V. (2007). *Non stationary variogram based on continuously varying weights*. Edmonton: CCG Report 9.
- Matheron, G. (1976). A simple substitute for conditional expectation: the disjunctive kriging. *Advanced geostatistics in the mining industry. Springer Netherlands*, 221-236.
- Parker, H. M. (1991). Statistical treatment of outlier data in epithermal gold deposit reserve estimation. *Mathematical Geology*, 23(2), 175-199.
- Rivoirard, J., Demange, C., Freulon, X., & Lécureuil, A. (2012). A Top-Cut Model for Deposits with Heavy-Tailed Grade Distribution. *Mathematical Geosciences*, 45(8), 967-982.
- Rossi, M. E., & Deutsch, C. V. (2013). *Mineral Resource Estimation*. London: Springer.
- Sinclair, A. J., & Blackwell, G. H. (2002). *Applied Mineral Inventory Estimation*. Cambridge University Press.
- Srivastava, R. M., & Parker, H. M. (1989). Robust measures of spatial continuity. *Geostatistics*, 1, 295-308.
- Wilde, B., & Deutsch, C. V. (2006). *Robust Alternative to the Traditional Variogram*. Edmonton: CCG Report 8.

# Appendix

## A.1 Probability plot

### A.1.1 Probability plot fitted model

The non-parametric lognormal fitted model is implemented in the GSLIB compatible `probplt_fit` which is developed from a GSLIB `probplt` and `probplt_pop` programs (Deutsch & Journel, 1998; Deutsch J. L., 2010). This program fits the non-parametric lognormal model to the data below the arbitrary threshold using the regression and simulated annealing methods. There is an option to identify the threshold, mean and standard deviation of two populations. If the user is not sure about the mean and standard deviations, these parameters can be optimized by choosing the `optimize` option in the parameter file.

The parameter file for `probplt_fit` is shown in Figure 44. The input data file (**Line 4**) and variable column and weight (**Line 5**) should be in the standard format of GSLIB. The output probability plot is in **line 7**. The scaling of the probability plot axis is defined in **Line 10** which also clarifies the lognormal or normal fit using in program. To fit a non-parametric model the number 1 should be placed in **Line 12**. In **Line 13** an output summary file including the parameters fit is given. The slopes of two populations and the difference between them are shown in a file in **Line 14**. If the difference between the first population and the second one is bigger than zero, it means the slope in second population is less than the slope in first population. In this situation the data points in the second population are considered as outliers and they should be fixed and reduced to the first population fitted model. After fitting the two populations, the identity of each sample and the fitted and adjusted value of the corresponding population fit is written in the output file in **Line 15**. By reducing the data values in the second population to the first population model, the final output file with the corrected outlier values is specified in **Line 16**. The number of populations should be two for this purpose of reducing the outliers (**Line17**). The temperature and iteration

for optimization are also listed in **Line 17**. The arbitrary cut off threshold to separate the populations, mean, standard deviation, option for optimization and the color of the fitted model line is specified in **Line 18**. Instead of drawing the second population fitted model, the deduction in outlier values to the first population fitted model is shown in the output probability plot. The deduction in the mean of the sample is also written in the output figure.

```

1           Parameters for PROBPLT_FIT
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat      -file with data
5   5  0                   -columns for variable and weight
6   -1.0  1.0e21          -trimming limits
7   probplt.ps            -file for PostScript output
8   0                     -number of points to plot(<0 for all)
9   1                     -0=arithmetic, 1=log scaling
10  0.01  100.0  10.0    -min,max,increment for labeling
11  Probability Plot      -title
12  1                     -fit populations
13  popsum.out           -file for population summary output
14  slop.out             -file for population slop
15  pop_data.out         -file for population fitted data output
16  fit_realization.out  -file for population fitted data output
17  2  1000  0.1         -n populations,maxiter,maxtemp(<1.0)
18  0.9  0.0  0.0  1  1  -Pop1:cutoff,mean,stdev,optimize,colour
19      0.0  0.0  1  7   -Last pop:mean,stdev,optimize,colour

```

**Figure 44: Parameter file of probplt\_fit**

## A.2 Pairwise Relative variogram

### A.2.1 Standard Pairwise relative variogram

To do the standardization for the pairwise relative variogram first the data should transform to the normal score with the `nscore` executable file from GSLIB library. In the second step, the data are transformed to the standard lognormal distribution. The lognormal algorithm is implemented in the FORTRAN program `lognorm_new`, which is developed from the `lognorm` program (Deutsch & Journel, 1998). The parameter file for `lognorm_new` is shown in Figure 45. The input data and the column to transform are entered in **Line 4** and **Line 5**. The mean and standard deviation of the output lognormal data are specified in **Line 6**. The output file includes all of the data from the original data file plus an additional column of the lognormal transformed data is shown in **Line 7**.

```
1           Parameters for Lognorm_new
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   5                           -columns for variable
6   1   1                       -mean and standard deviation
7   lognormal.out              -file for output
```

**Figure 45: Parameter file of `lognormal_new`**

The final step, which is computing the standard pairwise relative variogram is implemented in the `gam`, `gamv` and `gamv2004` programs from the GSLIB. The newly developed programs `gam_pair`, `gamv_pair` and `gamv2004_pair` have the same parameter files as the original programs. The only difference is on the value of the variogram type. In previous versions, the number (6) is used for

calculating the pairwise relative variogram. In developed versions, the number (-6) is used to calculate the standard pairwise relative variogram. As an example, the parameter file of gam is shown in Figure 46. The number (-6) for the standard pairwise relative variogram should be entered in **line 17**.

```

1           Parameters for gam_pair
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   2  1  2                     -number of variables, column numbers
6   -1.0e21    1.0e21          -trimming limits
7   gam.out                    -file for variogram output
8   1                           -grid or realization number
9   50  0.5  1.0                -nx, xmn, xsiz
10  50  0.5  1.0                -ny, ymn, ysiz
11  1   0.5  1.0                -nz, zmn, zsiz
12  2   10                       -number of directions, number of lags
13  1   0   0                    -ixd(1),iyd(1),izd(1)
14  0   1   0                    -ixd(2),iyd(2),izd(2)
15  1                             -standardize sill? (0=no, 1=yes)
16  1                             -number of variograms
17  1  1  -6                     -tail, head, variogram type

```

**Figure 46: Parameter file of gam\_pair**

### A.2.2 Correct unbiased estimate of traditional variogram

The standard pairwise relative variogram has a convergence issue which is fixed in the GSLIB compatible `correct_pair`. The basic idea is that this program is based on using the relationship between the normal and lognormal correlation coefficients. The parameter file of the `correct_pair` program is shown in Figure 48. The input file is the output of the `gam_pair` or `gamv_pair` programs that contains the standard pairwise relative variogram (**Line4**). The actual CV (correlation of variation) of data is entered in **Line5**. The output file specified in Line 6 is the correct unbiased estimate of the true variogram in standard GSLIB format.

```
1           Parameters for correct_pair
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   2                           -final CV
6   gam_correctpair.out        -file for variogram output
```

**Figure 47: Parameter file of correct\_pair**

### A.3 Cutting outliers

In the process of cutting and managing the outliers the `kt3d_new` program should be run twice and the `sgsim_new` program once. The sequence and how the parameter files should be set are explained hereby.

### A.3.1. Identify the outliers and volume of influence

To identify the outliers and specify the blocks being influenced by each outliers are implemented in the FORTRAN program `kt3d_new`, which is a developed version of the GSLIB program `kt3d` (Deutsch & Journel, 1998). The parameter file of the `kt3d_new` program is shown in Figure 48. This parameter file is almost the same as the `kt3d` parameter file except for the changes that have been made to lines 13 to 22. The parameter file used with `kt3d_new` can be used with `kt3d` by choosing zero for all the adjustment options. As the complete discussion about the `kt3d` program is available in (Deutsch & Journel, 1998), only the modified lines will be discussed here. For identifying the outliers, the option in **Line 13** should be greater than zero. In **Line 14**, based on the data values being studied, the user will be able to choose that above what threshold value the data should be considered as outlier. The user also can decide on the minimum weight assigned to an outlier in the estimation procedure in **Line15**. For example, by choosing the minimum weight of 0.05, all the grid blocks receiving weight more than this value from an outlier will be considered as the volume of influence for that specific outlier. The output file in **Line16** gives the volume of influence of each outlier that contains the outliers ID, outlier value, the outlier's weight assigned to the block, the coordinates of the blocks and the estimate value of the blocks. The other output file showing the local average of each outlier is specified in **Line 17**. In this first step, the option for managing and checking should be zero.

```

1           Parameters for KT3D_new
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   0 1 2 0 3 0                 -columns for DH,X,Y,Z,var,sec var
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  1                            -identify outlier (0-> not used)
14  18                          -max value to consider as outlier
15  0.05                        -min weight assigned to outlier
16  locations.out              -file for volume of influences
17  locations_ave.out          -file of local averages of ourtliers
18  0                            -manage outlier (0-> not used)
19  zcut.out                   -file with cutting levels and values
20  oms.out                    -output with managed outlier values
21  0                            -checking results (0-> not used)
22  checkfile.out              -output check file of local averages
23  50 0.5 1.0                 -nx, xmn, xsiz
24  50 0.5 1.0                 -ny, ymn, ysiz
25  1 0.5 1.0                  -nz, zmn, zsiz
26  1 1 1                      -x,y and z block discretization
27  4 8                         -min, max data for kriging
28  0                           -max per octant (0-> not used)
29  20.0 20.0 20.0            -maximum search radii
30  0.0 0.0 0.0               -angles for search ellipsoid
31  0 2.302                    -0=SK,1=OK,2=non-st SK,3=exdrift
32  0 0 0 0 0 0 0 0 0         -drift: x,y,z,xx,yy,zz,xy,xz,zy
33  0                           -0, variable; 1, estimate trend
34  extdrift.dat               -gridded file with drift/mean
35  4                           -column number in gridded file
36  1 0.2                      -nst, nugget effect
37  1 0.8 0.0 0.0 0.0         -it,cc,ang1,ang2,ang3
38  10.0 10.0 10.0           -a_hmax, a_hmin, a_vert

```

Figure 48: Parameter file of kt3d\_new

### A.3.2. Simulated uncertainty in local averages

The `sgsim_new` is the modified version of the `sgsim` program from GSLIB library. This program is modified to read the coordinates of the volume of influences related to each of the outliers and to compute the local averages for multiple realizations. The parameter file of the `sgsim_new` program is shown in Figure 49. The only difference between this parameter file and the one from the original program `sgsim` is the last four lines. The complete discussion about the `sgsim` program is available in (Deutsch & Journel, 1998) . The modified lines will be discussed here. The volume of influence file coming from the last step of `kt3d_new` is entered as an input file in **line 37**. The output file specified in **line 38** gives the uncertainty in local averages for each outlier for multiple realizations. Based on the uncertainty in local averages, the user can choose the arbitrary cutting level (e.g., median) in **line 39**. The final output in **line 40** contains the outlier ID, the outlier value, the mean of the local averages and the cutting values for each outlier based on an arbitrary cutting level.

```

1           Parameters for SGSIM_new
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   1 2 0 3 5 0                 -columns for DH,X,Y,Z,var,sec var
6   -1.0e21    1.0e21          -trimming limits
7   1                           -transform the data (0=no, 1=yes)
8   sgsim.trn                   -file for output trans table
9   0                           -consider ref. dist (0=no, 1=yes)
10  histsmth.out                -file with ref. dist distribution
11  1 2                          -columns for vr and wt
12  0.0  15.0                   -zmin,zmax(tail extrapolation)
13  1    0.0                    -lower tail option, parameter
14  1    15.0                   -upper tail option, parameter
15  1                           -debugging level: 0,1,2,3
16  sgsim.dbg                   -file for debugging output
17  sgsim.out                   -file for simulation output
18  1                           -number of realizations to generate
19  50  0.5  1.0                -nx, xmn, xsiz
20  50  0.5  1.0                -ny, ymn, ysiz
21  1  0.5  1.0                 -nz, zmn, zsiz
22  69069                       -random number seed
23  4  8                         -min and max original data for sim
24  12                          -number of simulated nodes to use
25  1                           -assign data to nodes (0=no, 1=yes)
26  1  3                         -multiple grid search(0=no,1=yes),num
27  0                           -maximum data per octant (0=not used)
28  10.0 10.0 10.0             -maximum search radii(hmax,hmin,vert)
29  0.0  0.0  0.0              -angles for search ellipsoid
30  51  51  11                  -size of covariance lookup table
31  0  0.60  1.0                -ktype: 0=SK,1=OK,2=LVM,3=EXDR,4=COLC
32  /data/ydata.dat            -file with LVM, EXDR,or COLC variable
33  4                           -column for secondary variable
34  1  0.1                      -nst, nugget effect
35  1  0.9  0.0  0.0  0.0      -it,cc,ang1,ang2,ang3
36  10.0 10.0 10.0             -a_hmax, a_hmin, a_vert
37  locations.out              -input file with volume of influences
38  average_sgs.out            -output of local averages of mult.rel
39  0.50                       -cutting level (median = 0.5)
40  z_cut.out                  -output of cutting level for outliers

```

**Figure 49: Parameter file of sgsm\_new**

### **A.3.3. Managing outliers based on cutting level**

The third step is focuses on running the `KT3D_new` program again. This time both the identify outlier (**Line13**) and the manage outlier (**Line18**) options should be more than zero (see Figure 50). The cutting level and values coming from the `sgsim_new` program are entered in **line19**. The output file specified in **line 20** gives the managed outlier value that corresponds to the cutting level chosen for uncertainty in the local averages.

```

1           Parameters for KT3D_new
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   0 1 2 0 3 0                 -columns for DH,X,Y,Z,var,sec var
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  1                            -identify outlier (0-> not used)
14  18                          -max value to consider as outlier
15  0.05                        -min weight assigned to outlier
16  locations.out               -file for volume of influences
17  locations_ave.out           -file of local averages of ourtliers
18  1                            -manage outlier (0-> not used)
19  zcut.out                    -file with cutting levels and values
20  oms.out                     -output with managed outlier values
21  0                            -checking results (0-> not used)
22  checkfile.out               -output check file of local averages
23  50 0.5 1.0                  -nx, xmn, xsiz
24  50 0.5 1.0                  -ny, ymn, ysiz
25  1 0.5 1.0                   -nz, zmn, zsiz
26  1 1 1                       -x,y and z block discretization
27  4 8                          -min, max data for kriging
28  0                           -max per octant (0-> not used)
29  20.0 20.0 20.0             -maximum search radii
30  0.0 0.0 0.0                -angles for search ellipsoid
31  0 2.302                     -0=SK,1=OK,2=non-st SK,3=exdrift
32  0 0 0 0 0 0 0 0 0          -drift: x,y,z,xx,yy,zz,xy,xz,zy
33  0                            -0, variable; 1, estimate trend
34  extdrift.dat                -gridded file with drift/mean
35  4                            -column number in gridded file
36  1 0.2                       -nst, nugget effect
37  1 0.8 0.0 0.0 0.0          -it,cc,ang1,ang2,ang3
38  10.0 10.0 10.0            -a_hmax, a_hmin, a_vert

```

**Figure 50: Parameter file of kt3d\_new**

#### **A.3.4. Check the results**

By reducing the values of outliers, the `KT3D_new` program can be run for the estimation with managed outlier values. The local averages can then be compared with the simulated local averages. For this purpose, just the option for checking in **line 21** should be greater than zero (see Figure 51), so the new outlier values would be used to do the kriging estimation. The output file in line 22 contains the original, simulated and OMS local averages for each outlier value.

```

1           Parameters for KT3D_new
2           *****
3   START OF PARAMETERS:
4   /data/cluster.dat           -file with data
5   0 1 2 0 3 0                 -columns for DH,X,Y,Z,var,sec var
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  1                           -identify outlier (0-> not used)
14  18                          -max value to consider as outlier
15  0.05                        -min weight assigned to outlier
16  locations.out              -file for volume of influences
17  locations_ave.out          -file of local averages of ourtliers
18  1                           -manage outlier (0-> not used)
19  zcut.out                   -file with cutting levels and values
20  oms.out                    -output with managed outlier values
21  1                           -checking results (0-> not used)
22  checkfile.out              -output check file of local averages
23  50 0.5 1.0                 -nx, xmn, xsiz
24  50 0.5 1.0                 -ny, ymn, ysiz
25  1 0.5 1.0                  -nz, zmn, zsiz
26  1 1 1                      -x,y and z block discretization
27  4 8                        -min, max data for kriging
28  0                          -max per octant (0-> not used)
29  20.0 20.0 20.0            -maximum search radii
30  0.0 0.0 0.0              -angles for search ellipsoid
31  0 2.302                   -0=SK,1=OK,2=non-st SK,3=exdrift
32  0 0 0 0 0 0 0 0 0        -drift: x,y,z,xx,yy,zz,xy,xz,zy
33  0                          -0, variable; 1, estimate trend
34  extdrift.dat              -gridded file with drift/mean
35  4                          -column number in gridded file
36  1 0.2                     -nst, nugget effect
37  1 0.8 0.0 0.0 0.0        -it,cc,ang1,ang2,ang3
38  10.0 10.0 10.0          -a_hmax, a_hmin, a_vert

```

**Figure 51: Parameter file of kt3d\_new**