The Piecewise Linear Model of Regionalization for Geostatistical Simulation

by

Fabio P. L. Pereira

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Quantifying uncertainty is key to rational decision-making in a geological context. Samples collected for Mineral Exploration are usually sparse and only represent a very small portion of the volume that might be mined in the future.

To characterize and quantify geological uncertainty, it is commonplace to use Stochastic Simulation. Uncertainty is quantified and propagated in several steps when modeling geology. Samples are separated into different domains, uncertainty is quantified in the domain boundaries, parameters for modeling are established, and domain models are created. Afterward, inside each domain, continuous variables such as grades are modeled. This thesis proposes a novel method for continuous variable simulation.

A common workflow to quantify uncertainty is to transform the data to a Gaussian distribution and use Sequential Gaussian Simulation (SGS) to generate realizations of the grades. However, this approach assumes that all spatial distributions have a Gaussian form, under the multivariate Gaussian assumption. Also, it is assumed that a single variogram model characterizes the spatial variability of the grade independent of the magnitude. High grades, however, are usually less continuous. Using SGS to model such variables, with a single variogram model will impose the same continuity for lows and high values which may be unrealistic.

The main contribution of this thesis is to propose a novel simulation framework for grades that have different continuity for low and high values. The Piecewise Linear Model of Regionalization (PLMR) defines different bins to the data distribution and imposes different spatial model to each bin. By doing so, the model can capture different spatial continuity of highs and low values in a consistent mathematical manner. The proposed framework considers indicator variograms as well as traditional variograms, which brings more spatial information to the simulated realizations. When comparing the PLMR to modeling under the multivariate Gaussian assumption the former tends to be, on average, more conservative regarding the influence of high values in nearby locations.

Acknowledgments

First, I would like to thanks Dr. Clayton Deutsch for his support, advice, and guidance in all phases of research. Being able to work with Clayton was an unparalleled experience in terms of intellectual and professional growth. This thesis was crafted during trying pandemic times, hence, I would also like to thank him for his kind understanding of the challenges of researching in such environment.

To my parents, brother and girlfriend: I'm beyond grateful for your unconditional love and support that carried me to the finish line of this thesis. Also, I would like to thanks my dear Brazilian friends that never left me feeling alone during isolation, and my UFMG friends that encouraged me to pursue a Masters in the first place. Without you folks I would not be able to complete this thesis!

I would like to thanks my dear CCG friends for the good moments and fruitful discussions, and CCG sponsors for making this study possible.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Description
(n)	Samples falling in location $\mathbf{u}$ neighborhood
$Acc.$	Accuracy
$C(\mathbf{h})$	Covariance of random variables spaced by vector $\mathbf{h}$
$C(\quad)$	Covariance
$D(\mathbf{u})$	Relative difference at location $\mathbf{u}$
$E\{\quad\}$	Expected value
$F(\quad)$	Cumulative distribution function
$F_{plmr}(u)$	PLMR distribution
$F_{rep}(u)$	Representative distribution of a geological variable
$I(\mathbf{u}; z)$	Indicator coding of the random function at threshold z
$N(0, 1)$	Normal distribution
$N(\mathbf{h})$	Number of pairs separated by $\mathbf{h}$.
$O(\quad)$	Objetive function
P_{accep}	Probability of accepting a trial set of PLMR parameters
S_i	Scaling factor of regionalization models
$Sill(z)$	Sill of the indicator variogram at threshold z
T_p	PLMR truncation point
$T_{i_{std}}$	Standardize truncation point of the Gaussian factor i
V	Stationary spatial volume
$Y_i(\mathbf{u})$	Gaussian factor i
$Y_i^{PL}(\mathbf{u})$	Piecewise linear transformed Gaussian factor i
$Z(\mathbf{u})$	Random function model
$Z_{sk}(\mathbf{u})$	Simple Kriging estimate at location u
Φ	Gaussian CDF
$\gamma(\mathbf{h})$	Variogram at lag $\mathbf{h}$
$\gamma(\mathbf{h}; z)$	Indicator variogram at lag $\mathbf{h}$ and threshold z
λ	Kriging weight
$\mathbf{w}_i$	Vector of normal deviates
Prob	Probability distribution function
$\overline{\gamma_{dl}}(h)$	l estimated variogram at direction d based on the trial set of PLMR parameters

Symbol	Description
ϕ	Gaussian PDF
ρ	Linear correlation
σ	Standard deviation
$\sigma_{Z_c}(\mathbf{u})$	Conditional standard deviation of variable $Z(\mathbf{u})$
θ	PLMR parameter set
$c0_i$	Nugget effect of Gaussian factor i
$e(\mathbf{u})$	Error at location u
f	Arbitrary function
l	Index of the four variogram used for fitting a PLMR
$m_Y(\mathbf{u})$	Mean of variable $Y(\mathbf{u})$
$m_{Zc}(\mathbf{u})$	Conditional mean of variable $Z(\mathbf{u})$
n_{dir}	Number of directions, i.e 1D, 2D or 3D
n_{lags}	Number of lags
$q_i(x)$	i quantile of variable x
r_{γ_i}	Variogram range of Gaussian factor i
t_k	Temperature (relative perturbation) used at iteration k of the fitting algorithm
$z_{emp}(\mathbf{u})$	Set of measurements made in a stationary domain

List of Abbreviations

Abbreviation	Description
CBK	Conditioning by Kriging
ccdf	Conditional Cumulative Distribution Function
Cd	Cadmium
CDF	Cumulative Distribution Function
CV	Cross-Validation
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
KDE	Kernel Density Estimation
LHSMDU	Latin Hypercube Sampling with Multidimensional Uniformity
LMR	Linear Model of Regionalization
MCS	Monte Carlo Simulation
MG	multiGaussian
MIK	Multiple Indicator Kriging
NS	Normal Score
PDF	Probability Distribution Function
PLMR	Piecewise Linear Model of Regionalization
PPMT	Projection Pursuit Multivariate Transform
Q-Q	Quantile-Quantile
RF	Random Function
RV	Regionalized Variable
SA	Simulated Annealing
SGS	Sequential Gaussian Simulation
SIS	Sequential Indicator Simulation
SK	Simple Kriging
Zn	Zinc

Chapter 1

Introduction

1.1 Spatial modeling and geostatistics

Numerical models of spatially distributed properties in the subsurface are often used in different industries such as mining or petroleum for resource assessment and engineering design. Geostatistics is a branch of applied statistics that aims at building models that takes into account spatial/temporal indexing of observations (Goovaerts et al., 1997). Even though relatively new, the field has gained attention and established itself as an important collection of methods used to build numerical models of the subsurface. Besides providing a solid theoretical background for spatial modeling of geological variables (Chiles & Delfiner, 2009; A. G. Journel & Huijbregts, 1978; Wackernagel, 2003), geostatistical methods became commonplace in day to day applications due to the importance given to implementation and transforming theory into sound practical methods. (C. V. Deutsch, 2021; C. V. Deutsch & Journel, 1998; A. G. Journel, 1989).

The variable under study is treated as a random variable that is indexed in space. In the mining industry, for example, sampling is widely spaced and inadequate to directly characterize the variable over the entire area. Hence, measurements are used to characterize a random mechanism, auto-correlated in space, that reflects desired statistics of the variable. Once the random mechanism is parametrized, inference about unsampled locations may be conducted. In the past, it was commonplace to use some form of spatial interpolation to have the best estimate that minimized error or estimation variance. However, a single best estimate is inadequate to characterize, jointly, uncertainty about several unsampled locations and the minimized error variance, being data value independent, is insufficient to be used as a reliable measure for risk assessment.

Quantifying joint uncertainty has become a cornerstone of modern geostatistical workflows. Stochastic simulation algorithms allow to jointly access the uncertainty of several locations being above a grade cut-off or belonging to a certain rock type. Instead of generating one single best estimate given the sample data, stochastic simulation aims at generating equiprobable alternative realizations instead of an interpolated map (A. G. Journel, 1989). These realizations will honor the global distribution, measures of spatial continuity such as variograms, and local observations referred to as the conditioning data (Rossi & Deutsch, 2013). With a set of realizations honoring input statistics and conditioning data, a transfer function, e.g. a pit optimizer, can be used to process the simulated maps and characterize uncertainty in the desired final application of the numerical models. All realizations should be processed all the time (C. V. Deutsch, 2018). Figure 1.1 shows schematically the concept of processing several realizations, instead of one, to access risk in the

response variable for risk-based decision making. Sequential Gaussian Simulation (SGS) is probably the most used algorithm to build uncertainty models of continuous variable (Rossi & Deutsch, 2013).

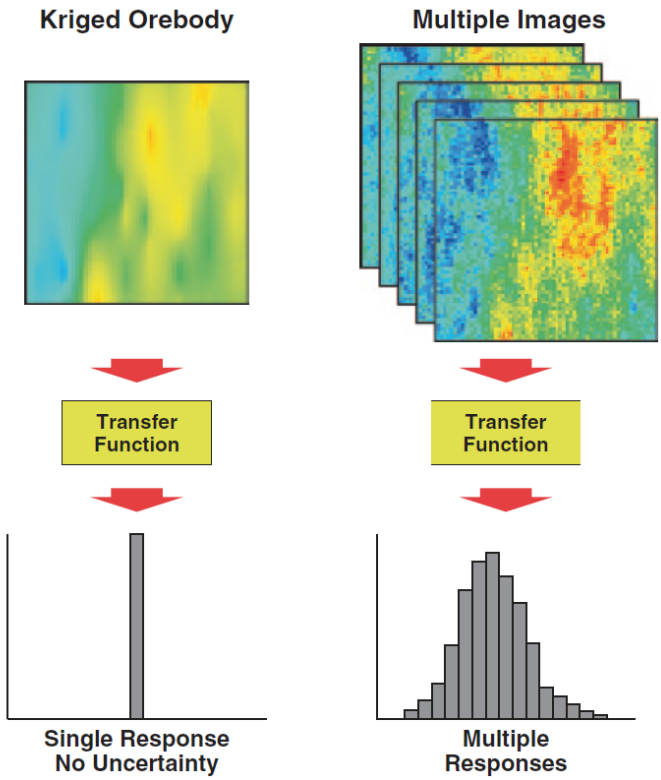

Figure 1.1: Tranfer function for estimated and simulated models. (Rossi & Deutsch, 2013)

1.2 Problem motivation

Modeling geological variables with high-valued samples is a challenge in the mining industry. Deposits, such as gold and diamonds, usually present samples that are considerably higher than central tendency measures such as the median or mean. Hence, the distribution of such variables will present a heavy right tail, high coefficient of variation, and are called positively skewed variables. This type of geological variable usually presents a significant challenge for spatial modeling. High grades in mineral deposits tend to be less continuous and more scattered in space. As a consequence, a destructuration effect may be present in the spatial continuity of high values. This effect is observed by looking at indicator variograms, where variograms at a high threshold value may be less continuous. Figure 1.2 shows an example, from the Swiss Jura dataset (Goovaerts et al., 1997), of a positively skewed distribution and indicator variograms of the same attribute. The variogram at a 0.9 quantile threshold has a range $\approx 60\%$ smaller than variograms at a 0.1 quantile threshold.

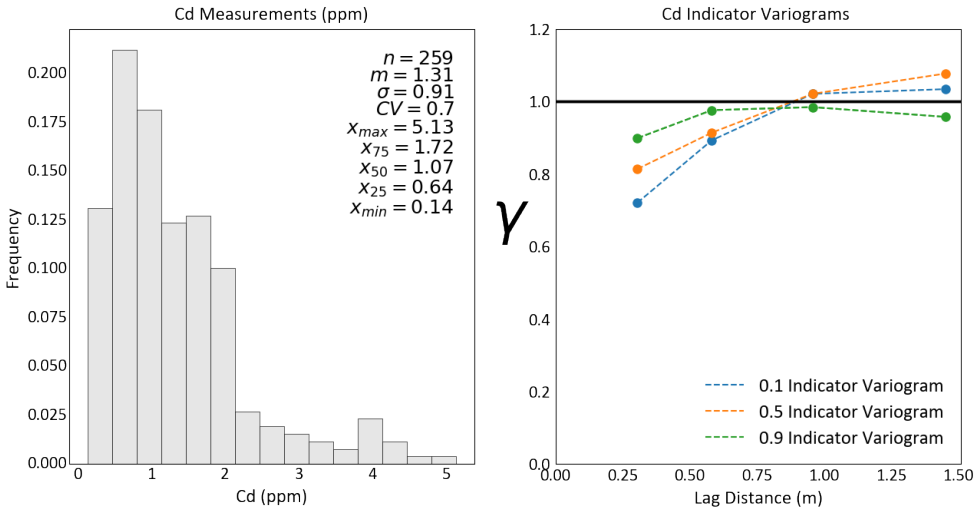

Figure 1.2: Example of a positively skewed geological variable and its indicator variograms.

When constructing an interpolated map of a positively skewed variable, there is a concern that a few high-grade samples might lead to over-estimation of areas near extreme high values. For mineral resource estimation, the impact of high-grades is mitigated by: 1) appropriate domaining of the region, 2) grade capping, and 3) limiting the influence of outliers (Leuangthong & Nowak, 2015). In an uncertainty quantification context, simulation might mitigate the impact of extreme values. Gaussian simulation algorithms (Rossi & Deutsch, 2013) mitigate the effect of outliers and asymmetric distributions by transforming the variable to a normal distribution. However, the de-structuration effect shown in Figure 1.2 brings challenges when trying to apply SGS or other Gaussian algorithms to simulate realizations of positively skewed variables. To use SGS, it is necessary to assume that the transformed variable is multiGaussian (MG), i.e. all spatial distributions will be Gaussian and solely characterized by a single covariance model, which makes it analytically simple but may inflate spatial disorder (A. G. Journel & Deutsch, 1993). Besides, under a MG model, the same pattern of spatial auto-correlation will be imposed to low and high values of the distribution. As shown in Figure 1.2, the continuity of highs and lows can be quite different making the Gaussian assumption inappropriate

The indicator formalism allows a non-parametric estimation of spatial distributions. The method maps the original variable into a binary one. This indicator function tells if a value is above or below a given threshold. Simulating under this framework reproduces better strings of low and high values and might be useful when dealing with variables presenting a skewed distribution (Rossi & Deutsch, 2013). However, Sequential Indicator Simulation (SIS) can be hard to use in practice. Variance inflation and order relation problems may happen (Emery, 2004; Rossi & Deutsch, 2013). This model requires fitting a variogram model at each threshold level where the binary transformation is applied, which may make the workflow tedious. Besides, fitting different variograms to different

thresholds does not take into account information of other cut-off levels, hence, the final model is not fully consistent.

1.3 Thesis outline

The present work proposes a new Non-Gaussian simulation framework, the Piecewise Linear Model of Regionalization (PLMR). The model proposes to extend the Linear Model of Regionalization (LMR) to allow non-linear combination of spatial factors, making the resulting model non-Gaussian. Also, the proposed framework uses indicator variograms as well as traditional variogram in a consistent mathematical manner, bringing more spatial information to the model. Chapter 2 presents the theoretical background of the novel PLMR. The chapter is mainly focused on the concept of the Random Function (RF) model and key assumptions/methods to parametrize it. Chapter 3 will introduce and define the PLMR showing the properties that make it relevant for modeling non-Gaussian geological variables. Chapter 4 will present an optimization framework to infer the model's parameters. Chapter 5 introduces a methodology to generate conditional realizations under a PLMR model. Conditioning is done by decomposing the model into its latent spatial (Gaussian) factor and simulating them. Chapter 6 will demonstrate the PLMR workflow and compare the results to a model built using the multiGaussian assumption.

Chapter 2

Theoretical background

2.1 Regionalized variables and random functions

In a geostatistical context, a Regionalized Variable (RV) $Z(\mathbf{u})$ can be defined as a set of functions $Z(\mathbf{u}); \forall \mathbf{u} \in V$ where u is a location coordinate vector and V is a volume in space relevant for the study (Wackernagel, 2003). In practice, information about the regionalized variables will come from a set of measurements of the same attribute, usually composited to some constant volume. This set of samples is then treated as a realization of the regionalized variable at the data locations $\mathbf{u}$.

The regionalized variable at a location is sometimes referred to as a random variable. The set of random variables is then treated as realization of a Random Function (RF). The term regionalized emphasizes that there is a structured component in the apparent randomness, i.e., samples close in space tend to be more similar than far away samples.

The RF model concept encompass the apparent randomness of a complex geological system and the regionalized aspect of the geological variable. The Cumulative Distribution Function (CDF) of a random variable, i.e. the cumulative proportion of values below a threshold, can be written as:

$$F(\mathbf{u}; z) = \text{Prob}(Z(\mathbf{u}) \leq z) \tag{2.1}$$

The CDF can be locally conditioned based on a set of measurements present in the neighborhood (n) centered at location $\mathbf{u}$. This distribution will be called Conditional Cumulative Distribution Function (ccdf) and can be defined as the following conditional distribution:

$$F(\mathbf{u}; z|(n)) = \text{Prob}(Z(\mathbf{u}) \leq z|(n)) \tag{2.2}$$

Under this framework, the spatial variability of the RF is fully characterized its K-variate CDFs where K are the relevant locations in the spatial volume V under study (C. V. Deutsch & Journel, 1998):

$$F(\mathbf{u}_1, \ldots, \mathbf{u}_K; z_1, \ldots, z_K) = \text{Prob}(Z(\mathbf{u}_1) \leq z_1, \ldots, Z(\mathbf{u}_K) \leq z_K) \tag{2.3}$$

In order to be able to proceed with any inference regarding the multivariate distribution shown in Equation 2.3 repetitive sampling is necessary. However, at each location $\mathbf{u}$, only one sample is available and replication of measurements is usually not possible. To overcome this issue, the RF model shown in Equation 2.3 should assume some form of stationarity over the domain V, i.e.

measurements at different locations $\mathbf{u_i}$, $i = 1, \ldots, K$ can be pooled together in order to make statistical inference possible. Geostatistical models usually assume that the RF defined over the volume V in space is second-order stationary, i.e. the first two moments of the regionalized variable are invariant under translation:

$$\begin{aligned} &\mathrm{E}\{Z(\mathbf{u})\} = m \\ &\mathrm{E}\{[Z(\mathbf{u}) - m(\mathbf{u}))][Z(\mathbf{u}+\mathbf{h}) - m(\mathbf{u}+\mathbf{h})]\} = C(\mathbf{h}) \end{aligned} \tag{2.4}$$

Where $\mathbf{h}$ is a separation vector with specific direction and length, $C(\mathbf{h})$ is the covariance of random variables spaced by $\mathbf{h}$, and m a constant stationary mean. Under the assumption of second-order stationarity, the covariance function only depends on the separation vector $\mathbf{h}$ and it measures the linear dependence between two locations separated by $\mathbf{h}$. The assumption of a constant stationary mean can be break by: filtering the mean at each location when estimating, i.e. ordinary kriging ; considering the mean as a deterministic component and using, e.g., universal kriging (Chiles & Delfiner, 2009) or calculating a trend and modelling with decorrelated residuals (Leuangthong & Deutsch, 2003; Qu & Deutsch, 2018).

Usually, simulation algorithms based on kriging will use a two-point covariance description of the variable to characterize the random mechanism and conduct inference regarding the K-variate distribution 2.3. Other methodologies such as Multi Point Statistics (Strebelle, 2002, 2012), or High-Order cumulants (Dimitrakopoulos, Mustapha, & Gloaguen, 2010; Mustapha & Dimitrakopoulos, 2011) have been proposed to use higher-order moments when characterizing the RF model.

Another widely used second moment in geostatistics is the variogram. The variogram arises by considering stationarity of increments $[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h}]; \forall \mathbf{u}, \mathbf{u}+\mathbf{h} \in V$, and it can be written as the variance of these increments:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= \mathrm{Var}\{Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\} \\ \gamma(\mathbf{h}) &= C(\mathbf{0}) - C(\mathbf{h}) \qquad , \\ C(\mathbf{0}) &= \mathrm{Var}\{Z(\mathbf{u})\} \end{aligned} \tag{2.5}$$

An empirical estimate of the variogram is calculated by:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \left[z\left(\mathbf{u_i}\right) - z\left(\mathbf{u_i}+\mathbf{h}\right) \right]^2 \tag{2.6}$$

The term $N(\mathbf{h})$ is the number of pairs in the volume V separated by $\mathbf{h}$. In practice, tolerance parameters will be added to $\mathbf{h}$ in order to have sufficient pairs to estimate the variogram. A valid model should be fitted to the empirical variogram values. By defining a valid model, one will characterize the two-point spatial dependency of the variable in space for any vector $\mathbf{h}$, hence defining the random mechanism of the RF. The parameters of the variogram model, e.g. range of spatial con-

tinuity, will control the weight that each sample will receive when any form of kriging is conducted for estimation/simulation.

2.2 Indicator coding of a random function

Knowledge about the two-point spatial distribution of a RF is obtained by calculations using the bivariate distribution $\mathrm{Prob}(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h}))$. Traditional variograms shown in Equation 2.5, capture the moment of inertia in the bivariate scattergram $Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h}); \forall \mathbf{u} \in V$ (A. G. Journel, 1983). Instead of defining the variogram by calculating a simple linear dependency between locations separated by a lag $\mathbf{h}$, it is possible to use the indicator coding of the RF to retrieve more information of the spatial bivariate distributions when modeling. The indicator function of the Random Function $Z(\mathbf{u})$ can be written as:

$$I(\mathbf{u}; z) = \begin{cases} 1, & Z(\mathbf{u}) \leq z \\ 0, & otherwise \end{cases} \tag{2.7}$$

Where z stands a specific value, e.g. the 0.1 quantile, of the random function $Z(\mathbf{u})$. The function maps values of $Z(\mathbf{u})$ that are above, or below, the threshold value z to binary indicator values. The indicator function can be thought of as a Binomial distributed variable (A. G. Journel, 1983). For example, if z is the median value of the variable $Z(\mathbf{u})$, $I(\mathbf{u}; z)$ the indicator function will follow a Binomial distribution with probability $p = 0.5$. The expected value and variance of the indicator function are shown in Equation 2.7 (A. G. Journel, 1983):

$$E\{I(\mathbf{u}; z)\} = 1 \cdot P(Z(\mathbf{u}) \leq z) + 0 \cdot P(Z(\mathbf{u}) > z) = P(Z(\mathbf{u}) \leq z) = F(z) \tag{2.8}$$

$$\mathrm{Var}\{I(\mathbf{u}; z)\} = F(z)(1 - F(z)) = F(z) - F(z)^2 \tag{2.9}$$

The mean and variance are a function of the threshold z and the CDF truncation defined by it, i.e $F(z)$. The non-centered covariance of the indicator function is:

$$C_{NC}(\mathbf{h}; z) = E\{I(\mathbf{u}; z), I(\mathbf{u}+\mathbf{h}; z)\} = P(Z(\mathbf{u}) \leq z, Z(\mathbf{u}+\mathbf{h}) \leq z) \tag{2.10}$$

The centered covariance of $I(\mathbf{u}; z)$, variogram and sill are:

$$C(\mathbf{h}; z) = P(Z(\mathbf{u}) \leq z, Z(\mathbf{u}+\mathbf{h}) \leq z) - F(z)^2 \tag{2.11}$$

$$\begin{aligned} &\gamma(\mathbf{h}; z) = C(\mathbf{0}; z) - C(\mathbf{h}; z) = F(z) - C_{NC}(\mathbf{h}; z) \\ &Sill(z) = F(z) - F(z)^2 \end{aligned} \tag{2.12}$$

From Equations 2.10 to 2.12, it can be seen that the indicator structural information of a regionalized random variable can be solely deduced from its bivariate distribution. In fact, the non-centered covariance is just an integration of the bivariate distribution below a given threshold (A. G. Journel, 1983). Therefore, a complete indicator structural analysis is richer than a simple $Z(\mathbf{u})$ variography. Besides providing more information, the rank order based analysis of the indicator approach makes its variograms robust regarding extreme values.

2.3 Multivariate Gaussian model

Inference about the K-variate distribution 2.3 under some form of stationarity assumption is necessary to achieve a joint model of uncertainty. A simplifying assumption is to assume that the RF model is multivariate Gaussian. This assumption makes all cumulative conditional distributions of the RF to have multivariate Gaussian. Under this assumption, any ccdf of the form $\text{Prob}(Z(\mathbf{u}) \leq z)|(n))$ is fully characterized by the conditional mean and variance calculated based on a covariance model and a set of conditioning data (n). In fact, under the MG assumption, Simple Kriging (SK) equations are equal to the normal equations (Leuangthong, Khan, & Deutsch, 2011). Hence, the two parameters that define the ccdf and quantify uncertainty in a given location $\mathbf{u}$ are the Simple Kriging mean and variance:

$$E\{Z(\mathbf{u})|(n)\} = \sum_{\alpha \in (n))} \lambda_\alpha \cdot Z\left(\mathbf{u}_\alpha\right) \tag{2.13}$$

$$\text{Var}\{Z(\mathbf{u})|(n)\} = 1 - \sum_{\alpha \in (n))} \lambda_\alpha C\left(\mathbf{u}, \mathbf{u}_\alpha\right) \tag{2.14}$$

$$\sum_{\beta \in (n)} \lambda_\beta C\left(\mathbf{u}_\alpha, \mathbf{u}_\beta\right) = C\left(\mathbf{u}, \mathbf{u}_\alpha\right); \forall \alpha \in (n) \tag{2.15}$$

From now on, the multivariate Gaussian distribution, when used in a spatial/geostatistical context, will be called MG model. Under the MG model, bivariate distributions used to characterize and calculate variogram values will have a Gaussian ellipse shape characterized by the first two marginal moments of $E\{Z(\mathbf{u}) \leq z\}$ and $E\{Z(\mathbf{u}+\mathbf{h}) \leq z\}$, and a correlation parameter $\rho(\mathbf{h})$ defining the degree of linear dependency between the two marginal distributions. Accepting the MG model to use SGS, all spatial distributions will be solely characterized by a single covariance model, which makes it analytically simple but will inflate spatial disorder (A. G. Journel & Deutsch, 1993). This loss of connectivity occurs symmetrically when moving away from the median, i.e it happens in low and high values. Figure 2.1 shows qualitatively the bivariate contours of the MG model and the theoretical 0.1, 0.5 and 0.9 threshold indicator variograms of a Spherical Variogram with range 10. The 0.1 and 0.9 indicator variograms are the same due to symmetry of the Gaussian bi-

variate distribution. Beside inflating spatial disorder beyond the covariance model (A. G. Journel & Deutsch, 1993), the MG model does not take into account the different continuity of extreme values in the time of modeling. The symmetrical behavior of indicator variograms imposes the same spatial continuity, derived from the model, to highs and low values. Since an attribute covariance is the average of all indicators covariance (Journal & Alabert, 1989), poor characterization of the low probability values may affect directly the model's response after being processed by a transfer function. In fact, the final image will not preserve strings of low and high values present in data, i.e. patterns of low entropy. Since it is quite common to be interested in these patterns, assuming an MG hypothesis may lead to inaccurate spatial continuity modeling and a bad characterization of uncertainty in the results (A. G. Journel & Deutsch, 1993).

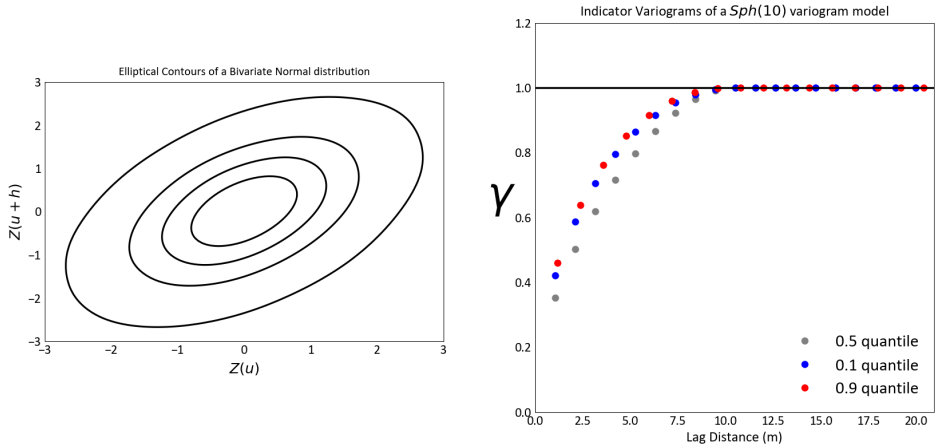

Figure 2.1: PLMR model and the LMRs embedded in it

2.4 Linear model of regionalization

Under the RF model concept, it is possible to think of the regionalized phenomena as a function of several independent random process acting in different spatial scales (Wackernagel, 2003). The simplest way to define this decomposition function is to assume it has a linear functional form. Hence, the RF $Z(\mathbf{u})$, assuming a stationary mean $m_Z(\mathbf{u}) = 0; \forall \mathbf{u}$, can be written as a sum of s independent spatial factor $Y_i(\mathbf{u})$, $i = 1, \ldots, s$:

$$Z(\mathbf{u}) = \sum_{i=1}^{s} S_i \cdot Y_i(\mathbf{u}) \tag{2.16}$$

The terms S_i controls the contribution of each spatial component to the final RF model. Under the decomposition shown in Equation 2.16, the variogram of $Z(\mathbf{u})$ may be written as a linear combination of the spatial components $Y_i(\mathbf{u})$ variograms, i.e. γ_i (Wackernagel, 2003), also know as a nested variogram model:

$$\gamma_z(\mathbf{h}) = \sum_{i=1}^{s} {S_i}^2 \cdot \gamma_i(\mathbf{u}) \tag{2.17}$$

The LMR makes it possible to break down the RF model and fit different aspects of the regionalized variable. When working under the MG model, each factor $Y_i(\mathbf{u})$ is a Gaussian RF with spatial variation defined by γ_i. Since a linear combination of Gaussian random variables is still Gaussian, $Z(\mathbf{u})$ will still be multiGaussian and Equation 2.17 can be used to to model the RF variograms and estimate the ccdfs parameters using Simple Kriging. Using RF multiGaussian model with a nested variogram model is commonly used for uncertainty estimation with stochastic simulation. However, the symmetric loss of spatial correlation around the median is still present even when decomposing the RF model into spatial components with differing scaling of continuity.

Chapter 3

The piecewise linear model of regionalization

One possible away to depart from the multiGausian model and its consequences is to rethink the decomposition of the Random Function model. The idea of fitting different sills is sound but the linear functional form of the RF decomposition makes the resulting model locked into a MG one. The piecewise linear model of regionalization aims at expanding the idea of decomposition of the Random Function to allow non-linear combination of spatial factors. In a general way, the decomposition of the RF model, assuming a stationary constant mean $m(\mathbf{u}) = 0$, into s spatial Gaussian factors $Y_i(\mathbf{u})$, $i = 1, \ldots, s$, can be written as:

$$Z(\mathbf{u}) = f(Y_i(\mathbf{u})) \tag{3.1}$$

As seen before the LMR assumes that $f(Y_i(\mathbf{u}))$ has a linear functional form, i.e $f(Y_i(\mathbf{u})) = \sum_{i=1}^{s} S_i Y_i(\mathbf{u})$, when decomposing the RF, and as consequence, defining symmetric indicator variograms around the median variogram. The LMR, by definition, imposes a constant contribution of each spatial factor to $Z(\mathbf{u})$ defined by the slope S_i. This multiplication of $Y_i(\mathbf{u})$ by a constant value can be thought of as a scaling that specifies the proportion that it contributes to the RF model. Hence, each factor scaling is a simple affine function $y = ax$.

A simple way to define a non-linear function is to define it as piecewise map of the domain $x$onto the co-domain y. This is done by defining bins in the domain, and at each bin imposing a different affine form to the map of x onto y. As more bins are added, the degree of non-linearity of the function will increase. This is shown qualitatively in Figure 3.3 where the Gaussian spatial factors are shown on the horizontal axes and their transformed values in the vertical axes.

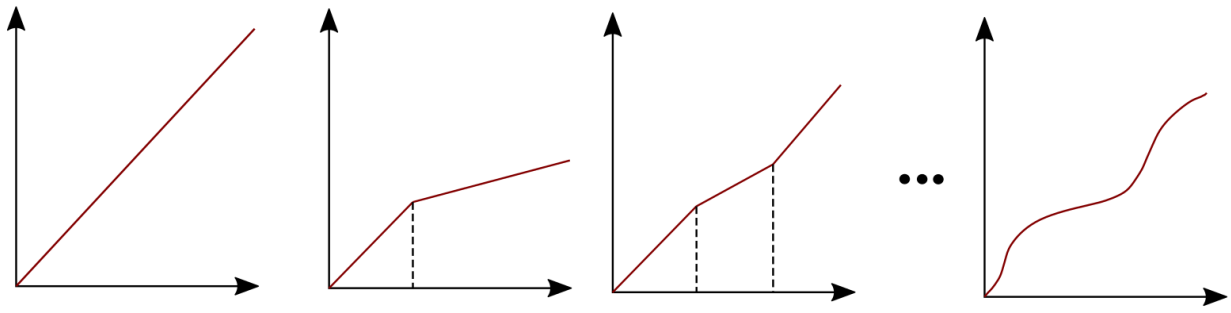

Figure 3.1: Process of transforming a linear function into an increasing non-linear one.

The PLMR aims at breaking this strict linear assumption in the simplest way possible, i.e, defin-

ing two bins in the domain space of spatial factors $Y_i(\mathbf{u})$. The next sections will explain the model in detail.

3.1 Model definition

Let's define an arbitrary RF model $Z(\mathbf{u})$ that has its spatial range of continuity characterized by two factors $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$. These two spatial factors are second-order stationary with variogram $\gamma_1(\mathbf{h})$ and $\gamma_2(\mathbf{h})$ and stationary mean of 0. These structures can be simulated independently reproducing variograms $\gamma_1(\mathbf{h})$, $\gamma_2(\mathbf{h})$ and having a global normal distribution. Hence, from now on, $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$ will be called Gaussian factors. As mentioned in the previous section, it is possible to define bins in the domain of a function and define a different affine map for each bin. Calling the point separating each bin a truncation point T_p, the piecewise linear transform of the Gaussian factors can be written as:

$$Y_1^{PL}(\mathbf{u}) = \begin{cases} S_1 Y_1(\mathbf{u}), & Y_1(\mathbf{u}) \leq T_p \\ S_2 Y_1(\mathbf{u}), & Y_1(\mathbf{u}) > T_p \end{cases} \tag{3.2}$$

$$Y_2^{PL}(\mathbf{u}) = \begin{cases} S_2 Y_2(\mathbf{u}), & Y_2(\mathbf{u}) \leq T_p \\ S_1 Y_2(\mathbf{u}), & Y_2(\mathbf{u}) > T_p \end{cases} \tag{3.3}$$

The pair of slopes is the same for each bin, however, the Gaussian factor being multiplied by each slope S_i is changed. After applying this piecewise linear function to $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$, the transformed values should be summed location wise, recomposing the RF $Z(\mathbf{u})$:

$$Z(\mathbf{u}) = \begin{cases} S_1 Y_1(\mathbf{u}) + S_2 Y_2(\mathbf{u}), & Y_1(\mathbf{u}) \leq T_p, Y_2(\mathbf{u}) \leq T_p \\ S_1 Y_1(\mathbf{u}) + S_1 Y_2(\mathbf{u}), & Y_1(\mathbf{u}) \leq T_p, Y_2(\mathbf{u}) > T_p \\ S_2 Y_1(\mathbf{u}) + S_2 Y_2(\mathbf{u}), & Y_1(\mathbf{u}) > T_p, Y_2(\mathbf{u}) \leq T_p \\ S_2 Y_1(\mathbf{u}) + S_1 Y_2(\mathbf{u}), & Y_1(\mathbf{u}) > T_p, Y_2(\mathbf{u}) > T_p \end{cases} \tag{3.4}$$

Equation 3.4 defines the decomposition of the RF model into its Gaussian factors. The idea behind this simple modification of the LMR is to control the contribution each variogram structure gives to highs and low values of the model. For example, setting $S_1 = 0.9$ and $S_2 = 0.1$ in Equation 3.4 would define a model that has more contribution of $Y_1(\mathbf{u})$ to values below T_p and $Y_2(\mathbf{u})$ contributing more to values above T_p. In this thesis, the PLMR will be restricted to two Gaussian Factors and one truncation point, as shown in Equation 3.4. However, there are no theoretical restrictions to expand the PLMR to an arbitrary number of Gaussian factors and truncation points.

3.2 PLMR parameters

Setting up a two-factor PLMR model requires, at least, inference of seven parameters:

$$\theta = (r_{\gamma_1}, r_{\gamma_2}, c0_1, c0_2, S_1, S_2, T_p) \tag{3.5}$$

Where $r_{\gamma_1}, r_{\gamma_2}$ are the ranges of the two variogram structures of $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$, $c0_1, c0_2$ are the nugget effects of the factors, and S_1, S_2 the two constant values that defines the scaling of the factors in the two bins defined by the truncation point T_p. In order to make the distribution of $Z(\mathbf{u})$ more well-behaved, the following restriction is imposed on the contribution parameters:

$$S_1^2 + S_2^2 = 1 \tag{3.6}$$

The truncation point is set to $T_p = 0.0$ to avoid a discontinuity point in the scaled Factors. This discontinuity point brings challenges at the time of conditionally simulating PLMR realizations. Hence, the final set of parameters becomes $\theta = (r_{\gamma_1}, r_{\gamma_2}, c0_1, c0_2, S_1)$ and the general two factor PLMR model equation can be written as:

$$Z(\mathbf{u}) = \begin{cases} \sqrt{S_1}Y_1(\mathbf{u}) + \sqrt{1-S_1}Y_2(\mathbf{u}), & Y_1(\mathbf{u}) \leq 0, Y_2(\mathbf{u}) \leq 0 \\ \sqrt{S_1}Y_1(\mathbf{u}) + \sqrt{S_1}Y_2(\mathbf{u}), & Y_1(\mathbf{u}) \leq 0, Y_2(\mathbf{u}) > 0 \\ \sqrt{1-S_1}Y_1(\mathbf{u}) + \sqrt{1-S_1}Y_2(\mathbf{u}), & Y_1(\mathbf{u}) > 0, Y_2(\mathbf{u}) \leq 0 \\ \sqrt{1-S_1}Y_1(\mathbf{u}) + \sqrt{S_1}Y_2(\mathbf{u}), & Y_1(\mathbf{u}) > 0, Y_2(\mathbf{u}) > 0 \end{cases} \tag{3.7}$$

Figure 3.2 shows a schematic representation of the steps to compute an unconditional PLMR realization. It can be seen how the piecewise linear transform accentuates specific features of each Gaussian factor realization. For example, high valued areas of the second Gaussian factor are considerably diminished while low values of the first factor become bigger in magnitude. Hence, the transformation is accentuate specific features of each variogram structure realization. With the general two-factor PLMR model defined above, the next sections will explore the relevant properties of the proposed framework to modeling non-Gaussian geological variables.

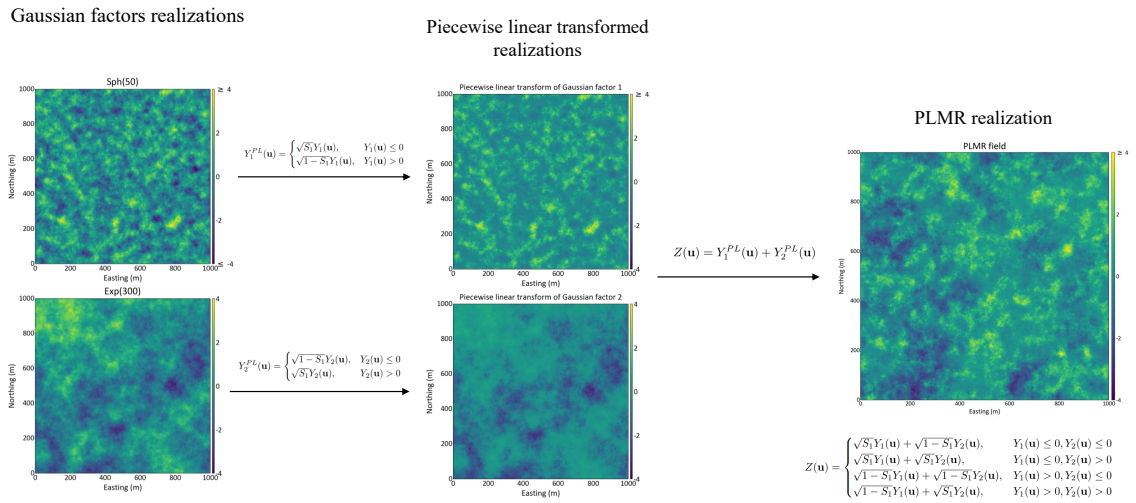

Figure 3.2: Schematic representation of the PLMR.

3.3 PLMR properties

To understand the properties of the PLMR and how it differs from the LMR a scatter plot of the scaled and unscaled Gaussian Factors using a linear and piecewise linear scaling is shown in Figure 3.3. The figure illustrates how a single PLMR model have elements of two LMRs embedded into it.

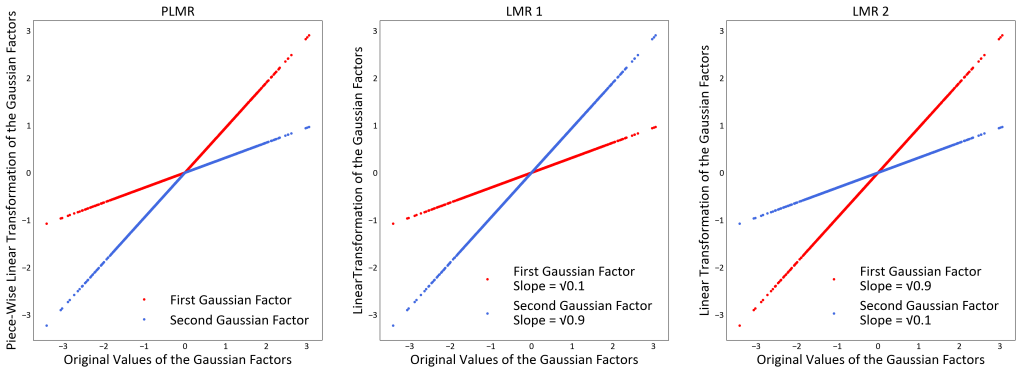

Figure 3.3: PLMR model and the LMRs embedded in it

When one of the scaling constant (S_1 and S_2) approaches 0, the other will approach 1 because of restriction 3.6. In this case, the contribution of each Gaussian factor to each bin of the PLMR will be very different. This set of parameters θ defines a PLMR that is able to produce more complex regionalization in comparison with the two LMRs. The piecewise scaling changes the shape of the spatial bivariate distributions in comparison with a linear scaling. This gives flexibility to model regionalized random variables that depart significantly from the multiGaussian assumption

approach. Figure 3.4 shows the PLMR bivariate distribution in comparison with a LMR and how its shape does not follow a bivariate Gaussian.

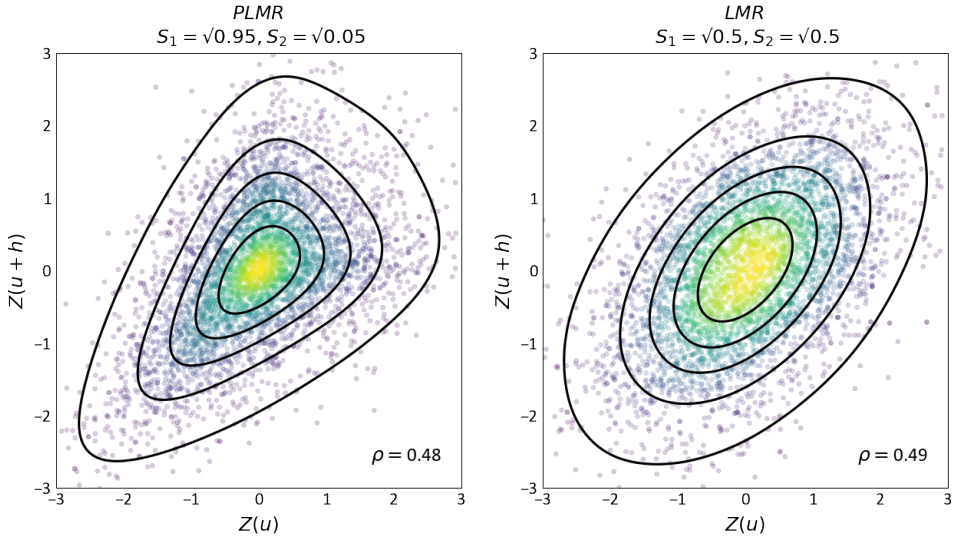

Figure 3.4: Bivariate distributions of a PLMR and a multiGaussian model.

Figure 3.5 shows two Gaussian factor realizations, with spatial continuities defined by a $Spherical(50)$ and $Exponential(300)$ variograms, two LMRs and a PLMR built from them, and the associated indicator variograms. The first LMR is defined by setting the scaling of the Gaussian factors $\sqrt{S_1} = 0.9$ and $\sqrt{S_2} = 0.1$, the second is defined by switching the contribution values. The PLMR is defined with the same scaling values of the first LMR and $T_p = 0.0$. Figure 3.5 shows how the PLMR is able to maintain aspects of the continuity of the two Gaussian factors, while the LMRs are more similar to the factor that is multiplied by the biggest scaling parameter. The consequence of applying a piecewise linear scaling to the factors, with very different contributions, is the possibility to define a model with non-Gaussian bivariate distributions, as shown in Figure 3.4. Consequentially, the RF defined with a PLMR will have asymmetrical indicator variograms. It is conventioned that the range of the first Gaussian factor variogram is the smaller between the two. Hence, by the model definition shown in Equation 3.7, values above the truncation point will have more contribution of the less continuous factor. As mentioned above, having less spatial continuity in high values is typical of non-Gaussian and positively skewed geological variables. The degree of non-gaussianity of the PLMR is controlled by the scaling constants S_1 and S_2. This behavior is illustrated in Figure 3.6 where four global distributions of different PLMR is shown. As the scaling constants diverges, the distribution differs more from a normal one.

It might be tempting to derive the mean and variance of $Z(\mathbf{u})$ assuming $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$ are Gaussian in each piece of Equation 3.7. It is known that multiplication of a normal variable by a constant can only re-scale its mean and variance; however, truncation of the factor's distribution will drastically change its shape and may make it non-Gaussian. This fact makes an analytical derivation

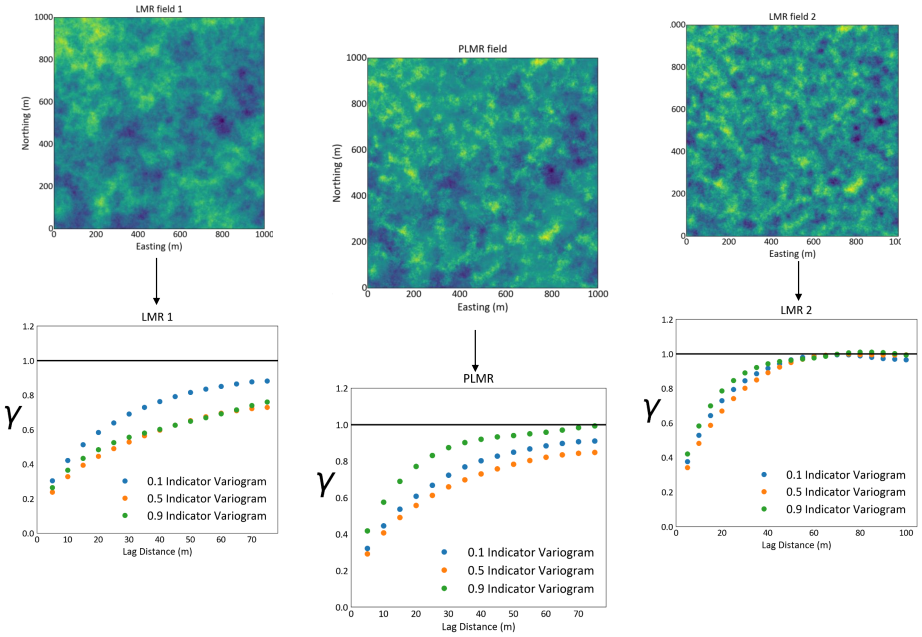

Figure 3.5: Realization of the PLMR and embedded LMRs.

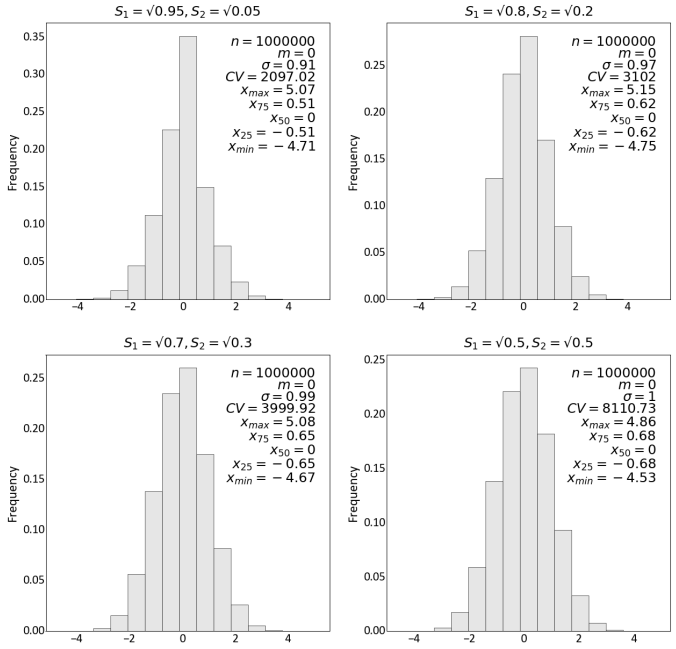

Figure 3.6: PLMR model global distribution as a function of the scaling parameters S_1 and S_2

of the PDF's shape difficult. Also, summing the factors after the truncation is a convolution of their PDFs (Krenek, Cha, & Cho, 2016) which adds extra complexity in the resultant shape. Some examples of the possible shapes can be found in (Krenek et al., 2016). Therefore, as demonstrated by Figure 3.6, the variable $Z(\mathbf{u})$ will not follow a normal distribution and some kind of variance inflation or deflation might occur.

It is possible, however, to write the mean and variance of each piece of Equation 3.4 as a function of the Gaussian CDF distribution. Doing that and treating the PDF of $Z(\mathbf{u})$ as a mixing of each piece of Equation 3.7, it is possible to derive the mean and variance of Equation 3.4. The tentative derivation of this result is deferred to Appendix A. An interesting property of applying the piecewise linear transform to Gaussian factors is that the mean and median of the resulting PLMR will still be the same as a normal distribution. Using more truncation quantiles and Gaussian factors may make the summing rule in Equation 3.7 excessively complex and more suitable to a numerical approach.

3.4 Calculating PLMR variograms

A simulation routine is needed to generate the PLMR's indicator and traditional variograms for any lag $\mathbf{h}$ and threshold z. The goal is to simulate a set of realizations of $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ and numerically calculate the indicator variograms using Equation 2.12. This section will outline the simulation routine to generate the bivariate spatial distribution of the simple PLMR model showed in Equation 3.7. First, let's write the realization of a set of $N(0,1)$ values as:

$$\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4]^T \tag{3.8}$$

Where each element i of the vector $\mathbf{w}$ stands for a vector of normal deviates:

$$\mathbf{w}_i = [w_{i_1}, w_{i_2}, ..., w_{i_n}] \tag{3.9}$$

Where n is the number of realizations and w_{i_j} the j realization of the standard Gaussian vector $\mathbf{w}_i$. Each vector of deviates $\mathbf{w}_i$, $i = 1, 2, 3, 4$, is independent of the others. Since the Gaussian factors are generated using pre-defined variogram models, Cholesky decomposition method can be applied to generate correlated variables $[Y_1(\mathbf{u}), Y_1(\mathbf{u}+\mathbf{h})]$ and $[Y_2(\mathbf{u}), Y_2(\mathbf{u}+\mathbf{h})]$ for a given lag $\mathbf{h}$. The covariance matrix of the two Gaussian Factors can be written as:

$$C(\mathbf{h}) = \begin{bmatrix} 1 & C_{Y_1}(\mathbf{h}) & 0 & 0 \\ C_{Y_1}(\mathbf{h}) & 1 & 0 & 0 \\ 0 & 0 & 1 & C_{Y_2}(\mathbf{h}) \\ 0 & 0 & C_{Y_2}(\mathbf{h}) & 1 \end{bmatrix} \tag{3.10}$$

The Cholesky decomposition of the above matrix can be defined as:

$$C(\mathbf{h}) = LL^T \tag{3.11}$$

It is well known that multiplication of the lower matrix L and the matrix of normal samples $\mathbf{w}$ generates a set of correlated variables that honors the covariance defined by Equation 3.10. Therefore, the correlated realizations of the Gaussian Factors in different locations can be calculated as:

$$L\mathbf{w} = \begin{bmatrix} Y_1(\mathbf{u}) \\ Y_1(\mathbf{u}+\mathbf{h}) \\ Y_2(\mathbf{u}) \\ Y_2(\mathbf{u}+\mathbf{h}) \end{bmatrix} \tag{3.12}$$

Now, samples have the correct spatial structure defined by the Gaussian factor's variograms. However, to retrieve the bivariate distribution of the PLMR variable $Z(\mathbf{u})$ it is necessary to apply the piecewise scaling rule, defined by the model's parameters, on the correlated simulated samples. Following the rule defined in Equation 3.7 values of $Y_2(\mathbf{u})$ and $Y_2(\mathbf{u}+\mathbf{h})$ below the truncation quantile T_p will be multiplied by $\sqrt{S_1}$ and above it by $\sqrt{1-S_1}$. The inverse should be done to $Y_1(\mathbf{u})$ and $Y_1(\mathbf{u}+\mathbf{h})$. Finally, the realizations of $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ are calculated by summing location-wise the piecewise transformed factors:

$$\begin{aligned} Z(\mathbf{u}) &= Y_1^{PL}(\mathbf{u}) + Y_2^{PL}(\mathbf{u}) \\ Z(\mathbf{u}+\mathbf{h}) &= Y_1^{PL}(\mathbf{u}+\mathbf{h}) + Y_2^{PL}(\mathbf{u}+\mathbf{h}) \end{aligned} \tag{3.13}$$

After simulating the PLMR bivariate distribution, indicator variograms are easily derived by integrating, numerically, the distribution $P(Z(\mathbf{u}) \leq z, Z(\mathbf{u}+\mathbf{h}) \leq z)$. As seen before, knowing the bivariate distribution and truncating it generates the non-centered covariance that can be easily be converted to indicator variogram by Equation 2.12. From the correct PLMR bivariate distribution, traditional variograms can be estimated by calculating the covariance between $[Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})]$ and then calculating the variogram from the relationship shown in Equation 2.5.

Even though the simulation steps to generate the bivariate distribution of the PLMR model is fairly straightforward, care must be taken defining the sampling strategies to ensure stability of the indicator variograms. Threshold the RF at extreme values will make one of the classes defined by indicator coding Equation 2.7 less likely to be sampled. For example, estimating the 0.1 indicator variogram requires a good characterization of the low probability event of $Z(\mathbf{u}) \leq q_{0.1}(Z(\mathbf{u}))$ and $Z(\mathbf{u}+\mathbf{h}) \leq q_{0.1}(Z(\mathbf{u}))$, where $q_{0.1}(.)$ represents the 0.1 quantile. A poor sampling of extreme values might impact the stability of the indicator variograms when the threshold moves away from the median value.

The current implementation of this thesis uses Monte Carlo Simulation (MCS) to sample standard univariate normal variables and process them using the Projection Pursuit Multivariate Transform (PPMT). The idea behind PPMT is to find the univariate projections of the data that are the most non-Gaussian and transform them into univariate Gaussian. If, for any direction in the multivariate space, the projection follows a univariate standard Gaussian distribution, the variables will follow a multivariate Gaussian distribution (Barnett, Manchuk, & Deutsch, 2014). Therefore, at least in an informal manner, PPMT can be seen as a convergence boost of the MCS samples to the Multivariate Gaussian distribution needed. For model 3.7, the bivariate distribution being estimated is a function of the realization of four random functions:

$$F(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})) = f(Y_1(\mathbf{u}), Y_1(\mathbf{u}+\mathbf{h}), Y_2(\mathbf{u}), Y_2(\mathbf{u}+\mathbf{h})) \tag{3.14}$$

Therefore, besides ensuring that each distribution on the right side of the above equation follows, marginally, a standard Gaussian shape, it is necessary to make them standard multivariate Gaussian. If the correct shape of the multidimensional distribution is preserved in the simulation, integration of $P(Z(\mathbf{u}) \leq z, Z(\mathbf{u}+\mathbf{h}) \leq z)$ will be more precise and the indicator variograms calculated from simulation will probably be more stable. This is even more necessary when a larger number of Gaussian factors are being used. Processing the MCS samples using PPMT makes them closer to the theoretical multivariate Gaussian distribution and improves precision of the correlation done by the Cholesky method. A comparison of the methods, alongside the Latin Hypercube Sampling with Multidimensional Uniformity (LHSMDU) (J. L. Deutsch & Deutsch, 2012), can be found in (Pereira & Deutsch, 2020).

Chapter 4

Inference of PLMR parameters

This chapter presents a semi-automatic optimization procedure to parametrize a PLMR given available information. The procedure is done by a stochastic optimization algorithm based on Simulated Annealing (SA) that randomly explores the parameter space keeping the best solution in terms of an objective function. The fitting takes into account traditional and indicator variograms simultaneously. The algorithm is not fully automated since it requires an initial parameter guess to be specified by the user.

Due to the stochastic nature of the algorithm, the optimization may converge to a near-optimal solution. For the purpose of fitting PLMR parameters a near-optimum solution is sufficient. It will be shown that small fluctuations in parameters do not severely impact the final variograms shapes. The next sections will go into more details about the fitting algorithm, show sensitivity results to changes in parameters and present a demonstration using the Swiss Jura dataset (Goovaerts et al., 1997).

4.1 Fitting procedure

The PLMR aims to take into account, in a consistent way, indicator variograms when defining a random function model. To accomplish this, the workflow defines different contributions of each variogram structure to different bins of the data. Setting up a second order PLMR model, imposing $T_p = 0$ and $S_1^2 + S_1^2 = 1$, is done by estimating the parameter set θ:

$$\theta = (r_{\gamma_1}, r_{\gamma_2}, c0_1, c0_2, S_1) \tag{4.1}$$

Where r_{γ_1} and r_{γ_2} are the ranges of the Gaussian factors variograms. The slopes S_1 and S_2 control the magnitude of each factor's contribution to different ranges of the data and, as a consequence, the degree of asymmetry in indicator variograms. The fitting procedure of a PLMR may be considered as an inverse problem. Given the available information, i.e., empirical variograms, it is necessary to estimate the parameters based on the minimization of an objective function. In practice, it means choosing a set of parameters that minimizes the difference between available information and the model state (simulated based on the parameter set θ). To capture the asymmetry in indicator variograms typical of non-Gaussian variables, the model is fit using indicator variograms at thresholds $q = 0.1, 0.5, 0.9$. Using the three thresholds, slopes S_1 and S_2 have information to be tuned. Fitting based on one indicator variogram would not provide any information on asymmetry and transition between thresholds. The traditional variogram is also used in the fitting procedure.

Figure 4.1 shows the traditional and indicator variograms of a PLMR. It is possible to see how the shapes are related and, in general, the range that is modeled by solely looking at the traditional variogram is the same as the range of the most spatially continuous Gaussian Factor in the model. Using it may seem redundant at first, but traditional variograms are used as input in the conditioning process, described in Chapter 5, hence accounting for them in the fitting procedure is necessary. If they are not used, there is no guarantee the traditional variogram based on parameters inferred solely from indicator variograms would match the empirical estimates.

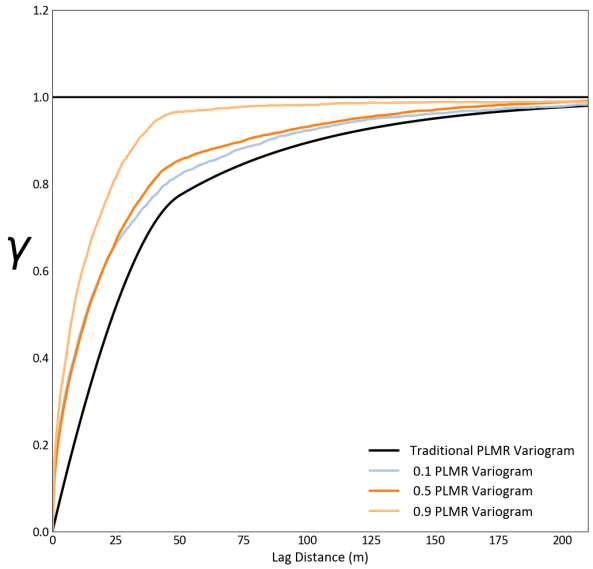

Figure 4.1: Traditional and indicator variograms of an arbitrary PLMR.

For a given set of parameters described θ traditional and/or indicator variograms can be easily simulated and accessed by the simulation procedure showed in Section 3.4. Optimization is done using a variation of SA (C. V. Deutsch, 1992; Glover & Kochenberger, 2006; Haykin, 2010). Defining the model parameters, the information used to tune them, and how to access the state of the model given a set of input parameters, the objective function for an estimate of θ' may be written as:

$$O(\theta') = \sum_{d=1}^{n_{dir}} \sum_{l=1}^{4} \sum_{h=1}^{n_{lags}} (\gamma'_{dl}(\mathbf{h}) - \gamma_{dl}(\mathbf{h}))^2 \tag{4.2}$$

Where:

n_{dir} = Number of directions of the problem, i.e 1D, 2D or 3D.

n_{lags} = Number lags where each variogram is estimated.

l = Index of the four variogram used for fitting.

$\gamma'_{dl}(\mathbf{u})$ = Estimated Variograms based on the trial set of parameters θ'.

The fitting algorithm starts with an initial set of parameters provided by the user and iterates over candidate solutions for the problem during a pre-defined number of iterations. The candidate solutions are defined by exploring the local neighborhood of the current set of parameters and keeping a candidate solution if it makes the objective function smaller. The size of this neighborhood is controlled by the temperature parameter t_k where k is the iteration count. In case the trial parameter set is not accepted based on a reduction of the objective function, it can still be accepted based on a probability defined by an exponential distribution proportional to the size of local search and the difference between current and trial objective function value. Hence, the trial parameter set at iteration $k+1$ is accepted if:

$$P_{accep}(\theta'^{(k+1)}) = \begin{cases} 1; \quad if \quad O(\theta'^{(k)}) \geq O(\theta'^{(k+1)}); \\ \exp \frac{(O(\theta'^{(k)}) - O(\theta'^{(k+1)}))}{t_k}; \quad if \quad O(\theta'^{(k)}) < O(\theta'^{(k+1)}); \end{cases} \tag{4.3}$$

In the PLMR context, the temperature t_k is a relative perturbation on the parameter defined by sampling a uniform distribution centered at the desired perturbation level. It starts centered at 7%, i.e $p = U(-7\%, 7\%)$. At each iteration, the parameters will be multiplied by sampled p values, the objective function updated and compared to the current best solution. The perturbation is decreased following the linear multiplicative annealing schedule:

$$t_{k+1} = \frac{t_0}{1 + ak} \tag{4.4}$$

Choosing a cooling schedule can be a sensitive choice depending on the size and complexity of the optimization problem. The goal when choosing the schedule is to ensure convergence while maintaining a reasonable computing time. The schedule shown in Equation 4.4 is used to make sure the temperature converges to a single value after all the iterations. Constant a changes the rate that the temperature decay over the iterations, larger values makes it decay faster and converge to a smaller temperature at the end of the algorithm. The value of the constant is chosen so the perturbations converge to $p = U(-3\%, 3\%)$ over a predefined number of iterations, usually set around 500 to 800. Perturbation values below this rate are too small for the scale of the problem. The temperature is changed 2 times during the fitting process.

Once the algorithm runs for a pre-defined number of iterations, the best solution found in the run is returned. The random acceptance chance showed in Equation 4.5 helps to avoid the algorithm getting stuck in a far from optimal solution. This may happen when the Gaussian factor ranges are similar.

The automatic fitting procedure is also useful to access if the data is prone to be modeled with a PLMR. When the two Gaussian factors are defined with the same variogram type, the model has the property to converge to a conventional LMR when one, or both, of the conditions happen:

$$S_1 \approx S_2 \approx 0.5$$
$$r_{\gamma_1} \approx r_{\gamma_2} \tag{4.5}$$

When S_1 and S_2 are ≈ 0.5 it indicates equal contributions of each structure to the two subsets of the data. In this case, even if ranges and structure type are different the model is equal to a LMR with equal contribution from each variogram. When the ranges and structures are equal the model will converge to a single variogram structure model. These two scenarios may be captured by the semi-automatic fitting procedure, indicating that the asymmetry in the indicator variograms is not large enough to model with a PLMR. Finally, from the author's experience with real datasets, the algorithm may converge to solutions that are good in terms of the cost function but are geologically unreasonable, e.g extreme large ranges. This scenario seems to be common when trend-like features can be seen in some indicator variograms.

4.2 Fitting the nugget effect

The nugget effect of a PLMR model is currently being treated as coming from the Gaussian Factor's realizations. In practice, this means that the nugget effect of each variogram will be a consequence of the piecewise linear mixing defined by the model and the inferred nugget effect of each Gaussian Factor. Hence, after being inferred, each Gaussian Factor will be simulated using the fitted nugget value. In this framework, it is not possible to control exactly the behavior of lag 0 for each variogram used to tune the model, but the nugget values fitted are consistent with the model. The semi-automatic fitting of the nugget effect is done in a step before the optimization of the remaining parameters. A nugget effect goal for each variogram used in fitting is specified by the user. The program proceeds, similarly to what was described in the previous section, to try different solutions and keep the one that minimizes the objective function. In this step, the function to be minimized can be written as:

$$O(nugg) = \sum_{d=1}^{n_{dim}} \sum_{l=1}^{4} (\gamma'_{dl}(0) - \gamma_{dl}(0))^2 \tag{4.6}$$

The equation is the same as Equation 4.2, but the squared error is computed only for $\mathbf{h} = 0$.

4.3 Setting up a initial guess for semi-automatic fitting

A general rule of thumb on how to set up an initial guess for PLMR parameters will be outlined in this section. The procedure will be described by modeling Cadmium (Cd) measurements in the Jura dataset. Figure 4.2 shows Traditional and Indicator Variograms of the Normal Score (NS) Cadmium measurements.

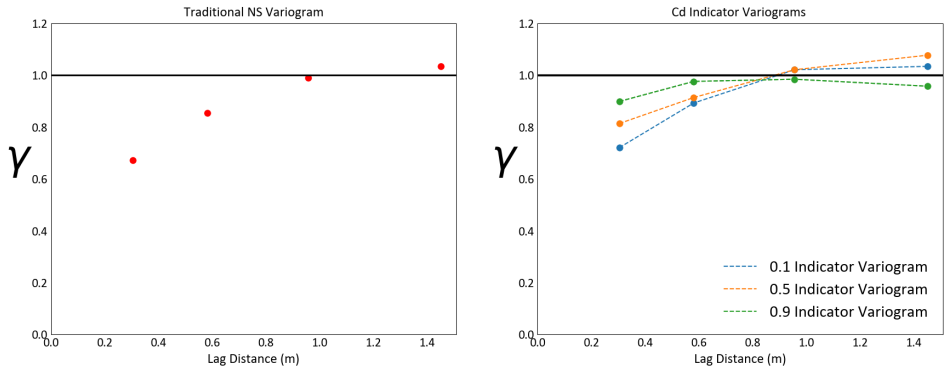

Figure 4.2: Empirical NS Cadmium variograms for the Swiss Jura dataset

Dots were connected in the Indicator Variograms to highlight the non-Gaussian asymmetrical behavior. The apparent range of the 0.1 and 0.5 variograms are close to 1 and they look fairly similar. The 0.9 threshold Variogram has a range of around 0.4 units. Traditional Variogram of the normal-scored transformed variable shows a range around 1. These apparent ranges will be set up as an initial guess. Initial slopes guesses will be $S_1 = 0.1$ and $S_2 = 0.9$ due to high asymmetry between 0.1 and 0.9 indicator variograms. The model will be built with two variograms of Spherical Variogram structures. Choosing variograms types is subjective but they should be able to reasonably fit all variograms used for tuning. The semi-automatic fitting was run and the range of the first Gaussian factor was slightly modified to provide a better fitting of the 0.9 threshold. Both sets of parameters are shown in table 4.1 and fitted variograms are shown in Figures 4.3 and 4.4.

Table 4.1: PLMR parameters for Cd measurements in the Jura data set

	Parameters from Semi-Automatic Fitting	Final Chosen Parameters
r_{γ_1}	0.40	0.45
r_{γ_2}	1.10	1.10
S_1	0.11	0.11
S_2	0.89	0.89

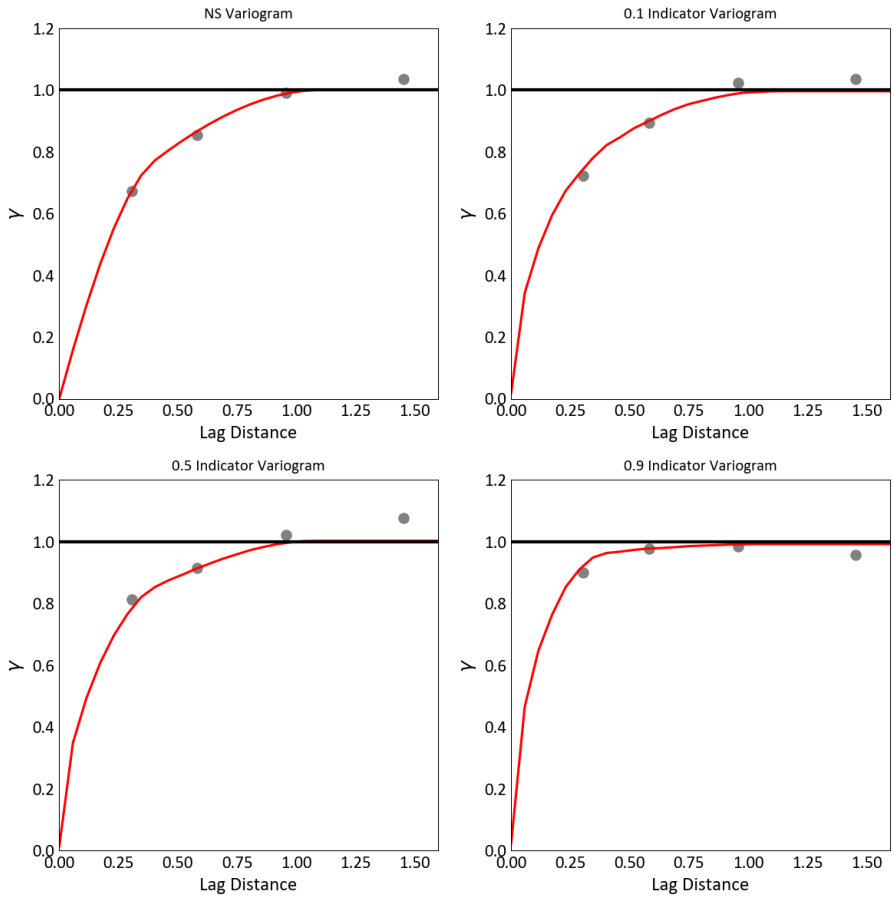

Figure 4.3: Fitted variograms - using values outputted from the optimization procedure.

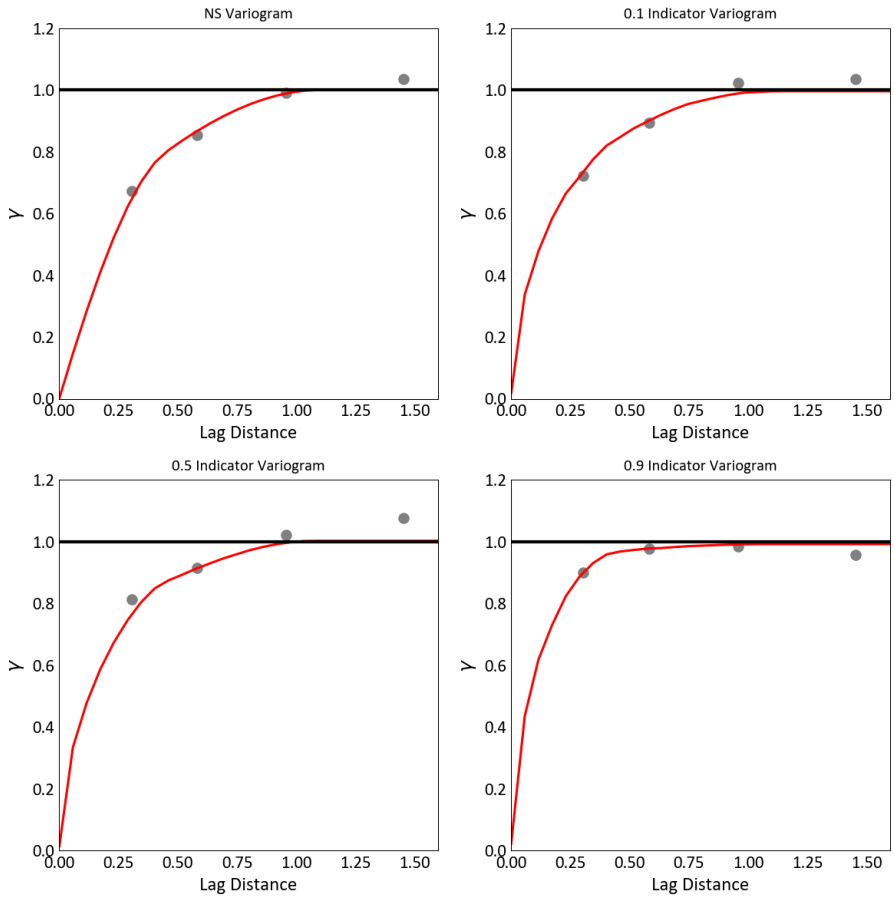

Figure 4.4: Fitted variograms with minor alteration in the range of the second Gaussian factor

The procedure used for the Jura data can be seen as a rule of thumb for setting up an initial guess. The two Gaussian factor are set to match the apparent ranges of the 0.1 and 0.9 indicator variograms. The first slope guess (S_1) is set to 0.1 when considerable asymmetry around the 0.5 threshold is present. If no asymmetrical behavior is seen, this value can be set higher, but it is worth noting that when the indicator variogram transition starts to become symmetric using a PLMR to model the variable must be evaluated.

4.4 Synthetic example and robustness to fluctuations in parameters

When the semi-automatic fitting procedure described above is run multiple times for a similar problem it is possible that results from each run will converge to slightly different parameters. This behavior arises from the stochastic nature of the optimization procedure. The perturbations p sampled from a uniform distribution can be thought of as Markov Chain. The memory of the process comes from the fact that each local search is a controlled perturbation centered on the current parameter estimate. Fluctuations in the chain may lead to slightly different final results. Also, near global optimum parameters, changes in the model become harder to accept due to the similarity of the evaluated objective function.

Fluctuation is more present in the final slopes S_1 and S_2 than in the ranges. This is demonstrated in a small synthetic example. Semi-automatic fitting is run 100 times and the solutions stored. Figure 4.5 shows a swarm plot of the relative difference between reference parameters and fitted ones, and Figure 4.6 the resulting PLMR variograms. The seemingly continuous grey line is the combination of the 100 resulting variograms plotted together, hence its width is proportional to the maximum fluctuation in each resulting variogram. This example was set up purposefully so the optimized parameters would not exactly match the reference ones. Even though there is a high deviation in the PLMR slopes, variograms still show a good visual fit. This happens because parameters interact in a non-linear fashion meaning that different solutions may yield equally good fits. Also, from the author's experience modeling real datasets with the model, small fluctuations in the PLMR slopes should not cause large deviations in the results. Inversely weighting the objective function by the variogram lag helps to stabilize results. Robustness to variations in the variogram ranges is similar to simulation with SGS. A subjective understanding of the data and geology define the acceptability of the final results.

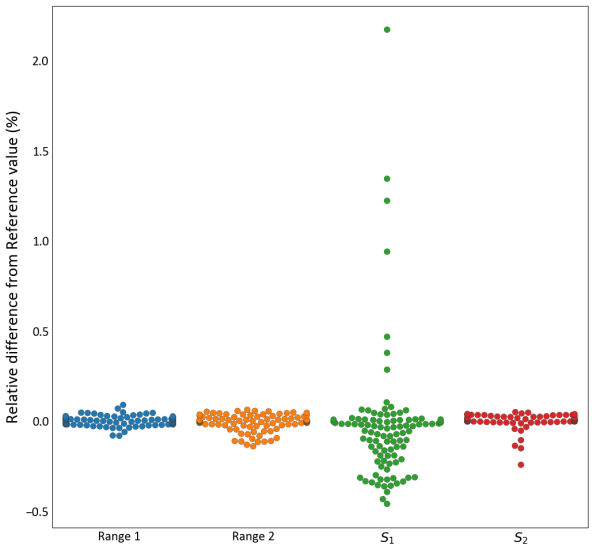

Figure 4.5: Swarm plot of the optimized parameters

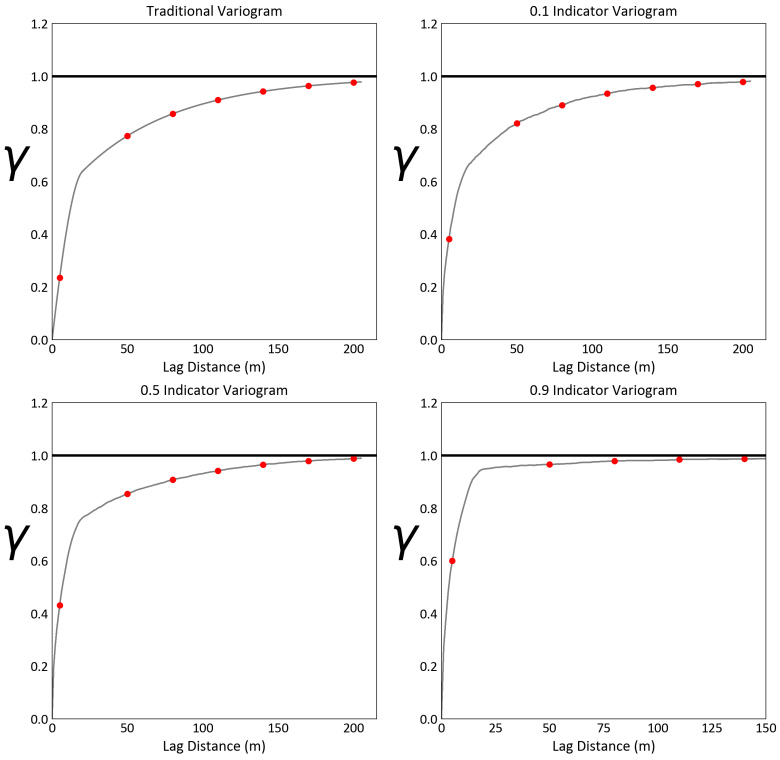

Figure 4.6: Resulting variograms

Chapter 5

Simulating conditional PLMR realizations

This chapter will discuss a methodology to simulate conditional realizations using a piecewise linear model of regionalization.

Conditional simulation is widely used to obtain models of uncertainty that reflect the true spatial variability of the variable being studied. Based on a second-order stationary Random Function model $Z(\mathbf{u})$, an unconditional realization is a realization from the set of all possible values of $Z(\mathbf{u})$ (Chiles & Delfiner, 2009).

For a correct assessment of uncertainty, the set of possible values of $Z(\mathbf{u})$ should be consistent with the conditioning data and previously simulated locations. This is achieved by the use of conditional simulation (spatially consistent Monte Carlo simulation), i.e. realizations that are randomly drawn from the subset of realizations that match the sample points (Chiles & Delfiner, 2009). Geostatistical conditional simulation algorithms are used to build models that reproduce global histogram, measures of spatial continuity such as the variogram and the variable represented by the conditioning data (Rossi & Deutsch, 2013).

Generating unconditional realizations of a PLMR model is easily achieved by simulating the Gaussian factors unconditionally using any Gaussian simulation technique, e.g., SGS, turning bands or spectral simulation, and applying the model's piecewise linear transform to the realizations. Conditional realizations using a PLMR are achieved by decomposing the model into its Gaussian factor's, simulating them and reapplying the piecewise linear transformation in the realizations.

5.1 Direct simulation

To directly generate conditional realizations, sequential simulation algorithms are commonly used in geostatistical workflows (Rossi & Deutsch, 2013). Sequential simulation algorithms work by factorizing the joint CDF of multiple random variables $Z(\mathbf{u})$, $i = 1, ..., K$, where K is the total number of random variables. Recalling the ccdf of a set of random variables:

$$F(Z(\mathbf{u}_1), \ldots, Z(\mathbf{u}_K)|(n) = \text{Prob}(Z(\mathbf{u}_i) \leq z_i; i = 1, \ldots, K|(n)) \tag{5.1}$$

The Sequential Simulation approach factorizes Equation 5.1 making the simulation a series of sequential inference about K univariate ccdf (C. V. Deutsch & Journel, 1998):

$$\begin{aligned} &\text{Prob}\left\{Z(\mathbf{u}_1) \leq z_1 \mid (n)\right\} \\ &\text{Prob}\left\{Z(\mathbf{u}_2) \leq z_2 \mid (n+1)\right\} \\ &\vdots \\ &\text{Prob}\left\{Z(\mathbf{u}_K) \leq z_K \mid (n+K-1)\right\} \end{aligned} \tag{5.2}$$

Under this framework, realizations of the ccdf at each location K will be drawn sequentially. The set of conditioning data (n) is composed of hard-data and previously simulated locations inside a neighborhood of the node being simulated. Calculation of the univariate is usually done in a non-parametric framework, i.e., using indicator techniques or using a parametric model such as the multiGaussian model. In the latter, the distributions would be simply parameterized by a mean, defined using Simple Kriging, and a variance defined by the kriging variance.

5.2 Conditioning by Kriging (CBK)

Instead of sampling from ccdfs in a sequential fashion, it is also possible to sample unconditional realizations of the RF model and then condition them on sampled data. The classical two-step conditioning approach would be to simulate unconditionally using Gaussian algorithms, e.g. Turning Bands (Chiles & Delfiner, 2009; Mantoglou & Wilson, 1982), and condition using Simple Kriging (Chiles & Delfiner, 2009; C. V. Deutsch & Journel, 1998; A. G. Journel & Huijbregts, 1978). When an unconditional realization is generated based on an estimated variogram, the global statistics of the random function being modeled are preserved up to fluctuations, however, the data sampled will not be reproduced and local features of the data are not going to be correctly placed in the final realization. Hence, since the unconditional map only reflects the global information provided by the variable, the process of conditioning will update the map and ensure that simulation matches and reflects the information from sample points (Chiles & Delfiner, 2009). Doing that, conditional statistics of the map will be representative of the underlying phenomena. Consider the RF $Z(\mathbf{u})$ as the sum of the estimator and the corresponding error (C. V. Deutsch & Journel, 1998):

$$Z(\mathbf{u}) = Z_{sk}(\mathbf{u}) + e(\mathbf{u}) \tag{5.3}$$

Where $Z_{sk}(\mathbf{u})$ is the Simple Kriging Estimator. The term $e(\mathbf{u})$ is simply $Z(\mathbf{u}) - Z_{sk}(\mathbf{u})$. Substituting $e(\mathbf{u})$ in equation 5.3:

$$Z(\mathbf{u}) = Z_{sk}(\mathbf{u}) + [Z(\mathbf{u}) - Z_{sk}(\mathbf{u})] \tag{5.4}$$

Therefore, to generate conditional realizations of a given random function (RF) that restore complete variance of the model, it is necessary to simulate the error term and add it to the values obtained from Kriging using the original data. If the spatial characteristics of the term $e(\mathbf{u})$ were

known, a direct simulation would be possible (C. V. Deutsch & Journel, 1998). Since this is not the case, error realizations will be constructed based on unconditional realizations of the RF. In the data locations $\mathbf{u}$: $Z_{sk}(\mathbf{u}) = Z(\mathbf{u})$, due to the exactitude property of Kriging.

When conditioning a PLMR realization by Kriging, hard-data will be reproduced and the map will represent local features of the domain. However, there might be some level of multiGaussian contamination in the results. Under the MG model, the Simple Kriging equations are equal to the normal equations (Leuangthong et al., 2011), indicator variograms show symmetric behavior around the 0.5 threshold and ccdfs are fully parametrized by its two first moments. Hence, simply conditioning the PLMR realizations using Simple Kriging is similar to assuming that the bivariate distribution between two locations is bivariate Gaussian. Under this assumption, the SK estimate and variance are respectively the conditional mean ($m_{Z_c}(\mathbf{u})$) and variance ($\sigma^2_{Z_c}(\mathbf{u})$) of the ccdf at location $\mathbf{u}$:

$$m_{Z_c}(\mathbf{u}) = Z_{sk}(\mathbf{u}) = \sum_{\alpha=1} \lambda_\alpha(\mathbf{u}) Z\left(\mathbf{u}_\alpha\right) \tag{5.5}$$

$$\sigma^2_{Z_c}(\mathbf{u}) = \sigma^2_{SK}(\mathbf{u}) = 1 - \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}) C\left(\mathbf{u} - \mathbf{u}_\alpha\right) \tag{5.6}$$

Conditioning an unconditional PLMR realization by Simple Kriging will make the resulting ccdf closer to the multiGaussian model. When the node being estimated is almost in a hard-data location, results will be similar to the expected conditional distribution. However, when the distance from a sample increases and is still smaller than the greatest Gaussian factor range, conditioning results will be contaminated by multiGaussian behavior. Besides being a function of the relative distance to data locations, Gaussian contamination is also a function of the value sampled at a location. Conditional covariance between two locations $\mathbf{u}$, $\mathbf{u}'$ in a stationary domain may be written as (Hadavand & Deutsch, 2020):

$$C(Z(\mathbf{u}), Z(\mathbf{u}') = C\left(\mathbf{u} - \mathbf{u}'\right) - \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha(\mathbf{u}) \lambda_\beta\left(\mathbf{u}'\right) C\left(\mathbf{u}_\alpha - \mathbf{u}_\beta\right) \quad \forall \mathbf{u}, \mathbf{u}' \tag{5.7}$$

The PLMR, due to asymmetrical indicator variograms, has data-dependent variograms, i.e the spatial continuity is a function of the data value. As mentioned before, using the MG model will entail the same pattern of spatial continuity around the median. As shown in Equation 5.7, applying CBK directly entails a conditional covariance that is data value independent. Hence, conditioning to extreme values of the PLMR distribution will accentuate the degree of Gaussian contamination.

This chapter will go over a conditioning algorithm proposed for the PLMR model. The main idea of the method is to decompose the model into its latent Gaussian factors in a way that honors the model piecewise linear transform, spatial auto-correlation and hard-data. This is achieved by imputing each factor in the data locations. The imputed values are then simulated using SGS,

and the piecewise linear transformation defined by the model's parameters is reapplied. Finally, it will be shown how this method reproduces indicator and traditional variograms and honors the asymmetry of non-Gaussian random variables.

5.3 PLMR transformation

Geostatistical workflows require building a representative distribution of the variable being modeled. This step starts with data visualization, cleaning, exploratory data analysis and may end with data processing steps like declustering and despiking (Rossi & Deutsch, 2013). Clustered samples, usually in areas of interest like high-grade zones, may lead to incorrect estimation of global histogram, local proportions, and uncertainty estimation. Spikes in the variable histogram, i.e multiple records with the same value, will lead to artifacts when data transformation like the Normal Score Transform is needed.

After a representative histogram is obtained, the variable under study should be transformed to the distribution defined by the PLMR parameters. Let's define a set of samples in a stationary domain V as $\{z_{emp}(\mathbf{u}); u \in V\}$ and it's representative CDF as $F_{rep}(z_{emp})$. A Quantile-Quantile (Q-Q) transformation will be applied to $F_{rep}(z_{emp})$ to transform it to the correct PLMR distribution. Even though the global distribution of a PLMR is fairly well behaved depending on how the model is defined, it is not a normal one. Hence, proceeding with simulation using NS values will lead to bias in the histogram once realizations of the Gaussian Factors are rebuilt into the final realizations in original units. Calling the PLMR cumulative distribution as $F_{plmr}(z)$, the transformation can be written as:

$$z = F_{plmr}^{-1}(F_{rep}(z_{emp})) \tag{5.8}$$

Since $F_{rep}(z_{emp})$ does not have an analytical representation, the samples are ordered and a cumulative probability assigned to it. For example, the value ranked at position q could have an empirical cumulative frequency of $\frac{q}{n}$, where n is the total number of sampled locations. Linear interpolation is used to extrapolate the CDF to unsampled values. Further details of Q-Q transformations and how to treat declustering weights, boundary probabilities, and back-transforms can be found in (C. V. Deutsch & Journel, 1998). The PLMR CDF, i.e. $F_{plmr}(z)$, is obtained by simulating independently two normal distributions and applying the transformation using the parameters inferred from the semi-automatic fitting procedure. Figure 5.1 shows a representative distribution obtained from the Jura dataset and a graphical representation of its transformation to PLMR units.

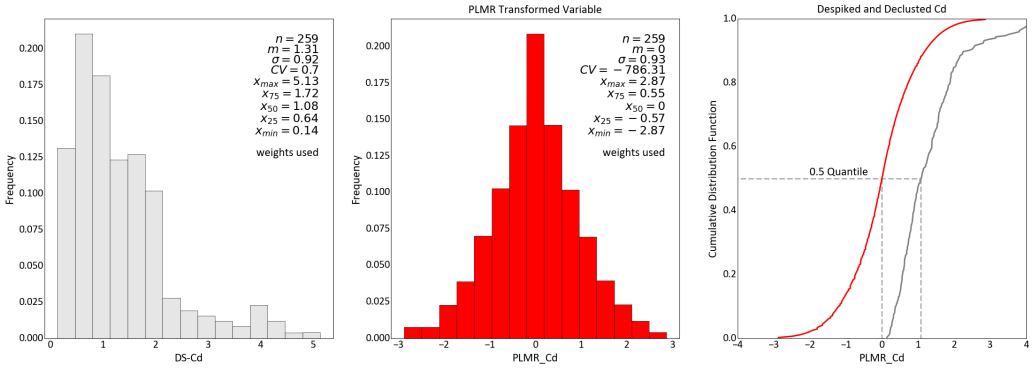

Figure 5.1: PLMR transformation example - CDF plots follows same color coding as histograms

5.4 Imputation of Gaussian factors

As mentioned before, naive use of the two-step Conditioning by Simple Kriging method adds undesired multiGaussian behavior to the final realizations. A methodology where the Gaussian factors are imputed at each data location and then simulated using SGS is proposed to mitigate this problem. The imputation framework aims to reconstruct the Gaussian factors from the information provided by the PLMR and sampled values. The methodology can be divided into different sub-problems:

1. Capture the relationship between Gaussian factors and the PLMR model, i.e., honoring the piecewise linear transform.
2. Reproduce each Gaussian factor's spatial behavior by honoring its variogram structure.
3. Ensure data reproduction.
4. Simulate the imputed values at each data location and re-apply the piecewise linear transform at each imputed Gaussian factor realization.

The imputation is done at each location where the attribute value is sampled. To make the text less cumbersome, when no distinction between the two Gaussian factor's is necessary, both distributions will be referred to as $Y(\mathbf{u})$. This means that the operations described should be conducted independently for each factor. Unless otherwise noted, the PLMR random function will be called $Z(\mathbf{u})$.

Given a random sequence to visit all nodes in the grid being simulated the procedure locally combines information from the distribution of $Y(\mathbf{u})|Z(\mathbf{u})$, simulated based on the model's parameter, and the spatial distribution of the factors given the previously imputed values, i.e $Y(\mathbf{u})|(n))$. The final goal of the imputation methodology is to sample from the following distribution:

$$\text{Prob}\left(Y(\mathbf{u}) \mid Z(\mathbf{u}), (n)\right) \tag{5.9}$$

The two pieces of information are combined using non-parametric Bayesian updating, and assuming conditional independence between $Y(\mathbf{u})|Z(\mathbf{u})$ and $Y(\mathbf{u})|(n)$:

$$\text{Prob}\left(Y(\mathbf{u}) \mid Z(\mathbf{u}), (n)\right) = \frac{\text{Prob}\left(Y(\mathbf{u})\right)\text{Prob}\left(Z(\mathbf{u}) \mid Y(\mathbf{u})\right)\text{Prob}\left((n) \mid Y(\mathbf{u}), Z(\mathbf{u})\right)}{\text{Prob}\left(Z(\mathbf{u}), (n)\right)} \tag{5.10}$$

$$\text{Prob}\left(Y(\mathbf{u}) \mid Z(\mathbf{u})\right) = \frac{\text{Prob}\left(Y(\mathbf{u}) \mid Z(\mathbf{u})\right)\text{Prob}\left(Y(\mathbf{u}) \mid (n)\right)}{\text{Prob}\left(Y(\mathbf{u})\right)} \tag{5.11}$$

The updated distribution is then sampled, and a realization of one of the Gaussian factors is generated. The other factor is computed afterward to ensure data reproduction. Each aspect of the methodology will be further explained in the following sections.

5.5 Distribution of the factors conditioned on sampled data

The distribution of the factors given a PLMR value is accessed by simulating a large number of samples from an independent normal distribution, computing the piecewise linear transformation and the final PLMR value. The outcome is the joint distribution $\text{Prob}(Y_1(\mathbf{u}), Y_2(\mathbf{u}), Z(\mathbf{u}))$, i.e. a table that stores the two independent normal realizations and the PLMR values that arise by applying the piecewise linear transformation at each pair of $y_1(\mathbf{u})$, $y_2(\mathbf{u})$. The lower case notation means a realization of the parent random function. The goal of this first Monte Carlo simulation is to model the distribution of $\text{Prob}(Y_1(\mathbf{u})|Z(\mathbf{u}))$ and $\text{Prob}(Y_2(\mathbf{u})|Z(\mathbf{u}))$. It is important to note that, for a realization of one of the factors, a given PLMR value z can be calculated by several values of the second factor.

This step is mapping the normal deviates into the PLMR space by a numerical procedure. The bivariate distributions $\text{Prob}(Y_1(\mathbf{u}), Z(\mathbf{u}))$ and $\text{Prob}(Y_2(\mathbf{u}), Z(\mathbf{u}))$ are sensitive to the quality of the sampling in Gaussian space. Hence, similarly to the procedure described to simulate PLMR variograms, PPMT (Barnett et al., 2014) is being used to pre-process the Monte Carlo samples. Figure 5.2 shows the scatter plot of the simulated factors and calculated PLMR values.

In case the truncation point is set to be different than zero, the distribution of $Y(\mathbf{u})|Z(\mathbf{u})$ will have a discontinuity. This is illustrated in Figure 5.3 where T_p is set to be the 0.1 quantile of a normal distribution, i.e. $T_p = -1.28$.

The bivariate relationships between the PLMR and Gaussian factors values depend on the parameters used to define the model. When both squared slopes equal 0.5, the scattered points will follow a Gaussian ellipse. As mentioned before, this model is equivalent to a LMR with equal contribution from both factors. Hence, the conditional expectation and variance would be:

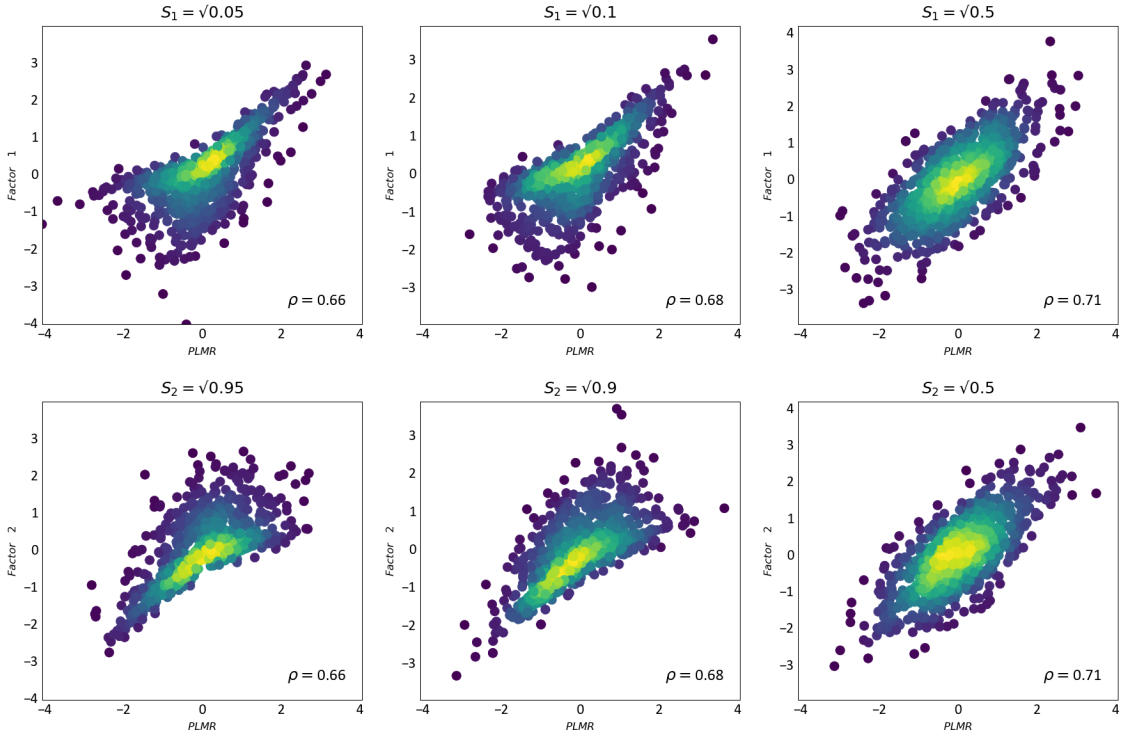

Figure 5.2: Bivariate scatter-grams of $\mathrm{Prob}(Y_1(\mathbf{u}), Z(\mathbf{u}))$ and $\mathrm{Prob}(Y_2(\mathbf{u}), Z(\mathbf{u}))$

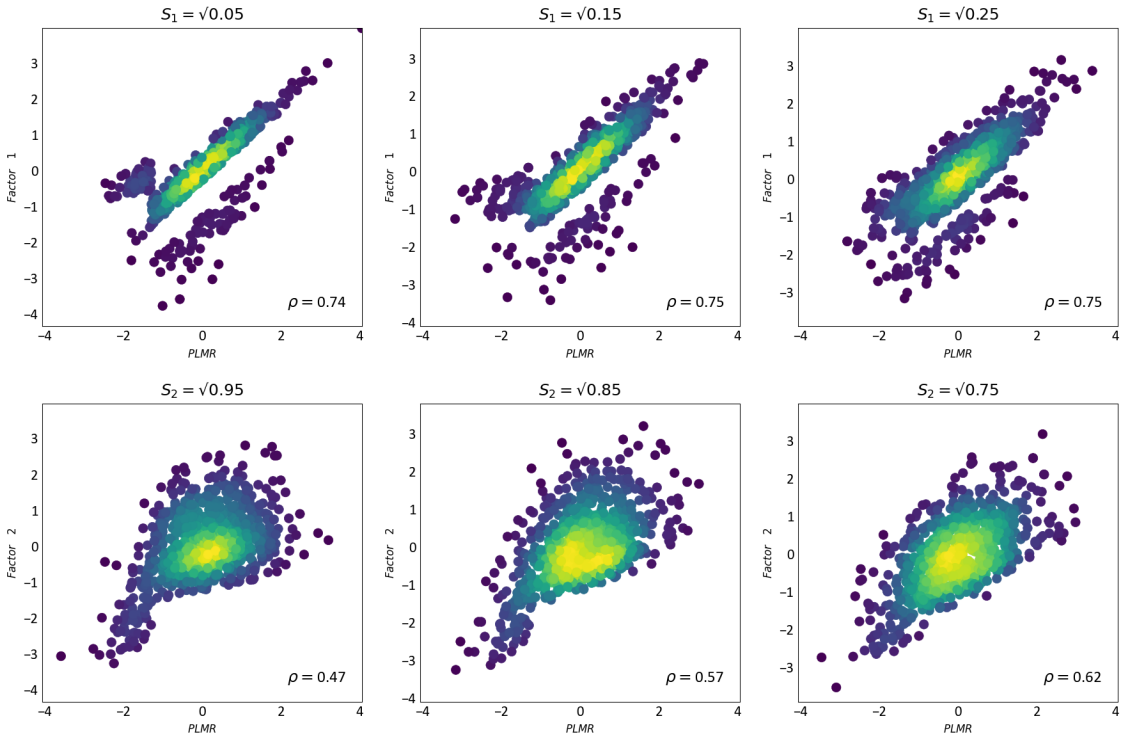

Figure 5.3: Bivariate scatter-grams of $\mathrm{Prob}(Y_1(\mathbf{u}), Z(\mathbf{u}))$ and $\mathrm{Prob}(Y_2(\mathbf{u}), Z(\mathbf{u}))$ with $T_p = -1.28$

$$E\{Y(\mathbf{u})|Z(\mathbf{u})\} = m_Y(\mathbf{u}) + \rho\frac{\sigma_Y(\mathbf{u})}{\sigma_Z(\mathbf{u})}(Z(\mathbf{u}) - m_Z(\mathbf{u})) \tag{5.12}$$

$$Var\{Y(\mathbf{u})|Z(\mathbf{u})\} = \sigma_Y(\mathbf{u})(1 - \rho^2_{Y(\mathbf{u}),Z(\mathbf{u})}) \tag{5.13}$$

Where $\rho_{Y(\mathbf{u}),Z(\mathbf{u})}$ is the correlation between the Gaussian factors and the PLMR values. However, for modeling continuous geological variables with non-Gaussian behavior, the most useful set of parameters are the ones where on squared slopes are significantly different. This highlights the spatial continuity of each factor in different ranges of the data, making it possible to define models with asymmetric indicator variograms. Figure 5.2 shows that the complexity of the bivariate shape increases as the slopes diverges and the conditional moments become progressively less linear.

To move on with the imputation framework, it is necessary to sample the distribution of $Y(\mathbf{u})|Z(\mathbf{u}) = z$. This is straightforward to implement in the presence of equal contributions since the conditional mean is linear, the conditional variance constant, and the conditional shape is Gaussian. However, in the case of different slopes, the non-linearity becomes pronounced, and the fitting assuming a linear relationship is unsatisfactory. A simple solution for this issue would be to calculate the conditional mean and variance non-parametrically but still assuming that the shape of $Y(\mathbf{u})|Z(\mathbf{u}) = z$ is Gaussian. Hence, this approach continues by calculating the conditional statistics using a Moving Window Search and using the estimated values as input in the MCS simulation. Figure 5.4 shows the non-parametric estimation of the two conditional moments.

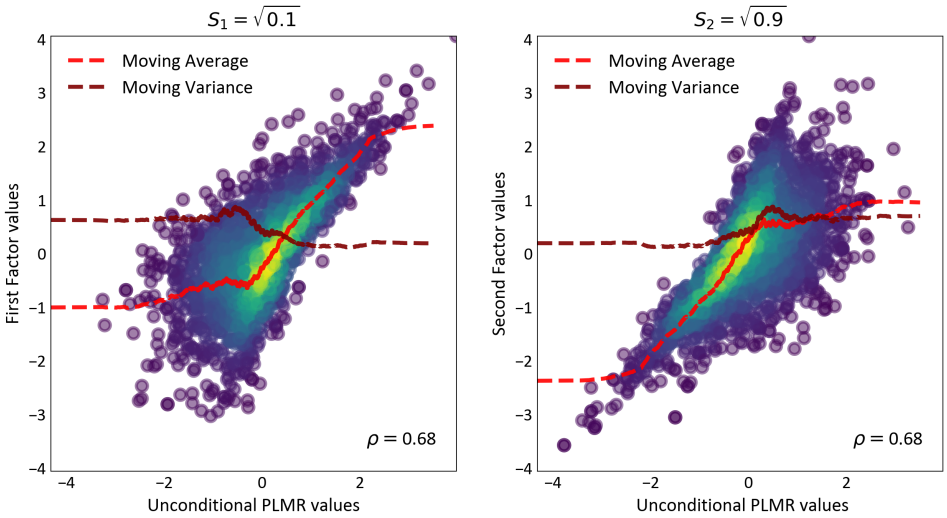

Figure 5.4: Moving average $E\{Y(\mathbf{u})|Z(\mathbf{u}) = z\}$ and variance $Var\{Y(\mathbf{u})|Z(\mathbf{u}) = z\}$ fitted to the distribution.

The slopes on top of the images in Figure 5.4 are the contribution of that factor to low values of the PLMR. For example, in the region where the factor's samples are multiplied by the smallest slope, $\sqrt{0.1}$, the cloud becomes more dispersed, the variance increases and the mean stays relatively

constant. On the other hand, when multiplied by $\sqrt{0.9}$, the PLMR values become more correlated with the Gaussian Factors, the mean increase or decrease almost linearly, and the variance is small.

The conditional statistics can be reasonably estimated using a Moving Window in the central part of the distribution. However, to have reliable samples, it is necessary to determine how well a Gaussian shape approximates the conditional distributions with the estimated conditional mean and variance. Figure 5.5 compares the distribution computed using Moving Window+MCS and the reference one from rejection sampling. Histograms and CDFs show how the results diverge from the reference distribution depending on the region of the bivariate distribution the values are being drawn from. When sampling from the more correlated part of the cloud, e.g, conditioning on a low PLMR value the second factor, and to a high value the first one, there is a significant bias in the mean, variance, and the reference histogram is considerably more skewed. However, when sampling from the more dispersed zone of the distribution, the mismatch decreases, but the variance is high.

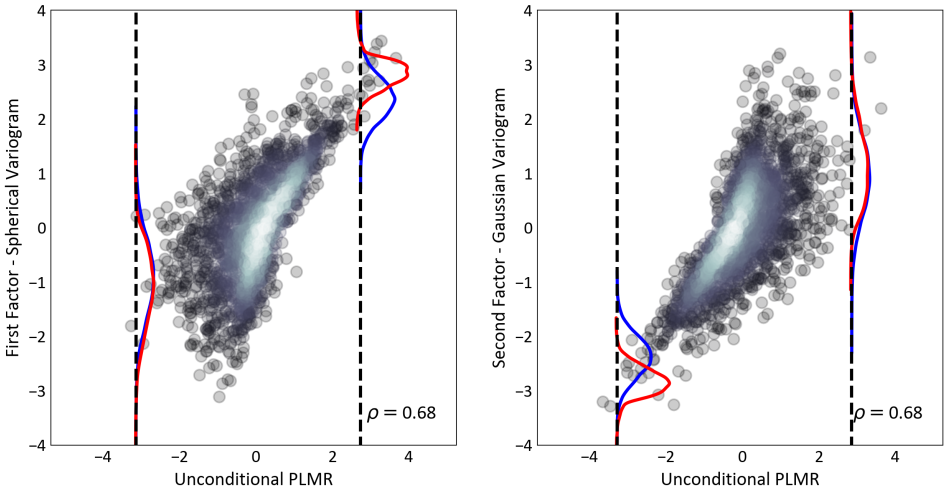

Figure 5.5: Conditional densities plotted on the bivariate space formed by the two distributions. Red is the reference CDF and blue is the one estimated with the non-parametric MCS procedure.

The mismatch between the estimated and reference conditional distributions, especially in the first central moment, may lead to bias in the histogram reproduction when the whole PLMR workflow is conducted. Hence, it is necessary to infer the conditional distributions $Y(\mathbf{u})|Z(\mathbf{u}) = z$ more reliably. Using a moving window to infer the mean presented problems when estimating highly correlated tail values and the conditional distributions are too skewed to be approximated by a Gaussian shape.

5.6 Fitting conditional distributions using Gaussian mixture models

Modeling the conditional distribution $Y(\mathbf{u})|Z(\mathbf{u}) = z$ by the non-parametric approach described above should be avoided. It is necessary to use a flexible method to capture the pronounced non-linear behavior that arises when one of the slopes approaches zero. In the imputation of geological variables literature, methods like Kernel Density Estimation (KDE) and Gibbs Sampler have been used before (Barnett & Deutsch, 2015). They were proposed as a tool to sample the secondary variable distribution, i.e., the distribution of the first variable being estimated given the outcomes of collocated secondary data. This approach is different than a full parametric one, where the conditional mean and variance are calculated using the normal equations.

Instead of using KDE or a Gibbs sampler, two techniques that can become computationally infeasible fairly quick, Silva and Deutsch (2018) proposed using a Gaussian Mixture Model (GMM) to fit the collocated secondary variable distribution. In their framework, the secondary distribution is modeled using a Gaussian Mixture Model and merged using a non-parametric Bayesian updating (Neufeld & Deutsch, 2006). Let each element of an arbitrary sample set $\mathbf{x} = ({x_1}^T, {x_2}^T, \ldots, {x_n}^T)$ represent a multi-dimensional vector. The estimated density using g Gaussian components can be expressed as:

$$f(\mathbf{x}, \Psi) = \sum_{i=1}^{g} \pi_i \phi(\mathbf{x}; m_i, \Sigma) \tag{5.14}$$

Where:

$$\begin{gathered} 0 \leq \pi_i \leq 1 \quad (i = 1, \ldots, g) \\ \sum_{i=1}^{g} \pi_i = 1 \\ \Psi = (\pi_1, \pi_2, \ldots, \pi_g, m_1, m_2, \ldots, m_3, \Sigma_1, \Sigma_2, \ldots, \Sigma_g) \end{gathered} \tag{5.15}$$

The mean, m_i, and covariance matrix, Σ_i, defines the $i = 1, \ldots, g$ Gaussian components densities present in the final mixture model. The mixing weights π_i controls the contribution of each component density to the mixture model. Therefore Equation 5.14 defines a valid probability density function (McLachlan, Lee, & Rathnayake, 2019). This PDF defined by the the GMM parameter vector Ψ can have highly non-Gaussian and non-symmetrical behavior. Figure 5.6 shows the flexibility of the method. In this example, $\Psi = (\pi_{1i}, \pi_{2i} = 1 - \pi_{1i}, m_1 = -1, m_2 = 1, \Sigma_1 = 1, \Sigma_2 = 1)$ where $\pi_{1i} = 0.1, 0.5, 0.9$.

Even in the simple homoscedastic case, i.e., all components share the same variance, example showed in Figure 5.6, it is possible to see that the methodology can model distributions that are non-symmetrical and non-Gaussian. Using more Gaussian components and different sets of parameters, it is possible to model PDFs that are multimodal, highly skewed, and disperse(McLachlan et al.,

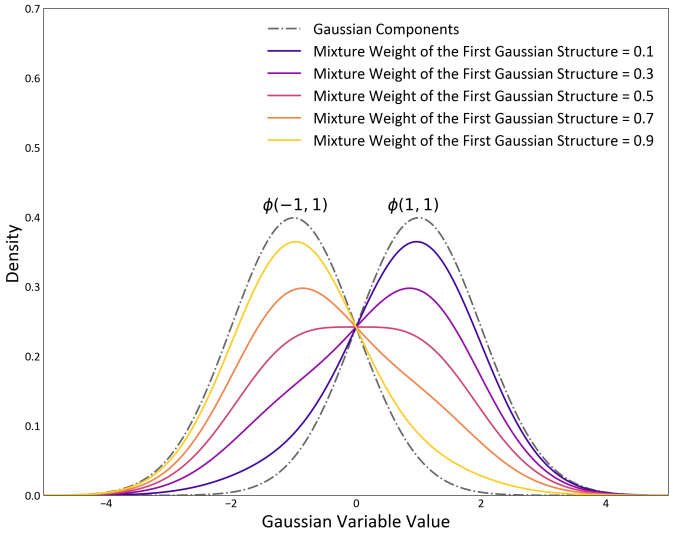

Figure 5.6: Different mixture distributions obtained using the same Gaussian components.

2019). In the context of the PLMR conditioning, the number of Gaussian Kernels is usually set around 4. As shown in Appendix A, the model can be seen as a mixture model between the 4 finite distributions that arise when applying the piecewise linear transformation. Hence, using a number close to that has some theoretical reasoning behind it.

The fitting of the GMM is done by Expectation-Maximization (EM). This method is well documented and applied to fit the model in several different areas. The method will not be explained in detail, interested readers are referred to (Friedman, Hastie, Tibshirani, et al., 2001; McLachlan et al., 2019). Giving the model equation 5.14, the log-likelihood $logL(\Psi)$ can be calculated giving a set of observations as (Silva & Deutsch, 2018):

$$\log L(\Psi) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i \phi \left(\mathbf{x}; m_i; \Sigma_i \right) \right\} \tag{5.16}$$

The EM works by maximizing the log-likelihood in two different steps. The central idea of the method is that each data comes from by one of the Gaussian Kernels. Since this information is unknown, the problem is formulated treating the label of each sample as missing data (McLachlan et al., 2019). In the first step in the algorithm, Expectation step (E-step), the label assignments' expected value is inferred based on the current set of parameters Ψ^t and the observed samples. Each data is assigned to a Kernel of the Gaussian Mixture Model at the end of this step. Giving the current assignment, a new set of parameters Ψ^{t+1} is calculated. This is the Maximization step. The algorithm ends when the difference between $logL(\Psi^t)$ and $logL(\Psi^{t-1})$ is good enough for the proposed application.

5.7 Imputing Gaussian factors

To achieve the goals outlined in Equation 5.4, i.e., respecting the piecewise linear transformation transformation and variogram of the Gaussian factors, the GMM should be used to infer the conditional distribution of $Y(\mathbf{u})|Z(\mathbf{u}) = z$. This conditional distribution calculated from the mixture model are then combined with the spatial information using non-parametric Bayesian updating (Neufeld & Deutsch, 2006). The method is outlined below:

1. Simulate two sets of independent Gaussian variables exhaustively.
2. Apply the piecewise linear transform to the outcome of item 1.
3. Fit a GMM to the Gaussian factors and the PLMR bivariate distribution. In this step, the relation between the model and Gaussian Factors is modeled.
4. Calculate the ccdf at the location being simulated by solving the Simple Kriging equations based on previously simulated values:

$$m_{y_c}(\mathbf{u}) = \sum_{\alpha \in (n)} \lambda_\alpha \cdot y\left(\mathbf{u}_\alpha\right) \tag{5.17}$$

$$\sigma^2_{y_c}(\mathbf{u}) = 1 - \sum_{\beta \in (n)} \lambda_\alpha C\left(\mathbf{u}, \mathbf{u}_\alpha\right) \tag{5.18}$$

$$\sum_{\beta \in (n)} \lambda_\beta C\left(\mathbf{u}_\alpha, \mathbf{u}_\beta\right) = C\left(\mathbf{u}, \mathbf{u}_\alpha\right); \forall \alpha \in (n) \tag{5.19}$$

5. Use the fitted GMM in item 3 to infer the conditional distribution of the factor being imputed given the transformed variable.
6. Combine results from 4 and 5 using non-parametric Bayesian updating (Neufeld & Deutsch, 2006).
7. Sample a realization from the updated distribution calculated in step 6.
8. Repeat steps 4 to 5 location-wise, adding each simulated value as conditioning data.

This methodology is the same algorithm Silva and Deutsch (2018) proposed. However, in the PLMR context, the GMM is fitted to a bivariate distribution that is completely synthetic, and the set of conditioning data when solving the Simple Kriging system is only composed of previously simulated locations.

The Bayesian updating step, i.e., when the spatial and model information are combined, is done by a non-parametric version of the method (Neufeld & Deutsch, 2006). By using a distribution-free

update scheme, no Gaussian relation is assumed between the spatial and factor's conditional distribution. However, there is still an assumption of independence between the merged information. This violates the value-dependence of variograms of the PLMR. In the future, options that use other probability merging schemes like conditional independence or Permanence of Ratios (A. Journel, 2002; Pyrcz & Deutsch, 2014) might be worth testing.

The methodology is applied independently to each Gaussian factor composing the PLMR model. Hence, the imputed values of each factor will reproduce the model's underlying variogram structures.

5.8 Data reproduction and simulation

Applying the methodology described in Section 5.7 is not sufficient to ensure data reproduction after the model is rebuilt. This is because imputation is conducted independently for each Gaussian Factor, so each pair of imputed values $y_1(\mathbf{u})$ and $y_2(\mathbf{u})$ will not converge to the correct PLMR value $z(\mathbf{u})$ once the piecewise linear transformation is applied. Hence, a modification is necessary to ensure data reproduction.

Honoring hard-data is achieved by, at each location $\mathbf{u}$, keeping one of the imputed factors back and calculating the other to match the transformed sampled value. For simplicity, let's suppose it is chosen to keep all imputed samples from the first Gaussian factor, $y_1(\mathbf{u})$, and the values of the second factor, $y_2(\mathbf{u})$, are back-calculated in all locations in the domain. Hence, to ensure data reproduction, it is necessary to invert the model as shown in Equation 5.21:

$$z(\mathbf{u}) = \begin{cases} \sqrt{S_1}y_1(\mathbf{u}) + \sqrt{1-S_1}y_2(\mathbf{u}), & y_1(\mathbf{u}) \leq 0, y_2(\mathbf{u}) \leq 0 \\ \sqrt{S_1}y_1(\mathbf{u}) + \sqrt{S_1}y_2(\mathbf{u}), & y_1(\mathbf{u}) \leq 0, y_2(\mathbf{u}) > 0 \\ \sqrt{1-S_1}y_1(\mathbf{u}) + \sqrt{1-S_1}y_2(\mathbf{u}), & y_1(\mathbf{u}) > 0, y_2(\mathbf{u}) \leq 0 \\ \sqrt{1-S_1}y_1(\mathbf{u}) + \sqrt{S_1}y_2(\mathbf{u}), & y_1(\mathbf{u}) > 0, y_2(\mathbf{u}) > 0 \end{cases} \tag{5.20}$$

$$y_2(\mathbf{u}) = \begin{cases} \frac{z(\mathbf{u})-\sqrt{S_1}y_1(\mathbf{u})}{\sqrt{1-S_1}}, & y_1(\mathbf{u}) \leq 0, y_2(\mathbf{u}) \leq 0 \\ \frac{z(\mathbf{u})-\sqrt{S_1}y_1(\mathbf{u})}{\sqrt{S_1}}, & y_1(\mathbf{u}) \leq 0, y_2(\mathbf{u}) > 0 \\ \frac{z(\mathbf{u})-\sqrt{1-S_1}y_1(\mathbf{u})}{\sqrt{1-S_1}}, & y_1(\mathbf{u}) > 0, y_2(\mathbf{u}) \leq 0 \\ \frac{z(\mathbf{u})-\sqrt{1-S_1}y_1(\mathbf{u})}{\sqrt{S_1}}, & y_1(\mathbf{u}) > 0, y_2(\mathbf{u}) > 0 \end{cases} \tag{5.21}$$

Back-calculating $y_2(\mathbf{u})$ given the PLMR value, i.e. $(z(\mathbf{u}))$, ensures that the data will be reproduced after the piecewise linear transformation is reapplied to the Gaussian Factors. Figure 5.7 shows the histograms of imputed and conditioned factors and how it exactly reproduces the data. It is worth noting, however, that the updated distribution of factor being back-calculated loses some

spatial correlation and approximate to the distribution of $Y(\mathbf{u})|Z(\mathbf{u})$ as highlighted in Figure 5.8. The Figure shows 30 realizations of the two imputed factors in an arbitrary location, where the values from factor 2 are being sampled from the correct updated distribution and values from factor 1 are solely being back-calculated. It can be seen how the distribution back-calculated is closer to the unconditional PLMR distribution in that location.

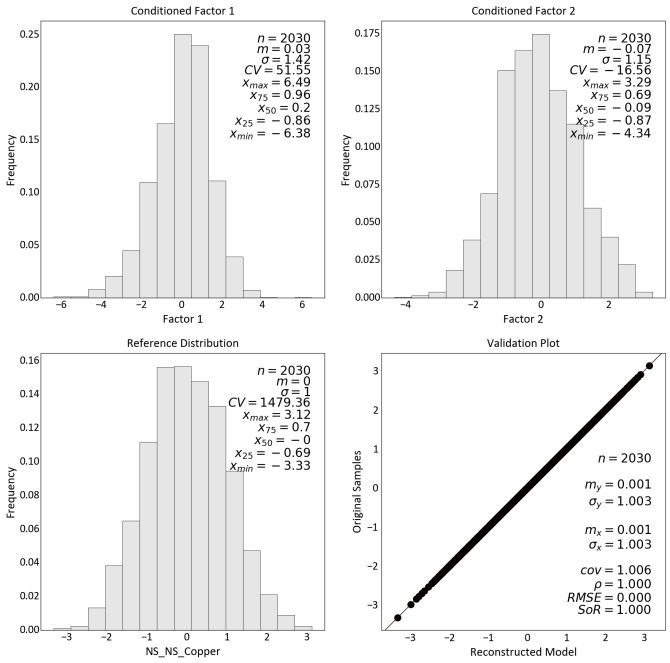

Figure 5.7: Histogram of the two imputed factors for one realization and reference PLMR histogram with validation plot.

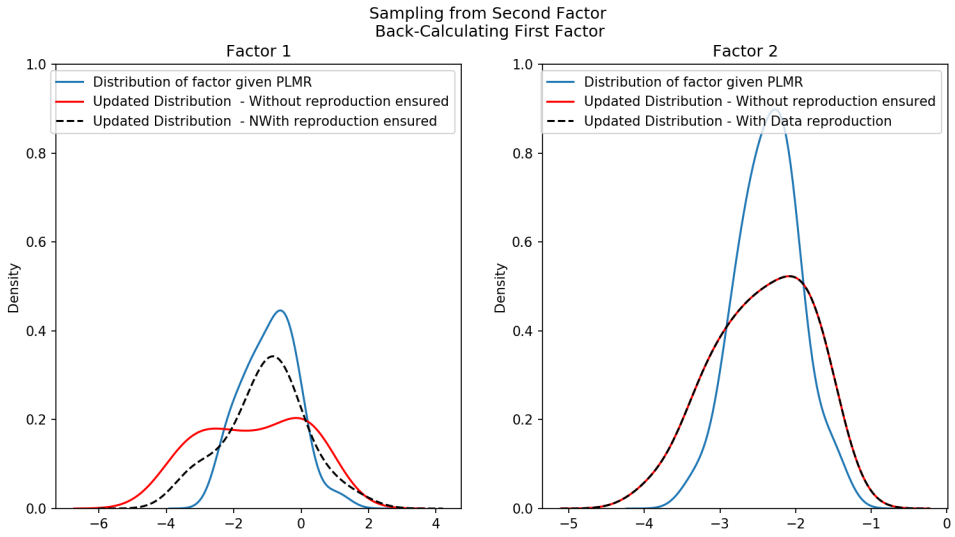

Figure 5.8: Merged distributions using a parametric Bayesian-updating scheme.

The method described leads to the question: which factor should be sampled and which back-calculated in a location $\mathbf{u}$? There is no clear answer to this question at this point. Sampling from the longest range makes its variogram reproduction better, since the factor with biggest range will be fully imputed. However, it also approximates the final model from a multiGaussian one. It is also possible to choose the factor to sample at each location with the smallest or greatest merged variance after the Bayesian Updated step. This seems appealing at first; however, doing that is similar to sampling the factor with smallest $Var\{Y(\mathbf{u})|Z(\mathbf{u})\}$ without accounting for any spatial information, which is similar to choosing the zone of the distribution that is more correlated in Figure 5.5. Hence, this sampling scheme may collapse into choosing based on the PLMR value and if it is above or below the truncation point. Figure 5.9 shows an example of this property and Figure 5.10 shows the loss of correlation in the longer-range structure from the back-calculation step.

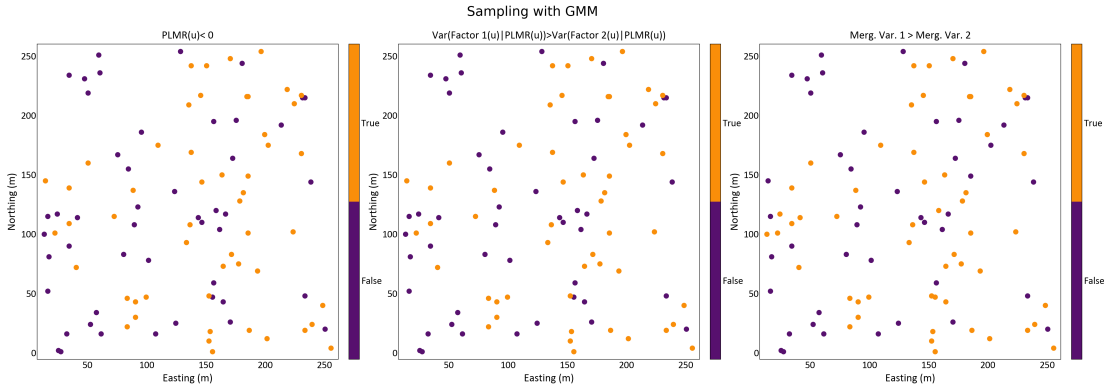

Figure 5.9: Indicator if: 1 - $Z(\mathbf{u})$ < 0, 2 - $\text{Var}\{Y_1(\mathbf{u})|Z(\mathbf{u})\}$ > $\text{Var}\{Y_2(\mathbf{u})|Z(\mathbf{u}),\}$, 3 - $\text{Var}\{Y_1(\mathbf{u})|Z(\mathbf{u}),(n)\}$ > $\text{Var}\{Y_2(\mathbf{u})|Z(\mathbf{u}),(n)\}$

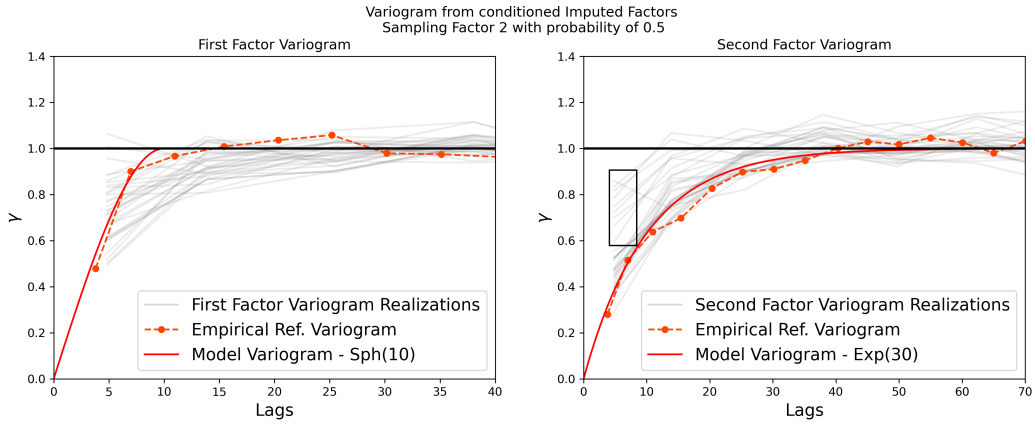

Figure 5.10: Variogram reproduction of the imputed factors. The subset of variograms inside the rectangle shows loss of correlation from back-calculation.

After the imputation and data conditioning steps, realizations of each Gaussian factor will be simulated using SGS. The algorithm performs Simple Kriging using the factor's variogram, a basic variogram model, hence no fitting is necessary. Simple Kriging ensures values spaced beyond the variogram range from conditioning points will regress to the global mean (C. V. Deutsch & Journel, 1998; A. G. Journel & Huijbregts, 1978). The Gaussian factor's realizations having a global normal distribution is a key assumption when building the model, or else bias in the mean is introduced to the final histogram. Simulating the imputed factors mitigates the loss of correlation in the imputed values, and the realizations will have the correct variogram reproduction. After the factor's realizations are simulated, the piecewise linear transform is reapplied, and realizations of the model in PLMR space are obtained.

5.9 Synthetic example

This subsection will present results for conditioning a synthetic PLMR dataset. The model used in this section is built using a $Spherical$ and $Exponential$ variogram structures with ranges of, respectively, 10 and 30. Slopes will be set to $\sqrt{0.1}$ and $\sqrt{0.9}$. Equation 5.22 describes the model used. Figure 5.11 shows the reference realization and sample locations. The reference map was generated in a 256 x 256 grid. Figure 5.12 shows the four PLMR variograms defined by the model.

$$Z(\mathbf{u}) = \begin{cases} \sqrt{0.1}Y_{Sph(10)}(\mathbf{u}) + \sqrt{0.9}Y_{Exp(40)}(\mathbf{u}), & Y_{Sph(10)} \leq 0, Y_{Exp(40)} \leq 0 \\ \sqrt{0.1}Y_{Sph(10)}(\mathbf{u}) + \sqrt{0.1}Y_{Exp(40)}(\mathbf{u}), & Y_{Sph(10)} \leq 0, Y_{Exp(40)} > 0 \\ \sqrt{0.9}Y_{Sph(10)}(\mathbf{u}) + \sqrt{0.9}Y_{Exp(40)}(\mathbf{u}), & Y_{Sph(10)} > 0, Y_{Exp(40)} \leq 0 \\ \sqrt{0.9}Y_{Sph(10)}(\mathbf{u}) + \sqrt{0.1}Y_{Exp(40)}(\mathbf{u}), & Y_{Sph(10)} > 0, Y_{Exp(40)} > 0 \end{cases} \tag{5.22}$$

This section will show results by sampling the factor with the greatest variance. The back-calculated factor distribution may diverge from the expected normal distribution behavior. Fluctua-

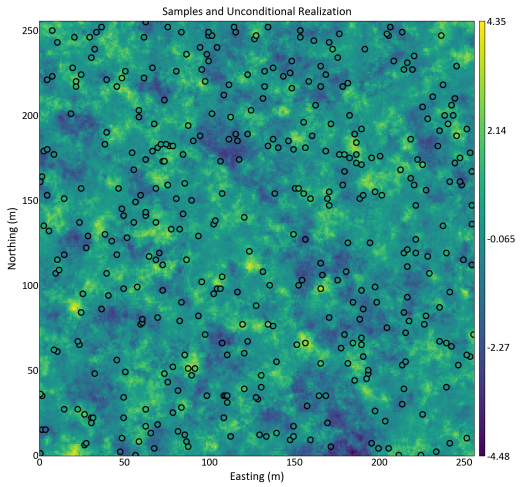

Figure 5.11: Samples and reference realization

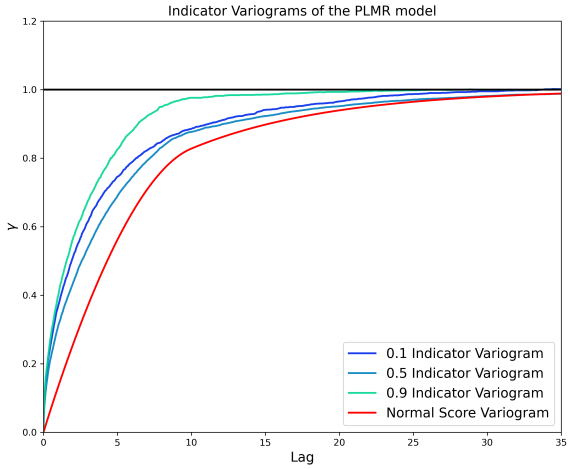

Figure 5.12: PLMR variograms.

tions in variograms and artifacts in histograms of the imputed values introduced by back-calculating one of the factors are mitigated simulated using SGS. Figures 5.13 and 5.14 shows variogram and histogram reproduction of the Gaussian factors. Simulation of the imputed factors can reproduce the correct distribution and spatial continuity of the Gaussian factors building the PLMR model.

Figures 5.15 shows the E-type realization of the simulated imputed factors. The map reproduces the spatial behavior seen in the Factor's reference map fairly well. However, this cannot be checked when working with a real dataset. In this scenario, checking if the realizations are reasonable normal and if the variograms of the factor's used to construct the PLMR model is fairly reproduced is sufficient.

Figure 5.16 shows histogram reproduction for the synthetic case example. Histogram repro-

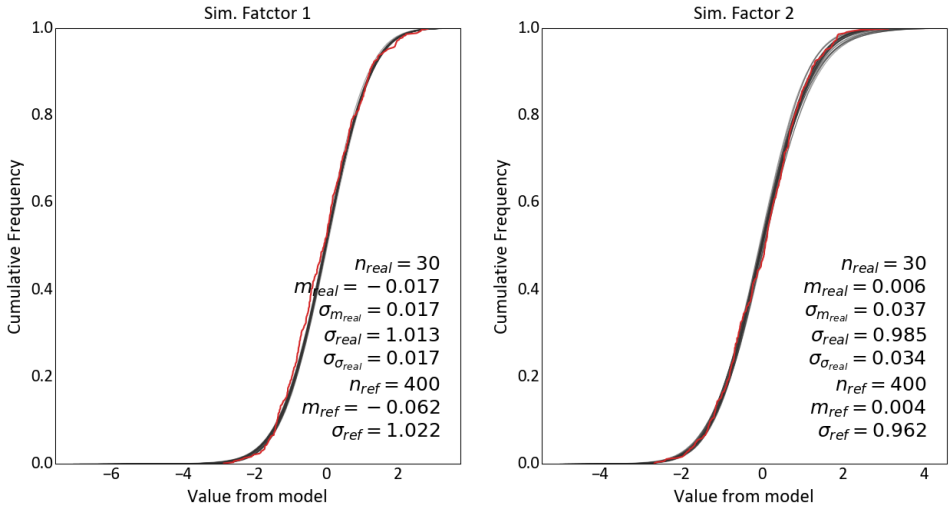

Figure 5.13: CDFs of the simulated Gaussian factors.

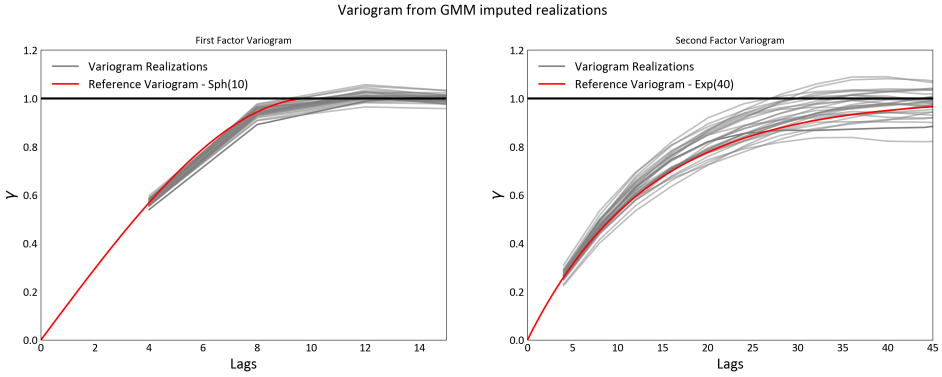

Figure 5.14: Variograms of the simulated Gaussian factors

duction is more robust to changes in the sampling scheme in the imputation step than variograms. After simulation of the imputed factors, the model is rebuilt, i.e the piecewise linear transformation is applied to the factor's realizations. When modeling with a real dataset, the final reconstructed model would be back-transformed to original units. Figures 5.16 and 5.17 show, respectively, histogram reproduction, and the E-type over 30 simulated results. The conditioning process ensures histogram reproduction and data reproduction. Simulated realizations show a similar spatial structure as the reference image and the E-type also looks reasonably similar to the reference. Figures 5.19 and 5.18 shows realizations and the model's reference variograms. As mentioned earlier, the loss of correlation is greatly reduced by simulation and variograms can be reproduced as demonstrated by the synthetic data example.

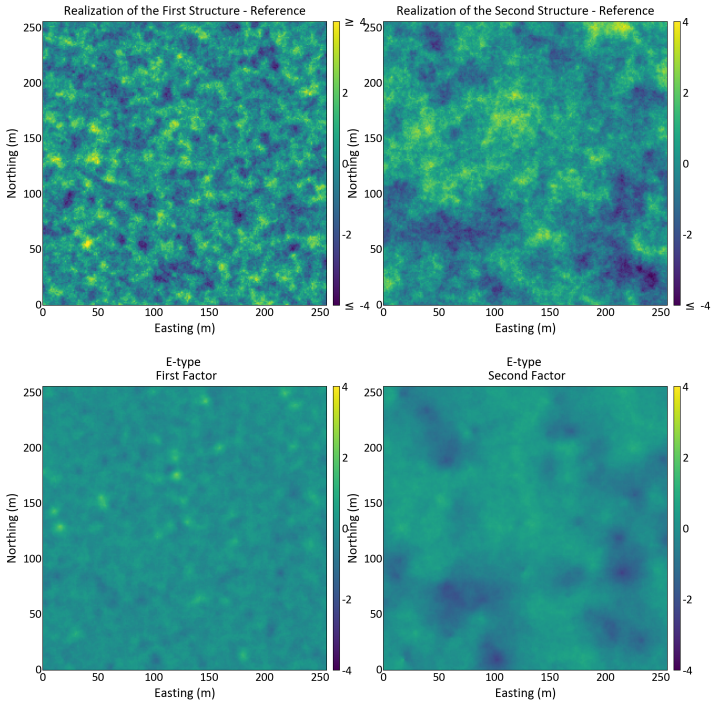

Figure 5.15: E-Type of the simulated Gaussian factors

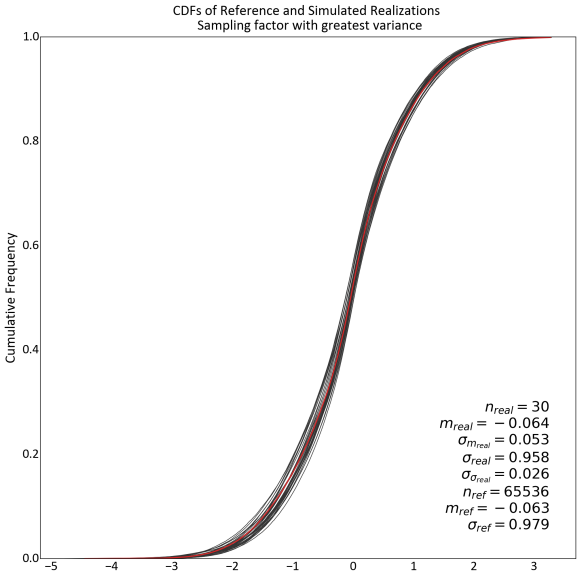

Figure 5.16: CDFs of the rebuilt PLMR model

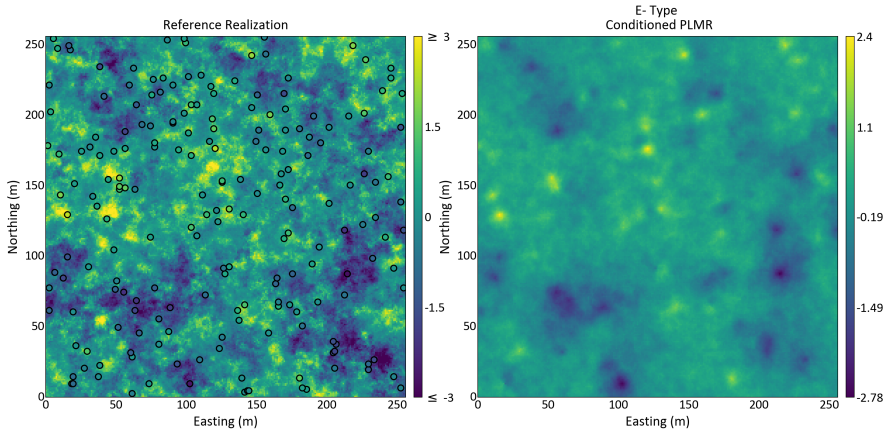

Figure 5.17: Reference image and E-type of the PLMR model

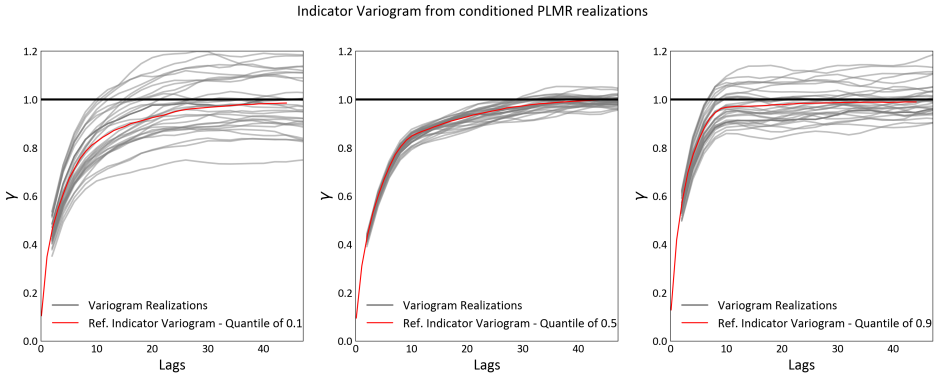

Figure 5.18: Indicator variogram reproduction

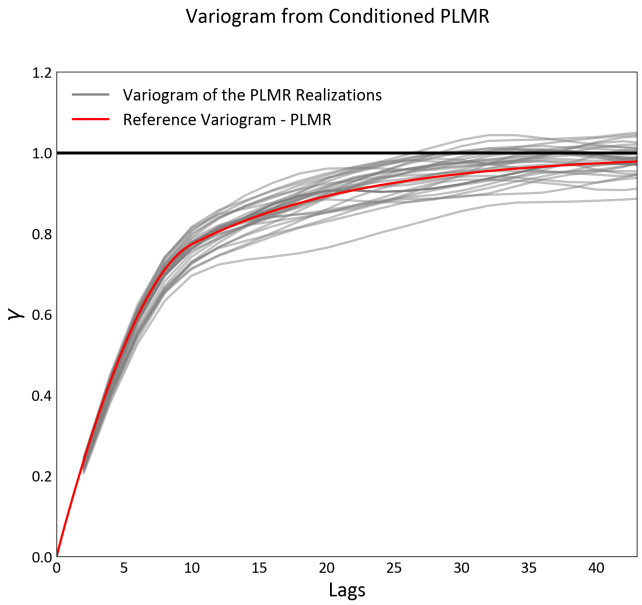

Figure 5.19: Traditional variogram reproduction

Chapter 6

Demonstration of the PLMR

This chapter presents an application of the PLMR to a real 3D dataset. The goal is to demonstrate the use of the PLMR, compare it to a model built using the multiGaussian assumption, and comment on future room for improvement in the methodology. The model referred to as the MG model will be generated using sequential Gaussian simulation with an LMR. For comparison purpose the PLMR will be called $Z_{plmr}(\mathbf{u})$ and the LMR, assuming a MG model, one will be called $Z_{lmr}(\mathbf{u})$.

Both models may benefit from a trend model; however, composites were considered stationary over the modeled area. Adding a trend might remove some of the non-Gaussian behavior seen in the Zinc (Zn) measurements, i.e. indicator variograms would become more symmetric. The chapter aims to explore the difference between the two models and how they handle the different continuity of highs and lows,.

6.1 Data pre-processing and variography

The dataset consists of Zinc measurements (%) composited to 2m intervals. The data comes from a Turkish underground mine. Drill Holes are mostly oriented with 90 degree azimuth. Sections of the composites are shown in Figures 6.1 and 6.2.

A representative Zinc distribution is calculated by despiking and declustering the attribute. Despiking is a crucial step to break ties when ranking samples and building empirical CDFs for Q-Q transformations. In the presence of ties between samples, i.e., same value for multiple composites, the Q-Q transform is not unique (Pyrcz & Deutsch, 2014), and the resulting histogram may present artifacts. Declustering is necessary to build a representative distribution that considers sampling bias, i.e, more data is usually collected near areas of interest in the deposit (Rossi & Deutsch, 2013).

Two data transformations will be applied to the representative distribution of the composited samples. First, a normal-score (NS) transform (C. V. Deutsch & Journel, 1998; Rossi & Deutsch, 2013) table was built considering the declustering weights and the variable transformed to a normal distribution. The NS measurements will be used to estimate variograms. Using the normal-scored values as traditional variogram, 0.1,0.5 and 0.9 quantile indicator variograms are calculated respecting directions of continuity of the deposit. Hence, four variograms at each principal direction of geological continuity will serve as input for a semi-automatic fitting procedure. In this demonstration of the PLMR, the deposit will be modeled OmniDirectionally. Modeling anisotropy with the PLMR is likely to be more explored in the future. The main draw-back of the model proposed in this thesis is the fact that each direction of continuity shares the same contributions S_1 and S_2. Hence, geological

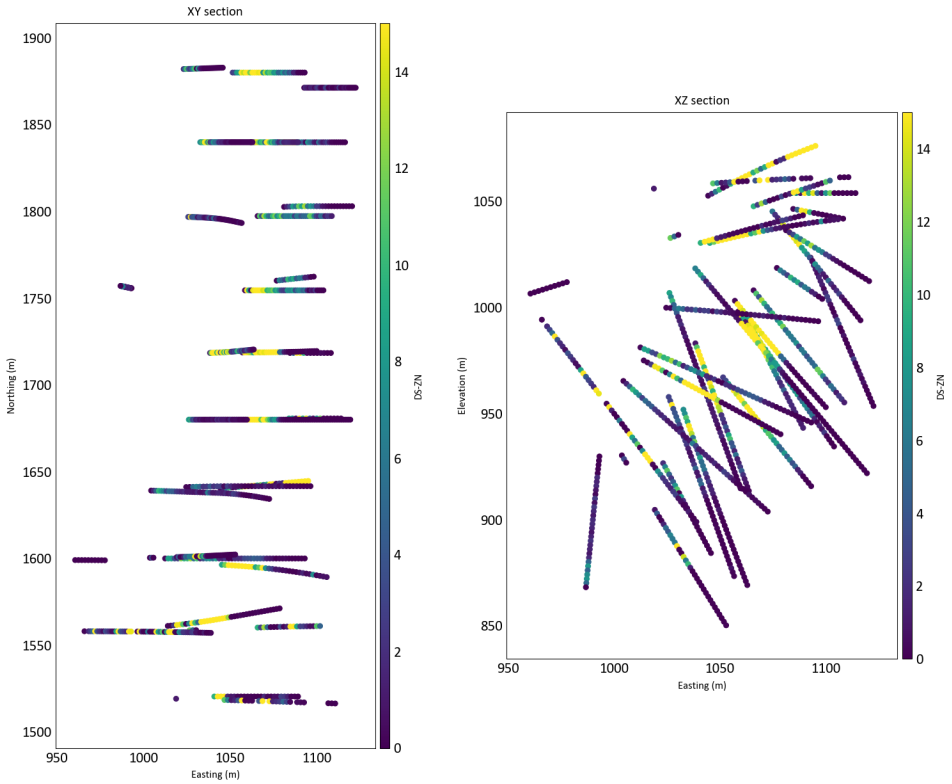

Figure 6.1: XY and XZ sections of the Zn composites

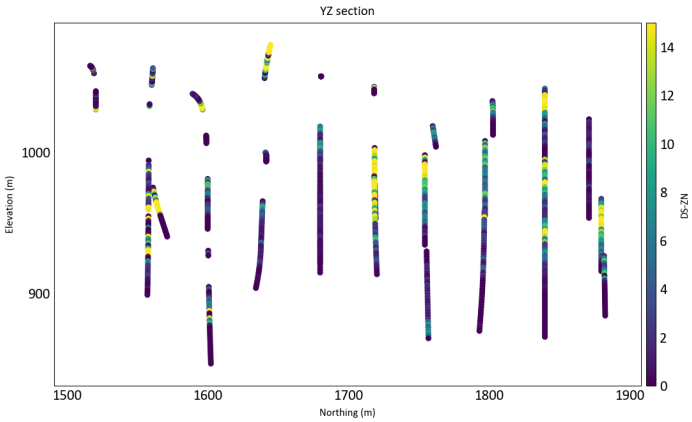

Figure 6.2: YZ section of the Zn composites

variables with asymmetry in the indicator variograms that are not the same across directions may bring challenges when applying with the PLMR. This is likely room for future improvement in the methodology. Therefore, for simplicity and more interpretability, the models used in this section were fitted OmniDirectionally.

Figure 6.3 shows the four variograms fitted for the NS Zinc variable. Table 6.1 shows the final parameters of the model. The indicator variograms indicate that the continuity of high values is considerably lower than that of low values. The model was fit using two Spherical variogram structures with ranges of 22 and 91 meters. The slopes of 0.115 and 0.885, and the ratio of approx 4:1 between each Gaussian factor range, indicate strong asymmetry of the indicator variogram. Results also show how the PLMR can simultaneously fit the four variograms while respecting the different spatial continuity of low and high values. Figure 6.4 and Table 6.2 shows, respectively, the LMR variogram that will be used for the MG model and it's parameters. The fit using each methodology, i.e, LMR and PLMR, resulted in variograms structures with similar ranges meaning the scale of spatial continuity is consistent between models; therefore, the difference between model is related to how the PLMR defines different contributions depending on the data value. In contrast, the LMR defines the contribution of each structure to be data-independent.

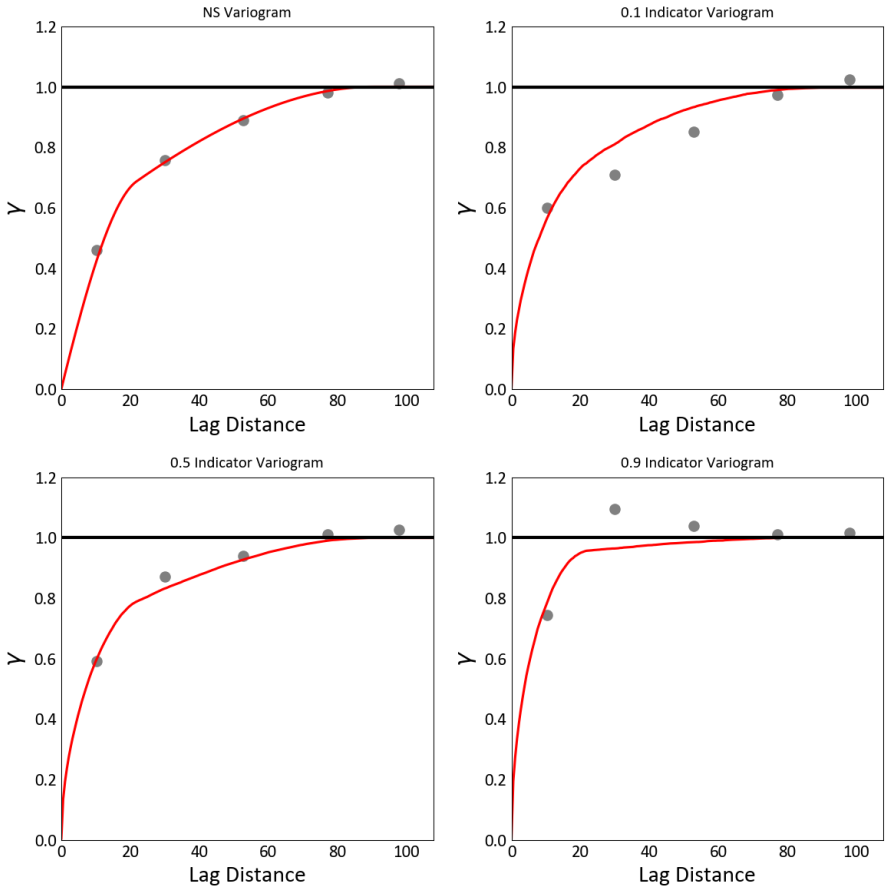

Figure 6.3: Variograms fitted using a PLMR.

Table 6.1: PLMR parameters

Parameter	Value
r_{γ_1}	22 (m)
r_{γ_2}	91 (m)
S_1	0.115
S_2	0.885

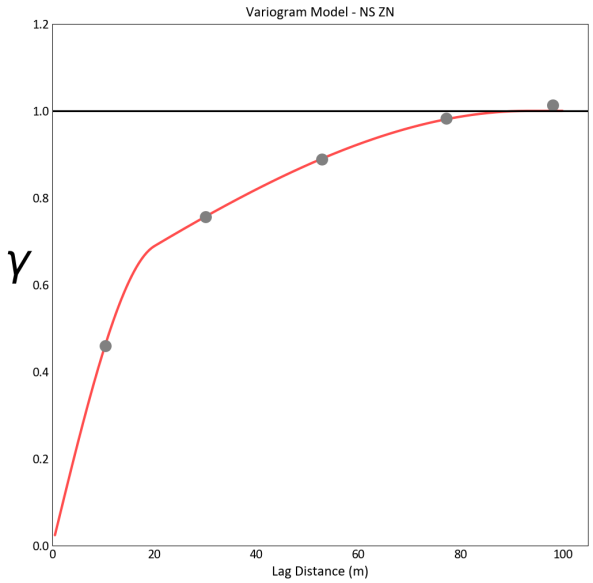

Figure 6.4: Variograms fitted using a LMR.

Table 6.2: LMR parameters

Variogram	Ranges	Contribution
γ_1	22 (m)	0.563
γ_2	91 (m)	0.437

After inference of the PLMR parameters, a second quantile-quantile transformation table will be applied to the Zn representative distribution using the reference PLMR distribution. This reference distribution is obtained using the simulation approach described in Chapter 5.

Even though the PLMR distribution is well-behaved and, depending on the model parameters, it might be similar to a normal distribution, caring on the PLMR workflow with NS samples will lead to bias in the final histogram reproduction. Transforming the data to exactly match the PLMR distribution will correct this issue. Figure 6.5 shows the distribution of the despiked clustered/declustered composites as well as the histograms after the two transformations are applied. The declustered mean is 15% lower than the naive average. The NS and PLMR distributions are relatively similar. However, the piecewise linear transform makes the resulting global histogram non-normal with a smaller variance.

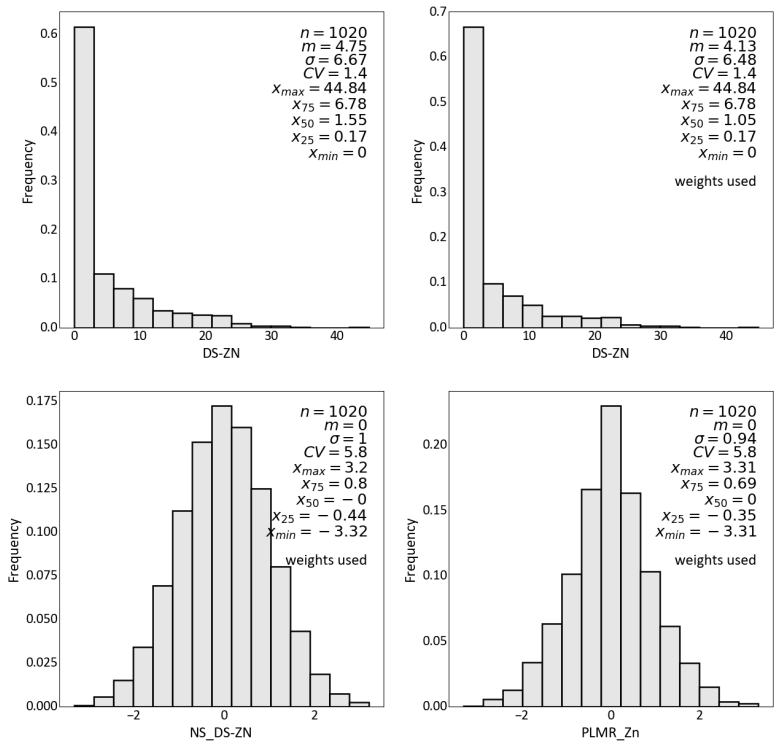

Figure 6.5: Top row: despiked histograms using/not using declustering weights. Bottom row: the two Q-Q transformations with weights applied

6.2 Conditioning by imputation of Gaussian factors

Once the parameters that define the model are inferred and the variable transformed to the correct units, decomposition of the data into the Gaussian factor's ($Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$) using the GMM imputation approach can be carried out. The GMM model fitted usually takes three to four Gaussian components. The GMM fitted to the conditional distribution of each Gaussian factor given the PLMR value in a location is shown below:

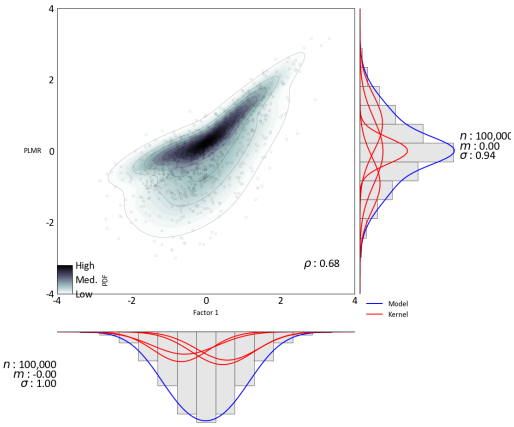

Figure 6.6: GMM fitting to $P(Y_1(\mathbf{u})|Z_{plmr}(\mathbf{u}))$

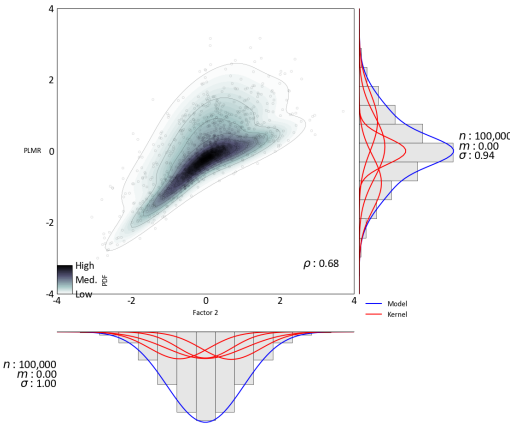

Figure 6.7: GMM fitting to $P(Y_2(\mathbf{u})|Z_{plmr}(\mathbf{u}))$

After the two Gaussian Mixture Models are fitted to the $P(Y_1(\mathbf{u})|Z_{plmr}(\mathbf{u}))$ and $P(Y_2(\mathbf{u}))|Z_{plmr}(\mathbf{u}))$ distributions, imputation of the Gaussian factors can be carried out. Imputation of the factors should be done using as much data as possible, ideally considering all data in the domain and independently for each factor. The most important decision the modeler needs to make in this step is which Gaussian factor to sample and which to reconstruct to ensure data reproduction. Histograms are shown sampling solely from the longest (top row) or smallest range (bottom row) variogram structure. Sampling only from one factor introduces artifact in the CDF of the factor being back-calculated. This behavior is more evident when back-calculating the most spatially continuous Gaussian factor.

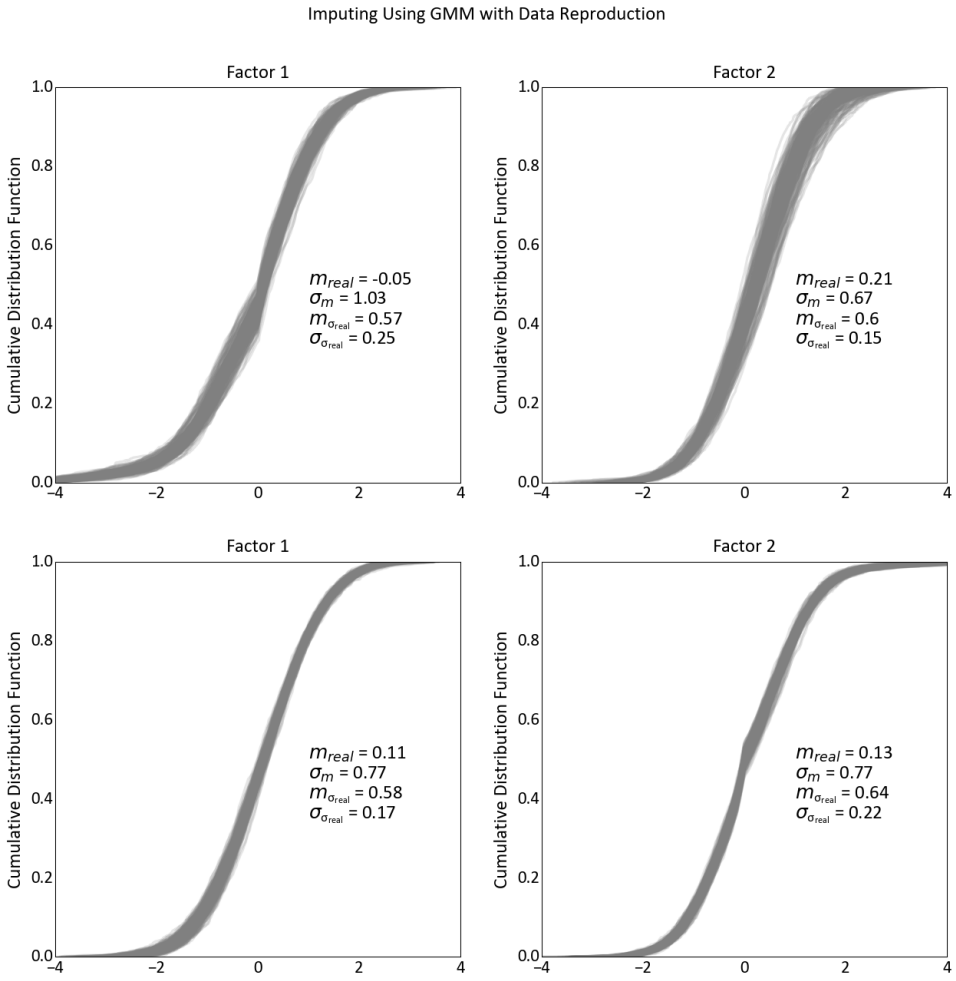

Figure 6.8: Distribution of the imputed values after reproduction of hard-data is ensured. CDFs are shown sampling solely from the longest (top row) or smallest range (bottom row) variogram structure.

6.3 Simulation

Once the factors are imputed and data reproduction ensured, simulation of the imputed values can be carried out. Simulation of the imputed factor's is done using SGS. No transformation should be applied since the factor's being simulated are already in Gaussian units.

Ideally, conditional simulation of the imputed factors should be conducted globally, but in practice, conditioning to a neighbor leads to reasonable results. It is necessary to check if the realizations are reasonably normally distributed and if the variograms of the factor's used to construct the PLMR model are fairly reproduced. This simulation step is the building block of uncertainty estimation with the PLMR, so a good reproduction of the factor's histogram and variograms usually means a fair reproduction when the model is rebuilt and back-transformed. The empirical CDF of the simulated factors are shown in Figure 6.9 with variogram reproduction in Figure 6.10. The data conditioning step makes it possible that some locations will have values outside the usual interval of normal deviates, i.e, $[-4, 4]$. This behavior can be seen in the CDFs of the first Gaussian factor. However, simulation mitigates the artifacts in the CDF of the factors seen in Figure 6.8.

Figure 6.11 shows the E-type over 200 realizations of the Gaussian factors. It is possible to see how each factor E-type shows structures in the maps related to the range that it gives more contribution to the PLMR model.

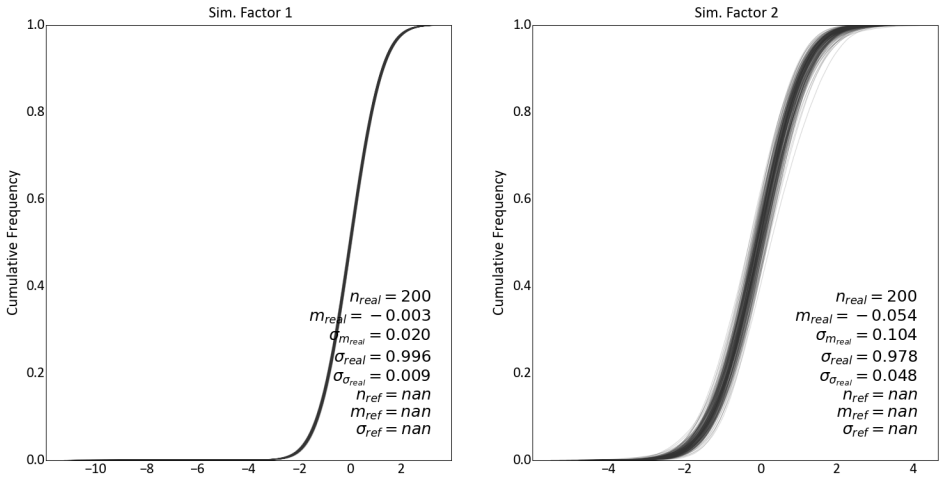

Figure 6.9: Distribution of the simulated imputed Gaussian factors - sampling from the longest range Gaussian factor.

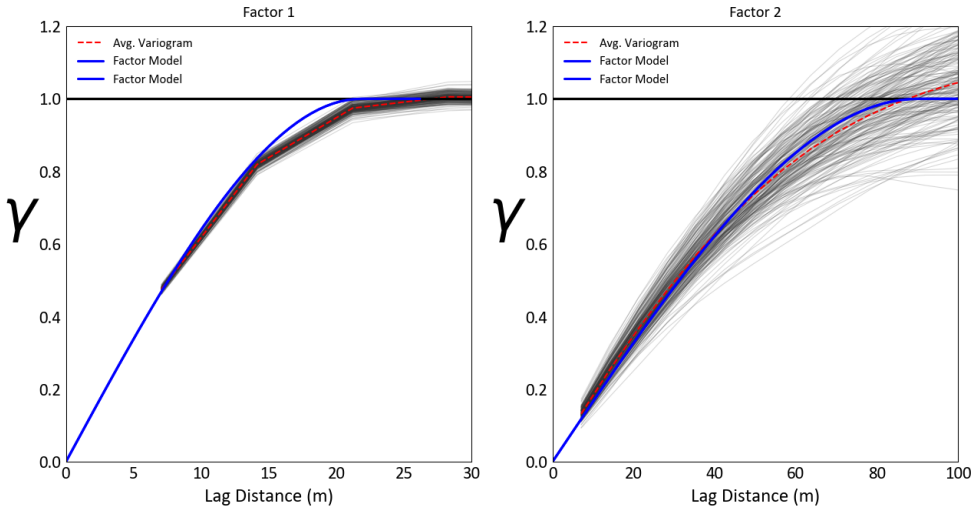

Figure 6.10: Variograms of the simulated imputed Gaussian factors - sampling from the longest range Gaussian factor.

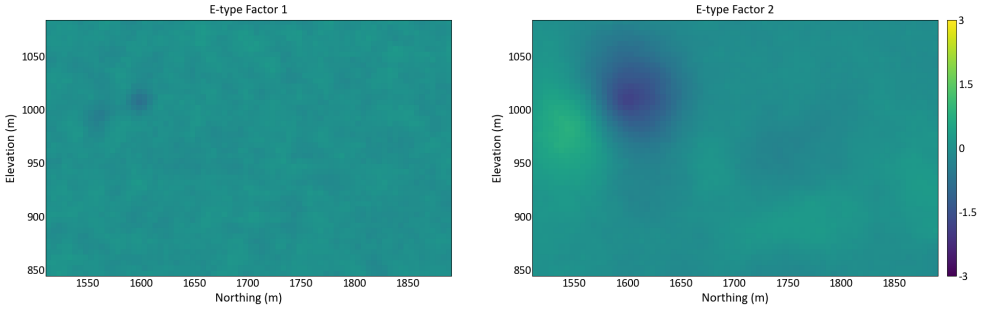

Figure 6.11: Slice of the E-type of the simulated imputed Gaussian factors

6.4 Results and comparison between models

This section presents results of simulating with a PLMR model and compare them to the results obtained by simulating using an LMR, i.e., fitting a nested variogram model and simulating using the SGS workflow. The same representative distribution is used with both models, and the PLMR will be simulated using two different sampling strategies: one sampling from the more continuous Gaussian factor and one sampling from the factor with smallest range.

After SGS is performed at the imputed Gaussian factors, the model is reconstructed by applying the piecewise linear transformation. The table used in the Q-Q PLMR transform is used to back-transform the final realizations to original units. Figure 6.12 shows histogram reproduction in original units for both models. The multiGaussian model presents slightly better histogram reproduction and more variability between the distributions.

Variograms should be checked in NS units. Therefore, applying an NS transformation to the

reconstructed model is necessary. It does not matter if the NS transformation is applied to final realizations in original or PLMR units. The four variograms used in the fitting procedure should be checked. Hence, it is possible to evaluate how the non-Gaussian behavior of the PLMR, i.e. asymmetrical indicators variograms, are affected by the conditioning process. Figure 6.13 shows variogram reproduction for the traditional variograms and 6.14 for indicator variograms for the two methodologies. Traditional variogram reproduction is similar between the two methodologies. However, reproduction of the Indicator variograms changes considerably between workflows. As expected, LMR realizations have more symmetric indicators variograms due to the MG model. The realizations' indicator variograms using the PLMR could better reproduce the different continuities of highs and lows.

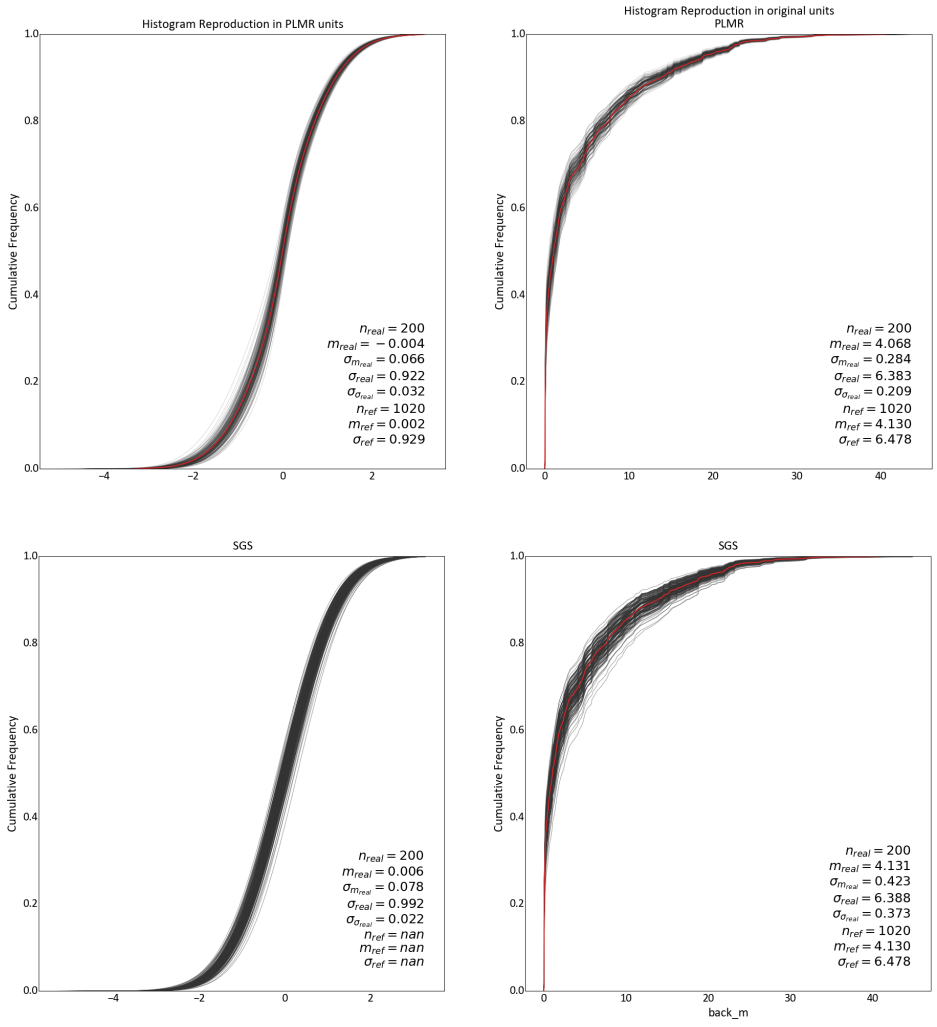

Figure 6.12: Histogram reproduction in PLMR and original units (top row). Histogram reproduction in NS original units (bottom row)

Slice plots of the models are shown in Figures 6.15 and 6.16 as well as relative difference between

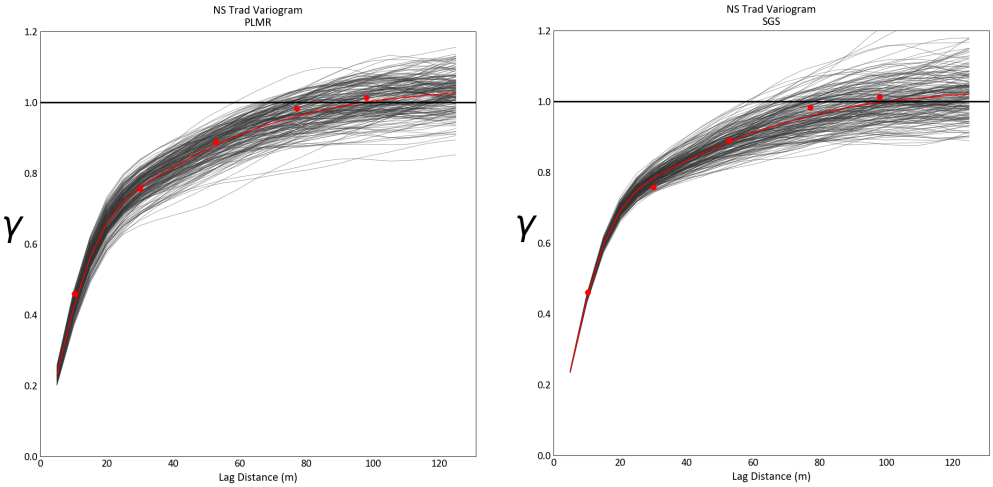

Figure 6.13: Traditional variogram reproduction

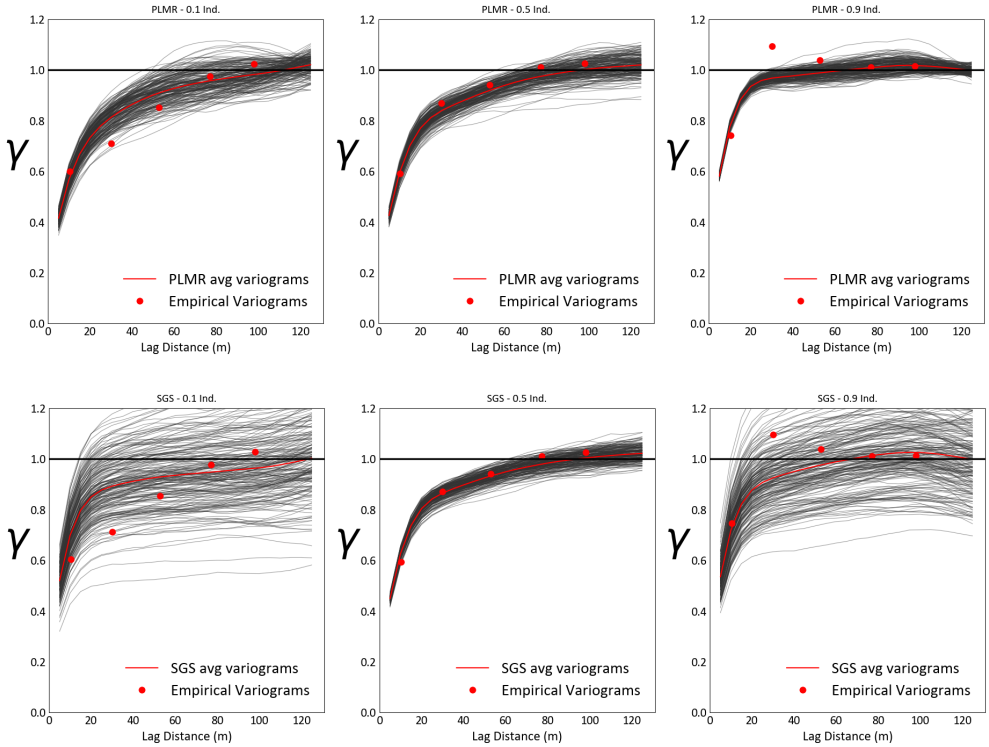

Figure 6.14: Indicator variogram reproduction

each PLMR model and the MG model, i.e:

$$D(\mathbf{u}) = \frac{E(z_{lmr}(\mathbf{u})) - E(z_{plmr}(\mathbf{u}))}{E(z_{lmr}(\mathbf{u}))} \tag{6.1}$$

Analyzing the slices shown in Figures 6.15 and 6.16 it is possible to see how the PLMR tends to estimate locations near high-grade zones lower than the MG model and low-grade zones higher than the conventional Gaussian simulation approach. The XZ sections of the models show how the MG model has the mean at nearby locations more influenced by the high-grade samples of the uppermost drill hole. This result can be also seen in the cross-plot between PLMR and MG model E-type estimates, shown in Figure 6.17. For values above the declustered mean, the MG model predicts higher values, and low values tend to be, in general, relatively lower. These results are in-line with the indicator variogram reproduction in each methodology, where the PLMR reproduces better the destructuration on higher thresholds variograms. The E-type using an LMR has higher continuity in the higher quantiles of the data, which is a consequence of the symmetric indicator variograms of multiGaussian assumption, meaning more smearing of high grades than the PLMR.

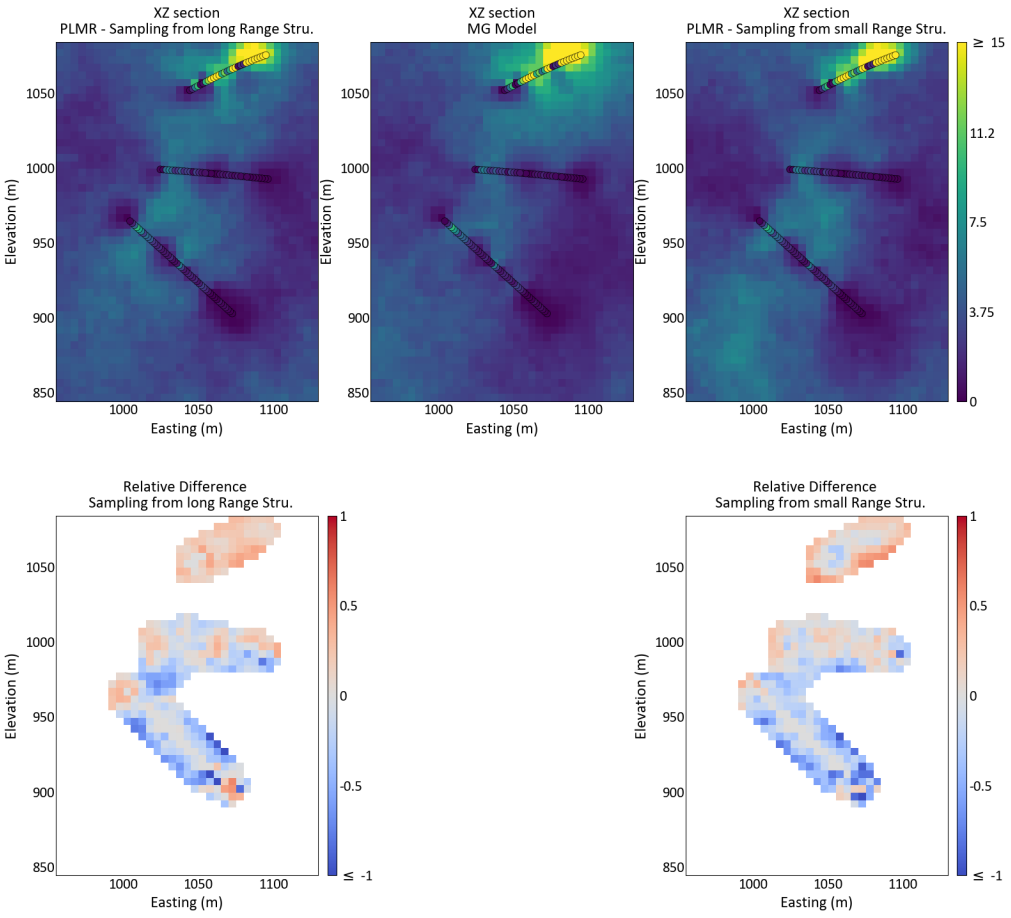

Figure 6.15: XZ section of the models and the relative difference to the MG model

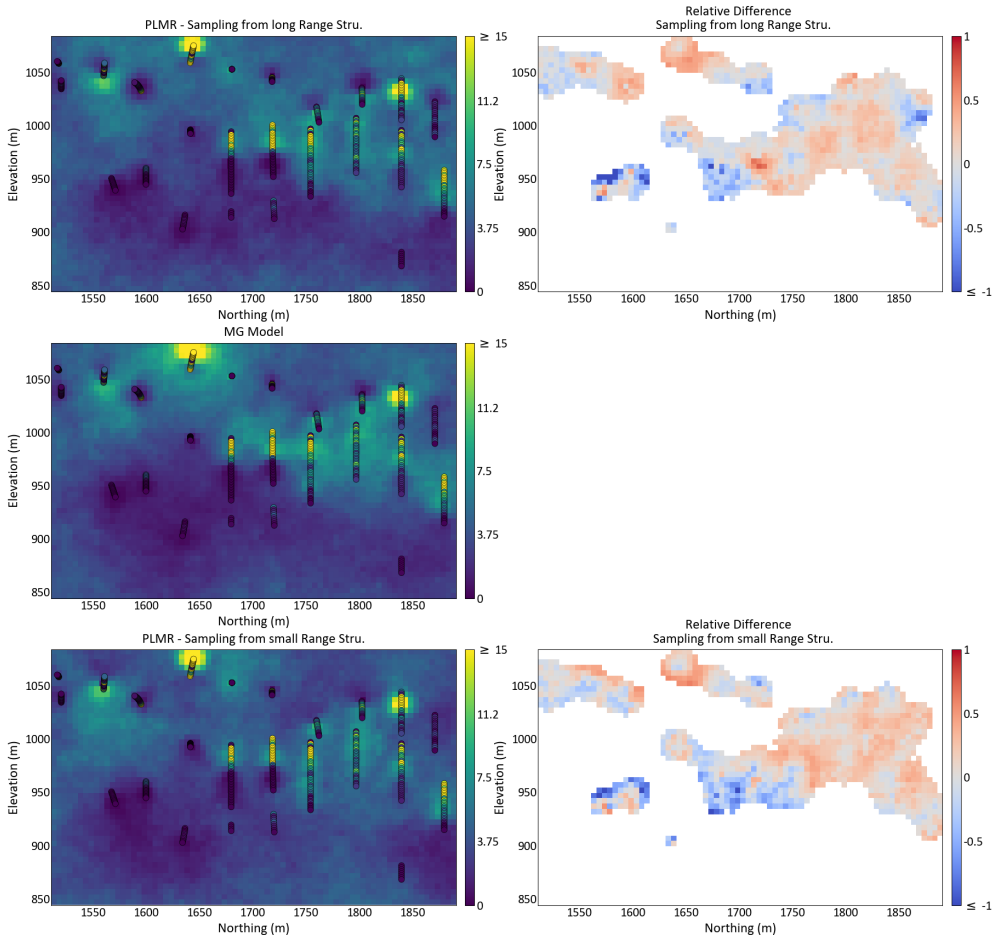

Figure 6.16: YZ section of the models and the relative difference to the MG model

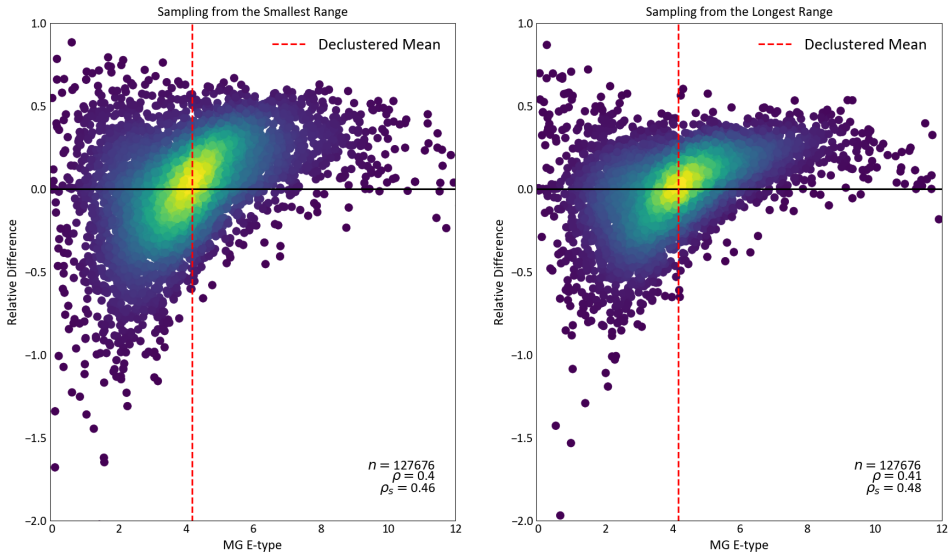

Figure 6.17: Scatter plot between difference ($D(\mathbf{u}$) and the SGS E-type.

6.5 Cross validation

Cross-Validation (CV) results were generated by leaving out 255 composites at random, 25% of the dataset, giving insight into how the models differ in locations heavily influenced by conditioning data. Figure 6.18 shows validation plots for the PLMR, sampling from one of the two factors, and the MG model. Results of the two methods are similar, independent of the PLMR conditioning strategy chosen. Sampling from the Factor with the smallest range has slightly better CV data reproduction.

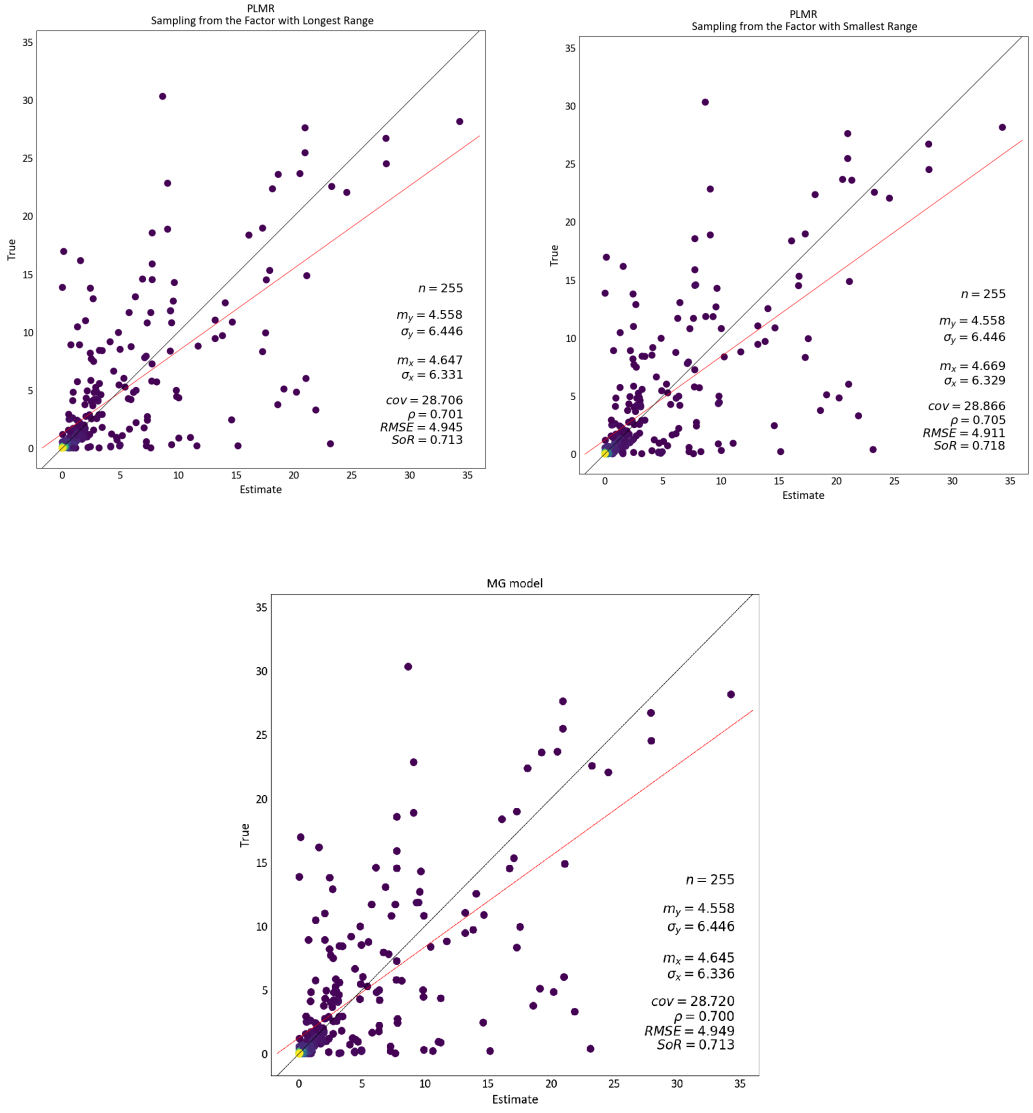

Figure 6.18: Validation plot between the three models.

In this case study, the PLMR tends to generate more conservative realizations. This was seen in the last section when comparing the E-type of the two methodologies, where the MG model tends

to smear more the high grades than the PLMR. Figure 6.19 shows the classification of the CV data as ore or waste for three cut-offs: 5,6 and 7 % Zinc. Accuracy of classification is also presented for each method in each cut-off value, being calculated as:

$$Acc. = \frac{TruePositive + TrueNegative}{n_{samples}} \tag{6.2}$$

Figure 6.19: Confusion matrixes - test data classification at cut-offs 5,6 and 7 % Zinc

Classification of the test data as ore or waste was similar using the PLMR or MG model. However, for a cut-off of 5 % , the MG model classified the removed composites with slightly increased accuracy. The higher accuracy of the LMR comes from a higher rate of True Negative that outweighs the decrease in True Positive relative to the PLMR. In the remaining cut-offs, 6 % and 7%, accuracy was, respectively, marginally better and the same when using the PLMR. However, even-though accuracy was the same for a threshold of 7 % Zn; the PLMR had slightly better performance regarding

True Positive and worse when classifying actual waste as waste, i.e, True Negative proportion.

The local difference in the estimation of high grades introduced by decomposing the RF using a piecewise linear function is likely to impact when the model is being processed by a transfer function. As mentioned in the introduction of this thesis, mitigating the impact of high grade samples is important when estimating mineral resources to avoid over-estimation. The PLMR, as well as Gaussian simulation, mitigates the effects of outliers by performing a Q-Q quantile transformation of the geological variable to a more symmetric distribution. As demonstrated in this chapter, by directing accounting for the loss of correlation in high values, the PLMR further mitigates, on average, the impact of high samples.

Chapter 7

Conclusion and future work

7.1 Conclusion

The PLMR is presented in this thesis as a novel simulation framework to deal with variables with non-Gaussian behavior. The model moves away from the multiGaussian model by decomposing the RF model using a piecewise linear function instead of a linear one. As a consequence, bivariate distributions under this framework will not have a Gaussian elliptical shape and indicator variograms will not be symmetric around the median. Besides proposing a new framework to deal with positively skewed variables, a conditioning method based on the decomposition of the PLMR into Gaussian factors and its imputation in every sampled location is also developed. A synthetic example showed how, under this framework, it is possible to reproduce all four variograms used in the PLMR fitting.

Comparing results from modeling with a PLMR and an LMR using a real dataset establishes the proposed methodology. Results show that, by capturing the spatial destructuration effect of a positively skewed variable, the final model is different from one using the multiGaussian approach, and smearing of high grade seems to be more contained. Cross-validation results removing composites at random show how, on average, classification into ore/waste is similar to one conducted with an MG model. This validates the PLMR and shows that, even if final realizations are considerably different, the proposed methodology does not yield results that are inconsistent with simulation using an LMR.

There are limitations to the methodology proposed in this thesis. The PLMR uses the same contribution parameters, i.e., S_1 and S_2, in all major directions of continuity. Implicitly, this is equivalent to assuming that all directions being modeled share the same asymmetry in indicator variograms which may not be the case for all deposits. Besides, using only two Gaussian factors restricts the flexibility of the PLMR to model different variogram shapes.

7.2 Future work

The PLMR is applicable for modeling real deposits, however, there is still room for future work to make the methodology more robust and flexible. First, it is necessary to better understand the conditioning step and how the sampling strategy affects final realizations. In the demonstration chapter, the final results were not greatly influenced by the choice of which factor to sample, as shown in the CV results. However, this is not necessarily true for every data configuration and set of PLMR parameter. A more robust sampling strategy that can be applied to most datasets

still needs to be devised. By doing that, it is possible to explore different implementations of the model to gain more computational efficiency. For example, if the factor to be imputed or back-calculated at each location is defined before imputation is conducted, the implementation could benefit from that and avoid imputing the two factors in all locations independently. Modeling of highly anisotropic regionalizations could also be explored in the future to access how restrictive using the same contribution parameters for all directions is.

The idea presented in this thesis opens room for several research paths that could make the PLMR more flexible. Expanding the PLMR to more than two Gaussian factors and one truncation quantile would give the workflow greater flexibility to tackle different types of deposits and variograms. More indicator variograms could be used to bring additional information regarding the transition between thresholds. The PLMR could also be truncated at point $T_p \neq 0$. Adding T_p as a free parameter to be inferred would likely give more flexibility regarding the pattern of indicator variograms asymmetry of the model. As mentioned in earlier sections, the proposed model aims at breaking the linear decomposition of the RF model in the simplest way possible, however, other types of functional forms may be explored and different types of non-linearity introduced.

Finally, it might be interesting to evaluate how the PLMR method will perform when a trend model is applied to the data before simulation. The PLMR assumes that composites are stationary over the area where the model is fitted. However, by defining different spatial continuity to different data quantiles, the model might tackle some of the trend-like features of the variable. Removing the trend also seems to remove part of the non-Gaussian behavior seen in positively skewed variables, i.e. it makes indicator variograms more symmetrical.

References

Barnett, R. M., & Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. *Mathematical Geosciences*, *47*(7), 791–817.

Barnett, R. M., Manchuk, J. G., & Deutsch, C. V. (2014). Projection pursuit multivariate transform. *Mathematical Geosciences*, *46*(3), 337–359.

Chiles, J.-P., & Delfiner, P. (2009). *Geostatistics: modeling spatial uncertainty* (Vol. 497). John Wiley & Sons.

Deutsch, C. V. (1992). *Annealing techniques applied to reservoir modeling and the integration of geological and engineering (well test) data* (Unpublished doctoral dissertation). stanford university.

Deutsch, C. V. (2018). All realizations all the time. In *Handbook of mathematical geosciences* (pp. 131–142). Springer, Cham.

Deutsch, C. V. (2021). Implementation of geostatistical algorithms. *Mathematical Geosciences*, *53*(2), 227–237.

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

Deutsch, J. L., & Deutsch, C. V. (2012). Latin hypercube sampling with multidimensional uniformity. *Journal of Statistical Planning and Inference*, *142*(3), 763–772.

Dimitrakopoulos, R., Mustapha, H., & Gloaguen, E. (2010). High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-gaussian and non-linear phenomena. *Mathematical Geosciences*, *42*(1), 65–99.

Emery, X. (2004). Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, *18*(6), 414–424.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Glover, F. W., & Kochenberger, G. A. (2006). *Handbook of metaheuristics* (Vol. 57). Springer Science & Business Media.

Goovaerts, P., et al. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Hadavand, Z., & Deutsch, C. (2020). Conditioning by kriging. , Retrieved from http://www.geostatisticslessons.com/lessons/conditioningbykriging.

Haykin, S. (2010). *Neural networks and learning machines, 3/e*. Pearson Education India.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions, volume 2* (Vol. 289). John wiley & sons.

Journal, A., & Alabert, F. (1989). Non-gaussian data expansion in the earth sciences. *Terra Nova*, *1*(2), 123–134.

Journel, A. (2002). Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical geology*, *34*(5), 573–596.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, *15*(3), 445–468.

Journel, A. G. (1989). *Fundamentals of geostatistics in five lessons* (Vol. 8). American Geophysical Union Washington, DC.

Journel, A. G., & Deutsch, C. V. (1993). Entropy and spatial disorder. *Mathematical Geology*, *25*(3), 329–355.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics*. New York: Academic Press.

Krenek, R., Cha, J., & Cho, B. R. (2016). Development of the convolutions of truncated normal random variables with three different quality characteristics in engineering applications. *Computers & Industrial Engineering*, *94*, 125–137.

Leuangthong, O., & Deutsch, C. V. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, *35*(2), 155–173.

Leuangthong, O., Khan, K. D., & Deutsch, C. V. (2011). *Solved problems in geostatistics*. John Wiley & Sons.

Leuangthong, O., & Nowak, M. (2015). Dealing with high-grade data in resource estimation. *Journal of the Southern African Institute of Mining and Metallurgy*, *115*(1), 27–36.

Mantoglou, A., & Wilson, J. L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research*, *18*(5), 1379–1394.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, *6*, 355–378.

Mustapha, H., & Dimitrakopoulos, R. (2011). Hosim: a high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. *Computers & geosciences*, *37*(9), 1242–1253.

Neufeld, C., & Deutsch, C. V. (2006). Data integration with non-parametric bayesian updating. *CCG annual report*, *8*(105).

Pereira, F. P., & Deutsch, C. V. (2020). Short note: A comparison between different methods to sample a multivariate gaussian distribution. *Center for Computational Geostatistics (CCG) Annual Report, 22(165)*.

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling*. OUP USA.

Qu, J., & Deutsch, C. V. (2018). Geostatistical simulation with a trend using gaussian mixture models. *Natural Resources Research*, *27*(3), 347–363.

Rossi, M. E., & Deutsch, C. V. (2013). *Mineral Resource Estimation*. Springer Netherlands.

Silva, D. S., & Deutsch, C. V. (2018). Multivariate data imputation using gaussian mixture models. *Spatial statistics*, *27*, 74–90.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point

statistics. *Mathematical geology*, *34*(1), 1–21.

Strebelle, S. (2012). Multiple-point geostatistics: from theory to practice. In *Ninth international geostatistics congress. springer, oslo, norway* (pp. 11–15).

Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Appendix A

Appendix - derivation of the PLMR global mean and variance

This section aims to show how to calculate the mean and variance of a two-factor, one truncation point PLMR. Taking the first piece of equation 3.4:

$$\begin{cases} Z_1(\mathbf{u}) = \sqrt{S_1}Y_1(\mathbf{u}) + \sqrt{1-S_1}Y_2(\mathbf{u}); \quad Y_1(\mathbf{u}) \leq 0, Y_2(\mathbf{u}) \leq 0 \end{cases} \tag{A.1}$$

The notation $Z_i(\mathbf{u})$, $i = 1, 2, ..., 4$, is introduced to keep clear which piece of the PDF is being analyzed. As mentioned before, it is well established that:

$$\begin{aligned} Y_1(u) &\mapsto N(0,1) \\ \sqrt{S_1}Y_1(u) &\mapsto N(0, S_1) \end{aligned} \tag{A.2}$$

Therefore, we need to determine the mean and variance of $\sqrt{S_1}Y_1(\mathbf{u})$ and $\sqrt{1-S_1}Y_2(\mathbf{u})$ after it is truncated at T_p. This problem is well studied in the literature. Following Johnson, Kotz, and Balakrishnan (1995), and defining two arbitrary truncation points $T_{1_{std}} = \frac{T_p - m_{Y_1}}{\sigma_{Y_1}(\mathbf{u})}$ and $T_{2_{std}} = \frac{T_p - m_{Y_2}(\mathbf{u})}{\sigma_{Y_2}(\mathbf{u})}$, it is possible to write the mean and variance of a truncated Gaussian distribution as:

$$E[Y_1(u)|T_{1_{std}} < y_1(\mathbf{u}) < T_{2_{std}}] = m_{Y_1}(\mathbf{u}) + \sigma_{Y_1}(\mathbf{u})\frac{\phi(T_{1_{std}}) - \phi(T_{2_{std}})}{\Phi(T_{2_{std}}) - \Phi(T_{1_{std}})} \tag{A.3}$$

$$Var[Y_1(u)|T_{1_{std}} < y_1(\mathbf{u}) < T_{2_{std}}] = \sigma_{Y_1}(\mathbf{u})^2(1 + \frac{T_{1_{std}}\phi(T_{1_{std}}) - T_{2_{std}}\phi(T_{2_{std}})}{\Phi(T_{2_{std}}) - \Phi(T_{1_{std}})} - (\frac{\phi(T_{1_{std}}) - \phi(T_{2_{std}})}{\Phi(T_{2_{std}}) - \Phi(T_{1_{std}})})^2) \tag{A.4}$$

Where ϕ = Gaussian PDF and Φ = Gaussian CDF. Putting together equations A.2 to A.4, it is possible to derive the needed statistics of the piecewise linear transform of a Gaussian Factor. Since they are independent by construction, we can calculate the mean and variance of the transformed factors and sum them to generate the results for each piece of equation 3.7. The workflow to do the calculations can be summarized as:

1. Use relation A.2 to calculate the mean of variance of each transform (e.g. $\sqrt{S_1}Y_1(u)$, $\sqrt{1-S_1}Y_2(u)$, etc.)
2. Use equations A.3 and A.4 to calculate the truncated mean and variance of each transformed variable from item 1.
3. Sum the truncated mean and variance following the piecewise linear rule.

Doing that, it is possible to treat its PDF as a mixture model of each piece of equation 3.7:

$$P(Z(\mathbf{u})) = w_1 P_1(Z_1(\mathbf{u})) + w_2 P_2(Z_2(\mathbf{u})) + w_3 P_3(Z_3(\mathbf{u})) + w_4 P_4(Z_4(\mathbf{u})) \tag{A.5}$$

The mean is going to be just the weighted sum of the mean of each piece:

$$E\{Z(\mathbf{u})\} = w_1 E\{Z_1(\mathbf{u})\} + w_2 E\{Z_2(\mathbf{u})\} + w_3 E\{Z_3(\mathbf{u})\} + w_4 E\{Z_4(\mathbf{u})\} \tag{A.6}$$

In the PLMR context, the weights w_i will have the same value, in this case 0.25. We can write the second moment of A.5 as:

$$\begin{aligned} E\{Z(\mathbf{u}))^2\} = w_1 \int_{SU_1} Z_1(\mathbf{u})^2 P_1(Z_1(\mathbf{u}))dZ_1 + w_2 \int_{SU_2} Z_2(\mathbf{u})^2 P_2(Z_2(\mathbf{u}))dZ_2 + \\ w_3 \int_{SU_3} Z_3(\mathbf{u})^2 P_3(Z_3(\mathbf{u}))dZ_3 + w_4 \int_{SU_4} Z_4(\mathbf{u})^2 P_4(Z_4(\mathbf{u}))dZ_4 \end{aligned} \tag{A.7}$$

Where SU_i is the support of each piece. Therefore:

$$\begin{aligned} E\{Z(\mathbf{u})^2\} = w_1(Var\{Z_1(\mathbf{u}) + m_{Z_1(\mathbf{u})}\} + w_2(Var\{Z_2(\mathbf{u}) + m_{Z_2(\mathbf{u})}\} + \\ w_3(Var(Z_3(\mathbf{u}) + m_{Z_3(\mathbf{u})}) + w_4(Var\{Z_4(\mathbf{u}) + m_{Z_4(\mathbf{u})}\} \end{aligned} \tag{A.8}$$

Therefore we can write the variance of equation 3.7 as:

$$Var\{Z(\mathbf{u})\} = \sum_{i=1}^{4} w_i Var\{Z_i(\mathbf{u}) + m_{Z_i(\mathbf{u})}\} - m_{Z(\mathbf{u})} \tag{A.9}$$

This approach seems to be extendable to any number of structures while using up to two truncation points.