University of Alberta

Application of Logratios for Geostatistical Modelling of Compositional Data

by

Michael Robert Job

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

Fall 2012
Edmonton, Alberta

Abstract

Numeric data for earth sciences often represent fractions or percentages of a whole, such as the chemical or mineralogical composition of a rock. The individual components are non-negative and have a constant sum of 100%. Satisfying these constraints at unsampled locations after estimation or simulation is a practical requirement, but not guaranteed by conventional mapping and modelling techniques. The components are constrained to the constant sum, which means that they are not free to vary independently, and there is at least one negative correlation. Standard statistical techniques are therefore not suited to compositional data, so transformations using the logarithms of the component ratios (logratios) are used to overcome these problems. Linear averaging and back-transformation of logratios results in a geometric rather than an arithmetic mean, which will result in a bias. A procedure using normal scores transformation of the logratio values and multiGaussian kriging was devised to overcome this bias. The key objective is to avoid estimating the logratios directly and then back-transforming into original data units. Instead, the conditional distributions of the components are modelled. Ordinary kriging, multiGaussian kriging and conditional simulation were used on data from the Alberta Oil Sands to assess the performance of the compositional geostatistics approach.

Acknowledgments

I would like to thank my supervisor Dr. Clayton Deutsch at the University of Alberta for his encouragement, advice and reminders to keep focused. Thank you also to the members of the Centre for Computational Geostatistics, who warmly welcomed me to Alberta, and especially to Jeff Boisvert and John Manchuk who helped me renew some long-forgotten mathematics.

Thank you to the directors of Quantitative Group in Fremantle, Australia – John Vann for enthusiastic discussions and manuscript review, and Scott Jackson for friendship and providing a workplace where knowledge is valued.

Thanks also to Georges Verly for very helpful discussions on multiGaussian kriging.

Most of all special thanks to my family, wife Fiona and sons Aidan and Ryan for their support and for displaying a great deal of patience and tolerance over these last few years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

CCG	Centre for Computational Geostatistics at the University of Alberta
CV	Coefficient of Variation
iid	Independent and identically distributed
LMC	Linear Model of Coregionalization
LU	Lower/Upper triangular matrix decomposition
MGK	MultiGaussian Kriging
MSE	Mean Squared Error
OK	Ordinary Kriging
OCK	Ordinary Cokriging
PCA	Principal Component Analysis
QKNA	Quantitative Kriging Neighbourhood Analysis
RCK	Rescaled Ordinary Cokriging
ReV	Regionalized Variable
RF	Random Function
RV	Random Variable
SGS	Sequential Gaussian Simulation
SK	Simple Kriging
SCK	Simple Cokriging
TBS	Turning Bands Simulation

Logratio transforms

alr	additive logratio transform
clr	centred logratio transform
ilr	isometric logratio transform
mlr	multiplicative logratio transform

Oil sands components

B	Bitumen
C	Coarse solid fraction
F	Fine solid fraction
W	Water

B:C	Bitumen/Coarse ratio
F:C	Fines/Coarse ratio
W:C	Water/Coarse ratio

Symbols

$\mathbf{u}$	a location in space
$Z(\mathbf{u})$	a random variable
$\mathbf{h}$	distance or lag vector
$C(\mathbf{h})$	covariance between points separated by distance $\mathbf{h}$
m	mean
σ^2	variance
E	expected value
$\gamma(\mathbf{h})$	variogram at lag $\mathbf{h}$
λ_i	weight assigned to data i
μ	Lagrange parameter
φ	Gaussian transformation
D	number of parts or components of composition
$\mathcal{S}^D$	D-part simplex
$\boldsymbol{\kappa}$	Closed sum constant
$\mathbb{R}^n$	n-dimension real Euclidean space
log	logarithm
exp	exponent
$g(\mathbf{x})$	geometric mean
$C(\mathrm{x})$	closure operation
$\oplus$	perturbation
$\odot$	power transformation
r	replaced component
t	number of rounded zeros
δ	imputed value for rounded zero
J	multinormal random vector
ω	order of variogram

Chapter 1 Introduction

1.1 Outline of Problem

Compositional data is defined as representing proportions of a whole – for example, a sample with whole rock geochemistry that sums to 100%, or the proportion of various mineral species in a rock, which sum to unity. Analysis of compositional data utilizes the understanding that the data contains information about the *relative* magnitudes of the components, not just the absolute magnitudes, and therefore these relationships can be expressed as ratios. Direct mathematical analysis of ratios is problematic due to the constant sum constraint, however, meaning that analysis of correlations of the raw components and other forms of standard multivariate statistical analysis, which are designed for unconstrained data, are not suitable (Aitchison, 1999).

Problems in applying standard statistical techniques to compositional data were recognized by Pearson in the late nineteenth century (Pearson, 1897). Chayes (1960) later recognized that the constant sum constraint suppressed positive and increased negative correlations between the components, although he did not propose a solution. Davis (1973, p.81) provides a very simple but effective example of how the 'closure problem' of a constant sum composition can result in a negative correlation of the relative magnitudes, even though the correlation between the absolute magnitudes was positive. At that stage, no completely satisfactory way had been developed for dealing with the closure problem according to Davis. It was not until Aitchison (1982, 1986) introduced the 'statistical analysis of compositional data' that methods were devised to address these problems.

Aitchison (1986, p. 65) noted that the logarithms of ratios are easier to handle mathematically and interpret statistically than the ratios themselves. For example, there is no relationship between the variance (var) of (x_i / x_j) and variance of (x_j / x_i), but a relationship exists between the logarithms of these ratios:

$$\mathrm{var}\{\ln(x_i / x_j)\} = \mathrm{var}\{\ln(x_j / x_i)\} \qquad \textbf{1-1}$$

Aitchison therefore proposed a number of transforms of the ratios using logarithms (referred to as 'logratios'). Atchley et al. (1976) had previously demonstrated that the direct use of ratios can result in very skewed and leptokurtic distributions as the coefficient of variation (CV) of the denominator increases. This leads to an increase in the covariances and spurious correlations between the ratio variables. Furthermore, standard unconstrained multivariate techniques can be applied to the logratio transformed data, with inferences translatable back into the original component space (Pawlowsky-Glahn et al., 2011). More information about the main transformations and operations used for compositional data analysis are given in Chapter 2.

The vast majority of work on compositional data in the last ten years has come from the following authors – Pawlowsky-Glahn, Martin-Fernandez and Barcelo-Vidal at Girona, Egozcue at Catalonia, Tolosana-Delgado at Gottingen as well as Aitchison himself at Glasgow. Most of this work has been published in Mathematical Geosciences (formerly Mathematical Geology), and various workshops/short courses under the 'CoDaWork' consortium.

Compositional data analysis in the earth sciences has proven to be successful in some non-spatial applications such as mineralogy and petrology (for example Thomas and Atchison, 2005 and Martin-Fernandez et al., 2005), where useful geological inferences can be made. Further examples are discussed in Section 2.7. The method has been controversial, however, and there are many critics of the techniques in the literature. Shurtz (2003) rejected the use of the proposed transforms as being 'unnecessary', and Baxter et al. (2005) regarded the theoretical basis as sound, but had concerns about the practical implementation and their empirical experience with the method – in particular, they found that principal component analysis (PCA) to recover cluster structure in the data does not work well with logratios.

Compositional data analysis techniques were extended to spatial data in the 1990's and 2000's, notably by Pawlowsky-Glahn and Olea (2004), although they have not found broad acceptance in geostatistics (Tolosana-Delgado et al., 2008), particularly in the mining industry. There are many situations in geostatistical modelling where the preservation of proportions of input data is of practical importance in spatial estimation or simulations. While a number of geostatistical estimation methods have been proposed (e.g., Pawlowsky-Glahn and Olea, 2004, Tolosana-Delgado et al., 2008) there have also been warnings in the literature (e.g., Lan et al., 2006), about the theoretical and practical reliability of these approaches, particularly where the estimation technique involves an averaging of log-transformed data.

Lan et al. (2006) and Lan (2007) discussed the benefits of applying the logratio transform for statistical analysis, but they concluded that using logratios to make spatial estimates using a linear approach such as kriging will result in a bias. This is due to the back-transform of the arithmetic average of logratio values returning the geometric mean of the proportions being studied, which is not the correct result for variables that average linearly.

Tolosana-Delgado et al. (2008), however, maintained that the back-transformed results after a 'standard' geostatistical (cokriging) approach are linear, unbiased (null expected error) and with minimal variance between the true and estimated value on a relative scale. They provided an example of their approach, which utilizes the Linear Model of Coregionalization (LMC; Journel and Huijbregts, 1978, p. 172) on the logratio transformed data, and ran a cokriging. They also suggested that conditional simulation would be possible by a straightforward application of sequential Gaussian simulation (SGS) or LU decomposition. Their method yielded positive and bounded compositions, but this is a feature of the back-transform. In summary, this method is claimed to produce robust results without the need to introduce constraining algorithms or posterior correction. Of course, linear unbiased estimates of logratios do not lead to unbiased estimates of the back transformed proportions.

1.2 Plan of Thesis

Basic relevant geostatistical theory and descriptions of techniques used in this thesis are briefly presented in Chapter 1. Theory and suggested application of compositional data, including transformations are presented in Chapter 2, with a small worked example of multiGaussian kriging in Chapter 3 The full case study is introduced in Chapter 4 and is based on oil sands data to explore the implications of various approaches to spatial modelling of compositional data. Linear kriging applications and their limitations for the

case study compositional data are discussed in Chapter 5; followed by multiGaussian kriging in Chapter 6 and conditional simulation applications in Chapter 7. Chapter 8 draws conclusions from the work and recommends areas of interest for further research.

1.3 Geostatistics

Geostatistics is a branch of applied statistics that deals with phenomena that fluctuate in space (Olea, 1991), such as gold grades in a vein, thicknesses of stratigraphic layers, or porosity and permeability of oil reservoirs. The term 'geostatistics' was first used by Matheron (1962), when he introduced his 'new science' that dealt with spatial aspects of variables.

Some of the main applications of geostatistics are to provide 'best' estimates of mineral/element grades at unsampled locations (and therefore estimates of the total mineral resource of a deposit), and to characterize the uncertainty and risk for these estimates. The terminology used throughout this thesis is drawn from the mining industry.

For more information on geostatistics the reader is referred to David (1977), Journel and Huijbregts (1978), Isaaks and Srivastava (1989), Deutsch and Journel (1998), Chiles and Delfiner (1999) and Sinclair and Blackwell (2002).

1.3.1 Random Functions and Regionalized Variables

A random variable (RV) is a variable that takes a range of values according to a probability distribution function. Of course, there is a real and defined value at the unsampled locations, but since this value is unknown, a probabilistic method using a distribution function is needed to estimate the value. This cumulative distribution function (cdf) is defined as:

$$F(\mathbf{u};z) = \text{Prob}\{Z(\mathbf{u}) \leq z\} \qquad \textbf{1-2}$$

Where **u** is a spatial location and z is an outcome (or threshold) of the random variable Z. A Random Function (RF) is the family or set of all RVs over an area of interest – in our case usually a geological 'domain'. This 'domain' generally implies relative homogeneity of some physical aspect of the geological environment, such as rock type or lithofacies. The concept of the Regionalized Variable (ReV) (Matheron, 1963) is an extension of the RV to spatial applications.

1.3.2 Stationarity

Strict stationarity requires all the moments of the distribution to be invariant under translation; i.e., exactly the same distribution at every point in the field considered. In mining applications this assumption is too strong, so the second-order stationarity is used – that is, the expected value (mean, m) of the variable is constant, and the covariance function ($C(\mathbf{h})$) between two points located at **u** and (**u** + **h**) where h is a distance vector, is independent of the locations of **u** and (**u** + **h**):

$$C(\mathbf{h}) = E[\{Z(\mathbf{u}) - m\}\{Z(\mathbf{u}+\mathbf{h}) - m\}] \qquad \textbf{1-3}$$

Stationarity is a property of the RF model, not a property of the underlying spatial distribution (Journel and Deutsch, 1998). The decision of stationarity cannot be proven or refuted *apriori*, but can be shown to be inappropriate after the initial decision is made. Therefore, the choice of domains can be an iterative process, and is a critical component of a geostatistical study. Even though geostatistical techniques may work for poorly chosen domains, the results will be less than optimal.

1.3.3 Declustering

Mineral exploration data (usually from drillholes) is often concentrated (or 'clustered') in zones of higher grades – therefore a raw histogram of the data for a particular geological domain (assuming equal sample support size) will be biased. Declustering is consequently required to estimate the underlying unbiased grade distribution. Declustering techniques assign a weight to each datum based on the proximity to other data. The main methods are polygonal and cell declustering (Journel, 1983; Isaaks and Srivastava, 1989, p.237 – 248) and kriging weight declustering (Isaaks and Srivastava, 1989, p. 510; Deutsch, 1989).

1.3.4 The Variogram

The variogram is the basic tool of geostatistics, as it characterizes a regionalized variable and is used for geostatistical estimation (kriging) and conditional simulation techniques. For a stationary RF, the variogram is defined as:

$$\gamma(\mathbf{h}) = \frac{1}{2} Var[Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})] \qquad \textbf{1-4}$$

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

with $C(\mathbf{h})$ being the stationary covariance and $C(\mathbf{0}) = \mathrm{Var}\{Z(\mathbf{u})\}$.

In practice, the experimental variogram is for a vector ($\mathbf{h}$) is calculated as:

$$\gamma(\mathbf{h}) = \frac{1}{2N} \sum_{i=1}^{N} [\{Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\}^2] \qquad \textbf{1-5}$$

Various tolerance parameters such as the lag tolerance and bandwidth can be applied if the sampling is not on a regular grid. Calculation of the experimental variogram should also consider anisotropy in the data, for example the major direction of continuity for a mineralized geological unit. The experimental variogram must then be modelled so that the variogram value can be estimated for all distances and directions, and to ensure that the model is positive definite; i.e., the variance of any linear combination of the variogram must be positive. There are a number of models that are admissible; e.g., nugget, spherical, Gaussian, exponential.

1.3.5 Kriging

Kriging is a linear estimator of the form:

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i [Z(\mathbf{u}_i) - m(\mathbf{u}_i)] \quad \textbf{1-6}$$

where $Z^*(\mathbf{u})$ is the estimate at location $\mathbf{u}$, $Z(\mathbf{u}_i)$ are the data values $i = 1,...,n$, $m(\mathbf{u})$ and $m(\mathbf{u}_i)$ are the means of $Z(\mathbf{u})$ and $Z(\mathbf{u}_i)$, and λ_i are the weights applied to the data values. For simplification, it is easier to work with the residuals, so the linear estimator becomes:

$$Y(\mathbf{u}) = Z(\mathbf{u}) - m(\mathbf{u})$$

$$Y^*(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{u}_i) \quad \textbf{1-7}$$

The goal is to determine the weights that will minimize the variance of the estimator:

$$Var\{Y^*(\mathbf{u}) - Y(\mathbf{u})\} = E[Y^*(\mathbf{u}) - Y(\mathbf{u})]^2 \quad \textbf{1-8}$$

Substituting equation 1-7 into this equation and expanding results in:

$$Var\{Y^*(\mathbf{u}) - Y(\mathbf{u})\} = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) - 2 \cdot \sum_{i=1}^{n} \lambda_i C(\mathbf{u}, \mathbf{u}_i) + C(\mathbf{0}) \quad \textbf{1-9}$$

where $C(\mathbf{u}_i,\mathbf{u}_j)$ is the covariance between data at points i and j, $C(\mathbf{u},\mathbf{u}_i)$ the covariance between the data at point i and the location to be estimated, and $C(\mathbf{0})$ the variance of the data. These covariances (for different distances and directions) are derived from the variogram model. To then minimize the kriging error variance, the partial derivatives with respect to the weights are set to zero:

$$\frac{\partial[C(\mathbf{0})]}{\partial \lambda_i} = 2 \cdot \sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) - 2 \cdot C(\mathbf{u}, \mathbf{u}_i) \qquad \textbf{i =1,…,n} \quad \textbf{1-10}$$

which results in the simple kriging (SK) system:

$$\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) = C(\mathbf{u}, \mathbf{u}_i) \quad \textbf{1-11}$$

with kriging variance:

$$\sigma^2_{SK} = C(\mathbf{0}) - \sum_{i=1}^{n} \lambda_i \cdot C(\mathbf{u}, \mathbf{u}_i) \quad \textbf{1-12}$$

For SK, the weights are not constrained, and since it works with the residuals from the mean, then the mean must be known. Where the mean is unknown (but constant in the

local neighborhood), then the weights are constrained to sum to one, leading to the ordinary kriging (OK) system of equations:

$$\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) + \mu = C(\mathbf{u}, \mathbf{u}_i) \qquad \mathbf{i = 1,\ldots,n} \tag{1-13}$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

where μ is the Lagrange parameter, and the kriging variance is:

$$\sigma_{OK}^2 = C(\mathbf{0}) - \mu - \sum_{i=1}^{n} \lambda_i \cdot C(\mathbf{u}, \mathbf{u}_i) \tag{1-14}$$

1.3.6 Cokriging

In cases where there are one or more secondary variables that are correlated with the primary variable, cokriging can be used. It is particularly useful if a variable of interest has been under-sampled with respect to other variables (i.e., heterotopic sampling).

Cokriging requires a joint model to describe the covariances between the variables – this is achieved by modelling the variograms for each variable ('direct variograms'), plus the cross-variograms for each pair of variables. The cross-variogram between each pair of variables (Z and Y) is defined as:

$$\gamma_{Z,Y}(\mathbf{h}) = \frac{1}{2} E\{(Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u}))(Y(\mathbf{u}+\mathbf{h}) - Y(\mathbf{u}))\} \tag{1-15}$$

For cokriging, linear combinations of variables are themselves treated as ReVs and their variances must be positive definite. The only practical model that ensures this condition is met is the LMC:

$$C_{ZY}(\mathbf{h}) = \sum_{l=1}^{L} b_{ZY}^{(l)} C_l(\mathbf{h}) \tag{1-16}$$

where l is the number of covariance structures 1,…,L, and b is the coefficient (variance contribution) for each structure. In practice, it is very difficult to model more than 3 or 4 variables with the LMC - if n is the number of variables, then n direct variograms and $n(n\text{-}1)/2$ cross-variograms are required. Fitting of models that are consistent with the experimental variograms can be very demanding – automated fitting is available in some software packages, but in many situations the model fitting must be a compromise.

The cokriging system is an extension of the kriging system, with the ordinary cokriging (OCK) estimator for two variables:

$$Z_{OCK}^{(1)*}(\mathbf{u}) = \sum_{i=1}^{n_1} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \sum_{j=1}^{n_2} \lambda_j(\mathbf{u}) Y(\mathbf{u}_j) \tag{1-17}$$

and kriging variance:

$$\sigma^2_{OCK} = C_{ZZ}(\mathbf{0}) - \mu - \sum_{i=1}^{n_1} \lambda_i(\mathbf{u}) C_{ZZ}(\mathbf{u}, \mathbf{u}_i) - \sum_{j=1}^{n_2} \lambda_j(\mathbf{u}) C_{ZY}(\mathbf{u}, \mathbf{u}_j) \quad \textbf{1-18}$$

A shortcoming of OCK is that the sum of the weights for the primary variable is one, so the sum of the weights for the secondary variables must be zero. This means that the weights of the secondary variables are small, but there is the risk of high negative weights (Isaaks and Srivastava, 1989, pp. 400-416 and Goovaerts, 1998). Therefore, 'rescaled' or 'standardized' cokriging (RCK) is recommended where a single unbiasedness constraint forces all primary and secondary weights to sum to one (Deutsch and Journel, 1998, p. 74). This is achieved by rescaling the secondary variable so that the mean is equal to the primary variable, and the cokriging estimator is:

$$Z^{(1)*}_{RCK}(\mathbf{u}) = \sum_{i=1}^{n_1} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \sum_{j=1}^{n_2} \lambda_j(\mathbf{u}) [Y(\mathbf{u}_j) - m_2 + m_1] \quad \textbf{1-19}$$

For simple cokriging (SCK), there remain no constraints on the weights. If the means of the variables have been standardized to zero, with a variance of one, then SCK is applicable for the multiGaussian kriging approach.

1.3.7 MultiGaussian kriging

MultiGaussian kriging (MGK) (Verly, 1983) is a non-linear estimation technique, where the conditional distribution (and conditional expectation) of a variable is modelled. The mean, probabilities to exceed a certain cut-off (and average values above and below that cut-off) and value that corresponds to a p quantile can then be calculated. These probabilities are generally not available with linear kriging.

Such non-linear kriging techniques (others include indicator kriging and lognormal kriging) are actually linear kriging applied to non-linear transform of the original data. Earth science data rarely have strictly Gaussian distributions, so for MGK, the variable (a realization of a random function Z_x) is transformed into a Gaussian random field Y_x with a mean of zero and variance of one, also known as a normal scores transform:

$$Y_{(x)} = \varphi^{-1}(Z_{(x)}) \quad \textbf{1-20}$$

The normal score (or Gaussian) transform φ is graphically defined by a one to one correspondence between the cdf of the RF $Z_{(x)}$ and a standard normal cdf (Journal and Huijbregts, 1978, p. 478).

They key and critical assumption is that the distribution of any value Y_x is multivariate Gaussian (Verly, 1983) i.e., every possible linear combination of the values has a normal distribution. For MGK, the conditional expectation (and conditional variance) at unsampled locations is required – SK is usually implemented to provide this. If the multiGaussian assumption holds, a Gaussian conditional probability density function where the mean is equal to the SK estimate y_x^{SK} and the variance is equal to the SK variance $(\sigma_x^{SK})^2$ is produced.

1.3.8 Conditional Simulation

Geostatistical simulation is a spatial extension of the concept of Monte Carlo simulation, where the declustered data histogram is reproduced, and the variance and spatial variability of data; i.e., from the variogram model is replicated.

Geostatistical simulations generate a *set* of images, or 'realizations' as opposed to estimates, which output a single image. The realizations constitute a range of spatial images that are consistent with the known statistical moments (variogram and histogram) of the declustered input data, and in the case of conditional simulations, the data themselves. Geostatistical simulations can be used to assess uncertainty over various scales or volumes (e.g., mining production intervals), and can assist in evaluating drill hole spacing, mining selectivity and blending, and mine financial modelling.

The two most commonly used methods for conditional simulation for continuous variables (e.g., metal grades) in the mining industry are:

1. Sequential Gaussian Simulation (SGS: Isaaks, 1990). The conditioning samples are migrated to the closest grid node, and a random path is defined through all the grid nodes. SK is used to construct the conditional Gaussian distribution at each node in the path using the conditioning and previously simulated data. A simulated value is drawn from this conditional distribution and added to the grid node. The next node on the random path is then simulated until all nodes are completed. This process is then repeated to generate *n* realizations; and
2. Turning Bands Simulation (TBS: Journel, 1974), which is efficient for generating non-conditional simulations. The method works by simulating on one-dimensional lines regularly spaced in three dimensions and then combining in three-dimensional space. The conditioning is performed by a separate kriging step.

Both of these simulation methods require the data to have a Gaussian distribution, which is extremely rare in earth sciences data, and a Gaussian transform of the data is therefore required. Checking that the assumption of a multivariate Gaussian distribution (as described above) holds is required for conditional simulation. After simulation of Gaussian values, the values are back-transformed into the original data space. Because SGS (and TBS) uses SK, the assumptions of stationarity (constant mean over the entire domain) are stronger than for OK (Deutsch and Journel, 1998, p. 145).

The results from all conditional simulation methods require extensive checking against the input data to ensure the histogram and variogram have been honoured.

In the multivariate case, relationships between variables are ideally preserved with joint simulation. For well-correlated variables, a Markov model approach can be adopted (Deutsch and Journel, 1998, p. 124) where the primary variable is simulated, and the secondary variables are simulated by collocated cokriging conditional on the simulated primary variable, and so on. However, the variance from the collocated cokriging can be inflated, which is a problem for simulations, since the variance is used directly to define the spread of the conditional distribution from which the random values are drawn. In such cases, full cosimulation, where modelling of direct and cross variograms for the Gaussian transformed data that conform to the LMC is recommended.

1.4 Concluding Comments for Chapter 1

The aim of the work presented here is to develop methods that allow the use of logratio transformed compositional data in geostatistics that will not introduce bias. Comparison of the use of logratio data using conventional linear geostatistical approaches (kriging and cokriging) against non-linear methods (multiGaussian kriging and conditional simulation) will demonstrate that estimation/simulation of local conditional distributions using non-linear techniques is the preferred option. In addition, there will be analysis of any advantages or disadvantages of the compositional data methods over conventional (i.e., not logratio transformed) methods.

Chapter 2 Compositional Data Theory

2.1 Introduction

The basic concepts of compositional data theory are presented in this chapter. For more details the interested reader can refer to Aitchison (1986), Pawlowsky-Glahn and Olea (2004), the Centre for Computational Geostatistics (CCG) Guidebook on Compositional Geostatistics (Manchuk, 2008), and the many other publications in Mathematical Geosciences as cited in Chapter 1.

Standard multivariate analysis is applicable for unconstrained vector data from real Euclidean space (Pawlowsky-Glahn and Egozcue, 2006). The sample space of compositions, however, is constrained to the restricted space of the simplex, a generalization of a triangle and tetrahedron (Aitchison et al., 2002). The D-part simplex, $\mathcal{S}^D$, is a subset of D dimensional real space, and for $D = 2$ it can be represented as a line, for $D = 3$ a triangle and for $D = 4$ a tetrahedron (Pawlowsky-Glahn and Egozcue, 2006). The simplex $\mathcal{S}^D$ is defined as:

$$\boldsymbol{\mathcal{S}^D} = \{[x_1, x_2,...,x_D] : x_j > 0; j = 1,2,...,D; x_1 + x_2 + ... + x_D = \kappa\}, \quad \textbf{2-1}$$

where $\boldsymbol{\kappa}$ can be 1, 100, 10^6 or any other constant (the 'closed sum').

Absolute values of the components in a composition are of limited meaning unless they are compared, by ratios, with other components (Aitchison and Egozcue, 2005). By applying a logratio transformation to the data in the original sample space (the simplex), the compositions are projected to multivariate real space, and statistical methods designed for multivariate normal distributions can be used (Aitchison, 1986, p. 114, Pawlowsky-Glahn and Olea, 2004, p. 26).

Real Euclidean space $\mathbb{R}^n$ is a linear vector space where vectors can be added and vectors multiplied by a scalar. For n =1, 2 and 3, the space can be represented geometrically (1 is the real number line, 2 is a two-dimensional plane, and 3 is ordinary three-dimensional space). If two vectors are added in $\mathbb{R}^n$, the resulting vector is again a vector in $\mathbb{R}^n$, and if a vector in $\mathbb{R}^n$ is multiplied by a scalar, then the result again is in $\mathbb{R}^n$ (Schneider et al., 1982, p. 66).

2.2 Logratio Transforms

There are four logratio transforms: the additive logratio (alr), centered logratio (clr), multiplicative logratio (mlr) (all Aitchison, 1982), and the isometric logratio (ilr), introduced by Egozcue et al. (2003). These logratio transforms are 'lossless', meaning that they do not lose any of the information from the whole composition. This is because '…there is a one-to-one correspondence between any D-part composition (i.e., consisting of D components, $x_1,\ldots,x_D$) and its logratio vector' (Aitchison, 1999).

The additive logratio (alr) transform is shown in Equation 2-2:

$$y_i = \log\left(\frac{x_i}{x_D}\right), \qquad i = 1,...,D-1 \qquad \textbf{2-2}$$

where the denominator (x_D) can be any of the components, with the conditions that:

1. The same component must be used as denominator for all data points; and
2. Each component must be strictly >0.

The choice of denominator does not affect the results of analyses (Aitchison, 1986, p. 142). The alr transformation results in one less transformed variable than the number of components considered. The alr back-transform, also known as the additive logistic transform (agl) (Aitchison, 1986, p.136), for the numerator components is shown in Equation 2-3 – since there is one less term in the alr transform compared to the original composition, the back-transform for the denominator is simply the difference between the sum of the D-1 components from the constant sum constraint (Equation 2-4):

$$x_i = \frac{\exp(y_i)}{\sum_{i=1}^{D-1}\exp(y_i)+1}, \qquad i = 1,...,D-1 \qquad \textbf{2-3}$$

$$x_D = \frac{1}{\sum_{i=1}^{D-1}\exp(y_i)+1} \qquad \textbf{2-4}$$

The centered logratio (clr) transform is shown in Equation 2-5:

$$y_i = \log\left(\frac{x_i}{g(\mathbf{x})}\right), \qquad i = 1,...,D \qquad \textbf{2-5}$$

where $g(\mathbf{x})$ is the geometric mean of all components. The clr back-transform is shown in Equation 2-6:

$$x_i = \frac{\exp(y_i)}{\sum_{i=1}^{D}\exp(y_i)}, \qquad i = 1,...,D \qquad \textbf{2-6}$$

The multiplicative logratio (mlr) transform is similar to the alr, but uses a 'filler' component (a component introduced to ensure the composition sums to unity, as discussed below in Section 2.3) as the denominator.

The isometric logratio (ilr) transform (Egozcue et al., 2003), with the concept of the transform shown in Equation 2-7:

$$ilr(x) = V.clr(x) \qquad \textbf{2-7}$$

where V is a matrix of D rows and (D – 1) columns such that $V \cdot V^t = I_{D-1}$ (identity matrix of D – 1 elements) and $V \cdot V^t = I_D + a1$, where a may be any value, and 1 is a matrix full of ones (Tolosana-Delgado, 2008). This can be written out as Equation 2-8, as expanded by Thio-Henestrosa and Martin-Fernandez, 2005:

$$y_i = \frac{1}{\sqrt{i(i+1)}} \log\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i} \right) \quad (i = 1, \ldots, D-1) \qquad \textbf{2-8}$$

with the ilr back-transform shown in Equation 2-9:

$$x = ilr^{-1}(y) = \left[\left(1 + \frac{\sum_{i=0, i\neq 1}^{D} f(i)}{f(0)} \right)^{-1}, \ldots, \left(1 + \frac{\sum_{i=0, i\neq D}^{D} f(i)}{f(D-1)} \right)^{-1} \right], \qquad \textbf{2-9}$$

where $f(i) = \left(\frac{1}{f(i-1)} \exp(\sqrt{i(i+1)}\, y_i) \right)^{-1/i}$ and $f(0) = 1$

The ilr transform returns one less variable than the number of components in the original data. Note that these equations conventionally refer to natural logarithms, but the transforms can be used with logarithms of any base.

2.2.1 Discussion

The choice of denominator for the alr transform can potentially be a problem, because the post-transform data sets using different denominators will generally be substantially different from each other. However, as Aitchison (1986, p. 142) has shown, this does not make any difference when using standard linear statistical methods. Nevertheless, due to the non-isometric character of the transformation, Pawlowsky-Glahn (2004) warns that care needs to be exercised when performing experimental data analysis, such as interpreting scatterplots.

In contrast, the clr transformation is symmetric, but the sum of the components after transformation is necessarily zero. This means that for random compositions the covariance matrix of the clr transformed vectors is singular; i.e., it has a determinant of zero (Pawlowsky-Glahn, 2004). For a geostatistical application, however, the covariance matrices required for kriging are drawn from a model of the covariances, not the actual experimental covariances, and as a result, none of the square matrices required are non-invertible, and the kriging equations can be solved.

The ilr transform has the advantage of conserving geometry (e.g., angles and distances) in both the simplex and real space. They are coordinates in an orthogonal system, and therefore classical multivariate techniques can be used (Pawlowsky-Glahn and Egozcue, 2006). A disadvantage is that the theory behind and the calculation of the transform is very complex. The back-transform in particular is difficult to implement. For this reason, the ilr transform has not been used for the case study presented here.

In conclusion, both the alr and clr transforms have been used in the case study presented in this thesis – the mlr transform is not used because the oil sands data already forms a full composition, and no filler component is required.

2.3 Subcompositions and Closure

There are difficulties when dealing directly with ratios of subcompositions, which is another advantage for introducing the logratio transform. A subcomposition is a subset extracted from a full composition, and normalized. Even though the ratio of any two components of a subcomposition is the same as the ratios in the full composition (Aitchison, 1986, p. 35), the covariance relationships between the variables in the subcomposition are not the same as those that exist between the same variables in the full composition, and there may be no relationship between the two covariance structures (Aitchison, 1986, p. 55).

For example, consider assay data for an iron ore deposit with Fe%, P%, SiO_2%, Al_2O_3% and LOI% (Loss on Ignition), and usually CaO%, MgO%, TiO_2%, S%, MnO% and K_2O%. Due to the presence of additional trace elements/compounds, the assays for these eleven components will rarely sum to 100%. Therefore, a residual (or 'filler') component can be used to complete the composition. However, if the residual part is of no interest, or indeed if only three of the variables (say Fe, Al_2O_3, SiO_2) are of interest, then the subcomposition can be taken for the components of interest using the closure operation:

$$x = C(x) = \left[\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right] \qquad \textbf{2-10}$$

where κ is the constant sum. In the above case, any analysis in logratio space for the eleven-part (or for that matter, the three-part) subcomposition will be consistent with the full twelve-part composition. The variance-covariance relationships between the ratios (and logratios) of the components will be the same for the subcomposition as they will be for the entire composition – this is not true for the raw components.

Note also that, for certain samples in an iron ore deposit, the sum of the assays could be greater than 100% due to laboratory or other procedural errors. In this case, some form of normalization would be required – whether proportional for each component using the closure operation (Equation 2-10), or if there is confidence for the results of some variables, then maintaining these values, and adjusting the values that have lower confidence.

2.4 Basic Operations

Two data manipulation procedures that are used in the compositional data framework are worth discussing. These operations are applied in the simplex, and logratio transformation is not required. The first, perturbation (denoted by $\oplus$), involves multiplying the components of a composition with the corresponding components of another composition and then applying the closure operation:

$$z = x \oplus y = C[x_1 \cdot y_1; ...; x_D \cdot y_D] \quad \textbf{2-11}$$

The second, power transformation (denoted by $\odot$), involves raising each component in a composition by a constant and the applying the closure operation:

$$z = \lambda \odot x = C[x_1^{\lambda}; ...; x_D^{\lambda}] \quad \textbf{2-12}$$

where λ is a real number.

In the earth sciences, perturbation can be used to analyse a system by using 'before' and 'after' compositions. For example, consider a three-part clay composition in weathered ultramafic rocks (vermiculite, kaolin, smectite). If x = [35, 50, 15] at one location (or point in time) and another composition y = [5, 90, 5], then perturbing results in:

$$\begin{aligned} z &= C[35 \cdot 5, 50 \cdot 90, 15 \cdot 5] \\ &= [3.7, 94.7, 1.6] \end{aligned}$$

This shows the difference in the proportional composition of the clays between x and y. Of course, more information is needed to know whether this composition was added to or subtracted from x. Powering, which essentially is perturbing a composition by itself λ times, can be used where a compositional process is cyclic (for example in depositional systems), and is useful for describing regression relationships for compositions (Aitchison, 2003). Perturbation is analogous to addition in real space, and powering is analogous to multiplication by a scalar in real space (Pawlowsky-Glahn et al., 2011).

One of the main uses of perturbation and powering is for graphical presentation, in cases where one or two components may dominate the composition. For example, Figure 2-1 below shows a 30 observation three-part clay compositions. It can be seen that kaolin is very dominant for about half of the data points.

Data centring can be used in this case to make visualization of any structure clear. Centring is a type of perturbation - the closure operation (Equation 2-10) is applied to the geometric mean of each of the components to create the centre of the dataset. The inverse of this centre is then taken and closed, and each observation can then be perturbed by this inverse vector, and closed. Figure 2-2 shows the three-part clay composition after centring. Trends in the data are more easily visualized, and the relationships between the clays and minerals of economic interest can be analysed.

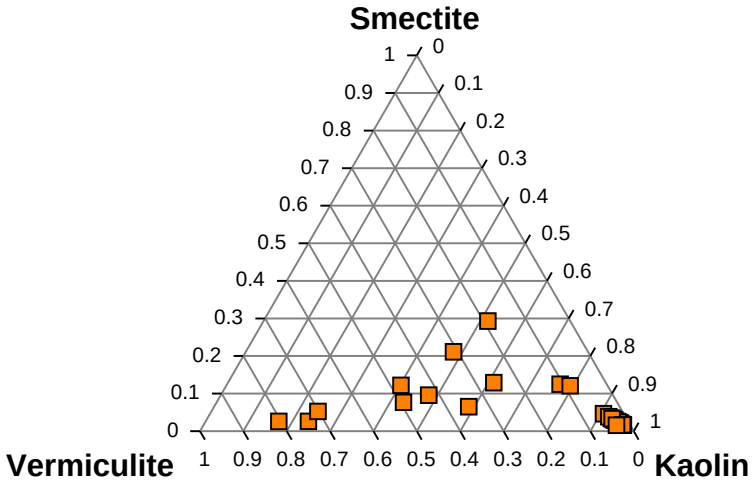

Figure 2-1. Ternary diagram for a three-part clay composition.

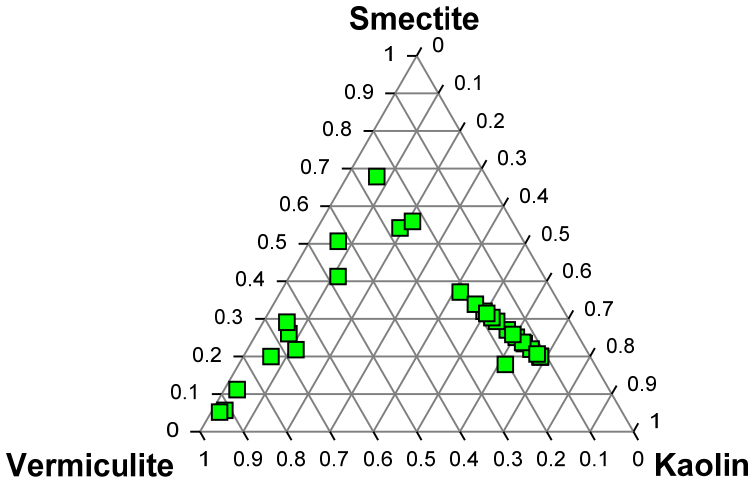

Figure 2-2. Ternary diagram for a centred three-part clay composition.

2.5 Rounded or Trace Zeros

It is possible in any given data set that some of the components have a zero value. Zeros are obviously problematic because the logarithm of zero is undefined and also cannot be used as a denominator which is required for some logratio transforms.

In many instances the zero could be due to the component being below the detection limit of the instrument used, or some other sampling problem. These instances are known as ‘rounded zeros’, and a number of replacement strategies for rounded zeros are discussed by Martin-Fernandez et al. (2003). These authors warn that the strategy chosen must not

distort the general structure of the data, in particular the covariance structure, as further analysis on the subpopulations will be misleading.

Such replacement techniques for dealing with rounded zeros can be divided into two categories: parametric and nonparametric. The former approaches rely on fully parametric multivariate models, which are useful when there are many missing values (Martin-Fernandez et al., 2003). Where there are few missing values (say, <10%), then the nonparametric methods are applicable, easier to apply (Manchuk, 2008), and are the 'most viable' methods (Aitchison and Egozcue, 2005) – the three main nonparametric methods are discussed below.

2.5.1 Additive Replacement

This method was proposed by Aitchison (1986, p. 269), where a composition x containing D components and t rounded zeros can be replaced by a new composition r by the following rule:

$$r_j = \begin{cases} \dfrac{\delta(t+1)(D-t)}{D^2}, & if\ x_j = 0 \\ x_j - \dfrac{\delta(t+1)t}{D^2}, & if\ x_j > 0 \end{cases}, \qquad \textbf{2-13}$$

where δ is a small value less than a specified threshold, such as the detection limit of the analytical method. Note that the compositions are not modified if there are no rounded zeros present in a given observation.

There are a number of problems with the additive replacement method, specifically the dependence between r, δ and the number of zeros. In addition, the covariance structure of subcompositions for parts that do have zeros is not preserved. Therefore any analysis obtained by multivariate methods based on the covariance structure could be distorted (Martin-Fernandez et al., 2003).

2.5.2 Simple Replacement

This is a common approach which essentially consists of assigning a small value to the rounded zero. Setting the missing value to half the detection limit of the instrument or analysis method being used is common practice in the minerals industry. Adding this small value causes the sum > 1.0 (assuming the composition was already closed), therefore the other components are restandardized to maintain the sum=1.0:

$$r_j = \begin{cases} \dfrac{c}{c + \sum_{k|x_k=0} \hat{\delta}_k} \cdot \hat{\delta}_j, if\ x_j = 0 \\ \dfrac{c}{c + \sum_{k|x_k=0} \hat{\delta}_k} \cdot x_j\ if\ x_j > 0 \end{cases}, \qquad \textbf{2-14}$$

where $\hat{\delta}$ is the small imputed value, and c is the constant sum constraint.

2.5.3 Multiplicative Replacement

This is very similar to the simple replacement, but leaves the small imputed value without modification to replace zeros, and then normalizes the other components. It is actually more straightforward than the simple replacement strategy:

$$r_j = \begin{cases} \delta_j, & if\ x_j = 0 \\ \left(1 - \dfrac{\sum_{k|x_k=0} \delta_k}{c}\right) x_j, & if\ x_j > 0 \end{cases}, \qquad \textbf{2-15}$$

The multiplicative method has many advantages over the additive method, as described by Martin-Fernandez et al. (2003). It is also recommended by Aitchison and Egozcue (2005). The advantages include the 'true' composition results if the δ values are the 'true' censored values, and that the ratios are preserved in all instances where there are no zeros; i.e., fully constrained compositions are not altered.

2.6 Essential Zeros

It is possible that a zero value implies a total absence of a component, such as the complete absence of sand-sized particles in very fine-grained sediment. These instances are known as 'essential zeros', and the replacement of these zeros with another value is therefore theoretically incorrect. One solution in this case, if plausible, is to separate the essential zeros from the rest of the population by domaining and considering a separate domain with *n*-1 components.

It could be difficult to separate a domain into zones where all the components are defined at every sample location and in this case amalgamation of similar variables (Martin-Fernandez et al., 2000) can be considered. This involves amalgamating components such that the resulting compositions no longer contain zeros. This strategy can only be used if:

1. There are more components measured than are needed for the study; and
2. The amalgamated components do not contain one of the primary variables of interest (or if they could be logically grouped).

It is also possible to combine the amalgamation and domaining approach, so that only selected zones or domains need to be subjected to amalgamation.

There has been work on alternative ways to treat essential zeros, other than excluding them. Aitchison and Kay (2003) proposed building a two-stage model. The 'first [stage] is to determine where the zeros will occur and the second [stage] on how the unit available is distributed among the non-zero parts' (Aitchison and Kay, 2003, p. 1). The method is computationally complex (although based on existing statistical theory), and is yet to be fully tested.

In summary, essential zeros cannot be dealt with by simple non-zero replacement, and despite Aitchison and Kay (2003) warning of the ad-hoc nature of amalgamation, it is clear that, at least for spatial data, that the judicious grouping of variables and robust domaining are the most workable solutions.

2.7 Compositional Data Framework Applications

The logratio transforms and basic compositional data operations as described above have been successfully used for non-spatial applications. Thomas and Aitchison (2005) use a suite of geochemical compositional data to determine if two Scottish limestone formations are compositionally different, and if any differences can be attributed to sedimentary or other geochemical processes. Martin-Fernandez et al, (2005) present a study for Cenozoic volcanic rocks in Hungary, and use the relationships between the sub-compositions to conclude that processes of magma differentiation in two different rock groups were similar. Buccianti and Pawlowsky-Glahn (2005) present a case study of water contamination over time around an active volcano in Sicily, and comment on the rapid alteration of water contamination due to degassing of the volcano compared to the slower processes of the intrusion of marine water and silicate weathering.

These three examples of compositional data analysis use techniques such as perturbation and powering to create graphs (including ternary diagrams), and principal component analysis of logratio transformed data to support their arguments. All the inferences and conclusions drawn are from data still in logratio space – none of these studies involve any averaging of the data in logratio space and then back-transformation to original data units.

However, in a geostatistical application, averaging of the composition at unsampled locations and subsequent back-transformation is required. There is a limited amount of published literature on the geostatistical application of compositional data - Pawlowsky-Glahn and Olea (2004, p. 123 - 164) and Boezio et al. (2011) show case studies where the logratio values are kriged, and then back-transformed to original data units.

Both these case studies conclude that the final results are unbiased – however, as shown in the next section (2.8), these conclusions cannot be correct. Comments on the bias, and reasons why the bias was not detected in these two case studies are shown at the end of the next section.

2.8 Linear Averaging of Logratios

Kriging is a linear estimator, where the estimated value is derived from the weighted sum of data values at neighbouring locations. However, the logratio transform is not a linear transformation of the original values:

$$a\log\left(\frac{x_i}{x_D}\right)+b\log\left(\frac{x_j}{x_D}\right)\neq\log\left(a.\frac{x_i}{x_D}+b.\frac{x_j}{x_D}\right) \qquad \textbf{2-16}$$

(after Lan, 2007), and the logratios are non-additive. Therefore, taking a linear average of the logratios will not result in a linear average in original units after back-transformation – there will be a bias, as briefly discussed in Section 1.1. This can be illustrated in a very simple two-point example, using four-component oil sands data. The components are bitumen (B), coarse (C) and fines (F) solid fraction, and water (W) - the values are expressed as proportions that sum to unity at each sample location. In the absence of any other information, a point exactly halfway between the two points will be estimated as the mean of the data values:

	Bitumen	Coarse	Fines	Water	Total
Point 1	0.0330	0.5800	0.3070	0.0800	1.0000
Point 2	0.0800	0.8000	0.1000	0.0200	1.0000
Mean	0.0565	0.6900	0.2035	0.0500	1.0000

Table 2-1. Two-point estimate, original data units.

If the alr transform (Equation 2-2) using the coarse fraction as the denominator is taken, then the mean of the logratios halfway between the data points is calculated as:

	alrB:C	alrF:C	alrW:C
Point 1	-2.8665	-0.6362	-1.9810
Point 2	-2.3026	-2.0794	-3.6889
Mean	-2.5846	-1.3578	-2.8349

Table 2-2. Two-point estimate, alr transform.

If the mean alr values shown in Table 2-2 are back-transformed into original units (from Equations 2-3 and 2-4), the results are:

	Bitumen	Coarse	Fines	Water	Total
Mean	0.0542	0.7187	0.1849	0.0422	1.0000

Table 2-3. Two point estimate after alr back-transform.

Clearly, this averaging does not agree with that shown in Table 2-1 – the coarse fraction is biased high, the bitumen, water and fines fraction biased low. The total of the components, by construction, sum to unity. In fact, these results for the alr back-transform are exactly equivalent to standardizing of the geometric means of the components. Table 2-4 shows the geometric mean for the two data points, and the standardized geometric mean (standardizing is the same as the closure operation shown in Equation 2-10).

	Bitumen	Coarse	Fines	Water	Total
Geo. Mean	0.0514	0.6812	0.1752	0.0400	0.9478
Standardised	0.0542	0.7187	0.1849	0.0422	1.0000

Table 2-4. Two-point estimate, geometric mean and standardized geometric mean.

From these results, directly applying kriging to the logratio values, and then back-transforming, would expectedly result in a relatively high bias for the dominant component, and lower biases for the other components. Where a component is only a small proportion of the total, however, this bias could go unnoticed.

For the previously mentioned case studies, in Pawlowsky-Glahn and Olea's case, two components were of low magnitude, and there was one very dominant component (~90%), so the results only *appear* unbiased. The dominant components in Boezio et al.'s case study, however, show clear signs of bias, and the estimates of most of the minority

components again only *appear* unbiased – the bias is present and identifiable, but has not been recognized by the authors.

Linear kriging of logratios *will* therefore result in bias, and non-linear methods, such as multiGaussian kriging as discussed in the next Chapter are required.

Chapter 3 MultiGaussian Kriging

3.1 Background and Theory

The advantages of using non-linear estimation techniques were briefly discussed in Section 1.3.7 where MultiGaussian kriging (MGK) was specifically introduced. Such methods are required for application to logarithmic transformed variables, in order that linear averaging and the implicit resultant bias be avoided. For MGK and other non-linear techniques the full conditional distribution is modelled

MGK was introduced by Verly (1983, 1984), as a non-linear kriging technique that utilized the multiGaussian model. The key to the method is to transform a RF $Z(x)$ into a Gaussian RF $Y(x)$ via the normal scores transform:

$$Y_{(x)} = \varphi^{-1}(Z_{(x)}) \qquad \textbf{3-1}$$

The conditional expectation and variance at unsampled locations is then required – this is usually performed by SK. It is then assumed that the distribution of any value Y_x is multivariate Gaussian, and is fully defined by a Gaussian conditional probability density function with the mean equal to the SK estimate y_x^{SK} and the variance equal to the SK variance $(\sigma_x^{SK})^2$.

This conditional probability density function is:

$$g_x(y \mid \text{data}) = \frac{1}{\sigma_x^{SK}} g\left(\frac{y - y_x^{SK}}{\sigma_x^{SK}}\right) \qquad \textbf{3-2}$$

where g is the standard Gaussian pdf. The estimate of a point-support function is obtained by calculating the expected value of the conditional distribution (from the conditional expectation and conditional variance), by 'multiGaussian kriging':

$$[\varphi(Y_x)]^{MGK} = \int \varphi(y) g_x(y \mid \text{data}) dy \qquad \textbf{3-3}$$

$$[\varphi(Y_x)]^{MGK} = \int \varphi(y_x^{SK} + \sigma_x^{SK} u) g(u) du \qquad \textbf{3-4}$$

In practice, an analytical expression of this equation is difficult to find, and a numerical integration is needed to construct the cdf. For a univariate problem, Saito and Goovaerts (2000) suggest discretizing the inverse of the cdf with say 100 quantiles yp(**u**) (for example, with probabilities 0.005 to 0.995). For a multivariate problem, discretizing the inverse of the cdfs independently will not work, since the correlations between the variables will not be honoured.

Monte Carlo simulation can therefore be used to generate random vectors (realizations) to generate the cdf (Verly, 1984). Where correlation exists in the multivariate case, random but correlated (not independent) vectors are required. Cholesky decomposition in the Monte Carlo method context (Rubinstein, 1981 p. 65-67) is one way to derive such correlated multinormal random vectors.

An advantage of MGK over other non-linear estimation techniques for the multivariate case (e.g., indicator kriging or disjunctive kriging) is that the correlations between the variables can be maintained. In addition, MGK does not produce order relation violations (Emery, 2006).

3.2 Multivariate Gaussian Distribution Assumptions

Verly (1984) discusses a number of apparently strong hypotheses that are used for the multiGaussian model; the existence of a normal score (Gaussian) transform, the multivariate Gaussian distribution of the RF $Y_{(x)}$ distribution, and strict stationarity of the RF $Y_{(x)}$.

The normal scores transform φ (and back-transform φ^{-1}) always exists in practice (Verly, 1984), and is easily implemented manually or by geostatistical software. Problems can arise with tied data (repeats of the same $Z_{(x)}$ value in the original data), but this can be dealt with by randomly assigning different Gaussian transformed cumulative frequencies, or by 'despiking', where the tied data are ordered according to the local averages (Verly, 1984).

The normal scores transform guarantees a univariate Gaussian distribution, but the second strong hypothesis i.e., the multivariate Gaussian distribution is difficult to verify. Departures from the multivariate Gaussian distribution can be caused by non-linearity, constraints and heteroscedasticity, which are common in earth sciences data (see Figure 3-1).

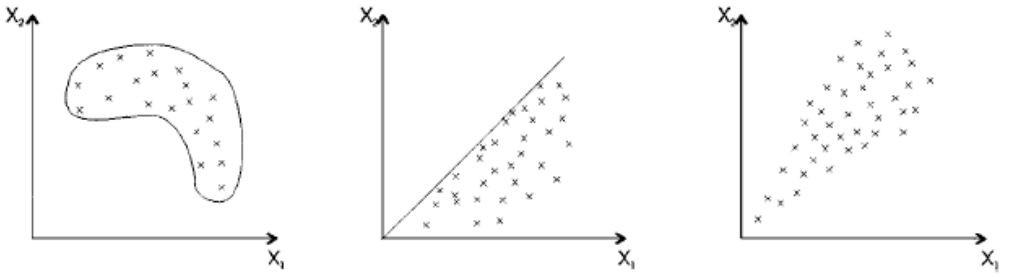

Figure 3-1. Bivariate distribution problems; non-linearity (left), constraints (middle), heteroscedasticity (right) (after Leuangthong and Deutsch, 2003).

Certain checks can be made to test that the bivariate distributions are Gaussian, which is a pre-requisite for a multivariate Gaussian distribution. Emery (2005a) recommends checks that include:

- Scatterplots of (Y(u), Y(u+h)) for different lags and directions should have an elliptical shape;
- The ratio of the square root of the variogram to the madogram should be constant at about $\sqrt{\pi}$;
- Checking the consistency of the theoretical indicator variograms that are derived from the Gaussian model to the actual experimental indicator variograms; and

- Verifying that no relationship exists for the mean and the variance of the transformed data at a local scale (i.e. data is homoscedastic).

These checks cannot prove a multivariate Gaussian distribution, but providing they do not reject the assumption outright, then from a pragmatic viewpoint the model can be applied.

The third hypothesis, that of strict stationarity, can be addressed by dividing the deposit into smaller, more homogenous domains, or as Emery (2005b) has suggested, substituting OK for SK to account for a locally varying mean. Emery (2005b) has obtained some encouraging results with the OK application to MGK, but OK is not a conditional expectation estimator, and the conditional estimation variance may not be minimized.

3.3 Practical Steps

The following is a step-by-step guide to the implementation of MGK for logratio-transformed variables, using the alr transformation. It assumes that the prerequisites for adopting a compositional data approach (i.e., that the variables are non-negative and sum to unity; and that zeros and null values in the data are dealt with appropriately) have been met. The steps are:

- Logratio transform of data using the additive logratio method (alr) with the form:

  $$y_i = \log\left(\frac{x_i}{x_D}\right), \qquad i = 1,...,D-1$$

  is recommended initially, since the number of resulting transformed variables is D - 1 compared to the original data, which is easier to deal with when modelling the LMC covariance function;

- Normal scores transform of the D - 1 data with random despiking, using suitable declustering weights:

  $$y_p = \varphi^{-1}(y_i), \qquad i = 1,...,D-1$$

- Assume that the D – 1 values form a multivariate Gaussian distribution – this can be checked as described in Section 3.1;

- Generate experimental direct and cross covariances for the normal scores data in the usual way (honouring any anisotropy, with appropriate lags, tolerances and bandwidths given the data configuration), and model appropriately with a LMC;

- Simple co-kriging of Gaussian transformed data into a point-scale grid. The conditional distribution is multivariate Gaussian at each estimated point, with the mean equal to the simple cokriging estimate (Y^{*SK}), and variance equal to the simple cokriging variance (σ^{2SK});

- Create a discretized cdf by generating $X_1,...,Xp$ as independent and identically distributed (iid) variables from N(0,1). Then derive the Cholesky decomposition of the correlation matrix of the normal score transformed variables (Σ), $\Sigma = BB^T$, where B is the lower triangular matrix and B^T is its transpose. Multiply each iid ($X_1,...,Xp$) by the Cholesky decomposition J = BX, which will result in p multinormal random vectors (J);

- Multiply the standard deviation of the simple kriging (σ^{SK}) by the multinormal random vector (J) and add the simple kriged estimate (Y^{*SK}) for each location (**u**) i.e. $yp(\mathbf{u}) = J.\sigma^{SK}(\mathbf{u}) + Y^{*SK}(\mathbf{u})$;
- Back-transform the resultant yp(**u**) quantiles from normal scores into the quantiles of the local probability distribution of the logratio transformed data z(alr) $p(\mathbf{u}) = F^{-1}[G(yp(\mathbf{u}))]$;
- Apply the inverse of the alr logratio transform (Equations 2-3 and 2-4) to each z(alr) p(**u**) quantile to get the conditional distribution in original data units zp (**u**);
- The kriged estimate is the conditional expectation of the cdf; i.e., the mean of zp (**u**), and the probabilities above certain cut-offs at the point scale are now easily calculated; and
- The MGK mean and the probabilities above cut-offs can then be calculated for larger blocks by taking the mean of these values from each point in the blocks.

These steps have been used for the following simple worked example, and also applied to the much larger case study in Chapter 6.

3.4 A Simple Worked Example

The following worked example has been performed in Microsoft Excel, with checking of many of the steps using the GSLIB suite of programs.

Consider the following two dimensional data layout, consisting of data at two locations (1 and 2), with compositional data (P1, P2, P3) informing each point – using multiGaussian kriging, determine the values for the three components at location 0 (Figure 3-2).

Location	Easting	Northing	P1	P2	P3
1	100	104	0.15	0.45	0.4
2	102	100	0.01	0.84	0.15
0	**100**	**100**	**?**	**?**	**?**

Figure 3-2. Data configuration.

A 1,000 point 'training image' (10 x 10 x 10) consisting of a valid three component data set was used to derive the normal scores transform and to calculate experimental covariances. The data was first logratio transformed (using alr), and then normal scores transformed, resulting in two variables (Y and Z). The normal scores transform table for the training image was used to transform the data at locations 1 and 2.

The non-ergodic covariances (Srivastava, 1987), and cross-covariances were then calculated for the two-dimensional data configuration shown in Figure 3-2. Since selected data in the training images can have exactly the same spatial configuration as the two-dimensional problem, the covariance values could be used directly in the kriging, and variogram or covariance modelling was not required. The covariances and cross covariances appear in Table 3-1:

Simple cokriging, using the covariances and the alr transformed data for locations 1 and 2, was used to obtain the estimate and estimation variance for both variables at location 0. The unsolved kriging matrix is shown in Table 3-2:

COV{Y1,Y1} = COV{Y2,Y2}	1.0000
COV{Z1,Z1} = COV{Z2,Z2}	1.0000
COV{Y1,Y2}	0.4232
COV{Y1,Z1} = COV{Y2,Z2}	0.7144
COV{Y1,Z2} = COV{Z1,Y2}	0.4170
COV{Z1,Z2}	0.4205
COV{Y1,Y0}	0.8516
COV{Y2,Y0}	0.4146
COV{Z1,Z0}	0.8007
COV{Z2,Z0}	0.4062
COV{Y1,Z0} = COV{Z1,Y0}	0.6531
COV{Y2,Z0} = COV{Z2,Y0}	0.4127

Table 3-1. Covariances and cross-covariances.

	LHS				Weights	RHS	RHS
	Y1	Y2	Z1	Z2		Y0	Z0
Y1	1.0000	0.4232	0.7144	0.4170	λ1	0.8516	0.6531
Y2	0.4232	1.0000	0.4170	0.7144	λ2	0.4146	0.4127
Z1	0.7144	0.4170	1.0000	0.4205	λ3	0.6531	0.8007
Z2	0.4170	0.7144	0.4205	1.0000	λ4	0.4127	0.4062

Table 3-2. Simple kriging matrix.

The weights are calculated by inverting the left hand side matrix and multiplying by the right hand side separately for Y0 and Z0. For the estimate and estimation variance for Z0, the left hand side is unchanged, but the right hand side is altered. The resulting weights for Y0 and Z0 are given in Table 3-3:

	Y0	Z0
Y1	0.7694	0.1448
Y2	0.0301	0.0553
Z1	0.0744	0.6625
Z2	0.0390	0.0278

Table 3-3. Simple kriging weights.

The associated estimates and estimation variances are given in Table 3-4:

	Y0	Z0
SK estimate	0.4884	1.1250
SK variance	0.2676	0.3409
SK std. dev.	0.5173	0.5839

Table 3-4. Simple kriging estimates and estimation variances.

These values now characterize a Gaussian distribution with the mean equal to the SK estimate and variance equal to the SK estimation variance. Two sets of five hundred random standard Gaussian values were generated – one each for variables Y and Z. The Cholesky decomposition was calculated for the correlation matrix for Y and Z (from the training image), shown in Table 3-5:

Correlation Matrix			Cholesky Lower		
	Y	**Z**		**Y**	**Z**
Y	1.0000	0.7154	**Y**	1.0000	0.0000
Z	0.7154	1.0000	**Z**	0.7154	0.6987

Table 3-5. Correlation and Cholesky matrices.

The two sets of the 500 random standard Gaussian variables were multiplied by the lower triangle of the Cholesky decomposition matrix to create the random but correlated Gaussian values. Each of these was multiplied by the SK standard deviation, and the SK estimated added – this resulted in five hundred random but correlated quantiles for both the Y and Z conditional distributions.

Each of the 500 quantiles was then back-transformed from Gaussian, and then through the alr back-transform, resulting in 500 quantiles in original data units. The mean of the quantiles is exactly the conditional expectation (and the MGK estimate), with the results at location 0 being:

P1 = 0.1884, P2 = 0.5924, P3 = 0.2192

The probabilities for each variable being above particular cut-offs (in 0.05 increments) are shown in Table 3-6:

Cut-off	P1	P2	P3
0.05	65.8%	97.0%	79.2%
0.10	55.0%	94.8%	70.2%
0.15	47.8%	91.6%	60.8%
0.20	37.6%	87.6%	49.0%
0.25	30.0%	82.4%	37.6%
0.30	23.4%	78.4%	26.0%
0.35	19.0%	74.8%	18.4%
0.40	14.6%	70.6%	14.0%
0.45	10.2%	67.0%	8.4%
0.50	8.2%	61.0%	6.8%
0.55	7.0%	56.2%	5.4%
0.60	5.6%	50.8%	4.8%
0.65	3.8%	46.8%	4.0%
0.70	2.4%	41.2%	2.8%
0.75	2.2%	36.8%	1.6%
0.80	1.6%	31.0%	0.8%
0.85	1.0%	25.8%	0.4%
0.90	0.8%	22.6%	0.2%
0.95	0.8%	16.6%	0.2%

Table 3-6. Probability to be above cut-off.

Given that there are only two points informing the multiGaussian kriging, validation is difficult – however, the results make sense given the informing data, and fall within the range of the input data.

MGK is a technique that is more complicated than OK or SK, and in the multivariate case requires the application Monte Carlo simulation. In this sense, MGK is more akin to conditional simulation than to kriging estimation. This of course has the advantage of resulting in a probabilistic model of the values of interest at the unsampled locations as opposed to a single expected value.

Chapter 4 Oil Sands Data Review and Statistics

4.1 Introduction

The data used for this case study comes from the Athabasca Oil Sands area in northern Alberta, Canada (Figure 4-1). The Alberta Oil Sands have the second-largest proven oil reserves in the world (after Saudi Arabia), but differ from this and most other oil producing regions in that the hydrocarbons occurs as bitumen, which is a viscous form of crude oil that will not flow unless heated or diluted with lighter hydrocarbons.

The geological setting of the Athabasca Oil Sands deposits is summarized in Mossop (1980). The bitumen is predominantly contained within the Lower Cretaceous McMurray Formation, which consists of medium to coarse grained sands mainly composed of quartz and feldspar in the lower part of the formation, through to variable sands and silts in the middle and upper part of the formation.

Figure 4-1. Location of Alberta Oil Sands.
(Source: http://en.wikipedia.org/wiki/Athabasca_Oil_Sands)

A generalized diagram of oil sand composition (such as would be viewed with a hand lens) is given in Figure 4-2. Some brief details of the differences between, high, medium and low grade material are in Table 4-1.

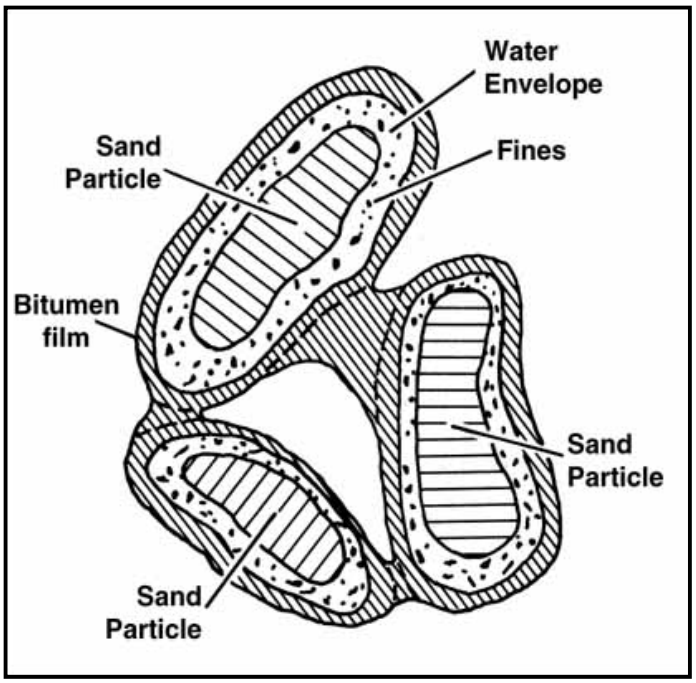

Figure 4-2. Generalized close-up view of oil sands (after Hennessey, 1990).

Quality	*Bitumen wt%*	*Water wt%*	*Solids wt%*	*Fines Propn. (<44µm)**
High	14.3	1.2	84.5	4.1
Medium	11.8	1.3	86.9	5.2
Low	8.6	5.3	86.1	35.4

* Fraction of fines in total solids

Table 4-1. Typical Oil Sands Composition (after Romanova et. al., 2003).

It can be seen that the low grade material has a much higher proportion of fines in the solids. For this reason the high grade material is much preferred during the processing phase due to higher bitumen content, lower fines and water therefore better recovery - the proportion of fines is a key driver of the bitumen recovery during processing (Wik et al., 2008).

Since the bitumen is in a near-solid state, and because much of the oil sands are less than 80 metres below surface, the deposits can often be extracted by 'conventional' mining techniques (i.e., truck and shovel) instead of the more usual well-based petroleum extraction methods. The other extraction methods used are referred to as 'in-situ methods', where steam (with or without other solvents or additives) is injected into the oil sands reservoir to reduce the viscosity of the bitumen, so it can then flow to other wells and be pumped to the surface.

In 2009, over half of the oil sands production from the Athabasca Oil Sands was from mining, but in-situ methods are expected to surpass production from mining by 2016 (CAPP, 2010). It is estimated that by 2016, about 70% of Western Canada's oil production will come from oil sands production (approximately 2,200 thousand barrels per day (CAPP, 2010)).

Understanding the geological characteristics of the oil sands is therefore critical for the mining and processing strategy.

4.2 Data Used

The data set used has four components, bitumen (B), water (W), coarse solids (C) and fine solids (F). The coarse solids can be considered as the 'sand' component (i.e., >44 μm) as shown in Figure 4-2. These four components complete a whole composition (i.e., sum to unity). Note that the data itself is in terms of proportions of one, but some of the description of data analysis presented here is in terms of percentages.

The data consists of vertical drill holes, with a maximum depth of 126 metres (m). A subset of the data was used for the case study within a 2,000m x 3,000m area that was drilled with a hole spacing of approximately 100m x 100m. Data is generally collected at 1.5m intervals down the hole, although there are some shorter sampled intervals where there is very high bitumen content – forty data points (from a limited number of drill-holes) had a bitumen value of one hundred percent, and zero for all the other components. It is likely that these small, very high bitumen grade samples have been given a 'nominal total bitumen' grade, as it is unlikely that there would be a true absence of solids or water.

Local upscaling (i.e., compositing with adjacent samples to produce samples that were all 1.5m in length) was therefore undertaken.

In the resulting composited data set there are almost two hundred samples where bitumen is zero, and six samples where bitumen is one (and the other components are zero), but for all the other sample points the components are valid and sum to unity. Note that because this data is across multiple facies, it is likely that there are samples that are from zones or domains that are not of particular interest – data within a bitumen-bearing zone was therefore separately domained. It is unlikely that the samples of 100% bitumen are really total bitumen (they may also be possible database errors). These zero or 100% bitumen samples were not within the domain that was ultimately chosen for the case study – they were vertically above or below the study domain. The domain definition is discussed in the next section, and the location of the zero and 100% bitumen composites is shown in Figure 4-4.

4.2.1 Domaining

Drillhole data was initially selected from the main bitumen-bearing horizon, as defined by the 'rock-type' code in the drillhole file. This rock-type has by far the most bitumen, although not all composite samples are high in hydrocarbon. In addition, the rock-type vertically below this also has locally high bitumen grades. Whether this is due to incorrect geological logging (assuming the rock-types were defined on logging as opposed to some geophysical method) or not is unclear. It could be the case that the bitumen occurs locally over different strata/horizons.

The selected horizon therefore may be geologically valid, but it does not really constitute a ‘stationary domain’. Exploratory Data Analysis (EDA) showed that there were significant trends in the vertical direction within this rock-type domain. Figure 4-3 shows a swath plot for bitumen, with the bitumen grades averaged over 5m vertical intervals across the entire selected domain (the count of samples per vertical interval is also shown). It can be seen that the bitumen content averages 2% at the 240mRL, increasing to 13% at the 190mRL, before decreasing to 4% at the 160mRL.

The other variables show similar trends in the vertical direction, although there are no significant trends apparent in the horizontal directions. A severe trend such as this violates the assumptions of stationarity required for the application of geostatistical methods (Leuangthong and Deutsch, 2004). OK with a restricted vertical search neighbourhood may perform satisfactorily in this situation, but MGK will probably perform poorly due to the stricter assumptions of stationarity required. Verly (1984) and Emery (2005a, 2005b) discuss the problems of a locally varying mean and trends for MGK.

Trend modelling such as that discussed by Isaaks and Srivastava (1989, p. 531) and Leuangthong and Deutsch (2004) is an option to account for this trend. This would introduce another level of complexity to the process, and is not a core component of the research. A simpler solution is to change the decision of domain stationarity by adjusting the boundary geometry.

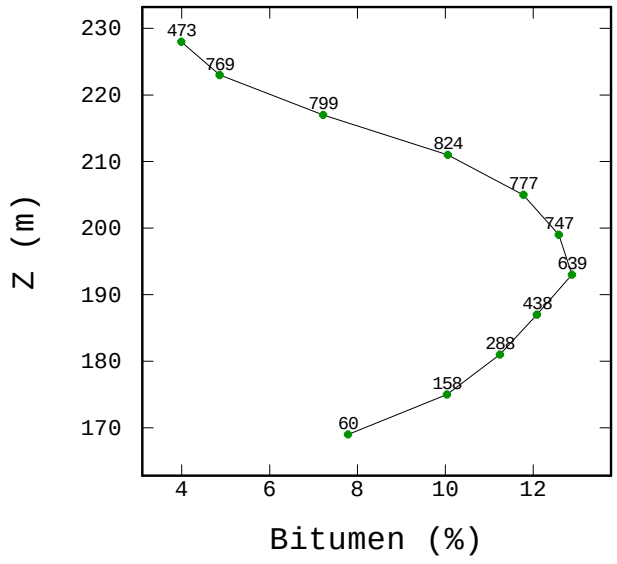

Figure 4-3. Vertical swath plot for bitumen, rock-type domain.

To attempt to manage the non-stationarity discussed above, a bitumen ‘grade-based domain’ was constructed, using a cut-off grade of approximately 7 to 8% bitumen, which is the cut-off grade used in many of the operations currently (e.g., see Devenny, 2010). There are numerous warnings in the literature about using approximate economic cut-off grades as the basis for domain boundaries (e.g., Emery and Ortiz, 2005). Indeed, for an industrial application for oil sands modelling, a much more thorough stratigraphic/facies

analysis would be required to assist in domaining decisions. However, the purpose of this thesis is to demonstrate geostatistical modelling of compositional data. Therefore, if it can be established that there is a pronounced break in the grade distribution at this point, then the use of a 'hard boundary' domain is arguably justified. A 'hard boundary' is when only samples from within a domain are used for geostatistical modelling – samples from other domains are not used to inform the model.

A simple test for a hard boundary domain decision is to take the means of the grades for the first sample on either side of the grade boundary (often a three-dimensional wireframe in practice), and then the take means of the first two samples either side of the boundary etc. and chart these values. Figure 4-5 shows the grade change across the 7% bitumen domain boundary – Domain 0 overlies the domain considered, Domain 1 is the 7% bitumen domain and Domain 2 is below Domain 1 (see Figure 4-4). The mean grades above the 7% bitumen domain boundary are about 3.5%, but the mean grades within the 7% bitumen domain jumps to almost 12%, and then drops to about 2.5% below the 7% bitumen domain boundary. The number of data at each location is annotated in the figure.

It is concluded that the use of the 7% domain boundary for the purposes of this study is justified. Note also that the grade change across the boundaries of Domain 1 for the other variables is similarly abrupt – from outside to inside Domain 1 the coarse fraction goes from 45% to 70%, the fines fraction from 40% to 15% and water from 9% to 4.5%.

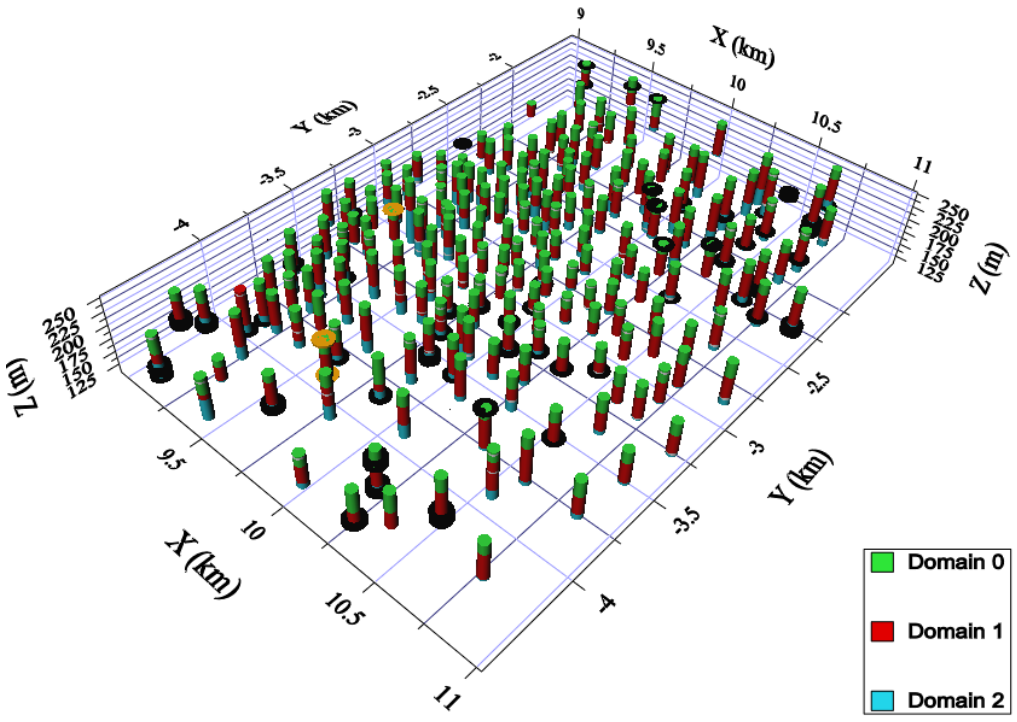

Figure 4-4. Domain locations showing 0% bitumen (black) and 100% bitumen (orange), 3x vertical exaggeration.

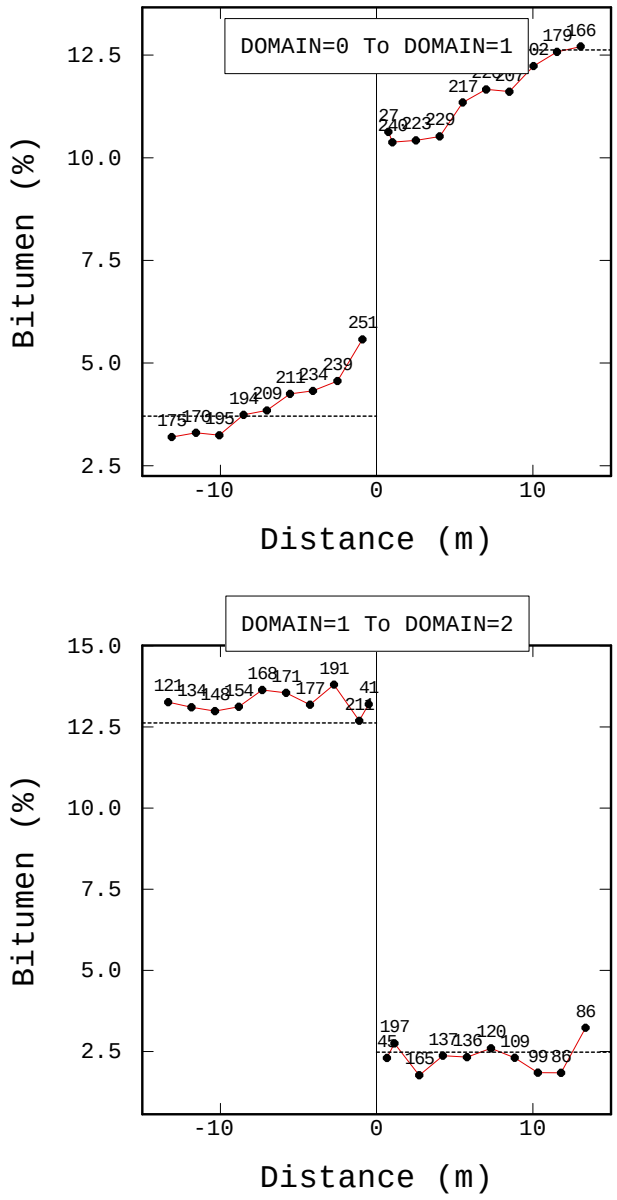

Figure 4-5. Bitumen domain boundary analysis.

The vertical swath plot for Domain 1 is shown in Figure 4-6 – there is an overall increase in bitumen grade with depth, but it is not as pronounced as was seen for the previous domain based solely on rock-type (cf. Figure 4-3). Note that the horizontal axes of bitumen units for Figure 4-3 and Figure 4-6 are the same.

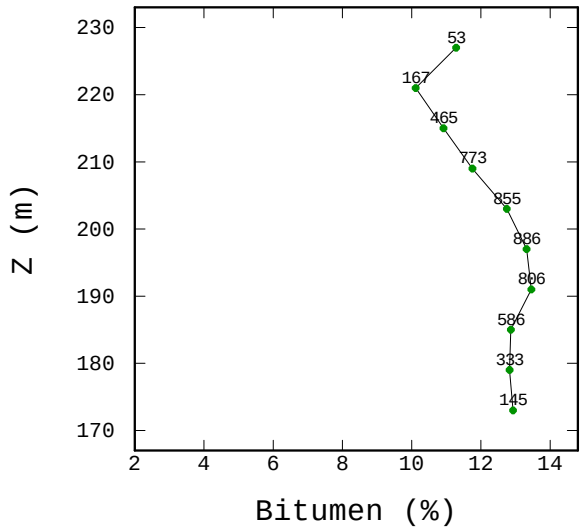

Figure 4-6. Vertical swath plot for bitumen, grade-based domain.

4.2.2 Basic Statistics

The location of the Domain 1 data set is shown in Figure 4-7, with declustered basic statistics in Table 4-2 and histograms shown in Figure 4-9.

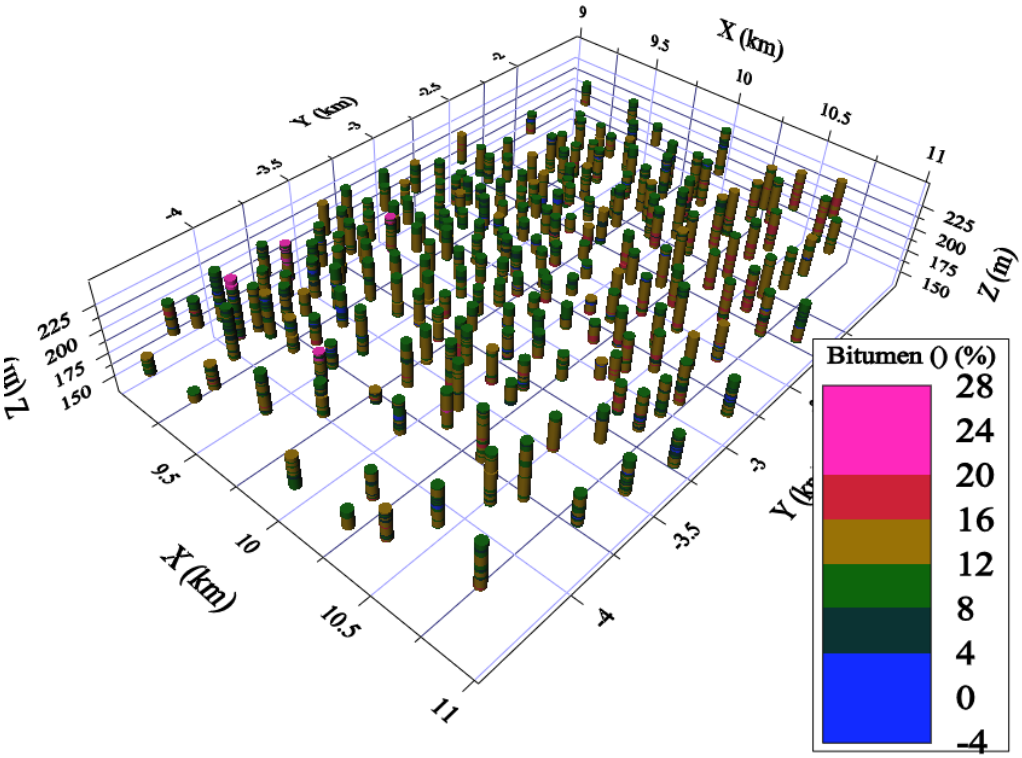

Figure 4-7. Location of selected drill holes, 5x vertical exaggeration.

Cell declustering with a grid size of 450mE x 450mN x 1.5mRL was selected for weighting - Figure 4-8 below shows the declustered means for bitumen for a range of cell sizes. Cell sizes range from 100mE x 100mN x 1.5mRL (Cell Size 100) to 1000mE x 1000mN x 1.5mRL (Cell size 1000) on the chart horizontal axis.

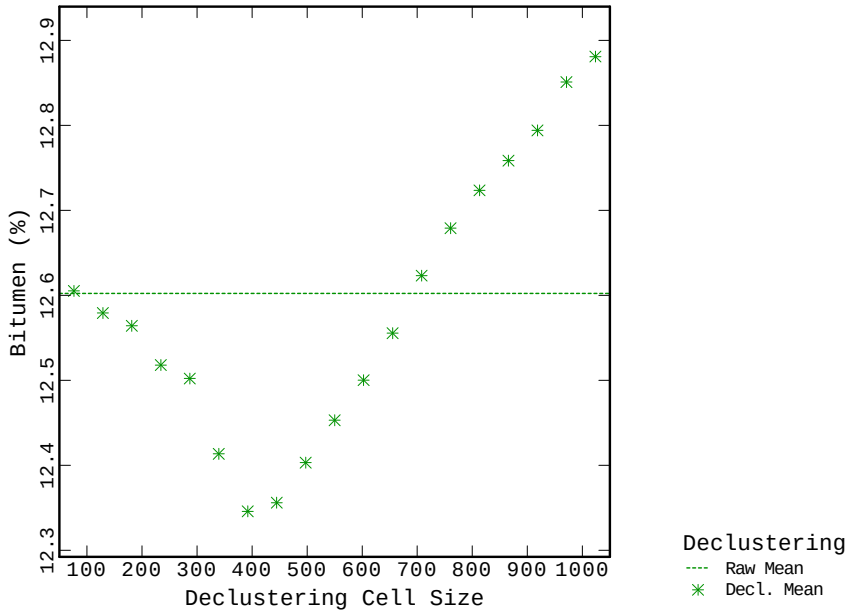

Figure 4-8. Cell declustering - bitumen.

Variable	Count	Minimum	Maximum	Mean	Std. Dev.	Variance	CV	Geo. Mean
Bitumen	5124	0.0019	0.6901	0.1238	0.0436	0.0019	0.3522	0.1160
Coarse	5124	0.0092	0.8544	0.7029	0.1182	0.0140	0.1682	0.6883
Fines	5124	0.00080	0.9105	0.1283	0.1203	0.0145	0.9376	0.0864
Water	5124	0.0045	0.1730	0.0451	0.0238	0.00055	0.5276	0.0394
Total	5124	0.9999	1.0001	1.0000	0.0000	0.0000	0.0000	1.0000

Table 4-2. Declustered basic statistics, >7% bitumen domain.

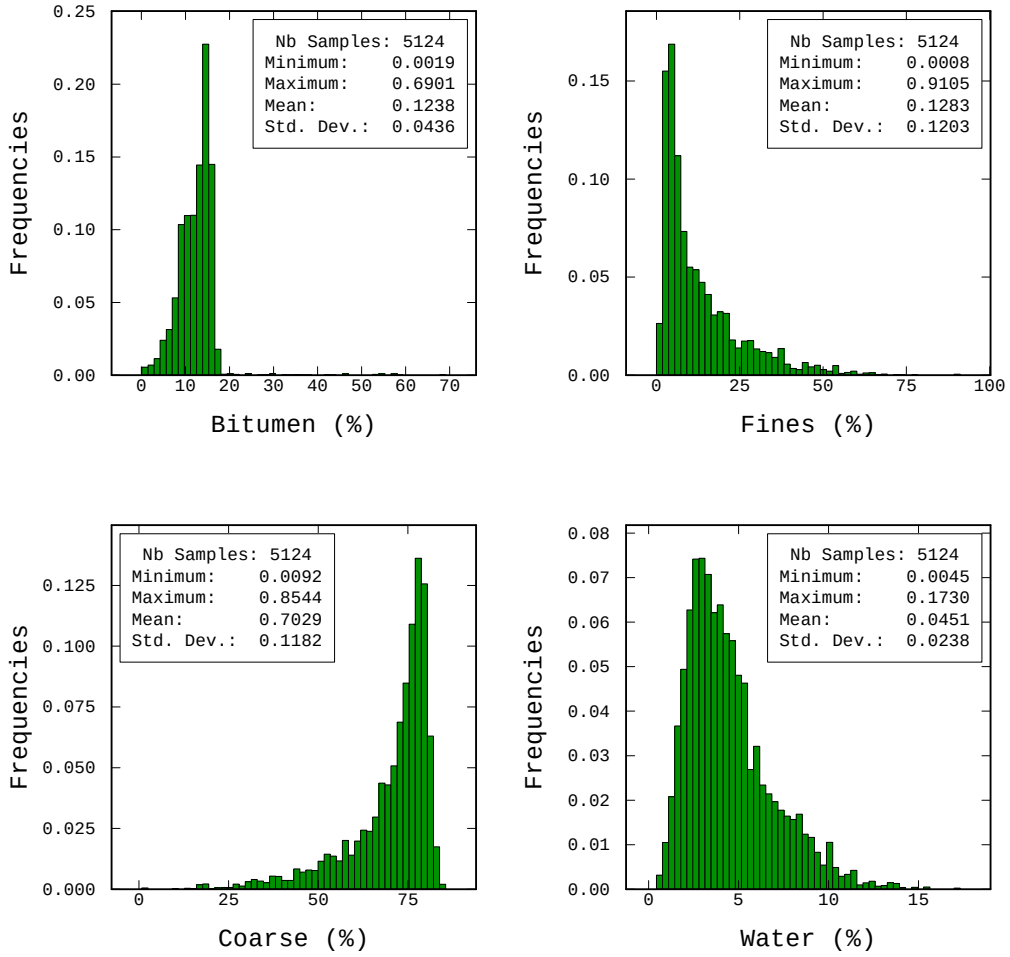

Figure 4-9. Declustered basic statistics and histograms, original data.

Scatterplots between the variables are shown in Figure 4-10. There are moderate positive correlations for B & C and for F &W, and strong negative correlations for B & F, B & W, C & F and for C & W.

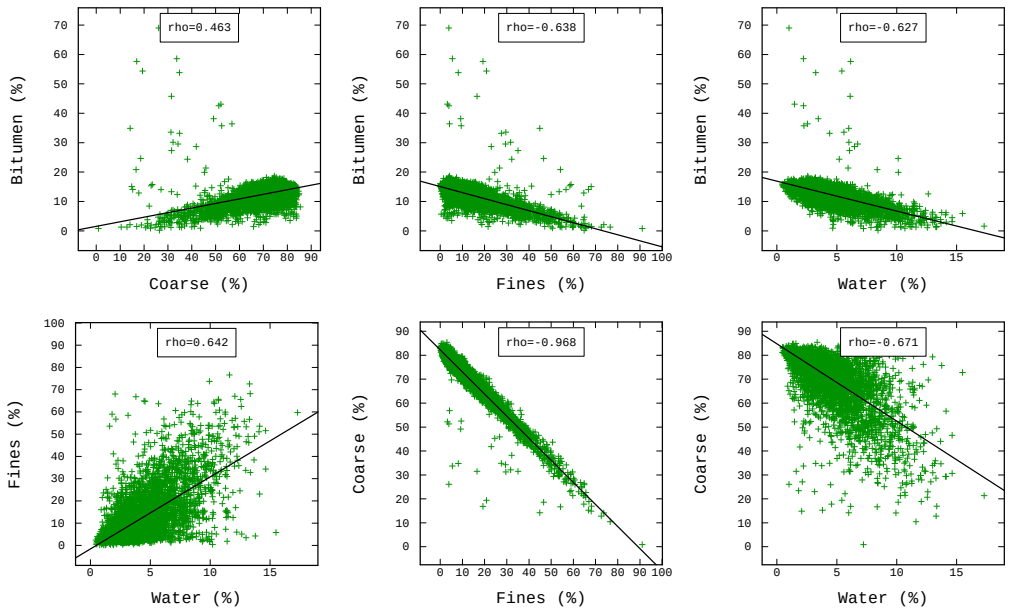

Figure 4-10. Scatterplots for original data units.

4.2.3 Logratio Transforms

The alr (Equation 2-2) and clr (Equation 2-5) transforms were calculated, with the coarse fraction selected as the denominator for alr because it is the major component of the composition.

Statistical analysis of the three alr variables (alrB:C, alrF:C and alrW:C) and the four clr variables (clrB, clrC, clrF and clrW) was undertaken, with the basic statistics shown in Table 4-3 and histograms for the transformed data shown in Figure 4-11 and Figure 4-12. The transformed data in most cases approximates normal distributions, although there are a small number of extreme values that can cause a skewed distribution (for example, the high value of 2.06 for alrW:C).

Variable	Count	Minimum	Maximum	Mean	Std. Dev.	Variance
alrB:C	5124	-4.9375	1.2266	-1.7628	0.2799	0.0784
alrF:C	5124	-6.9411	4.5948	-2.1485	1.0526	1.1080
alrW:C	5124	-5.1957	2.0616	-2.9438	0.6560	0.4303
clrB	5124	-3.6411	2.0187	-0.0490	0.4995	0.2495
clrC	5124	-1.6292	3.2221	1.7138	0.3866	0.1495
clrF	5124	-3.7189	2.9656	-0.4347	0.6936	0.4811
clrW	5124	-2.6068	0.4325	-1.2300	0.3725	0.1387

Table 4-3. Basic statistics, logratio transformed data.

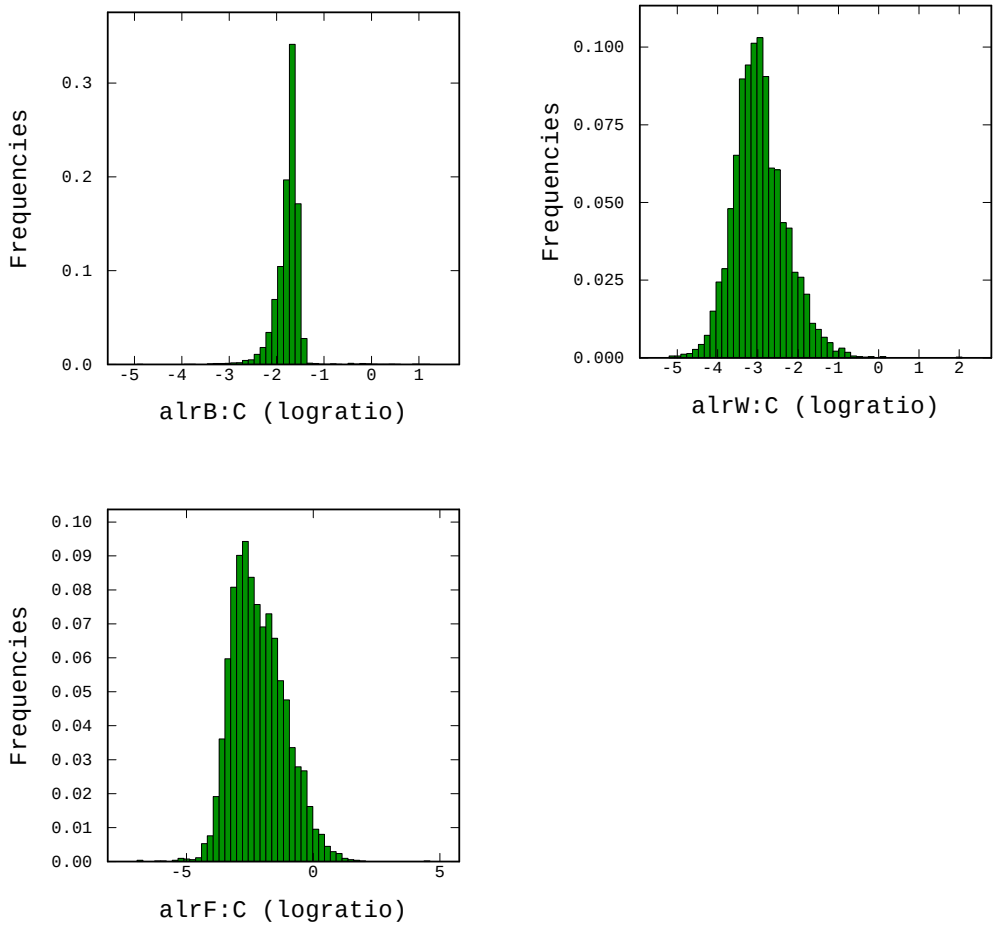

Figure 4-11. Histograms, alr transformed data.

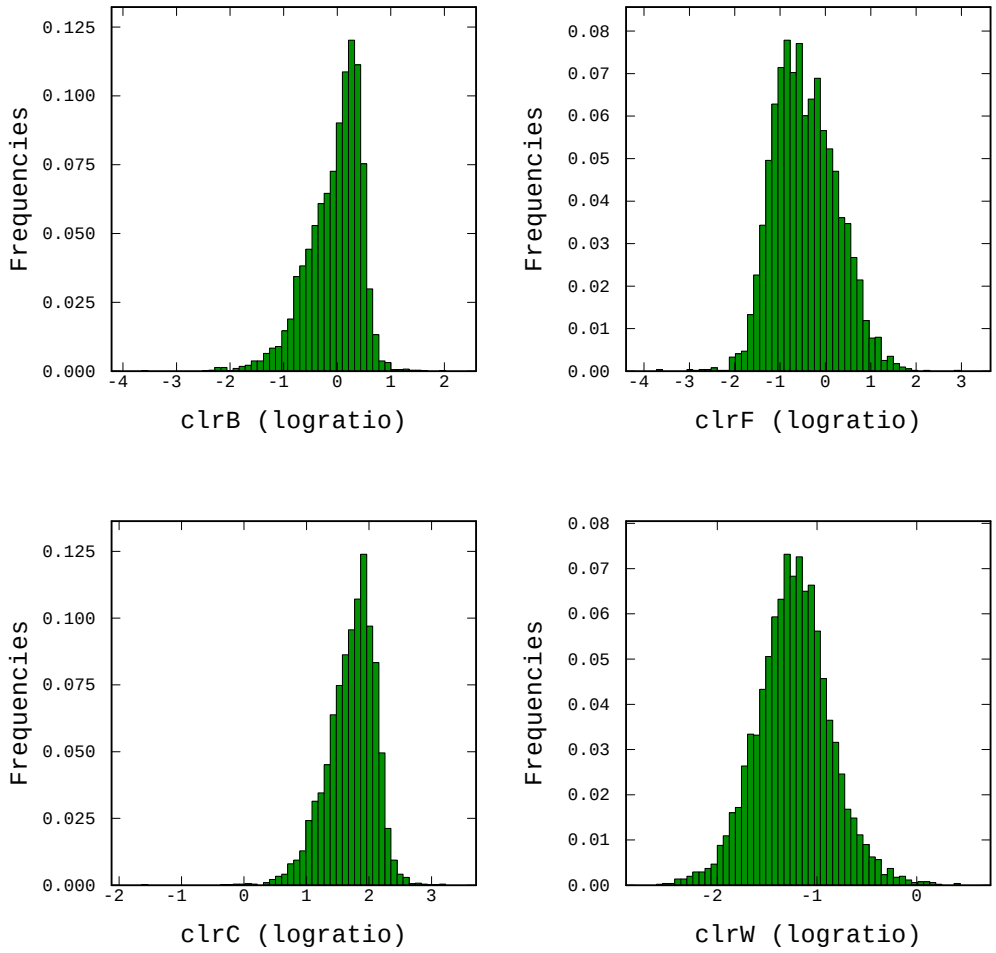

Figure 4-12. Histograms, clr transformed data.

Correlation matrices for the alr and clr transformed data are in Table 4-4, and scatterplots for the alr transformed data are shown in Figure 4-13.

	alrB:C	alrF:C	alrW:C
alrB:C	1.000	-0.218	-0.313
alrF:C	-0.218	1.000	0.737
alrW:C	-0.313	0.737	1.000

	clrB	clrC	clrF	clrW
clrB	1.000	0.830	-0.844	-0.632
clrC	0.830	1.000	-0.890	-0.493
clrF	-0.844	-0.890	1.000	0.193
clrW	-0.632	-0.493	0.193	1.000

Table 4-4. Correlation matrices, alr and clr transformed data.

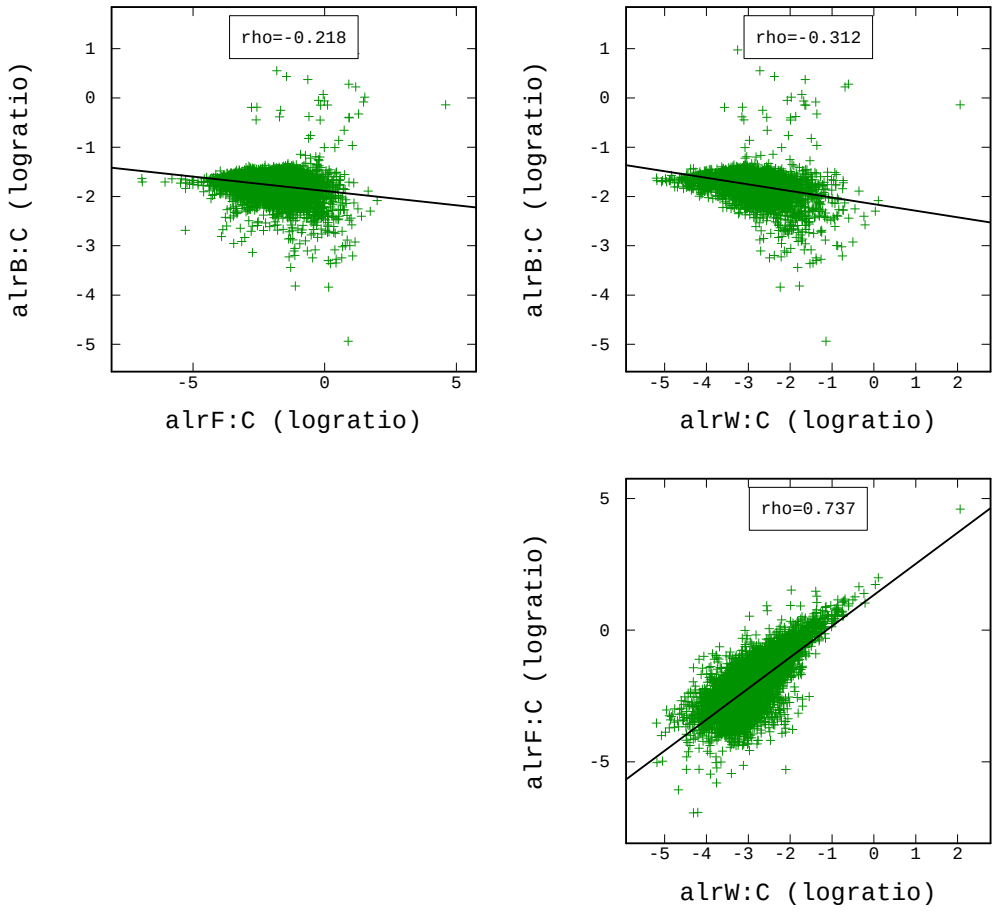

Figure 4-13. Scatterplots, alr transformed data.

There is a strong positive correlation between F:C & W:C, as there was for F & W in the original data, but only moderate negative correlations between B:C & F:C and between B:C & W:C, whereas B & F and B & W in the original data have stronger negative correlations. The clr transforms, however, show similar correlation patterns to those seen in the original data.

Geostatistical applications for the oil sands data set are discussed in the next three chapters – comparisons are made for different geostatistical techniques using the data in original units and with the logratio transforms applied.

Chapter 5 Linear Estimation

5.1 Linear Kriging and Cokriging

The results from Section 2.8 show that directly kriging logratio transformed data and then back-transforming the mean will result in bias, although detection of the bias depends on the relative magnitudes of the components. For the oil sands data set, this bias can be demonstrated by comparing the performance of estimates with and without the logratio transformations.

The original data unit components were estimated directly by independent OK, OCK and RCK, and then compared to OK, OCK and RCK using the logratio transformation methods. Where the constant sum constraint was not satisfied for the original data unit estimates, the estimated values were normalized to the constant sum. Of course, this is simply an approximation to get an acceptable result and has no theoretical basis. There is also a potential problem with negative estimates (from possible high negative kriging weights) – setting these estimates to a small positive value is arbitrary and subjective (Pawlowsky-Glahn and Olea, 2004). The logratio methodology will, by construction, result in the correct constant sum.

Note that cokriging is not particularly widely used in mining industry practice, due to the extra covariance functions that must be defined, which means there is significantly more variogram/covariance modelling required (for n variables, $n(n+1)/2$ variogram models are needed). In addition, the demands of the LMC must be considered -for example, where the different variables within a domain are well correlated, but have different spatial ranges, fitting an adequate model to the experimental direct and cross-variograms can be very difficult, and often compromises must be made. The consequences of 'forcing' a model that does not reflect the underlying experimental variogram can result in the modelled covariances not matching the actual data configuration adequately.

Unless the variable of interest (primary) has been significantly under sampled compared to the other variables (secondary), then the weights given to the secondary variables by cokriging tend to be negligible, meaning that any improvements in the estimation variance are also small. In cases where all the variables have been sampled equally (isotopic sampling), and the direct variograms/covariances are proportional to the cross variograms/covariances, Wackernagel (1995, p. 150) showed that OK and OCK perform almost identically, and cokriging for estimation is unnecessary; this is the condition of 'autokrigeability' (see also Goovaerts, 1998).

There is no general consensus that cokriging in the isotopic case is unnecessary, however. Rivoirard (1994, p. 9) argues that even "when the values for all variables are available at all sample points, cokriging will improve the coherence between the estimated values by taking account of the relationships between the variables". For a case involving compositional data, then cokriging must be at least considered – the correlations between the variables for the oil sands data are relatively strong, and the preservation of these correlations is important to the problem at hand.

5.2 Variography

Variographic analysis was performed for the direct and cross-variograms for the three separate data sets (for data in original units, alr-transformed and clr-transformed). All variables, both before and after logratio transformation, were essentially isotropic in the horizontal direction, but with much shorter ranges in the vertical (downhole) direction, reflecting the near-horizontal stratigraphy. Therefore, the experimental variograms were generated with no horizontal anisotropy (major range = semi-major range), with the minor direction in the vertical.

The experimental variograms were calculated with a lag of 150m in the horizontal (i.e., isotropic) and 1.5m in the vertical direction, with a lag tolerance half of the lag spacing. An angular tolerance of 90° was used for the horizontal variograms, and 45° for the vertical. Testing of various horizontal bandwidths and angular tolerances made little difference to the appearance of the experimental variograms, but the use of a 10m vertical bandwidth for the horizontal direction resulted in more robust variograms. This is logical, because there is a drift in the vertical direction (emphasizing the stratigraphic nature of the deposit), so restricting the search for sample pairs vertically improved the experimental variogram structure.

The experimental variograms were modelled with a nugget effect and two or three spherical structures, and for the cross-variograms, the rules of the LMC were followed. The behaviour of the variogram at the origin has the most influence on the outcome of a kriging (Chiles and Delfiner, 1999, p. 175), so particular care was taken to model the nugget effect from the variogram in the vertical direction and the short-range structures for the vertical and horizontal directions.

The variogram models for the original unit variables are shown in Table 5-1 and the cross-variogram model parameters appear in Table 5-2 to Table 5-4. The direct experimental and model variogram for bitumen is shown in Figure 5-1. A significant number of variograms have been modelled for this study, but only key examples will be shown in the body of the thesis.

Variogram Models - original data units						
			Variance	Range (metres)		
Variable	Structure	Type	(sill)	Major	Semi	Minor
Bitumen	1	Nugget	0.00050			
	2	Spherical	0.00037	200	200	5
	3	Spherical	0.00047	3000	3000	20
Coarse	1	Nugget	0.0026			
	2	Spherical	0.0041	40	40	4
	3	Spherical	0.0026	220	220	14
	4	Spherical	0.0025	900	900	40
Fines	1	Nugget	0.0026			
	2	Spherical	0.0055	35	35	3.5
	3	Spherical	0.0019	200	200	15
	4	Spherical	0.0030	800	800	24
Water	1	Nugget	0.00012			
	2	Spherical	0.00016	40	40	5
	3	Spherical	0.00013	150	150	20
	4	Spherical	0.00010	2000	2000	50

Table 5-1. Variogram model parameters, original data units.

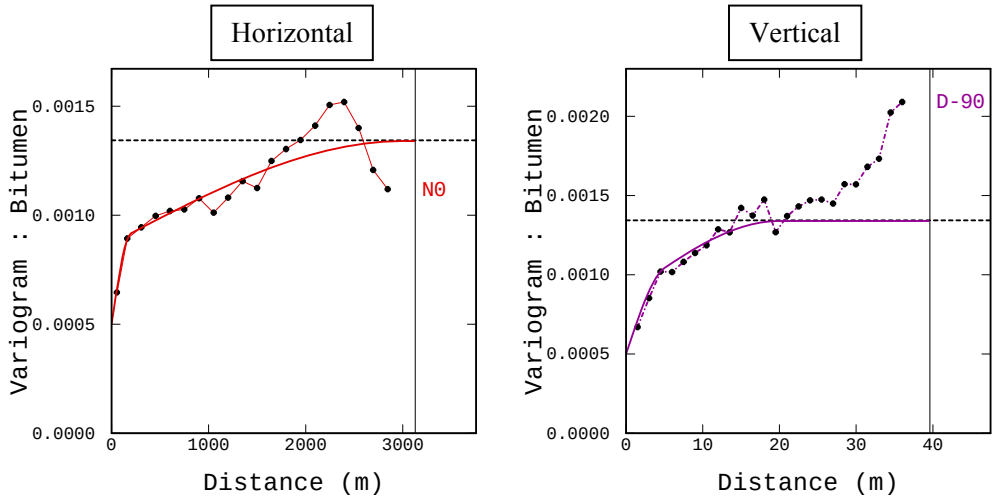

Figure 5-1. Variogram model for original unit bitumen.

Cross Variogram Models - original data units									
Structure	Type	Variance-Covariance matrix					Range (metres) Major	Semi	Minor
1	Nugget		Bitumen	Coarse	Fines	Water			
		Bitumen	0.3815	0.1172	-0.0790	-0.0899			
		Coarse	0.1172	0.2269	-0.0849	-0.0434			
		Fines	-0.0790	-0.0849	0.1386	0.0823			
		Water	-0.0899	-0.0434	0.0823	0.1994			
2	Spherical		Bitumen	Coarse	Fines	Water	110	110	10
		Bitumen	0.4264	0.3509	-0.4085	-0.4294			
		Coarse	0.3509	0.6566	-0.6736	-0.4532			
		Fines	-0.4085	-0.6736	0.7286	0.4355			
		Water	-0.4294	-0.4532	0.4355	0.6240			
3	Spherical		Bitumen	Coarse	Fines	Water	1050	1050	30
		Bitumen	0.1921	0.1005	-0.1497	-0.1192			
		Coarse	0.1005	0.1165	-0.1104	-0.1434			
		Fines	-0.1497	-0.1104	0.1328	0.1336			
		Water	-0.1192	-0.1434	0.1336	0.1766			

Table 5-2. Cross-variogram model parameters, original data units.

Cross Variogram Models - alr transformed								
Structure	Type	Variance-Covariance matrix				Range (metres) Major	Semi	Minor
1	Nugget		alrB:C	alrF:C	alrW:C			
		alrB:C	0.0229	-0.0133	-0.0142			
		alrF:C	-0.0133	0.1365	0.0651			
		alrW:C	-0.0142	0.0651	0.0532			
2	Spherical		alrB:C	alrF:C	alrW:C	250	250	8
		alrB:C	0.0226	-0.0300	-0.0200			
		alrF:C	-0.0300	0.6412	0.3000			
		alrW:C	-0.0200	0.3000	0.2514			
3	Spherical		alrB:C	alrF:C	alrW:C	950	950	21
		alrB:C	0.0265	-0.0400	-0.0400			
		alrF:C	-0.0400	0.3440	0.1400			
		alrW:C	-0.0400	0.1400	0.1232			

Table 5-3. Cross-variogram model parameters, alr transformed data.

Cross Variogram Models - clr transformed									
Structure	Type	Variance-Covariance matrix					Range (metres) Major	Semi	Minor
1	Nugget		clrB	clrC	clrF	clrW			
		clrB	0.1401	0.0836	-0.1602	-0.0636			
		clrC	0.0836	0.0704	-0.1241	-0.0300			
		clrF	-0.1602	-0.1241	0.2603	0.0240			
		clrW	-0.0636	-0.0300	0.0240	0.0695			
2	Spherical		clrB	clrC	clrF	clrW	110	110	10
		clrB	0.0326	0.0356	-0.0458	-0.0225			
		clrC	0.0356	0.0443	-0.0606	-0.0192			
		clrF	-0.0458	-0.0606	0.1058	0.0006			
		clrW	-0.0225	-0.0192	0.0006	0.0411			
3	Spherical		clrB	clrC	clrF	clrW	1050	1050	30
		clrB	0.0670	0.0398	-0.0820	-0.0248			
		clrC	0.0398	0.0312	-0.0505	-0.0204			
		clrF	-0.0820	-0.0505	0.1082	0.0243			
		clrW	-0.0248	-0.0204	0.0243	0.0209			

Table 5-4. Cross-variogram model parameters, clr transformed data.

The ranges for bitumen and water in the original unit models were greater than those for coarse and fines (approximately 2,000m cf. 900m), with bitumen having a higher nugget than the other variables (~40% cf. ~20%).

The cross-variograms for the original unit data have a low to moderate relative nugget (~15 to 35% of the total sill). For the horizontal directions there is a steep first structure with a range of 110m that reaches approximately 80% to 90% of the total variance. The variogram then flattens out, before reaching the total sill at 1050m. The range in the vertical direction has been modelled at 30m. The relative nugget effects for the alr transformed variables are similar to the original data units, but they are significantly higher for the clr transformed data (from 50% to 60%).

5.3 Cross-Validation and Block Estimation

Cross-validation is a technique that can be used to compare estimation methods and to check the validity of the variogram models. Clark (1986) provides an overview of the application of cross-validation, also known as the 'leaving-one-out' method (Davis, 1987):

- Remove one sample from the data set;
- Use the remaining data set to estimate at the removed sample location; and
- Calculate the error (estimated value (Z^*) – true value (Z)).

If the estimation is unbiased, then the mean error should be zero, and have a minimal standard deviation. Useful visual checks (Deutsch and Journel, 1998 p. 94, Clark, 1986) include:

- Scatterplots of true vs. estimated values should show high correlation and few outliers;
- Histograms of the errors should be symmetric and centred at zero; and
- Scatterplots of the errors vs. estimated value should be centred at zero error (conditional unbiasedness), and show no obvious pattern, as the error and estimated value should be independent.

Note that cross-validation does not remove the subjective element from model fitting (Clark, 1986), and interpretation of the results is best done on a comparative basis. The absolute 'best' model cannot be defined by the technique, but poor models and methods can be identified. Cross-validation has only been undertaken in this instance for OK and OCK, as the software packages used for this study cannot perform cross validation for RCK.

5.3.1 Original Unit Cross Validation

Cross-validation was run for OK and OCK using the variogram models shown in Table 5-1 and Table 5-2, and the search neighbourhoods shown in Table 5-7. Example cross-validation plots for bitumen estimated by OK are shown in Figure 5-2. The upper left diagram shows the scatterplot for estimated value Z* vs. true value Z, the upper right diagram the histogram of the estimation error and the lower left diagram the scatterplot of the estimation errors vs. the estimated values Z*.

The scatterplot of Z* vs. Z shows that the true high grade values have generally not been reproduced by OK, and the estimates are closer to the overall mean grade of the domain; this smoothing is typical of OK. The points highlighted in red are those where the estimation error is less than -0.1 and greater than 0.1.

The histogram of the estimation errors (only the range between -0.15 and 0.15 is shown), although having a mean very close to zero (0.0002), is not quite symmetric, and has a negative skew which is also due to the poor reproduction of the true bitumen high grades.

The scatterplot of Z* vs. the estimation error is centred at zero error, and shows that there is no strong relationship between the error and the estimated value (conditional unbiasedness), although the obvious outliers (highlighted in red) are the same data points as the mostly true high grades that have been smoothed during estimation shown in the scatterplot of Z* vs. Z.

The points showing the high bias were from single samples from a number of holes spread across the deposit. For the 17 locations with low bias though, 11 of these samples were from contiguous high bitumen zones from two drillholes in the south-western part of the study area, with the other samples from three other holes. If these samples are removed for cross validation, then the correlation between the true and estimated values increases to 0.7, and the mean error, which is already negligible, increases by 0.00001. This is not unexpected due to the smoothing from OK.

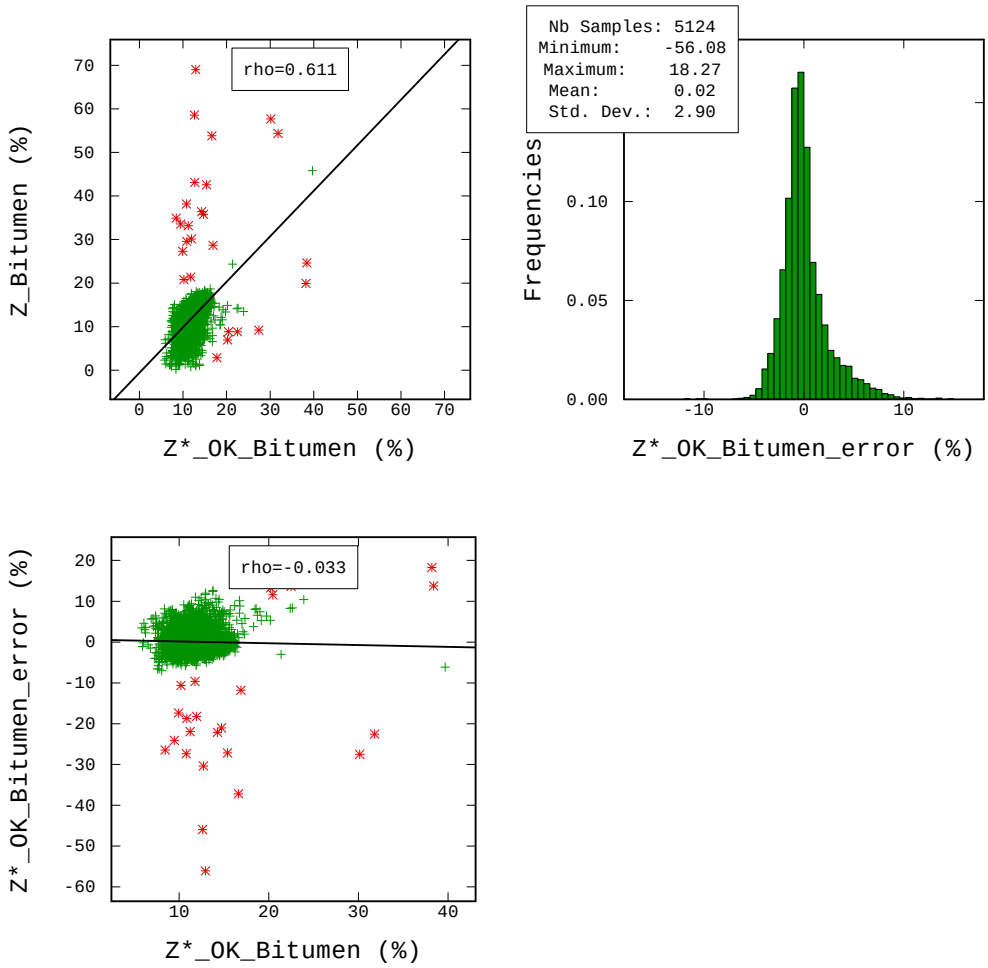

Figure 5-2. Cross-validation for bitumen, OK.

Table 5-5 shows the statistics for the estimation error, including the mean squared error (MSE) which uses the mean (bias) and variance of the errors:

$$MSE = E\{[z^* - z]^2\} = error\ \text{var} + bias^2 \qquad \textbf{5-1}$$

and the percentage bias (% bias):

$$\%\ bias = 100 * \left(\sum (z^* - z) / \sum z\right) \qquad \textbf{5-2}$$

The performance of OK and OCK are very similar, and essentially unbiased. The cross-validation shows that with the exception of the bitumen true high grades, the estimation methods and variogram models are suitable, and unbiased.

VARIABLE	Minimum	Maximum	Mean	Variance	% Bias	MSE
Z*_OK_Bitumen_error	-0.5608	0.1827	1.93E-04	8.43E-04	0.156%	8.43E-04
Z*_OK_Coarse_error	-0.3130	0.5970	4.76E-04	6.02E-03	0.068%	6.02E-03
Z*_OK_Fines_error	-0.6635	0.3298	-6.36E-04	6.77E-03	-0.496%	6.77E-03
Z*_OK_Water_error	-0.0986	0.0743	-9.50E-05	2.28E-04	-0.211%	2.28E-04
Z*_OCK_Bitumen_error	-0.5575	0.2127	7.50E-05	8.17E-04	0.061%	8.17E-04
Z*_OCK_Coarse_error	-0.3469	0.6019	4.55E-04	6.00E-03	0.065%	6.00E-03
Z*_OCK_Fines_error	-0.6527	0.3839	-3.62E-04	6.86E-03	-0.282%	6.86E-03
Z*_OCK_Water_error	-0.0958	0.0803	-6.80E-05	2.29E-04	-0.151%	2.29E-04

Table 5-5. Estimation error statistics from cross-validation OK and OCK.

As an aside, the correlations between OK and OCK for bitumen is 0.94, and 0.99 for the other components, backing the argument of Wackernagel (1995) and Goovaerts (1998) that OK and OCK perform almost identically.

The OK and OCK cross validation estimates for the components in original units do not sum to one – Figure 5-3 shows the histograms for the summed totals.

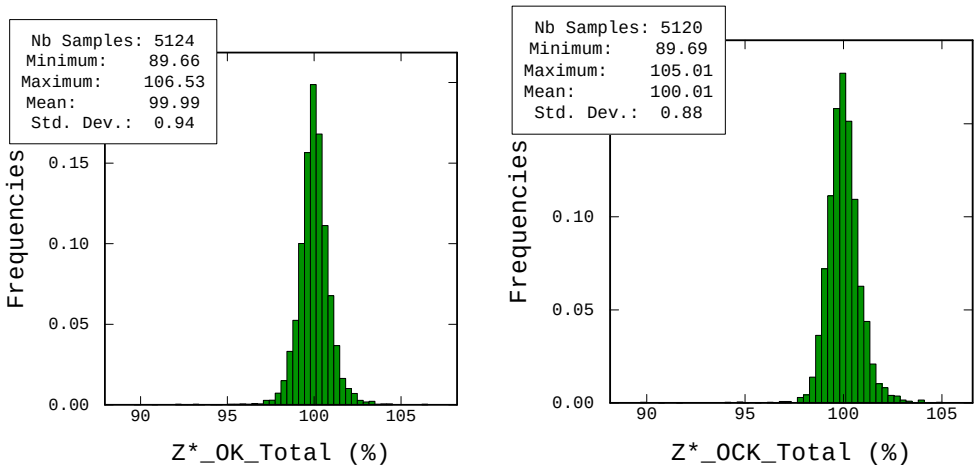

Figure 5-3. Histograms of summed components for original data unit cross-validation.

To be consistent with the constant sum, the component estimates were normalized – the graphical analysis for the normalized values are very similar to the non-normalized values, but the errors (Table 5-6) are different to those for the non-normalized values shown in Table 5-5. In particular, the biases for the normalised values are always greater than the non-normalized values.

This can be seen on a local basis in the scatterplot of the non-normalized and normalized data for bitumen shown below (Figure 5-4) – there is a clear divergence from the bisector line above 15% bitumen. The normalized values are biased high compared to the non-normalized values. The issue of normalization introducing bias is explored more fully in Section 5.3.2.

VARIABLE	Minimum	Maximum	Mean	Variance	% Bias	MSE
Z*_OK_Bitumen_normerr	-0.5598	0.2134	2.61E-04	8.40E-04	0.211%	8.40E-04
Z*_OK_Coarse_normerr	-0.3163	0.5943	5.94E-04	5.98E-03	0.085%	5.98E-03
Z*_OK_Fines_normerr	-0.6646	0.3292	-7.31E-04	6.79E-03	-0.570%	6.79E-03
Z*_OK_Water_normerr	-0.0995	0.0727	-1.24E-04	2.28E-04	-0.275%	2.28E-04
Z*_OCK_Bitumen_normerr	-0.5585	0.2403	1.42E-04	8.16E-04	0.115%	8.16E-04
Z*_OCK_Coarse_normerr	-0.3576	0.5903	5.81E-04	6.03E-03	0.083%	6.03E-03
Z*_OCK_Fines_normerr	-0.6576	0.3714	-6.69E-04	6.84E-03	-0.521%	6.84E-03
Z*_OCK_Water_normerr	-0.0958	0.0811	-5.60E-05	2.28E-04	-0.124%	2.28E-04

Table 5-6. Estimation error statistics for normalized cross-validation OK and OCK.

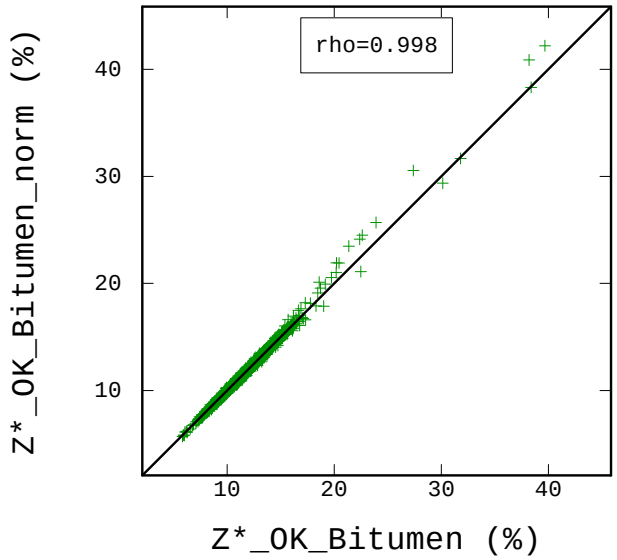

Figure 5-4. Scatterplot, normalized and non-normalized OK cross-validation for bitumen.

5.3.2 Original Unit Block Estimation

The previous discussion on all forms of kriging, including cross-validation has been to consider *point* kriging. However, in a mining application, larger volumes must be considered. For example, a selective mining unit (SMU) is the smallest volume on which selection of ore or waste (or some other destination such as a marginal product stockpile) can be made. Therefore, *block* kriging, where block averages are calculated from point-scale data, is required.

It is possible to discretize the blocks with many points, perform point kriging, and average the results over the whole block. However, this may require very intensive computation due to the number of kriging systems to solve. The computation can be reduced, however, by solving only one kriging system per block using volume averaged covariances. See Isaaks and Srivastava (1989, pp. 323 – 337) for further details.

A three-dimensional block model with a cell size of 50mE x 50mN x 1.5mRL (with 40 cells in E, 60 in N and 44 vertically) over the same area as the selected drilling data and restricted spatially to the 7% bitumen domain was constructed. OK, OCK and RCK estimates were run for the original and transformed data, using the variogram models shown in Table 5-1 and Table 5-2. In addition, an estimate using inverse distance squared (ID2) was run as a check – the same search parameters as the OK were used. The estimates for the original data units are useful to benchmark the results for the logratio transformed estimation (Section 5.3.4).

Quantitative Kriging Neighbourhood Analysis (QKNA, see Vann et al., 2003) was undertaken for a few variables to select a suitable kriging search neighbourhood. The use of a search neighbourhood that is too restrictive may result in serious conditional bias, conversely a neighbourhood that is too large may be computationally heavy, and can result in significant negative kriging weights for OK (depending on the covariance function employed). This can lead to estimates which are outside the ranges of the input data – this in itself may not be a problem, but can often result in negative estimates for variables that by definition must be positive.

The QKNA showed that the use of more than 20 samples resulted in very high negative kriging weights. The search neighbourhood parameters used for the original data units and those used later for the logratio transforms are shown in Table 5-7. No octant search settings or restrictions on the maximum number of samples from a single drillhole were applied.

		Search Ellipse (m)		Number of samples		
Data	**Variables**	**Horizontal**	**Vertical**	**Min.**	**Max.**	**Discretisation**
Non-transformed variables	**Bitumen**	2000	20	10	20	10 x 10 x 1
	Coarse	900	30	10	20	10 x 10 x 1
	Fines	800	20	10	20	10 x 10 x 1
	Water	2000	20	10	20	10 x 10 x 1
Non-transformed	**Cokriged**	1000	20	10	20	10 x 10 x 1
Logratio transformed	**alr cokriged**	900	20	10	20	10 x 10 x 1
	clr cokriged	1000	20	10	20	10 x 10 x 1

Table 5-7. Kriging search neighbourhood parameters.

The results for the estimation of the original data units by OK, OCK and RCK are shown in Table 5-8 – the individual components for each block were added together ('Total'). The results for the individual components estimated by all three methods (if viewed in isolation) appear reasonable when compared to the input data, with similar means and less dispersed ranges than the drillhole data, and no negative estimates.

	Component	Minimum	Maximum	Mean	Variance
OK	Bitumen	0.0451	0.4099	0.1267	3.89E-04
	Coarse	0.3931	0.8104	0.7104	2.92E-03
	Fines	0.0212	0.4040	0.1197	3.23E-03
	Water	0.0140	0.0933	0.0430	1.23E-04
	Total	0.9175	1.0902	0.9998	1.50E-04
OCK	Bitumen	0.0199	0.3819	0.1268	3.57E-04
	Coarse	0.3692	0.8172	0.7115	2.59E-03
	Fines	0.0174	0.4918	0.1187	3.12E-03
	Water	0.0103	0.1159	0.0431	1.53E-04
	Total	0.9750	1.1144	1.0000	1.20E-05
RCK	Bitumen	0.0184	0.3276	0.1268	2.62E-04
	Coarse	0.3886	0.8115	0.7112	1.20E-03
	Fines	0.0219	0.4836	0.1192	1.50E-03
	Water	0.0094	0.1168	0.0425	1.38E-04
	Total	0.9248	1.0450	0.9997	2.50E-05

Table 5-8. Results for OK, OCK and RCK estimates for the original data units.

Figure 5-5 below shows a series of swath plots in three directions comparing the bitumen in the drilling with the estimates. Swath plots were used in Chapter 4 to show drifts or trends – here, in addition to showing any trends, the performance of the estimates can be assessed. In general, an unbiased estimate will be a linear smoothing of the more heterogeneous areas.

The swath plots show that the all the estimators perform globally reasonable and across the domain-wide slices, with the estimated values showing smoothing as expected compared to the input drilling. Swath plots for the other components also show smoothing and lack of bias compared to the input data, which is consistent with the cross-validation results.

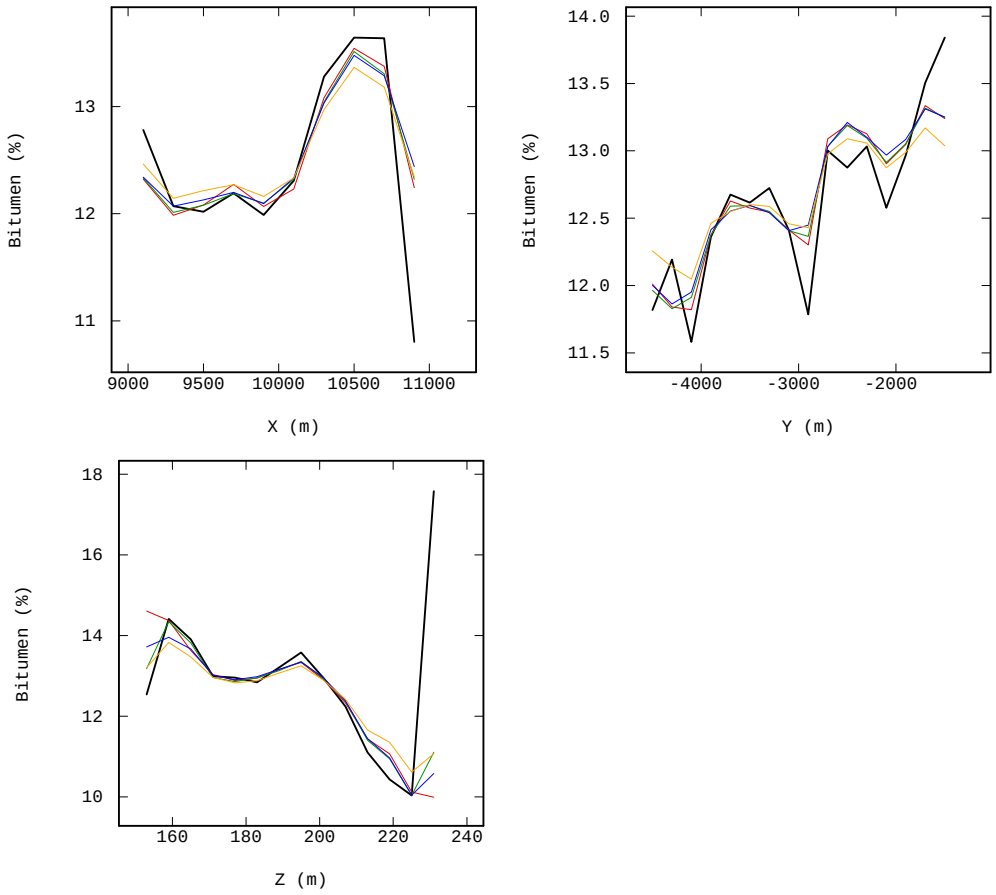

Figure 5-5. Swath plots for bitumen estimates vs. drilling (easting– top left, northing – top right, RL – bottom left. Colour scheme: black = drilling, red = ID2, green = OK, blue = OCK, orange = RCK.).

The OK and both cokriging methods do not result in the estimated components summing to the required constant of one (Figure 5-6). Normalization to the constant sum was performed, and Table 5-9 below shows the relative differences in percentages between the normalized and non-normalized estimates.

While the global means appear relatively unbiased, this is not the case for the tails of the distributions. For example, from Table 5-9 for OK, the minimum value for the normalized bitumen value is 2.35% higher relative to the non-normalized value, but the maximum normalized bitumen value is 6.345% lower. Figure 5-7 shows scatterplots for the normalized and non-normalized estimates for OK for bitumen and coarse. The high-grade tail (>21% bitumen non-normalised) has been highlighted in red, and the corresponding points for the coarse scatterplot are also highlighted.

The normalization has introduced a bias – the high or low tails of a distribution are often critical (e.g., a variable of economic value, or a driver for metallurgical processing), so this distortion of the distribution of an estimate by normalization is incorrect, and not advised.

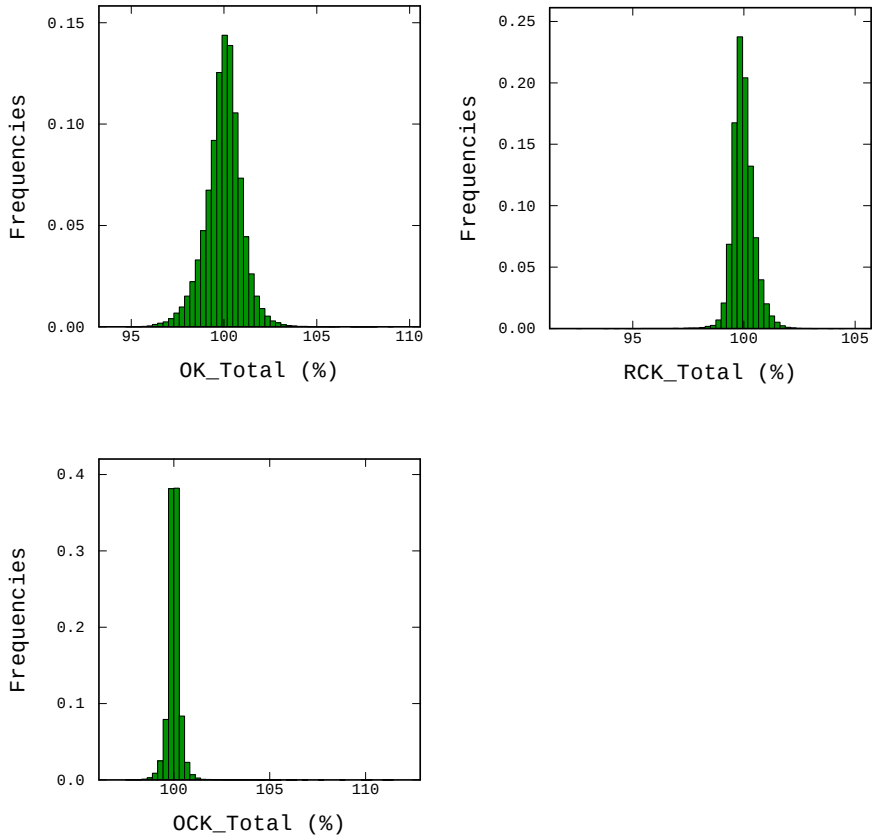

Figure 5-6. Histograms for total of added components in each block, OK, OCK, RCK.

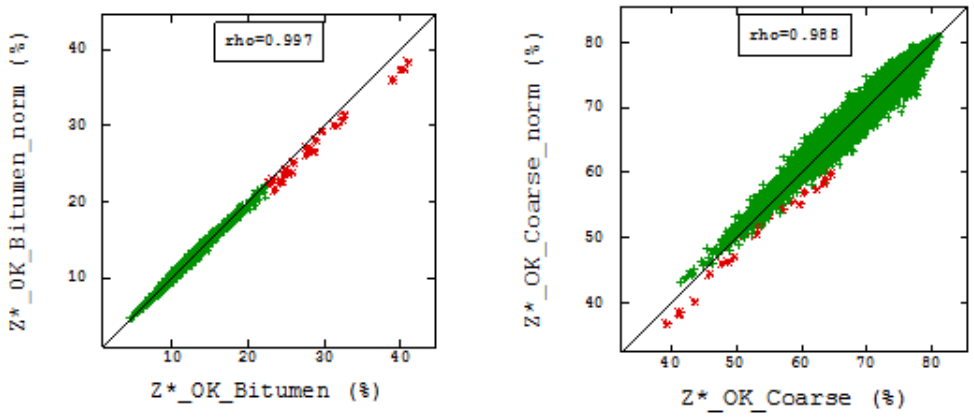

Figure 5-7. Scatterplots, non-normalized vs. normalized OK estimates for bitumen and coarse.

	Component	Minimum	Maximum	Mean	Variance
OK	Bitumen	2.350%	-6.345%	0.0079%	-2.564%
	Coarse	-6.344%	-0.031%	0.0127%	-5.822%
	Fines	0.424%	1.015%	0.0919%	1.238%
	Water	-0.858%	1.929%	0.0698%	0.000%
OCK	Bitumen	0.202%	-9.876%	-0.0079%	-2.778%
	Coarse	-0.520%	0.625%	0.0014%	1.544%
	Fines	0.860%	0.183%	-0.0169%	-0.321%
	Water	0.487%	-0.984%	-0.0232%	0.000%
RCK	Bitumen	0.163%	1.218%	0.0552%	3.846%
	Coarse	-0.126%	0.691%	0.0394%	5.000%
	Fines	-0.411%	0.182%	-0.0252%	-1.333%
	Water	0.106%	-1.858%	-0.0235%	0.000%

Table 5-9. Relative differences between normalized and non-normalized component estimates.

The magnitude of the bias is also dependent on the choice of kriging algorithm – from Table 5-9, the normalization bias for RCK is lower compared to the other estimates. This is not unexpected, as the range and skewness of the distribution of the summed components (Figure 5-6) is lowest for RCK.

5.3.3 Logratio Cross-Validation

Cross-validation was undertaken for the variables in logratio space, using the variogram models shown in Table 5-3 and Table 5-4 and the search parameters shown in Table 5-7. Statistics for the mean estimation error are shown in Table 5-10. The mean for alrB:C is zero, but the mean of the errors for the other alr variables diverge from zero. However, the percentage bias is still low, given the values of the logratio variables. The means for the clr transformed variables are also farther from zero than those shown for the estimates in original data units, and the percentage bias for clrB appear significant. In this case however, the mean value for clrB is relatively close to zero, so the percentage bias measure is very sensitive in this case. It therefore appears that the kriged estimates in logratio space are relatively unbiased. The significance of these values can be assessed against the results once the logratio values are back-transformed into original data units, which is shown in Table 5-11.

In Table 5-11, the mean error for bitumen is slightly positive, but there are significant biases for the other components, including a pronounced negative bias for the fines fraction. For both the alr and clr back-transforms, similar magnitudes of bias are seen for all the components. Note that the MSE is not particularly high, as the variance of the errors is quite low, but showing consistent biases.

VARIABLE	Minimum	Maximum	Mean	Variance	% Bias	MSE
Z*_alr_B:C_error	-2.8086	2.2549	0.0000	0.0475	0.000%	0.0475
Z*_alr_F:C_error	-5.9264	3.4350	-0.0034	0.5079	0.158%	0.5079
Z*_alr_W:C_error	-5.0011	2.7086	-0.0022	0.1889	0.107%	0.1889
Z*_clr_B_error	-2.4052	3.0928	0.0011	0.1470	-2.245%	0.1470
Z*_clr_C_error	-1.2298	3.2126	0.0008	0.0752	0.047%	0.0752
Z*_clr_F_error	-2.9666	2.5722	-0.0013	0.2664	0.299%	0.2664
Z*_clr_W_error	-1.8264	1.0528	-0.0007	0.0717	0.057%	0.0717

Table 5-10. Estimation error statistics from cross-validation for logratio units.

VARIABLE	Minimum	Maximum	Mean	Variance	% Bias	MSE
Z*_alr_Bitumen_error	-0.5732	0.1733	3.94E-04	8.11E-04	0.318%	8.11E-04
Z*_alr_Coarse_error	-0.4620	0.6590	1.56E-02	6.22E-03	2.226%	6.47E-03
Z*_alr_Fines_error	-0.7341	0.4367	-1.51E-02	7.02E-03	-11.776%	7.24E-03
Z*_alr_Water_error	-0.0934	0.0912	-9.29E-04	2.38E-04	-2.060%	2.39E-04
Z*_clr_Bitumen_error	-0.5802	0.1313	8.39E-04	9.28E-04	0.678%	9.29E-04
Z*_clr_Coarse_error	-0.3418	0.6963	2.14E-02	6.98E-03	3.046%	7.44E-03
Z*_clr_Fines_error	-0.7765	0.3232	-2.09E-02	8.17E-03	-16.321%	8.61E-03
Z*_clr_Water_error	-0.1089	0.0776	-1.31E-03	2.65E-04	-2.900%	2.67E-04

Table 5-11. Estimation error statistics from cross-validation for back-transformed logratio variables to original data units.

These results are consistent with those shown in the simple example in Section 2.8, and the relative biases are as expected given that the logratio methods have been demonstrated to result in means that approach the standardized geometric mean of the original data. Table 5-12 shows the naive i.e., not declustered arithmetic and geometric means of the original data, the standardized geometric mean (i.e., the global 'Expected Result'), the global Z^* estimates for the logratio methods and the global percentage biases.

Note that the biases are based on the raw arithmetic mean for the components, and therefore they will not exactly match the biases shown in Table 5-11. The actual estimates show that the coarse fraction is not as positively biased high, or the fines fraction as negatively biased as the standardized geometric mean, but the other components are very similar to the expected results. Local variations in data configuration, and therefore in the kriging system, mean that it is unlikely that the expected and actual logratio Z^* estimates will match exactly. The sum of the components, of course, is unbiased.

	Bitumen	Coarse	Fines	Water	Total
Arith. Mean	0.1263	0.7115	0.1201	0.0421	1.0000
Geom. Mean	0.1201	0.6998	0.0816	0.0369	0.9384
Expected Result	0.1279	0.7458	0.0870	0.0393	1.0000
Expected % Bias	1.3%	4.8%	-27.6%	-6.8%	0.0%
Z* alr Results	0.1266	0.7272	0.1050	0.0412	1.0000
Z* alr % Bias	0.3%	2.2%	-12.6%	-2.2%	0.0%
Z* clr Results	0.1271	0.7329	0.0992	0.0408	1.0000
Z* clr % Bias	0.7%	3.0%	-17.4%	-3.1%	0.0%

Table 5-12. 'Expected' results from direct kriging of logratios.

5.3.4 Logratio Block Estimation

Block estimates were made into the same three-dimensional model used for the estimates in original data space, with the variogram models shown in Table 5-3 and Table 5-4 and search parameters shown in Table 5-7. The basic statistics of the OCK and RCK estimates for the logratio data are similar to the declustered input data indicating no bias in the logratio transformed space (see Table 5-13 for RCK results).

	Drilling			RCK Estimates		
Variable	**Minimum**	**Maximum**	**Mean**	**Minimum**	**Maximum**	**Mean**
alrB:C	-4.9375	1.2266	-1.7628	-3.2141	0.3337	-1.7620
alrF:C	-6.9411	4.5948	-2.1485	-4.9742	1.3495	-2.1645
alrW:C	-5.1957	2.0616	-2.9438	-4.8359	-0.5495	-2.9133
clrB	-3.6411	2.0187	-0.0490	-1.0176	0.7193	-0.0445
clrC	-1.6292	3.2221	1.7138	0.6191	2.3827	1.7124
clrF	-3.7189	2.9656	-0.4347	-1.5952	0.7432	-0.4510
clrW	-2.6068	0.4325	-1.2300	-1.9563	-0.6536	-1.2178

Table 5-13. Transformed drilling data v. rescaled cokriged estimate.

Figure 5-8 below shows the swath plot for the alrB:C estimators vs. the drilling, which shows an acceptable smoothing and lack of bias for each of the logratio estimators. Swath plots for the estimates of the other logratio variables show similar lack of bias in logratio space, which is consistent with the cross-validation results.

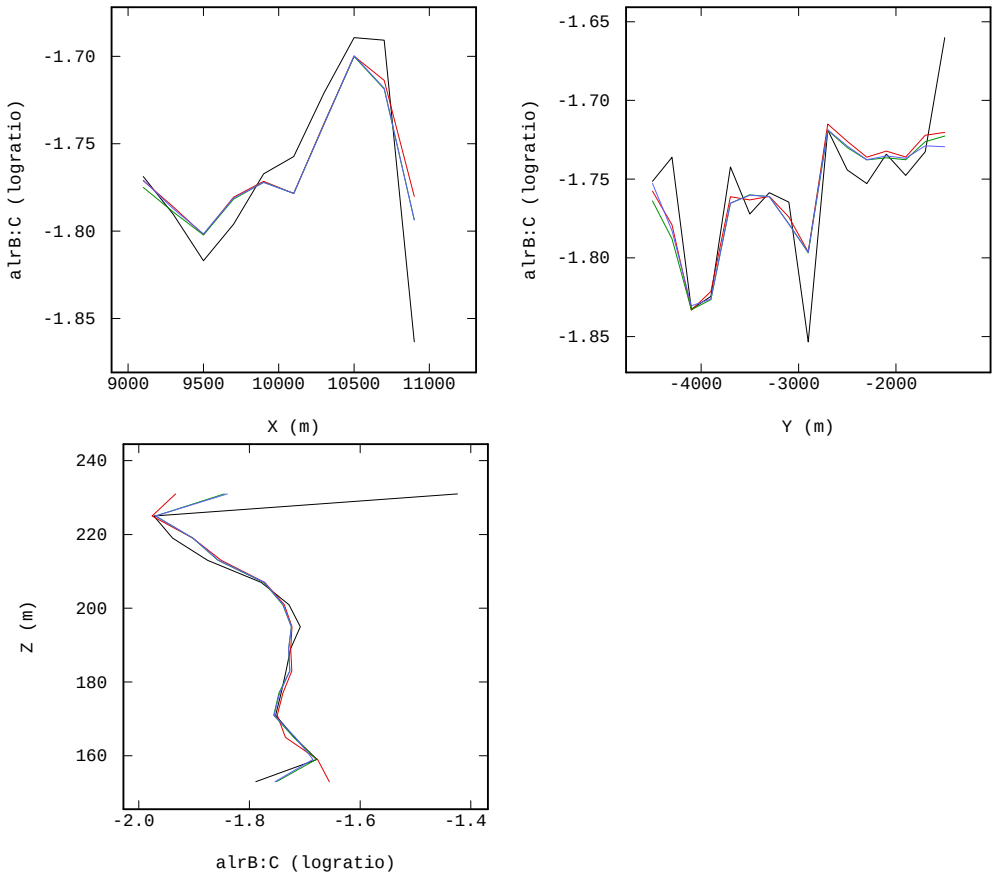

Figure 5-8. Swath plots for alrB:C estimates vs. drilling (easting– left, northing – middle, RL – right. Colour scheme: black = drilling, red = ID2, green = OCK, blue = RCK).

The logratio estimates were then back-transformed into original data units. Initial validation of the results involved comparison with the input data (see Table 5-14 and Table 5-15). The maxima and minima for all components are within the range of the input data. Other observations include:

- The means for bitumen in the estimates are slightly higher than the declustered mean of 0.124, but the maximum value for the clr method is significantly lower than that for the alr method;
- The means for the coarse estimates are much higher than the declustered sample mean of 0.703, indicating a positive bias;
- The means for the fines estimates are below the declustered sample mean of 0.128, indicating a negative bias, and the maximum value for the clr method is significantly lower than that for the alr method; and
- The means for the water estimates are slightly lower than the declustered sample mean of 0.045.

	Declustered Drilling			Back-transformed OCK				Back-transformed RCK			
Variable	Min.	Max.	Mean	Min.	Max.	Mean	Mean %	Min.	Max.	Mean	Mean %
Bitumen	0.002	0.690	0.124	0.019	0.411	0.127	102.6%	0.018	0.411	0.127	102.7%
Coarse	0.009	0.854	0.703	0.176	0.834	0.730	103.9%	0.175	0.833	0.731	104.0%
Fines	0.001	0.911	0.128	0.006	0.675	0.101	78.4%	0.006	0.676	0.100	77.7%
Water	0.005	0.173	0.045	0.007	0.135	0.042	94.0%	0.006	0.134	0.042	93.3%
Total	1.000	1.000	1.000	1.000	1.000	1.000	100.0%	1.000	1.000	1.000	100.0%

Table 5-14. Comparison of input data and cokriged models, alr method.

	Declustered Drilling			Back-transformed OCK				Back-transformed RCK			
Variable	Min.	Max.	Mean	Min.	Max.	Mean	Mean %	Min.	Max.	Mean	Mean %
Bitumen	0.002	0.690	0.124	0.056	0.263	0.128	103.5%	0.059	0.275	0.129	103.9%
Coarse	0.009	0.854	0.703	0.339	0.827	0.737	104.8%	0.417	0.826	0.739	105.1%
Fines	0.001	0.911	0.128	0.014	0.489	0.094	73.2%	0.017	0.401	0.092	71.8%
Water	0.005	0.173	0.045	0.010	0.109	0.041	91.6%	0.011	0.103	0.040	89.7%
Total	1.000	1.000	1.000	1.000	1.000	1.000	100.0%	1.000	1.000	1.000	100.0%

Table 5-15. Comparison of input data and cokriged models, clr method.

The coarse and fines fractions are biased as expected, and this is confirmed by the swath plots of the back-transform for the alr method (Figure 5-9). It can be seen that the estimates for bitumen and water have no obvious bias visually, but the bias for the coarse (positive) and fines (negative) fractions are clear to see in all three directions.

Plots for the clr method show a very similar pattern to the alr method. These results agree with those from the cross-validation – the positive bias occurs for the dominant component. The alr and clr methods result in similar biases.

Aitchison (1986) asserts that the choice of denominator does *not* have an effect on the process, but it is worth testing whether using the coarse fraction as the denominator for the alr method has an effect on the bias. To test this, the estimates were re-run for the alr method using the fines as the denominator.

The logratios were re-calculated and direct and cross-variograms modelled for the new data. The variogram models were similar to those for the initial alr transform, although slight editing of ranges and sills was required to get a better fit for the experimental variogram. The search neighbourhood used was the same as for the initial alr estimate. Rescaled cokriging was run for the three alr variables, and then back-transformed into original data units.

The results were almost identical to those from the initial alr transform that used the coarse fraction as the denominator – the bitumen and water *appear* acceptable, but coarse was positively biased and the fines negatively biased. These results confirm that choice of denominator does not have an effect on the alr method, or for that matter any output statistics.

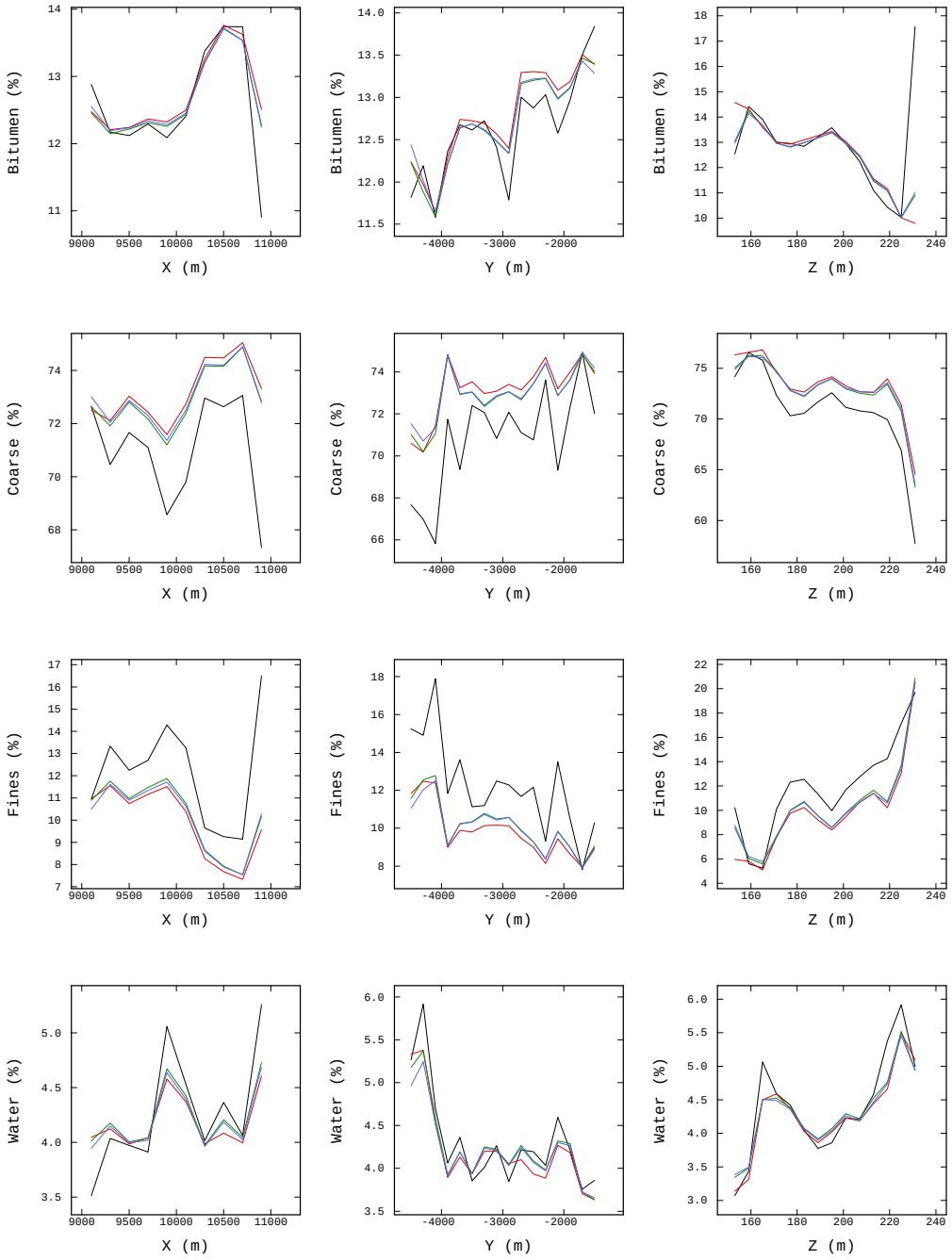

Figure 5-9. Swath plots, original data units back-transformed from alr estimate (Colour scheme: black = drilling, red = ID2, green = OCK, blue = RCK).

5.4 Concluding Comments for Chapter 5

Linear estimation of compositional data has a number of problems – direct estimation of the original data does not result in closed compositions at the unsampled locations, and does not preclude the possibility of negative estimates. Normalization of the estimates to close to the constant sum will result in bias.

Direct kriging and back-transformation of (smoothed) estimates of logratios is also problematic, since the linear averaging of log transformed variables when back-transformed does not result in the linear averaging in the original units, and consequently the results are biased. This bias is not due choice of logratio transform (or choice of denominator for the alr transform) or to the estimation method – the approximate same results were seen regardless of whether ID2, OK, OCK or RCK was used.

The bias as expected from the simple example was confirmed by application of the techniques to the oil sands data set. The estimated values tend towards the standardized geometric means of the original data – the extent of the bias depends upon the relative magnitudes of the original data.

Therefore, it is concluded that a non-linear technique will be required to correctly model the conditional distribution of the logratios before back-transformation into original data units. In particular, quantiles used to discretize the conditional distribution are passed through the back-transforms – the whole distribution rather than just the expected value is available through each of the back-transformation steps.

Chapter 6 MultiGaussian Kriging

6.1 Normal Scores Transform

Cell declustering with a grid size of 450mE x 450mN x 1.5mRL was selected for weighting during the normal scores transform. Declustering analysis for the logratio variables showed very similar patterns to those of the original data unit components discussed in Section 4.2.2. All of the logratio transformed components were transformed to Gaussian distributions via the normal scores method with these declustering weights applied.

6.2 Checks for Bivariate Gaussian Distributions

Application of Gaussian-based algorithms for *Z(x)* relies heavily on an assumption of multivariate Gaussian distribution for *Y(x)*. The normal scores transformation only ensures that the marginal distribution of *Y(x)* is Gaussian; it does not guarantee a multivariate Gaussian distribution. There are several checks (summarized in Chapter 3 and in Emery, 2005a) that can be made for bivariate Gaussian distributions (i.e., binormality), which is a consequence of the multivariate Gaussian assumption. Checking for higher order distributions is possible (assuming enough data were available), but Verly (1984) points out that 'it seems unlikely that a deposit could fail the check on trinormality after having passed the one on binormality'. These checks are not definitive (indeed, they can be quite subjective), and they are not formal statistical tests, but they are recommended to validate the use of the assumed multi-Gaussian model.

6.2.1 H-Scatterplots and Bivariate Scatterplots

H-scatterplots are a means of visualizing the information from a variogram or spatial covariance at a single lag; they are scatterplots between a variable and itself at some given vector separation – distances, directions and tolerances can be manipulated. The H-scatterplot of a Gaussian transformed variable should be elliptical in shape if it has a bivariate Gaussian distribution. Figure 6-1 shows the H-scatterplot for Gaussian-transformed alrB:C for a vertical direction, with a lag equal to the downhole sample interval. The red ellipse outlines on the plot are constant probability density contours for 25%, 50% and 95% as described by Deutsch and Deutsch (2011). The density contours are visually elliptical, and it could be concluded that the distribution is approximately bivariate Gaussian.

Deutsch and Deutsch (2011) however propose a quantitative check that compares the fraction of points falling within the contours with the theoretical fraction, and a second step then compares the fraction of points falling within the quadrants of the constant density ellipses with the theoretical value. The differences from the expected number in each quadrant for each probability contour are summed, averaged and standardized to give a single measure of deviation from the perfect bivariate Gaussian distribution. The likelihood of the distribution to not be bivariate Gaussian is then determined from this measure of deviation – in Figure 6-1, the value of >99.99% indicates that the distribution is not bivariate Gaussian.

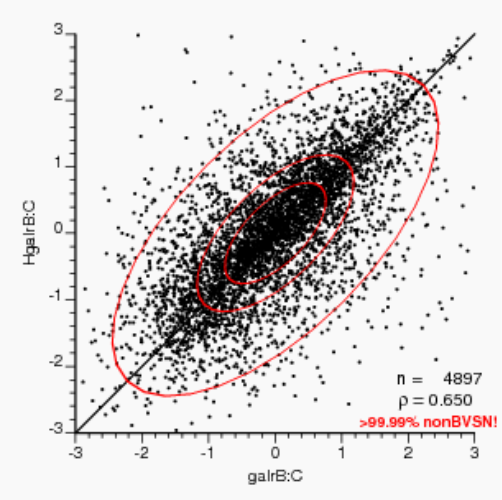

Figure 6-1. H-scatterplot, Gaussian-transformed alrB:C values, vertical direction.

Bivariate scatterplots between the three Gaussian transformed alr variables are shown in Figure 6-2. As with the H-scatterplot, the measure of deviation shows that the distributions are very likely not bivariate Gaussian, even though the scatters are approximate elliptical.

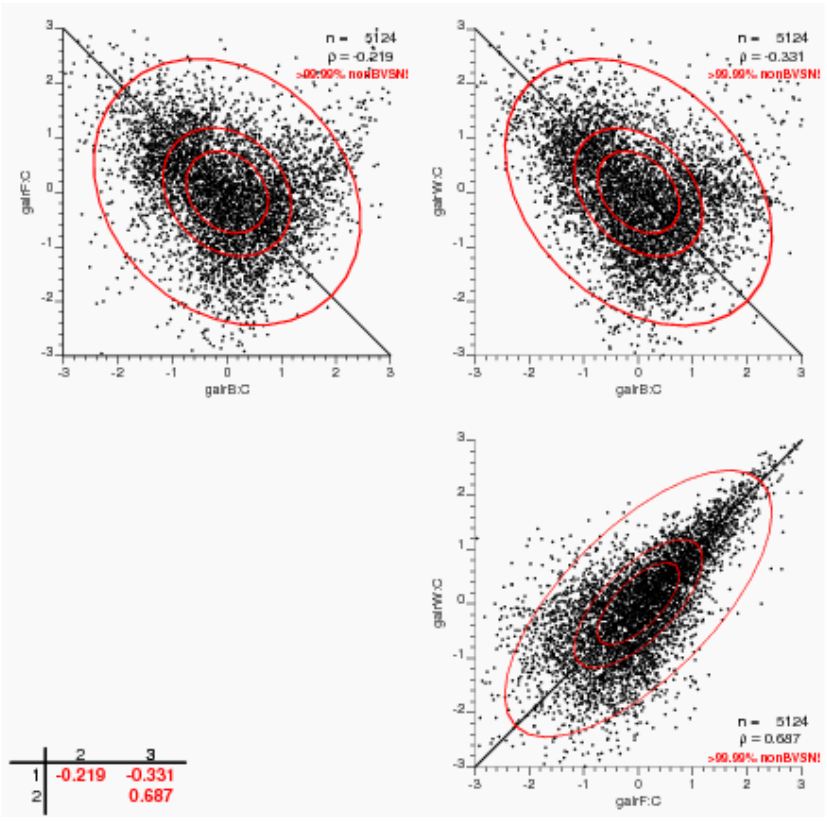

Figure 6-2. Bivariate scatterplots for Gaussian alr values.

The scatterplot method is only one of the possible checks for a bivariate Gaussian distribution – other checks are described below.

6.2.2 Square Root of Variogram vs. Madogram

The ratio of the square root of the variogram to the madogram i.e. the first-order variogram:

$$\gamma_1(\boldsymbol{h}) = \frac{1}{2} E\{|\, Y(\mathbf{u}+\boldsymbol{h}) - Y(\mathbf{u})\,|\} \qquad \textbf{6-1}$$

should be constant ($=\sqrt{\pi}$ i.e., approximately 1.7725) under the bivariate Gaussian assumption (Emery, 2005a). Figure 6-3 below shows these plots for the omni-directional horizontal case, for the alr data. Plots for the clr method and for the downhole direction all show that the ratio is close to the theoretical value beyond the first lag, with the differences in percentage terms between the experimental and theoretical values for each lag distance shown in Table 6-1.

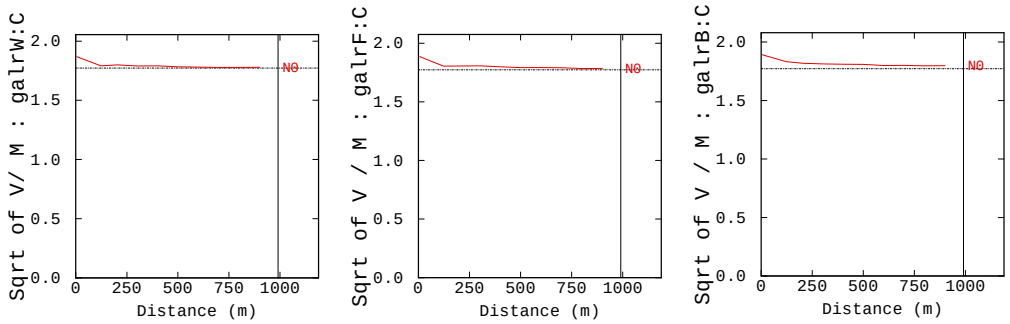

Figure 6-3. Ratio of square root of variogram vs. madogram, normal scores of alr transformed data, horizontal direction. Left to right alrW:C, alrF:C, alrB:C.

Lag (m)	3.3	122.7	205.5	302.3	401.0	500.5	600.4
galrB:C	109.9%	103.8%	103.6%	102.9%	102.9%	102.6%	102.1%
galrF:C	108.9%	102.3%	102.3%	102.4%	101.9%	101.4%	101.4%
galrW:C	107.4%	101.4%	101.7%	101.3%	101.2%	100.7%	100.5%
Lag (m)	**700.2**	**800.9**	**900.2**	**998.7**	**1099.5**	**1200.8**	**1299.2**
galrB:C	102.1%	101.7%	101.8%	101.4%	101.1%	101.2%	101.0%
galrF:C	101.3%	100.7%	100.8%	100.9%	100.5%	100.5%	100.1%
galrW:C	100.3%	100.3%	100.4%	100.6%	100.7%	100.6%	100.8%
Lag (m)	**1399.0**	**1498.7**	**1597.7**	**1700.6**	**1798.9**	**1895.8**	
galrB:C	100.6%	100.4%	100.7%	100.6%	100.7%	100.6%	
galrF:C	99.7%	100.0%	99.8%	100.1%	99.9%	99.5%	
galrW:C	100.5%	100.4%	100.4%	100.0%	100.2%	99.8%	

Table 6-1. Difference between experimental and theoretical values, ratio between square root of variogram and madogram.

It is unclear, however, what differences are *acceptable* – Emery (2005a) only mentions that the ratio must be independent of the lag, and Verly (1984) only discusses the values as being 'very close'. This check is therefore somewhat subjective, and is only a partial test for binormality – a more rigorous check is discussed in the next section.

6.2.3 Variograms of order ω

Another check is to compare the experimental variogram of some order ω to the theoretical variogram via the relationship for a fixed ω and varying **h**:

$$\frac{\gamma_\omega(\boldsymbol{h})}{\gamma_\omega(\boldsymbol{h_0})} = \left[\frac{\gamma(\boldsymbol{h})}{\gamma(\boldsymbol{h_0})}\right]^{\frac{\omega}{2}} \qquad \textbf{6-2}$$

where $\mathbf{h_0}$ is a reference lag distance where the variogram is reliable (often near the sill). A scatterplot between these two functions at various lags should plot on the bisector line if the bivariate Gaussian assumption is not violated.

The rodogram is the variogram of order ½:

$$\gamma_{0.5}(\boldsymbol{h}) = \frac{1}{2} E\left\{\sqrt{|\, Y(\mathbf{u}+\boldsymbol{h}) - Y(\mathbf{u})\,|}\right\} \qquad \textbf{6-3}$$

The relationship between the rodogram and theoretical variogram to the power of 0.25 from Equation 6-2 was checked downhole, with an example scatterplot shown for the Gaussian alrB:C component in Figure 6-4. The lag chosen as 'being reliable' was 25.5m.

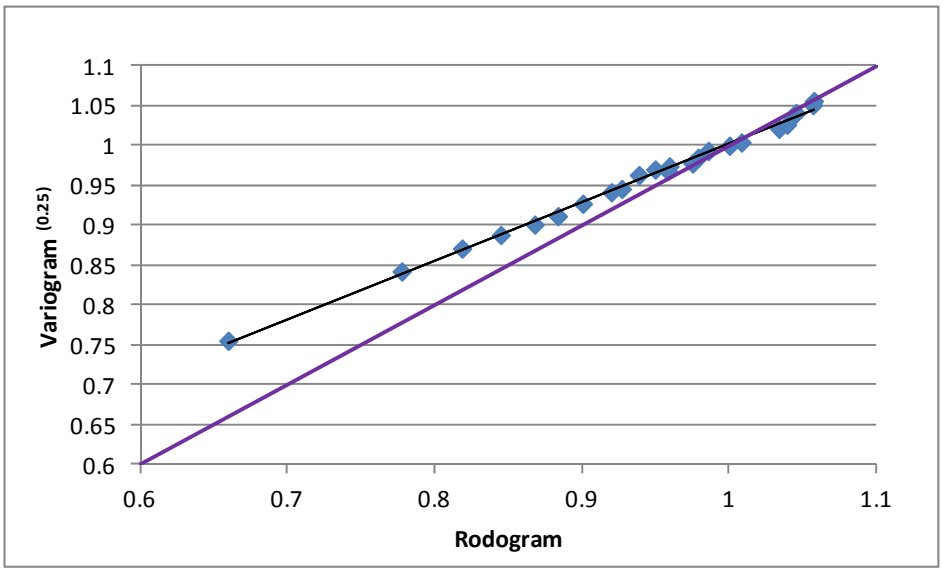

Figure 6-4. Scatterplot, rodogram vs. variogram to power of 0.25, Gaussian transformed alrB:C.

The regression line (black) indeed *looks* linear, but it is biased, only converging with the theoretical vs. experimental bisector line (magenta) above a value of one. Plots for the other variables are very similar, and changing the 'reliable lag' distance does not significantly alter the appearance of the plots.

It therefore seems that the bivariate Gaussian assumption is not met for this data set, most likely due to heteroscedasticity. It is possible that alternatives to the multivariate Gaussian framework could be used, for example Indicator kriging, but the assumptions associated with them (e.g., modelling of the tails of the distributions, order relation corrections) can also be strong. Further suggestions for alternative methods and transformations are discussed in Section 8.2.

So, even though the data departs from the multivariate Gaussian assumption, the lack of suitable alternatives means that MGK is still the best option of estimating logratio transformed compositional data.

6.3 Variography

Experimental variograms and cross-variograms were generated for the seven Gaussian transformed logratio components - the directions are the same as those used for the untransformed data. These experimental variograms were fitted with a model that consisted of a nugget and two spherical structures (Table 6-2 and Table 6-3. Figure 6-5 shows the direct and cross variogram in the horizontal direction for the Gaussian transformed alr variables). The parameters (lags, tolerances) used for the experimental variograms were essentially the same as for the non-Gaussian transformed components (see Chapter 5).

Cross Variogram Models - Gaussian alr transformed								
						Range (metres)		
Structure	Type	Variance-Covariance matrix				Major	Semi	Minor
1	Nugget		galrB:C	galrF:C	galrW:C			
		galrB:C	0.2712	-0.0275	-0.5350			
		galrF:C	-0.0275	0.2009	0.0685			
		galrW:C	-0.5350	0.0685	0.1858			
2	Spherical		galrB:C	galrF:C	galrW:C	150	150	8
		galrB:C	0.4488	-0.0471	-0.1918			
		galrF:C	-0.0471	0.5954	0.4537			
		galrW:C	-0.1918	0.4537	0.5913			
3	Spherical		galrB:C	galrF:C	galrW:C	950	950	20
		galrB:C	0.2800	-0.1784	-0.0977			
		galrF:C	-0.1784	0.2037	0.1926			
		galrW:C	-0.0977	0.1926	0.2229			

Table 6-2. Gaussian variogram model parameters, alr transformed data.

Cross Variogram Models - Gaussian clr transformed									
Structure	Type	Variance-Covariance matrix					Range (metres) Major	Semi	Minor
1	Nugget		gclrB	gclrC	gclrF	gclrW			
		gclrB	0.5239	0.4360	-0.4628	-0.2992			
		gclrC	0.4360	0.4552	-0.4631	-0.1914			
		gclrF	-0.4628	-0.4631	0.5634	0.1001			
		gclrW	-0.2992	-0.1914	0.1001	0.4610			
2	Spherical		gclrB	gclrC	gclrF	gclrW	110	110	10
		gclrB	0.2547	0.2795	-0.1913	-0.1969			
		gclrC	0.2795	0.3733	-0.2617	-0.1886			
		gclrF	-0.1913	-0.2617	0.2489	0.0011			
		gclrW	-0.1969	-0.1886	0.0011	0.4215			
3	Spherical		gclrB	gclrC	gclrF	gclrW	1050	1050	30
		gclrB	0.2214	0.1827	-0.2029	-0.1184			
		gclrC	0.1827	0.1715	-0.1640	-0.1311			
		gclrF	-0.2029	-0.1640	0.1878	0.1034			
		gclrW	-0.1184	-0.1311	0.1034	0.1175			

Table 6-3. Gaussian variogram model parameters, clr transformed data.

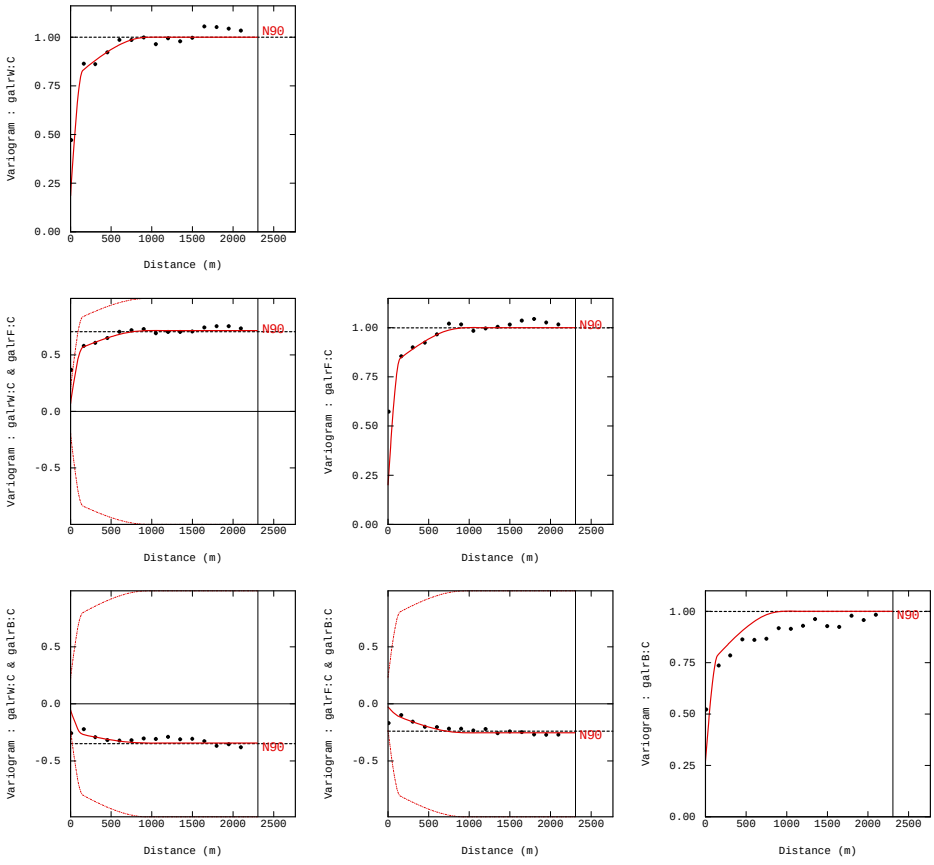

Figure 6-5. Direct and cross-variograms for Gaussian-transformed alr data, horizontal direction.

6.4 MGK Estimate (alr)

A grid was set up to cover the same area as the kriging grid, but with smaller block sizes (10mE x10mN x1.5mRL, 200 blocks E, 300 blocks W, 44 blocks RL).

Simple cokriging of the three normal scores transformed alr variables was performed using the variogram parameters shown in Table 6-2 and the same search parameters as used for OCK and RCK (Chapter 5). Three sets of fifty independent random standard normal values were generated at each grid node (one set for each variable), and then random but correlated Gaussian values were generated using the Cholesky lower triangle matrix shown in Table 6-4.

Correlation Matrix				Cholesky Lower Matrix			
	galrB:C	**galrF:C**	**galrW:C**		**galrB:C**	**galrF:C**	**galrW:C**
galrB:C	1	-0.2189	-0.3307	**galrB:C**	1	0	0
galrF:C	-0.2189	1	0.6868	**galrF:C**	-0.2189	0.9757	0
galrW:C	-0.3307	0.6868	1	**galrW:C**	-0.3307	0.6297	0.7030

Table 6-4. Correlation and Cholesky lower triangle matrices for Gaussian transformed alr variables.

The standard deviation of the SCK estimate for each variable was multiplied by the correlated random Gaussian value for that variable, and the estimated value for the SCK was added, resulting in fifty quantiles of the conditional distribution for each variable. Each quantile was then back-transformed through the inverse normal scores and alr transforms, as described in Chapter 3.

Fifty realizations for each of the four variables in original data units were thus available – the mean of the realizations represents the multiGaussian kriged value, and probabilities to be above or below selected cut-offs can be calculated. These quantile values were then averaged into the larger blocks used for the initial kriging estimates (50m x 50m x 1.5m), and the results are shown in Table 6-5.

The advantage of MGK over OK or OCK is that the conditional distribution is modelled. For data transforms that are non-linear (both the logratio and normal scores transforms are non-linear), the critical feature is that the quantiles are modelled, rather than just the expected value, and these quantiles can be passed through non-linear transformations.

	Bitumen	Coarse	Fines	Water
Mean Value	0.1262	0.7082	0.1212	0.0444
Probability > 0.05	97.1%	100.0%	65.7%	27.9%
Probability > 0.1	79.2%	99.9%	39.8%	4.7%
Probability > 0.15	22.6%	99.9%	26.5%	1.0%
Probability > 0.2	0.8%	99.8%	18.2%	0.3%
Probability > 0.25	0.6%	99.6%	13.0%	0.1%
Probability > 0.3	0.4%	99.1%	9.3%	0.1%
Probability > 0.35	0.3%	98.4%	6.5%	0.0%
Probability > 0.4	0.3%	97.4%	4.4%	0.0%
Probability > 0.45	0.2%	95.7%	2.8%	0.0%
Probability > 0.5	0.2%	93.3%	1.8%	0.0%
Probability > 0.55	0.1%	90.0%	1.2%	0.0%
Probability > 0.6	0.1%	85.4%	0.7%	0.0%
Probability > 0.65	0.0%	78.7%	0.4%	0.0%
Probability > 0.7	0.0%	67.6%	0.1%	0.0%
Probability > 0.75	0.0%	48.5%	0.1%	0.0%
Probability > 0.8	0.0%	12.3%	0.1%	0.0%
Probability > 0.85	0.0%	0.5%	0.1%	0.0%
Probability > 0.9	0.0%	0.0%	0.0%	0.0%
Probability > 0.95	0.0%	0.0%	0.0%	0.0%

Table 6-5. MGK estimate results.

6.4.1 MGK validation

The mean values for the MGK conditionally distributed estimate compare well with the declustered sample data, as shown in Table 6-6. The values of individual realizations at the point scale can fall outside the range of the input data, but they do not do so when averaged at the block scale.

	Drilling			MGK Estimate (mean)		
	Min.	**Max.**	**Mean**	**Min.**	**Max.**	**Mean**
Bitumen	0.0019	0.6901	0.1263	0.0305	0.3882	0.1262
Coarse	0.0092	0.8544	0.7115	0.3017	0.8281	0.7082
Fines	0.0008	0.9105	0.1201	0.0146	0.5222	0.1212
Water	0.0045	0.1730	0.0421	0.0084	0.1211	0.0444

Table 6-6. MGK estimate compared to drilling data, original data units.

Swath plots (Figure 6-6) show that the mean of the MGK has performed well globally and reasonably locally, although there appears to be a slightly high bias for water between 170 - 210mRL. Unlike the directly kriged estimates for the logratio transformed variables (Chapter 5), there is no significant bias across the entire domain for the coarse and fines fractions (compare Figure 6-6 with Figure 5-8).

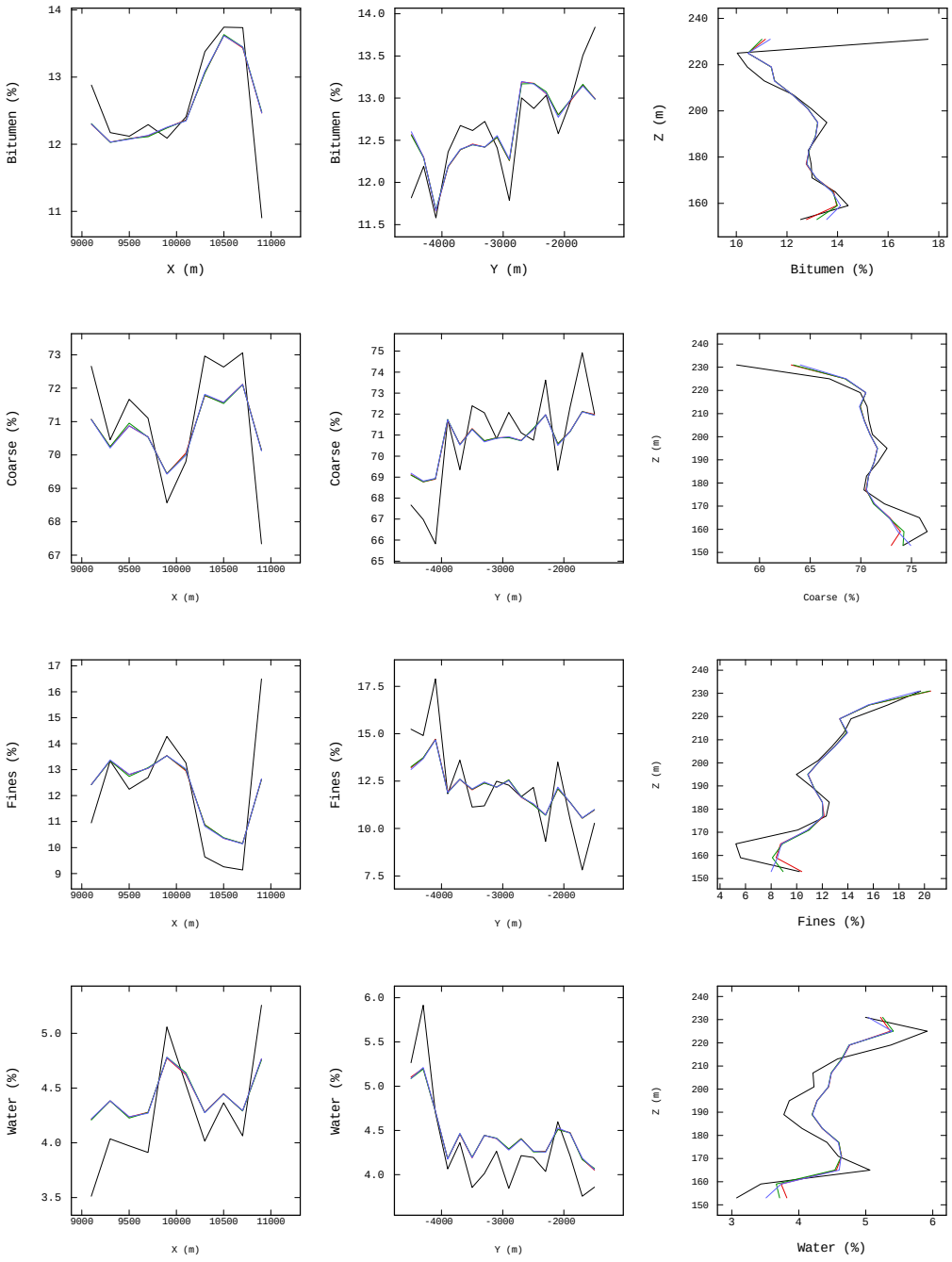

Figure 6-6. Swath plots for MGK alr estimates vs. drilling (easting– left, northing – middle, RL – right. Colour scheme: black = drilling, red = mean of realizations, green = realization 01, blue = realization 25).

Cross validation was also used to establish if there was any bias. MGK estimates were made at each sample data point using the procedure described above, and scatter plots comparing the mean of the distributions to the true values are shown below in Figure 6-7.

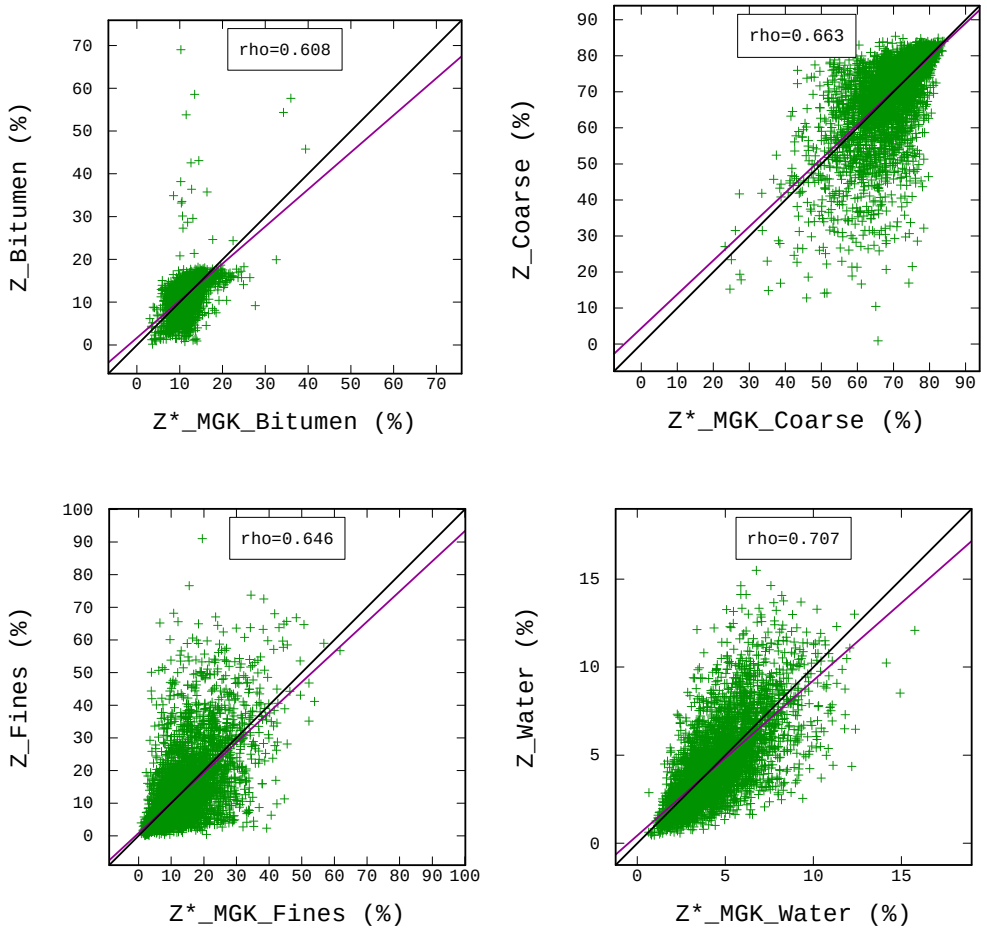

Figure 6-7. Cross-validation scatterplots, bitumen, coarse, fines, water (black line = bisector, magenta line = linear regression).

As for the OK and OCK of the original data units (Section 5.3), the results are reasonable with the exception with the higher-grade bitumen values. Statistics of the errors for the mean MGK values are shown in Table 6-7 and histograms shown in Figure 6-8. The distributions of the errors show that there is no significant global bias, although the relatively poor performance of the water estimate as previously noted for the block estimates is also seen here.

VARIABLE	Minimum	Maximum	Mean	Variance	% Bias	MSE
Z*_MGK_Bitumen_error	-0.5872	0.1847	-6.80E-05	8.58E-04	-0.055%	8.58E-04
Z*_MGK_Coarse_error	-0.3414	0.6479	-1.64E-03	6.63E-03	-0.233%	6.63E-03
Z*_MGK_Fines_error	-0.7152	0.3693	9.68E-04	7.62E-03	0.754%	7.62E-03
Z*_MGK_Water_error	-0.0916	0.0782	7.39E-04	2.60E-04	1.639%	2.61E-04

Table 6-7. Statistics of estimation errors, mean of MGK distribution.

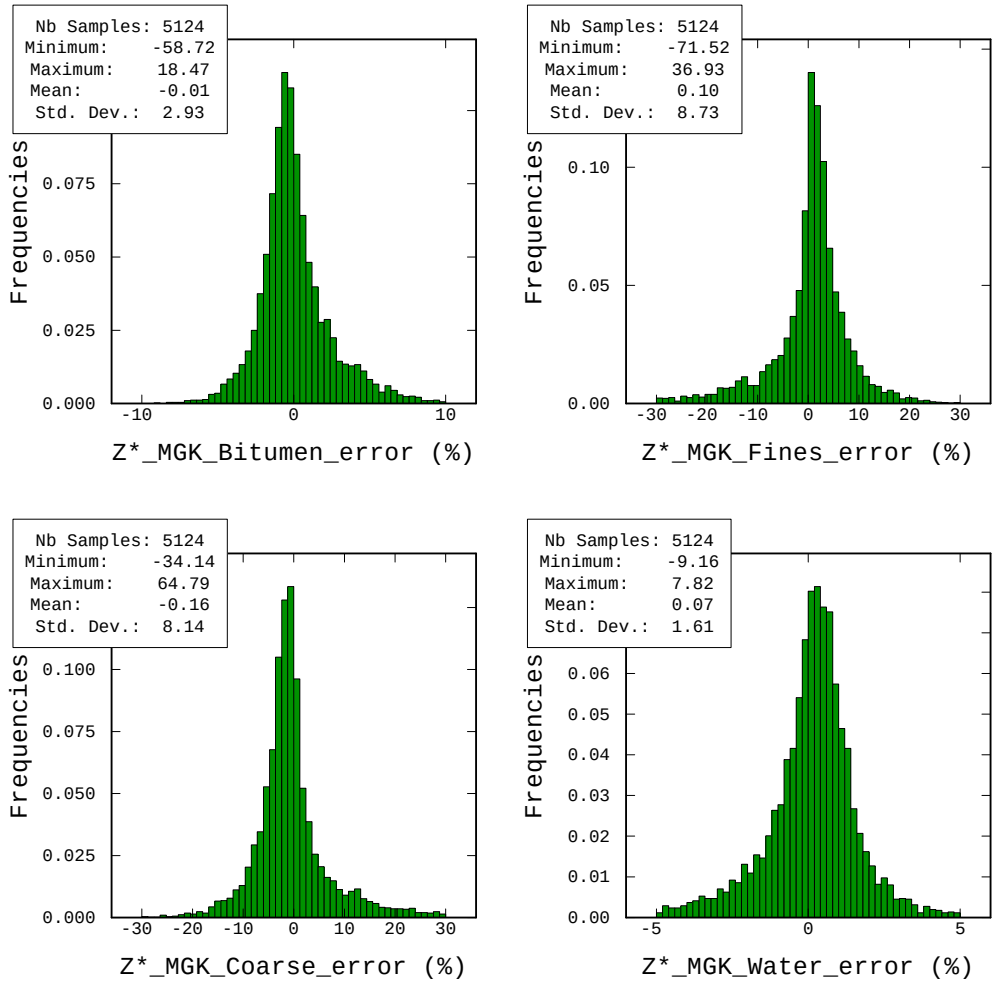

Figure 6-8. Histograms of estimation errors, mean of MGK distribution.

6.5 Concluding Comments for Chapter 6

MultiGaussian kriging as implemented here appears to be able to adequately model the conditional distribution of the logratio transformed data. The performance of MGK is far superior to the direct estimate and back-transform for the logratio values, which produces a bias. MGK, by contrast, produces results that are not systematically biased.

The quantiles of the logratio distributions are passed through the normal scores transform and back via the inverse of the normal score and logratio transformations into original data units. It is only when the distribution of the components has been finally back-transformed into original data units that averaging can occur to produce non-biased results.

MGK results in a much richer model than OCK or RCK, as the probabilities to be above (or below) any particular cut-off can be obtained from the final conditional distribution.

Chapter 7 Conditional Simulation

7.1 Introduction

The performance of conditional simulation for compositional data was compared with and without the logratio transform, as was done for cokriging in Chapter 5. Sequential Gaussian simulation (SGS) is the most widely used form of Gaussian conditional simulation for continuous variables, and has been applied for this study.

7.2 Conditional Simulation without Logratio Transform

To perform a conventional multivariate conditional simulation study without considering the compositional data paradigm, the data in original space was transformed to a Gaussian distribution, using the same grid declustering weights as described in Chapter 6.

7.2.1 Checks for Bivariate Gaussian Distributions

Some of the checks for the bivariate Gaussian distribution discussed in Chapters 3 and 6 were applied - Figure 7-1 shows the vertical H-scatterplot for bitumen, which indicates that the distribution is not bivariate standard normal. The vertical H-scatterplots for the other values have similar results, with moderate correlation coefficients, but a poor measure of deviation value. The bivariate scatterplots between the different components at the same sample location also show that the distribution is very unlikely to be bivariate standard normal.

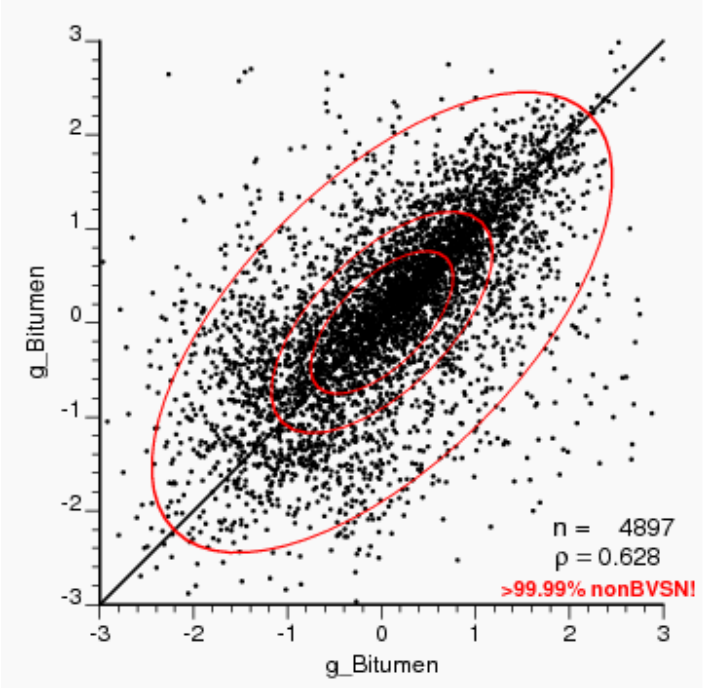

Figure 7-1. H-scatterplot for Gaussian transformed bitumen, vertical direction.

The checks for the ratio between the square root of the variogram and the madogram as described in Section 6.2.2 have been conducted. These checks *appear* to be adequate (visually and numerically), but the lack of a quantitative definition of acceptability means that this check must be deemed as inconclusive. The variogram of order ω checks as described in Section 6.2.3 show that the distribution is not likely to be bivariate Gaussian - Figure 7-2 shows the scatterplot for Gaussian transformed bitumen, and it is clear that the scatter, although linear, is not on the bisector line.

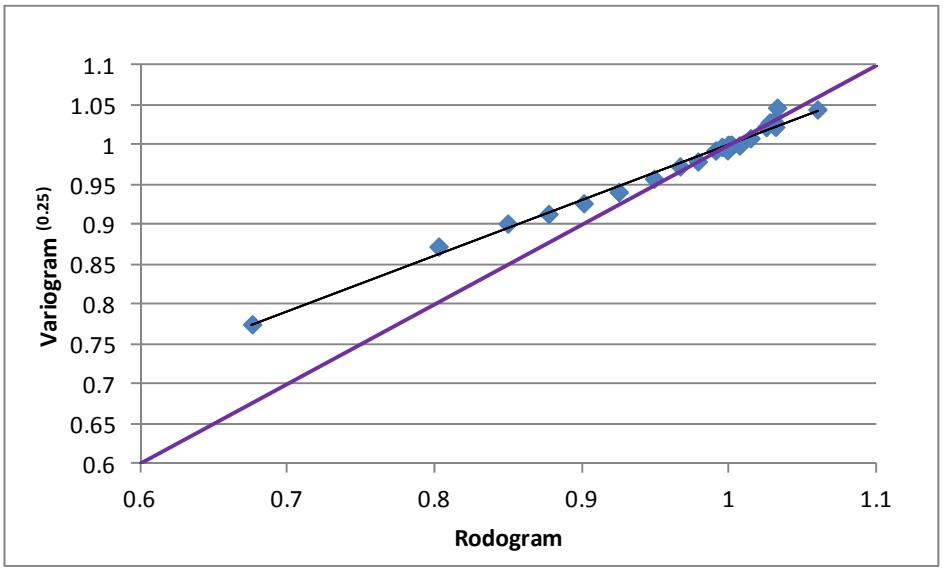

Figure 7-2. Scatterplot, rodogram vs. variogram to power of 0.25, Gaussian-transformed bitumen.

Therefore, the checks on the Gaussian-transformed data from the original data units do not appear to have a bivariate standard normal distribution, similar to the logratio transformed data described in Section 6.2. However, there is no real suitable alternative for a probabilistic method such as conditional simulation for continuous variables, so sequential Gaussian simulation (SGS) was performed for the original data and logratio values after the normal scores transformation.

7.2.2 Conditional Simulation Parameters

The grid used for the conditional simulations was the same as that used MGK (10mE x10mN x1.5mRL, 200 nodes E, 300 nodes W, 44 nodes RL). After testing various search neighbourhoods, twenty five realizations utilizing full cosimulation were run using SGS. The search neighbourhood used 100 nodes E, 100 nodes N, 15 nodes RL, using 20 real data points and 20 simulated points.

The variogram model parameters for the Gaussian-transformed components are shown below in Table 7-1. SK with a constant mean of zero was used and independent random paths were used for each realization.

Cross Variogram Models - Gaussian original units									
Structure	Type	Variance-Covariance matrix					Range (metres) Major	Semi	Minor
1	Nugget		gBitumen	gCoarse	gFines	gWater			
		gBitumen	0.2666	0.1482	-0.1400	-0.0882			
		gCoarse	0.1482	0.2079	-0.1208	-0.0827			
		gFines	-0.1400	-0.1208	0.2179	0.1436			
		gWater	-0.0882	-0.0827	0.1436	0.1746			
2	Spherical		gBitumen	gCoarse	gFines	gWater	200	200	5
		gBitumen	0.4415	0.2432	-0.2948	-0.4132			
		gCoarse	0.2432	0.6818	-0.6161	-0.4297			
		gFines	-0.2948	-0.6161	0.6157	0.3465			
		gWater	-0.4132	-0.4297	0.3465	0.6780			
3	Spherical		gBitumen	gCoarse	gFines	gWater	1300	1300	18
		gBitumen	0.2919	0.1257	-0.2128	-0.1382			
		gCoarse	0.1257	0.1104	-0.1168	-0.1274			
		gFines	-0.2128	-0.1168	0.1664	0.1311			
		gWater	-0.1382	-0.1274	0.1311	0.1474			

Table 7-1. Variogram model parameters for Gaussian transformed original data, without logratio transform.

7.2.3 Validation Checks - Basic Statistics in Gaussian Space

The variances for the realizations were slightly below one (overall about 0.95, although some variables had a mean variance of 0.92). Nowak and Verly (2004) suggest that to bring the variance of the normal scores realizations closer to one, the sill of the variogram model can be rescaled upwards. To test the rescaled variogram method, 25 realizations were re-run with the variance-covariance values rescaled by a factor of 1.1 (variogram model ranges unchanged).

The statistics in Table 7-2 show that the resulting variances of the realizations are close to one. It was noted, however, that the means did *not* approximate zero – Nowak and Verly (2004) propose a number of reasons for this phenomenon, such as the influence of large portions of the models that are well away from conditioning data, or incorrect declustering parameters used during the normal scores transform. This is not the case here, as conditioning data extends throughout the domain being simulated, and the declustering has been thoroughly studied (although the weights chosen are, by nature, subjective). Nowak and Verly (2004) suggest that in this case the simulated values should be adjusted, using a 'progressive correction that depends on the distance of the simulated node from the conditioning data'.

This adjustment involves the calculation of a 'maximum' adjustment factor that is then scaled by the ratio of standard deviations for each node over the maximum standard deviation for all nodes (see Nowak and Verly, 2004 p.393 for full details). This adjustment was applied to the normal scores values for the simulated data set, with the results shown in Table 7-3.

	Transformed Drilling				Simulations			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Var.**
gBitumen	-3.3968	3.6148	6.70E-05	0.9997	-5.5516	5.5190	0.1056	1.0138
gCoarse	-3.2140	3.7223	1.53E-04	0.9993	-5.4269	5.6719	0.0516	1.0052
gFines	-3.5166	3.2064	-1.31E-04	0.9992	-5.5582	5.4848	-0.0758	1.0081
gWater	-3.6614	3.4212	-4.70E-05	0.9996	-5.3834	5.4582	-0.0678	1.0032

Table 7-2. Statistics, drilling vs. realizations, Gaussian transformed original unit components.

	Transformed Drilling				Simulations			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Var.**
gBitumen	-3.3968	3.6148	6.70E-05	0.9997	-5.6042	5.4131	0.0531	1.0071
gCoarse	-3.2140	3.7223	1.53E-04	0.9993	-5.4439	5.6384	0.0271	1.0029
gFines	-3.5166	3.2064	-1.31E-04	0.9992	-5.4893	5.5168	-0.0376	1.0055
gWater	-3.6614	3.4212	-4.70E-05	0.9996	-5.3462	5.5058	-0.0347	1.0022

Table 7-3. Statistics, drilling vs. realizations, Gaussian transformed original unit components with adjustment.

It can be seen that the adjusted means are closer to zero than the non-adjusted means, by about half the difference from zero, and the variances are also closer to one.

There are numerous possible underlying reasons why the normal scores simulated data do not follow the expected distribution – the presence of trends, incorrect assumptions about the declustering and therefore the distribution of the sample data, spatial distribution of the data, modelling of variograms, choice of search neighbourhoods, and importantly, the violation of the multiGaussian assumptions. Ideally, the adjustments described above (including the re-scaling of the variogram sills) would not be needed; however, the practice of geostatistical simulation is often more complex than the simple random function models used by the technique.

7.2.4 Validation Checks -Basic Statistics in Original Units

The basic statistics of the final back-transformed original unit data were very well reproduced (see Table 7-4). The minima and maxima for the realizations are the same as the input data, and the histograms for the 25 realizations were similar to the input data. Figure 7-3 shows an example for bitumen – histograms for the other components show similar reproduction of declustered input data.

	Declustered Drilling				Original Unit Realizations			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Var.**
Bitumen	0.0019	0.6901	0.1238	0.0019	0.0019	0.6901	0.1254	0.0017
Coarse	0.0092	0.8544	0.7029	0.0140	0.0092	0.8544	0.7057	0.0133
Fines	0.0008	0.9105	0.1283	0.0145	0.0008	0.9105	0.1250	0.0140
Water	0.0045	0.1730	0.0451	0.0006	0.0045	0.1730	0.0439	0.0005

Table 7-4. Drilling vs. realizations basic statistics, adjusted original units.

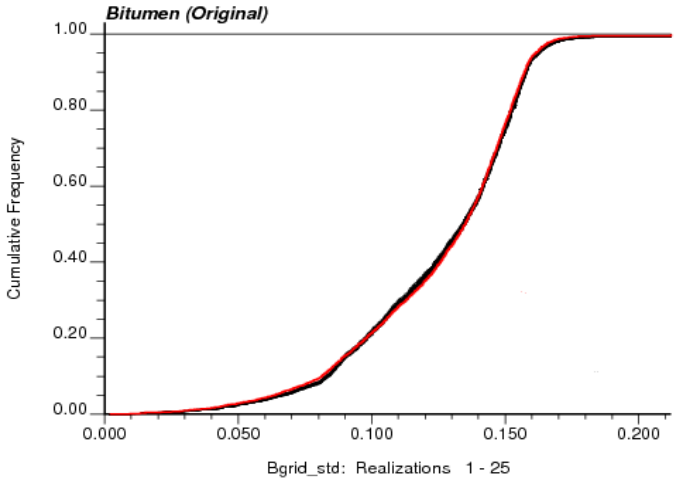

Figure 7-3. Histogram reproduction, bitumen (original units). (Colour scheme: black = realizations, red = drilling data).

Each component, when viewed in isolation, therefore appears to meet the basic criterion of statistical reproduction. However, when the variables for each realization are added together, the sum of the components is not one. Figure 7-4 shows the histogram of the summed components for a realization selected at random (#19) – clearly, the requirement for the constant sum of one has not been met for conditional cosimulation – even though the mean of the distribution is 1, the sums range from 0.5 to 1.5. All of the other realizations show a similar pattern

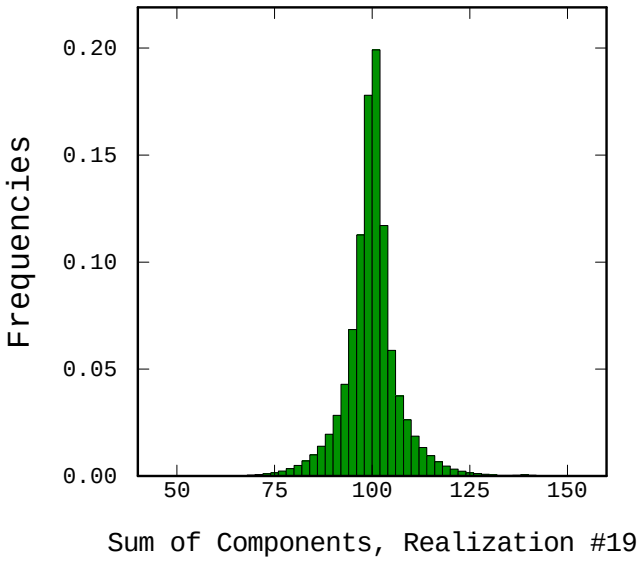

Figure 7-4. Histogram of summed components, original data units, realization #19.

A post-processing normalization of the components was performed – the proportions for each component were simply rescaled using the closure operation so that they summed to one. The basic statistics for these data are shown in Table 7-5 - it results in the ranges of the simulated data being greater than the input data, but the means and variances appear relatively non-biased comparable to the non-normalized data.

Therefore, normalization of the conditional simulations will result in biases to the tails of the distributions, as was the case for the kriged estimates discussed in Section 5.3.

	Declustered Drilling				Raw - Normalised			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Std. Dev.**
Bitumen	0.0019	0.6901	0.1238	0.0019	0.0012	0.7094	0.1273	0.0013
Coarse	0.0092	0.8544	0.7029	0.0140	0.0078	0.9185	0.7112	0.0111
Fines	0.0008	0.9105	0.1283	0.0145	0.0005	0.9158	0.1184	0.0119
Water	0.0045	0.1730	0.0451	0.0006	0.0029	0.2959	0.0431	0.0005

Table 7-5. Drilling vs. realizations basic statistics, normalized original units.

7.2.5 Validation Checks – Variograms in Gaussian Space

Comparison of the normal scores cross-variogram models and the variograms for the realizations (still in Gaussian space) shows that the reproduction of the variogram is good. Figure 7-5 shows the checks in the horizontal direction - note that the model shown here does not have the variances and covariances adjusted by a factor of 1.1. The cross-variograms for the vertical direction perform adequately, although the realizations have a slightly longer range and lower variance than the input model.

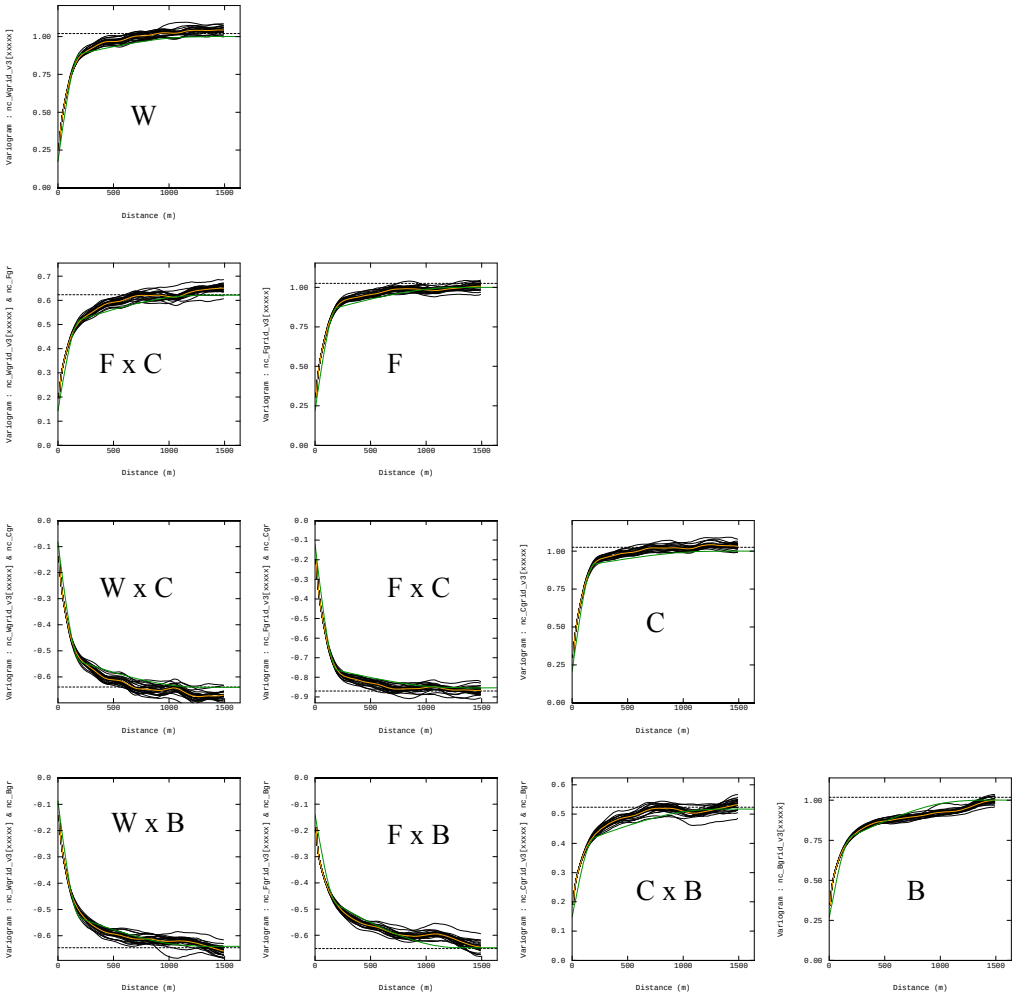

Figure 7-5. Normal scores cross-variogram validation (green = model, orange = mean of realizations).

7.2.6 Validation Checks – Scatterplots in Original Units

Comparison of bivariate scatterplots between the input data and back-transformed realization data is another useful validation step. Figure 7-6 shows scatterplots for bitumen and coarse – the grade ranges have been discretized into 100 bins, and coloured by the count of pairs per discretised bin. The legends have been scaled according to the number of data points in the drilling and for the realizations.

The upper left chart shows the drilling with the constraint. The upper right chart shows the randomly selected SGS realization (#19), which is not constrained, and the sum of the bitumen and coarse alone often exceeds the constant sum. The normalized realization (lower right chart) *appears* a better approximation, showing the same constraint as the input data.

However, normalization introduces a bias to the tails of the distribution, as previously discussed. To therefore produce a set of realizations that honour the spatial distributions of a compositional data set and meet the constant sum requirement, the logratio approach will next be demonstrated.

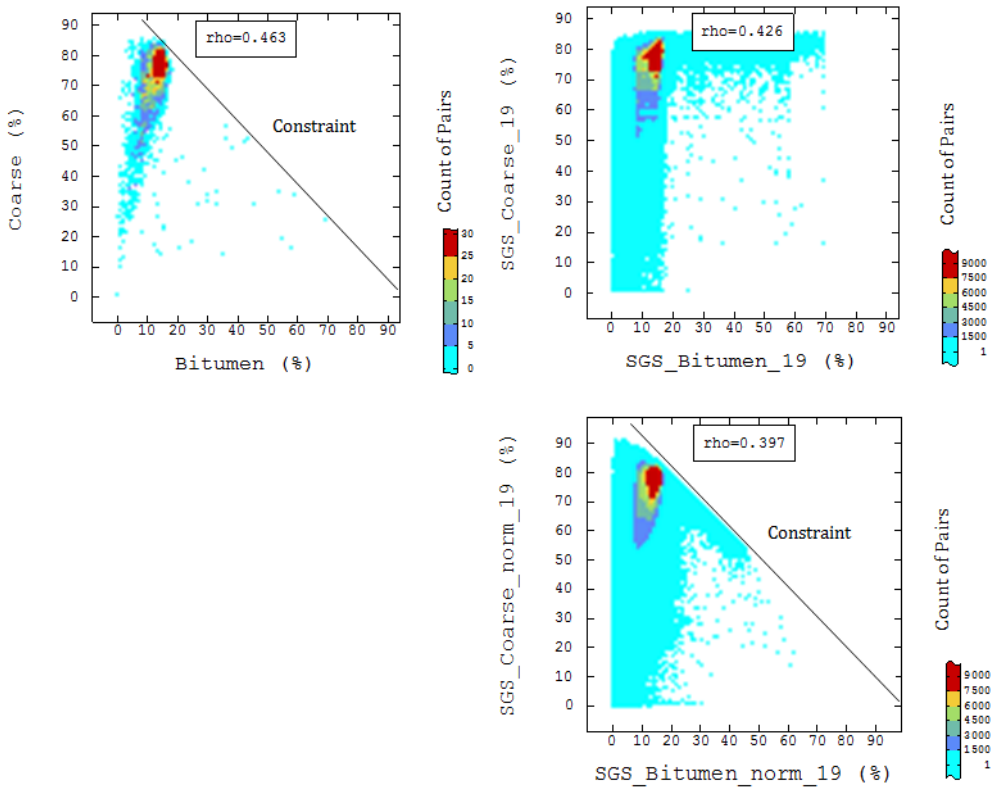

Figure 7-6. Scatterplots for bitumen (x-axis) and coarse (y-axis). Upper left = drilling, upper right = realization 19, lower right = normalized realization #19.

7.3 Conditional Simulation with Logratio Transform

A summary of the transformations and procedure for the conditional simulations using the logratio transform is shown below (Figure 7-7), similar to the methodology of Boisvert et al. (2009). The logratio transform, normal scores transform and experimental/model variograms are exactly the same as those used for MGK (see Chapter 6).

Twenty-five realizations were generated for both the alr and clr components, using SGS and the same search parameters as described for non-logratio transformed components as described in Section 7.2.2.

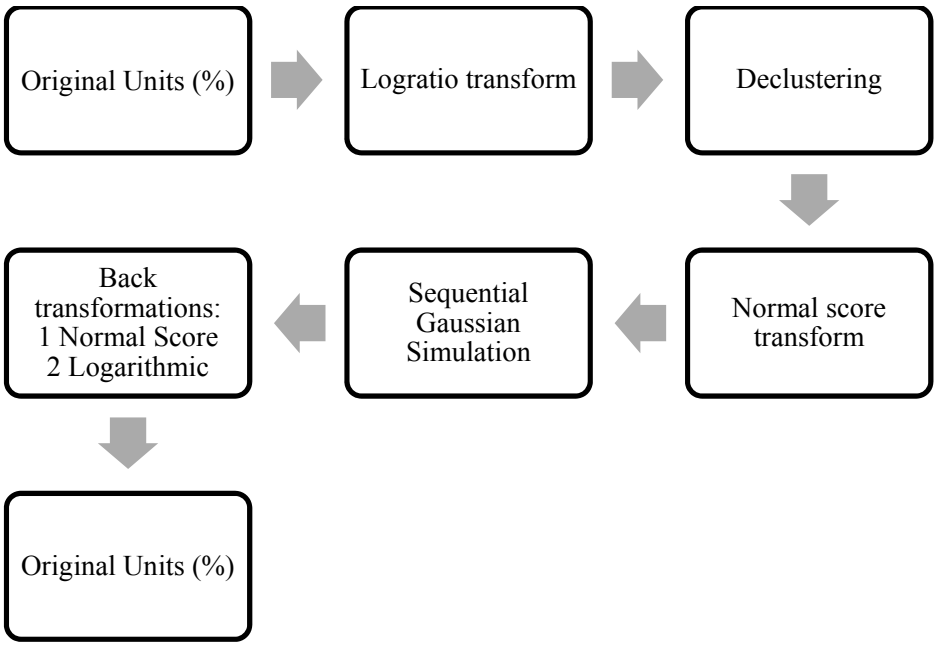

Figure 7-7. Generalized process flow (modified after Boisvert et al., 2009).

7.3.1 Validation Checks -Basic Statistics in Gaussian Space

The variances of the realizations were lower than those of the input data, therefore rescaling of the variogram model variances and covariances by a factor of 1.1 was again applied. The statistics of the realizations show that the rescaling of the input variogram results in a variance usually slightly greater than one for the logratio variables (Table 7-6), and the means are closer to zero than the previous run, although insignificantly.

	Transformed Drilling				Realizations			
VARIABLE	Min.	Max.	Mean	Var.	Min.	Max.	Mean	Var.
galrB:C	-3.2848	4.0861	0.0001	0.9992	-5.5971	5.4105	0.0116	1.0346
galrF:C	-3.4524	3.5083	0.0000	0.9992	-5.4666	6.0496	0.0072	0.9825
galrW:C	-3.8251	3.5083	0.0000	0.9996	-5.4512	5.2579	-0.0095	0.9920
gclrB	-3.2871	3.4729	0.0001	0.9986	-5.6663	5.7632	-0.0033	1.0311
gclrC	-3.5053	3.4282	0.0000	0.9990	-5.6324	5.5860	-0.0066	1.0239
gclrF	-3.4297	3.4569	0.0000	0.9990	-5.7112	5.8119	0.0083	1.0372
gclrW	-3.819	3.5161	0.0000	0.9996	-6.4002	5.5912	-0.0072	1.0339

Table 7-6. Normal scores statistics, drilling vs. realizations, logratio variables.

7.3.2 Validation Checks - Basic Statistics in Logratio Space

The normal scores realizations were back-transformed into the logratio values - the basic statistics were very well reproduced (see Table 7-7 and Figure 7-8 for an example histogram).

	Transformed Drilling				Realizations			
VARIABLE	Min.	Max.	Mean	Var.	Min.	Max.	Mean	Var.
alrB:C	-4.9375	1.2266	-1.7628	0.0783	-4.9376	1.2266	-1.7605	0.0812
alrF:C	-6.9411	4.5948	-2.1485	1.1080	-6.9412	4.5949	-2.1600	1.1522
alrW:C	-5.1957	2.0616	-2.9438	0.4303	-5.1958	2.0617	-2.9175	0.4372
clrB	-3.6411	2.0187	-0.0490	0.2495	-3.6412	2.0187	-0.0527	0.2772
clrC	-1.6292	3.2221	1.7138	0.1495	-1.6292	3.2222	1.7094	0.1597
clrF	-3.7189	2.9656	-0.4347	0.4811	-3.7190	2.9657	-0.4536	0.5297
clrW	-2.6068	0.4325	-1.2300	0.1388	-2.6069	0.4325	-1.2063	0.1452

Table 7-7: Drilling vs. realizations basic statistics, logratio space.

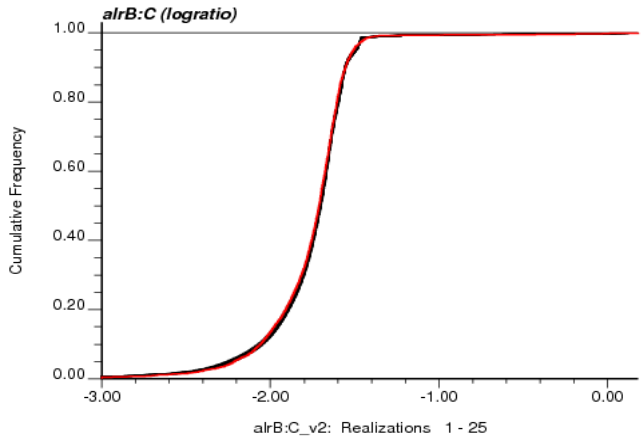

Figure 7-8. Histogram reproduction, alrB:C. (Colour scheme: black = realizations, red = drilling data).

7.3.3 Validation Checks - Basic Statistics in Original Units

The logratio variables were transformed back into the original data units. Comparison of these with the raw input data shows that, similar to the kriging (see Chapter 6), the means and variances of the individual components have been adequately reproduced, but the ranges of the simulations exceed those of the input data for all the variables, with the exception of bitumen for the clr method, which is significantly lower (see Table 7-8). Figure 7-9 shows the histogram of the realizations compared to the input data for bitumen via the alr method – these results show that the realizations do not exactly reproduce the original input data. The sum of the components at each node is correct, summing to unity.

	Declustered Drilling				alr Realizations			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Var.**
Bitumen	0.0019	0.6901	0.1238	0.0019	0.0001	0.7720	0.1263	0.0016
Coarse	0.0092	0.8544	0.7029	0.0140	0.0093	0.9650	0.7086	0.0126
Fines	0.0008	0.9105	0.1283	0.0145	0.0002	0.9890	0.1209	0.0145
Water	0.0045	0.1730	0.0451	0.0006	0.0006	0.8675	0.0442	0.0009

	Declustered Drilling				clr - Realizations			
VARIABLE	**Min.**	**Max.**	**Mean**	**Var.**	**Min.**	**Max.**	**Mean**	**Var.**
Bitumen	0.0019	0.6901	0.1238	0.0019	0.0012	0.5650	0.1260	0.0012
Coarse	0.0092	0.8544	0.7029	0.0140	0.0093	0.9241	0.7084	0.0124
Fines	0.0008	0.9105	0.1283	0.0145	0.0007	0.9771	0.1217	0.0149
Water	0.0045	0.1730	0.0451	0.0006	0.0022	0.4275	0.0439	0.0007

Table 7-8: Drilling and realizations basic statistics, final logratio back-transforms (alr top, clr bottom).

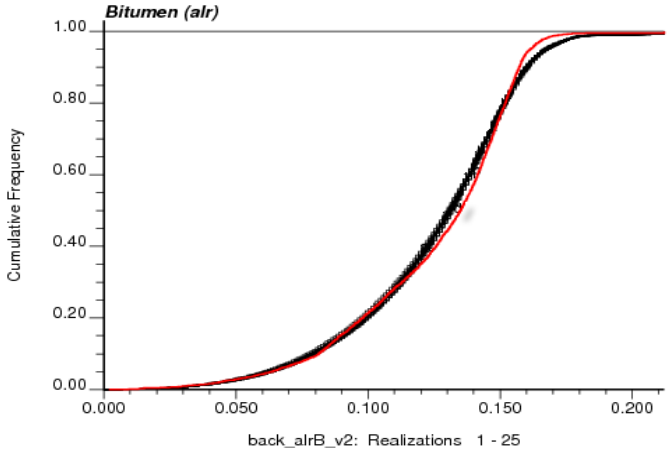

Figure 7-9. Histogram reproduction, back-transformed bitumen via alr. (Colour scheme: black = realizations, red = drilling data).

7.3.4 Validation Checks – Variograms in Gaussian Space

Figure 7-10 shows the normal scores cross-variogram checks for the alr transformed data in the horizontal direction, which indicates good to reasonable variogram reproduction of the original variogram models, but not of the rescaled variograms. The realization variograms in the vertical direction typically show a slightly longer range and lower variance compared to the input model.

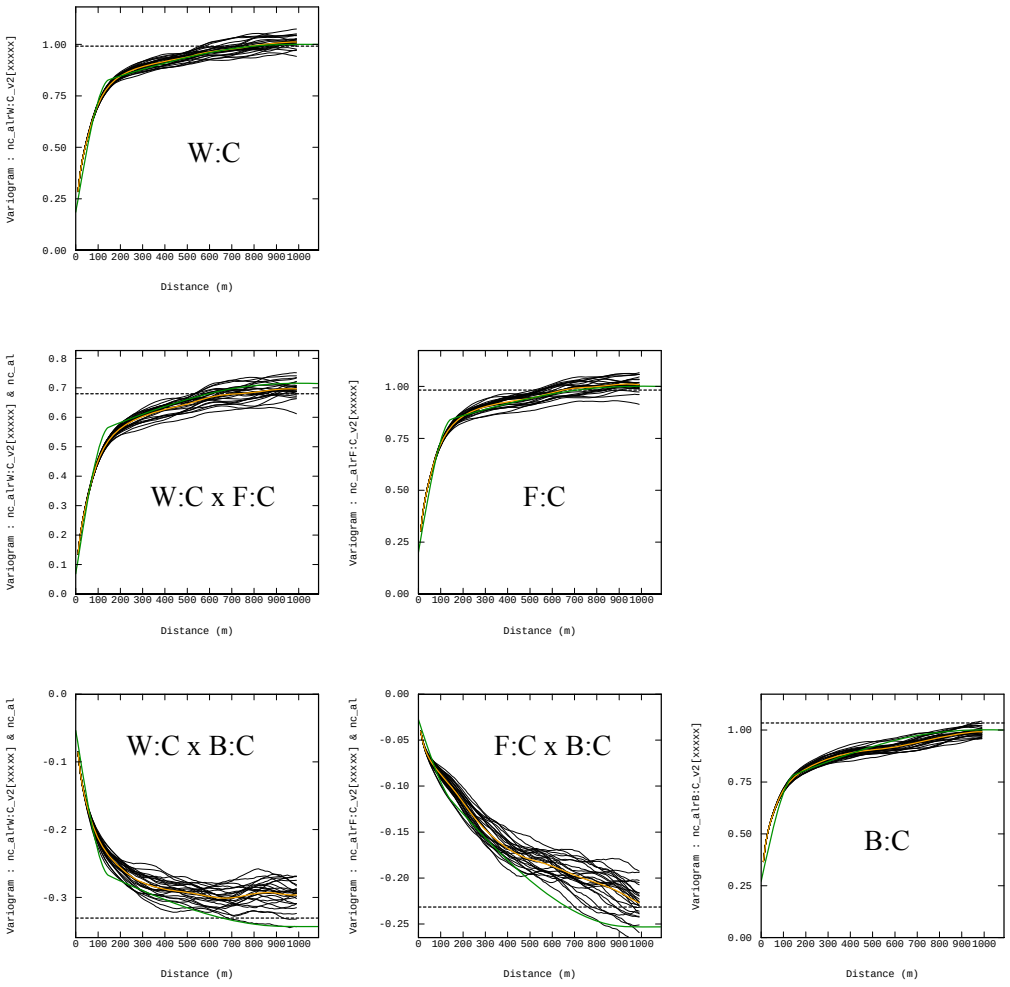

Figure 7-10. Normal scores cross-variogram validation for alr method, horizontal direction (green = model, orange = mean of realizations).

The clr method variogram reproduction is not as convincing when compared to the non-rescaled variogram, since the realizations have significantly shorter ranges than the model. However, if the realizations are compared to the rescaled variograms, the comparison is good up to about 80% of the range, but they do not reach the rescaled sill of 1.1. Figure 7-11 below shows the realizations compared to the original variogram model (left) and to the rescaled variogram model (right) for the clr values. In summary,

the rescaling results in the realizations not matching the input variogram, but results in the variances matching the normal score variance.

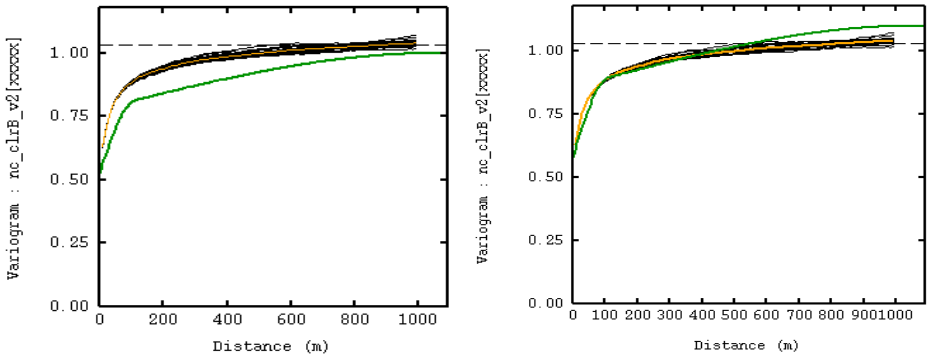

Figure 7-11. Normal scores cross-variogram validation for clr method, horizontal direction, bitumen only. (green = model, orange = mean of realizations). Left = original variogram model, right = rescaled variogram model.

7.3.5 Validation Checks – Scatterplots in Original Units

Figure 7-12 shows the scatterplot of bitumen vs. coarse for realization #19 for the alr and clr methods – the colour scales are the count of pairs, as described for Figure 7-6. Both show the same constraint as the input data, although the truncated maximum bitumen value for the clr method results in less dispersion and a higher correlation coefficient. The scatterplot comparison is subjective, but it is clear that the alr method results are more consistent with the input data than the clr method.

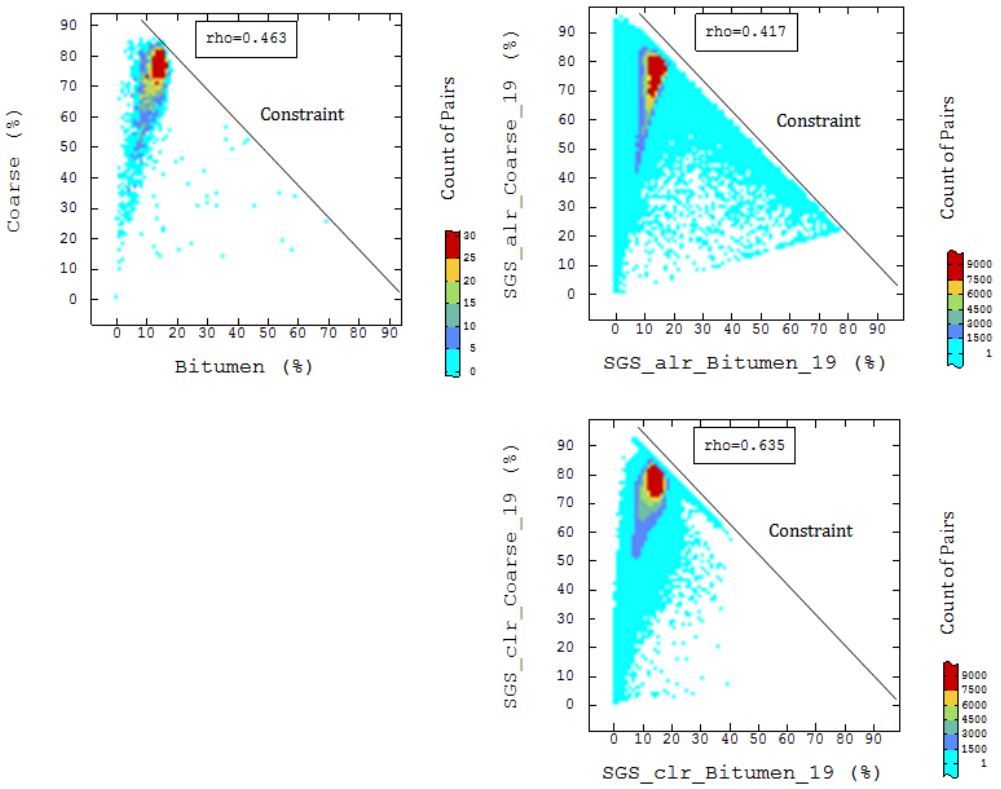

Figure 7-12. Scatterplots for bitumen (x-axis) and coarse (y-axis). Upper left = drilling, upper right = alr realization 19, lower right = clr realization 19.

7.3.6 Validation Checks - Trend Analysis in Original Units

Swath plots comparing the drilling and the means of the realizations for both logratio methods are shown in Figure 7-13. Both methods show reasonable comparison in the horizontal directions, but the alr method is much closer to the drilling data than the clr method for the vertical direction.

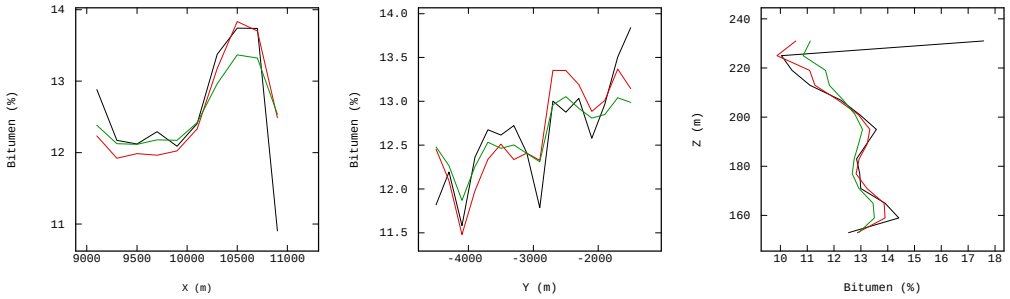

Figure 7-13. Swath plots for mean of bitumen realizations (left = Easting, middle = Northing, right = RL. Colour scheme: black = drilling, red = alr, green = clr).

7.3.7 Validation Checks - Comparison with multiGaussian kriging

A comparison was made (in original data units for the alr transform method) between the mean of the MGK estimate with the mean of the conditional simulation realizations, and for the proportions of the model above certain cut-offs The results shown in Table 7-9 are remarkably similar, and confirm that MGK is a robust method for working with compositional data.

	multiGaussian kriging				Conditional Simulation			
	Bitumen	**Coarse**	**Fines**	**Water**	**Bitumen**	**Coarse**	**Fines**	**Water**
Mean Value	0.1262	0.7082	0.1212	0.0444	0.1263	0.7086	0.1209	0.0442
Probability > 0.05	97.1%	100.0%	65.7%	27.9%	97.3%	100.0%	65.7%	27.5%
Probability > 0.1	79.2%	99.9%	39.8%	4.7%	79.3%	99.9%	39.7%	4.7%
Probability > 0.15	22.6%	99.9%	26.5%	1.0%	22.4%	99.9%	26.3%	1.0%
Probability > 0.2	0.8%	99.8%	18.2%	0.3%	0.9%	99.8%	18.1%	0.3%
Probability > 0.25	0.6%	99.6%	13.0%	0.1%	0.6%	99.6%	12.9%	0.1%
Probability > 0.3	0.4%	99.1%	9.3%	0.1%	0.4%	99.1%	9.2%	0.1%
Probability > 0.35	0.3%	98.4%	6.5%	0.0%	0.4%	98.4%	6.4%	0.0%
Probability > 0.4	0.3%	97.4%	4.4%	0.0%	0.3%	97.4%	4.3%	0.0%
Probability > 0.45	0.2%	95.7%	2.8%	0.0%	0.2%	95.8%	2.8%	0.0%
Probability > 0.5	0.2%	93.3%	1.8%	0.0%	0.2%	93.4%	1.8%	0.0%
Probability > 0.55	0.1%	90.0%	1.2%	0.0%	0.1%	90.1%	1.2%	0.0%
Probability > 0.6	0.1%	85.4%	0.7%	0.0%	0.1%	85.5%	0.7%	0.0%
Probability > 0.65	0.0%	78.7%	0.4%	0.0%	0.0%	78.8%	0.4%	0.0%
Probability > 0.7	0.0%	67.6%	0.1%	0.0%	0.0%	67.7%	0.2%	0.0%
Probability > 0.75	0.0%	48.5%	0.1%	0.0%	0.0%	48.6%	0.1%	0.0%
Probability > 0.8	0.0%	12.3%	0.1%	0.0%	0.0%	12.5%	0.1%	0.0%
Probability > 0.85	0.0%	0.5%	0.1%	0.0%	0.0%	0.6%	0.1%	0.0%
Probability > 0.9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Probability > 0.95	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 7-9. Comparison between MGK and conditional simulations, alr transform method.

7.4 Concluding Comments for Chapter 7

Conditional cosimulation of the original data units (after the normal scores transform and back-transform) will result in a set of simulations that are valid for the components individually, but the constant sum constraint is not honoured. Normalization of the simulated data, as with the kriged estimates, will result in bias of the tails of the component distributions.

Conditional cosimulation using logratio transforms of compositional data is a promising method. The use of the normal scores transform and back-transform of the quantiles overcomes the inherent bias in estimating logratios directly. The work required before running the simulations is no more onerous than is required for multiGaussian kriging, but the extensive checking and validation required (of which only a small portion is presented here) adds to the workload considerably.

Both the alr and clr methods produce broadly comparable results - on a practical note, the alr method has an advantage, since one less variable than the number of components is produced, and therefore the cross-variogram modelling and statistical checks in normal scores space are less onerous. In addition, the validation checks for the alr method are more convincing than those for the clr method.

Chapter 8 Conclusions and Further Work

8.1 Conclusions

Compositional data is very common in the earth sciences, and in many cases for mineral resource estimation, full or sub-compositional data will be available. The data in a composition must be greater than zero, and sum to a constant. As one component increases, one or more of the other components must decrease – the components are not free to vary independently. Therefore, there must be at least one negative correlation between the components, and there generally is a bias towards negative correlations – it is possible for correlations between the components to thus be spurious.

Any application of standard statistical (including geostatistical) techniques of the data in its raw form can be misleading due to these spurious correlations. Of course, in some instances the negative correlations could have some underlying paragenetic basis (such as mineral replacement), but it may not be possible to separate inferences and results from these genetic relationships and those due to spurious correlations.

For compositional data, the relative magnitudes between the components are important, not the absolute values. The relative magnitudes can be described in terms of ratios, but ratios are difficult to deal with mathematically and statistically. Logarithms of ratios (logratios) have convenient statistical properties, and they also transform the data from the constrained space of the simplex to unconstrained real space. By using logratios, a range of multivariate statistical techniques can be used with compositional data.

It has been shown that directly kriging this logratio-transformed data, with a direct back-transform applied results in estimates that are biased, even though the results are non-negative and meet the constant sum constraint. This is because a linear averaging of logratios, followed by application of the back-transform, will result in values that approach the global standardized geometric mean of the components in original data units, not the arithmetic mean.

In cases where the components are of similar magnitude, or if there is one very dominant component, then these biases may not be noticed from cursory statistical checks. When the components are of different magnitudes, the biases are obvious and easily detected. In summary, to avoid bias, direct kriging of logratios must be rejected from the set of valid geostatistical techniques.

The alternative to direct kriging of logratios is to apply a non-linear approach, in which the conditional distributions of the components are modelled instead of unique values from a linear kriging. It has been shown that both multiGaussian kriging and conditional simulation, where in both cases the logratio values are transformed into Gaussian values, are valid techniques for dealing with compositional data. Averaging of data values that have been subjected to non-linear transformation, and then back-transformed into original data units is incorrect. Only after the distributions have been modelled (via quantiles) through the transformation and back-transformation processes can averaging in original data units can occur.

Spatial modelling of compositional data utilizing logratio transformations is yet to find widespread acceptance in geostatistics. This is mainly due to:

1. The complexity of the numerous transforms (normalizing, logratio, Gaussian);
2. The difficulty of dealing with zeros (which are very common); and
3. The biased results when directly kriging and back-transforming the logratio values.

The approach advocated here results in the highly desirable result of the relative proportions of the components being maintained at unsampled spatial locations for estimation and simulation *without bias*. It is hoped that application of this approach will lead to an increase in the popularity of the use of logratios for geostatistical modelling of compositional data.

8.2 Future Work

Although the logratio transform is central to the compositional data methodology, it might be possible to eliminate this transformation step for MGK or Gaussian conditional simulation. This is because both these methods rely on a Gaussian transform of the input values, and consequently a direct Gaussian transform of the ratios could be considered instead of log transformations. It is recognized that the Gaussian transformation does not have many of the 'desirable' properties of logarithmic transforms, such as the relationship between the variances of the logratios shown in Equation 1-1. However, since the logarithmic transform of ratios was designed to release compositional data from the simplex, and into 'unconstrained' space, then the Gaussian transform is also capable of this.

The MGK and conditional simulation procedures described rely on the assumptions of a multivariate Gaussian distribution. However, when it is difficult to show that the data even conforms to a bivariate Gaussian distribution, then this assumption may be flawed, and alternative techniques may be required. It is possible that other non-linear techniques such as indicator methods may be used in this case, although accounting for correlations between variables with these methods might be difficult. In addition, the implications of using multiGaussian methods when the data does not strictly conform to this model would be an area well worth investigating.

The MGK as applied to the logratio transformed data in this study is performed on the 'point-scale' data, and then 'blocked-up' to the block-scale. Implementation of MGK for compositional data at the block-scale selective mining unit is recommended.

Several approaches were made in this work to considering how acceptable the assumption of the multivariate Gaussian distribution following a standard normal scores transformation is. It is noted that these approaches are likely to reject many practical mining data sets. There is probably a use for more flexible testing approaches that can exclude serious departures but give criteria where some (less departing) data sets can be accepted as 'fit for purpose'. This could be a pragmatic avenue for further research.

To avoid multivariate Gaussian distribution assumptions, the stepwise conditional transformation (SCT, Leuangthong and Deutsch, 2003) could be applied. This transformation ensures that the data are multivariate Gaussian with zero correlation, with the correlation between the variables captured in the transformation and back-transformation.

Another technique that could be used to transform the data to a multivariate Gaussian distribution is the projection pursuit multivariate transform (PPMT, Barnett et al., 2012). "The idea is to identify linear projection vectors in the data that are the most complex (non-Gaussian) – once these projection vectors have been determined, the individual high dimension points are transformed to normalize the projection ('Gaussianize'). By iterating this 'identify and Gaussianize' process, the high dimensional data is gradually transformed to a multivariate Gaussian distribution" (Barnett et al, 2012).

The alr transformed data was passed through the PPMT algorithm – bivariate scatterplots are shown in Figure 8-1. These transformed variables are highly bivariate Gaussian, in contrast to the basic normal-scores transformed data shown in Figure 6-2. In addition the correlation between the variables has been removed – for MGK, this would eliminate the use of the Cholesky decomposition to create the random but correlated values required to discretize the cdf.

After the MGK or conditional simulation approach described previously, the multivariate standard normal transformation (MSNT, Deutsch, 2011) back-transform can then be applied. This back-transformation uses an interpolation method to weight the estimates based on their proximity to the observed data – essentially a type of inverse distance weighting.

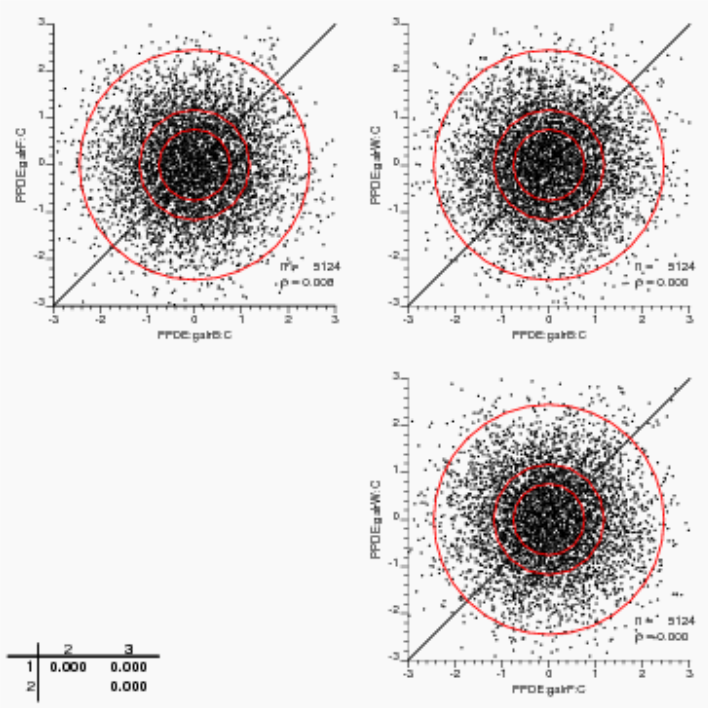

Figure 8-1. Bivariate scatterplots for PPMT alr values.

Therefore, there are techniques under development that address the non-Gaussian behavior of data when Gaussian-based geostatistical algorithms provide a convenient and tractable solution.

An interesting practical application for further compositional geostatistical investigation might be an iron ore deposit that has eight or more components. It is possible that there are only three of the components are of interest for a particular study (perhaps Fe, Al_2O_3 and SiO_2 for product characterization). If so, then the subcomposition of these three variables can be taken, and the multiGaussian kriging or conditional simulation procedure applied.

Bibliography

Aitchison J., 1982, The statistical analysis of compositional data, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 44, No. 2, pp. 139 - 177.

Aitchison, J., 1986, The statistical analysis of compositional data, Chapman and Hall, London.

Aitchison, J., 1999, Logratios and natural laws in compositional data analysis, Mathematical Geology, Vol. 31, No. 5, pp. 563 - 580.

Aitchison, J., 2003, A concise guide to compositional data analysis, Short course notes, CoDaWork '03.

Aitchison, J., Barcelo-Vidal, C. and Pawlowsky-Glahn, V., 2002, Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis, Archaeometry, Vol. 44, No. 2, pp. 295 - 304.

Aitchison, J. and Egozcue, J.J., 2005, Compositional data: where are we and where should we be heading?, Mathematical Geology, Vol. 37, No. 7, pp. 829 - 850.

Aitchison, J. and Kay, J.W., 2003, Possible solutions of some essential zero problems in compositional data analysis, in Compositional data analysis workshop – CoDaWork '03, eds. Thio-Henestrosa, S. and Martin-Fernandez, J.A.

Atchley, W.R., Gaskins, C.T. and Anderson, D., 1976, Statistical properties of ratios. I. Empirical results, Systematic Zoology, Vol. 25, pp. 137 - 148.

Barnett, R.M., Manchuk, J.G. and Deutsch, C.V., 2012, Projection pursuit multivariate transform, fourteenth annual report of the Centre for Computational Geostatistics, University of Alberta, pp. 103-1 – 103-18.

Baxter, M.J., Beardah, C.C, Cool, H.E.M. and Jackson, C.M., 2005, Compositional data analysis of some alkaline glasses, Mathematical Geology, Vol. 37, No. 2, pp. 183 - 196.

Boezio, M., Costa, J. and Koppe, J., 2011, Ordinary cokriging of additive log-ratios for estimating grades in iron ore deposits. CoDaWork 2011 (Proceedings of the 4th international workshop on compositional data analysis), eds. Egozcue, J., Tolosana-Delgado, R. and Ortego, M.

Boisvert, J.B., Rossi, M.E. and Deutsch, C.V., 2009, Multivariate geostatistical simulation of proportions and nonadditive geometallurgical variables, eleventh annual report of the Centre for Computational Geostatistics, University of Alberta, pp. 303-1 – 303-8.

CAPP (Canadian Association of Petroleum Producers), 2010, Crude oil: Forecast, Markets & Pipelines. June 2010. (http://www.capp.ca/getdoc.aspx?DocId=173003)

Chayes, F., 1960, On correlation between variables of constant sum, Journal of Geophysical Research, Vol. 65, No. 12, pp. 4185 – 4193.

Chiles, J-P. and Delfiner, P., 1999, Geostatistics: modeling spatial uncertainty, John Wiley & Sons, New York.

Clark, I., 1986, The art of cross validation in geostatistical applications, Proceedings of APCOM 19, Society for Mining Metallurgy and Exploration, Inc. Littleton, Colorado, pp. 211 – 220.

David, M., 1977, Geostatistical ore reserve estimation: Developments in Geomathematics 2. Elsevier, Amsterdam.

Davis, B.M, 1987, Uses and abuses of cross-validation in geostatistics, Mathematical Geology, Vil. 19, No. 3, pp. 241 – 248.

Davis, J.C., 1973, Statistics and Data Analysis in Geology, John Wiley and Sons, New York.

Deutsch, C. V., 1989, DECLUS: A Fortran 77 program for determining optimum spatial declustering weights. Computers & Geosciences, Vol. 15, No. 3, pp. 325 – 332.

Deutsch, C.V., 2011, Multivariate standard normal transformation, thirteenth annual report of the Centre for Computational Geostatistics, University of Alberta, paper 101.

Deutsch, C.V. and Journel, A.G., 1998, GSLIB Geostatistical software library and user's guide, second edition. Oxford University Press, New York.

Deutsch, J.L. and Deutsch C.V., 2011, Plotting and checking the bivariate distributions of multiple Gaussian data, Computers and Geosciences, Vol. 37, pp. 1677 – 1684.

Devenny, D.W., 2010, Oil sands tailings technologies and practices, Alberta Energy Research Institute. http://eipa.alberta.ca/media/40994/report%20b%20%20integrated%20oil%20sands%20tailings%20treatment%20technologies%20march%202010.pdf

Emery, X., 2005a, Variograms of order ω: A tool to validate a bivariate distribution model, Mathematical Geology, Vol. 37, No. 2, pp. 163 - 181.

Emery, X., 2005b, Simple and ordinary multiGaussian kriging for estimating recoverable reserves. Mathematical Geology, Vol. 37, No.3, pp. 295 - 319.

Emery, X., 2006, MultiGaussian kriging for point-support estimation: incorporating constraints on the sum of the kriging weights. http://captura.uchile.cl/jspui/bitstream/2250/5661/1/Emery_Xavier_Multigaussian.pdf

Emery, X. and Ortiz, J., 2005, Estimation of mineral resources using grade domains: critical analysis and a suggested methodology, The Journal of The South African Institute of Mining and Metallurgy, Vol. 105, pp. 247 - 255.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis, Mathematical Geology, Vol. 35, No. 3, pp. 279 - 300.

Goovaerts, P., 1998, Ordinary cokriging revisited, Mathematical Geology, Vol. 30, No. 1, pp. 21 - 42.

Hennessy, J.A., 1990, Tar sands exploration and geology, surface mining (2nd Edition), in; Kennedy, B.E., (ed.), Society for Mining Metallurgy and Exploration, Inc. Littleton, Colorado.

Isaaks, E.H., 1990, The application of Monte Carlo methods to the analysis of spatially correlated data, PhD thesis, Stanford University.

Isaaks, E.H., and Srivastava, R.M., 1989, Applied geostatistics, Oxford University Press, New York.

Journel, A. G., 1974, Geostatistics for conditional simulation of ore bodies, Economic Geology, Vol. 69, pp. 673 - 687.

Journel, A. G., 1983, Nonparametric estimation of spatial distributions. Mathematical Geology, Vol. 15, No. 3, pp. 445 – 468.

Journel, A.G., and Huijbregts, Ch.J., 1978, Mining geostatistics, Academic Press, London.

Lan, Z., 2007, Modeling the volume-dependent distribution of categorical variables, MSc. thesis, University of Alberta.

Lan, Z., Leuangthong, O. and Deutsch, C.V., 2006, Why logratios are a bad idea for multiscale facies modeling, eighth annual report of the Centre for Computational Geostatistics, University of Alberta, pp. 211-1 – 211-11.

Leuangthong, O. and Deutsch, C.V., 2003, Stepwise conditional transformation for simulation of multiple variables, Mathematical Geology, Vol. 35, No. 2, pp. 155 - 173.

Leuangthong, O. and Deutsch, C.V., 2004, Transformation of residuals to avoid artifacts in geostatistical modelling with a trend, Mathematical Geology, Vol. 36, No. 3, pp. 287 - 305.

Manchuk, J. G., 2008, Guide to geostatistics with compositional data, Centre for Computational Geostatistics (CCG) Guidebook Series, University of Alberta, Edmonton, Vol. 7, 34p.

Matheron, G., 1962, Traite de geostatistique applique, tome I, Memoires du Bureau de Recherches Geologiques et Minieres, No. 14, (Editions Technip : Paris).

Matheron, G., 1963, Principles of geostatistics, Economic Geology, Vol. 58, pp. 1246 - 1266.

Martin-Fernandez, J.A., Barcelo-Vidal, C. and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets, Studies in Classification, Data Analysis, and Knowledge Organization (eds. Kiers, H., Rasson, J., Groenen, P. and Shader, M.), Springer-Verlag, Berlin, pp. 155 - 160.

Martin-Fernandez, J.A., Barcelo-Vidal, C. and Pawlowsky-Glahn, V., 2003, Dealing with zeros and missing values in compositional data sets using nonparametric imputation, Mathematical Geology, Vol. 35, No. 3, pp. 253 - 278.

Martin-Fernandez, J.A., Barcelo-Vidal, C., Pawlowsky-Glahn, V., Kovacs, L.O. and Kovacs G.P., 2005, Subcompositional patterns in Cenozoic volcanic rocks of Hungary, Mathematical Geology, Vol. 37, No. 7, pp.729 - 752.

Mossop, G.D., 1980, Geology of the Athabasca oil sands, Science, New Series, Vol. 207, No. 4427, pp. 145 - 152.

Nowak, M and Verly, G, 2004, The practice of sequential Gaussian simulation, in Geostatistics Banff 2004, Proceedings of the 7th International Geostatistical Congress, (eds. Leuangthong, O. and Deutsch, C.V.), pp. 387 - 398 (Springer).

Olea, R.A. (ed.), 1991, Geostatistical glossary and multilingual dictionary. International Association for Mathematical Geology, Studies in Mathematical Geology Volume 3, Oxford University Press, New York.

Pawlowsky-Glahn, V., 2004, Lecture notes on compositional data analysis, University of Girona short course.

Pawlowsky-Glahn, V. and Egozcue, J.J., 2006, Compositional data and their analysis: an introduction, in Compositional data analysis in the geosciences: from theory to practice (eds. Buccianti, A., Mateu_Figueras, G. and Pawlowsky-Glahn, V.), Geological Society, London, Special Publications, 264, pp. 1 - 10.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado R., 2011, Lecture notes on compositional data analysis, March 2011. Short course notes. http://www.compositionaldata.com/material/others/Lecture_notes_11.pdf

Pawlowsky-Glahn, V. and Olea, R., 2004, Geostatistical analysis of compositional data, Oxford University Press, New York.

Pearson, K., 1897, Mathematical contributions to the theory of evolution – On a form of spurious correlation which may arise when indices are used in the measurement of organs, Proceedings of the Royal Society of London, Vol. 60, pp. 489 - 498.

Rivoirard, J., 1994, Introduction to disjunctive kriging and non-linear geostatistics, Clarendon Press, Oxford.

Romanova, U.G., Yarranton, H.W. and Schramm, L.L., 2003, Towards the improvement of the efficiency of oil sands froth treatment, Canadian International Petroleum Conference, 2003. Paper 2003-010.

Rubinstein, R.Y., 1981, Simulation and the Monte Carlo Method, John Wiley & Sons, New York.

Saito, H. and Goovaerts, P., 2000, Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site, Environmental Science & Technology, Vol. 34, No. 19, pp. 4228 – 4235.

Schneider, D.M., Steeg, M., and Young, F.H., 1982, Linear algebra: A concrete introduction, Macmillan Publishing, New York.

Shurtz, R.F., 2003, Compositional geometry and mass conservation, Mathematical Geology, Vol. 35, No. 8, pp. 927 - 938.

Sinclair, A.J. and Blackwell, G.H., 2002, Applied mineral inventory estimation, Cambridge University Press, Cambridge.

Srivastava, R.M., 1987, A non-ergodic framework for variograms and covariance functions, M.Sc. thesis, Stanford University.

Thio-Henestrosa, S. and Martin-Fernandez, J.A., 2005, Dealing with compositional data: the freeware CoDaPack, Mathematical Geology, Vol. 37, No. 7, pp. 773 - 793.

Thomas, C.W. and Aitchison, J., 2005, Compositional data analysis of geological variability and process: A case study, Mathematical Geology, Vol. 37, No. 7, pp. 753 - 772.

Tolosana-Delgado, R., 2008, Compositional data analysis in a nutshell, University of Gottingen on-line reference, http://www.sediment.uni-goettingen.de/staff/tolosana/extra/CoDaNutshell.pdf

Tolosana-Delgado, R., Egozcue, J.J. and Pawlowsky-Glahn, V., 2008, Cokriging of compositions: log-ratios and unbiasedness, Proceedings of the Eighth International Geostatistics Congress (eds. Ortiz, J.M and Emery, X.), pp. 299 - 308.

Verly, G.W, 1983, The multiGaussian approach and its applications to the estimation of local reserves, Mathematical Geology, Vol. 15, No. 2, pp. 259 - 286.

Verly, G.W., 1984, Estimation of spatial point and block distributions: the multiGaussian model, PhD thesis, Stanford University.

Vann, J., Jackson, S. and Bertoli, O., 2003, Quantitative kriging neighbourhood analysis for the mining geologist – A description of the method with worked case examples, Proceedings Fifth International Mining Geology Conference, pp. 215 - 223 (The Australasian Institute of Mining and Metallurgy: Melbourne).

Wackernagel, H., 1995, Multivariate geostatistics: An introduction with applications. Springer-Verlag, Berlin.

Wik, S., Sparks, B.D., Ng, S., Tu, Y., Li, Z., Chung, K.H., and Kotlyar, L.S., 2008, Effect of bitumen composition and process water chemistry on model oilsands separation using a warm slurry extraction process simulation, Fuel, Vol. 87, Issue 7, pp. 1413 - 1421.