Geometallurgical Modeling with Data Imputation and Response Surface Methodology

by

Paolo Christmasdato Kumara

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

Geostatistical modeling used to focus on the grade of the main commodity or metal being mined and sold for profit. As mining has developed, the metallurgical characteristics of the rock have become important. Geometallurgy tests are developed to understand the processing characteristics of the mined rock. Yet, geometallurgy tests are expensive and takes longer than geology grade assays. Multivariate geostatistical techniques are used to model the geometallurgy variables. Yet, many of the geometallurgy variables do not average linearly and are compositional, that is, they sum to unity. These complexities make modeling geometallurgy variables challenging.

Missing geometallurgical data may degrade the quality of prediction. There are two evident modeling frameworks that can be used. The first framework is an imputation framework which calculates the spatial continuity and relationships to grade variables to predict the missing data. The second framework is a response surface methodology (RSM) framework that accounts for the relationship to grade variables. There are different RSM techniques including (1) linear regression, (2) Alternating Conditional Expectations (ACE), and (3) random forest; that are compared to understand their advantages and disadvantages. The two frameworks perform differently in different circumstances. Considerations for the best framework are developed in this thesis.

Two new imputation techniques that account for data spatial continuity and complex multivariate relationships are developed. The first proposed technique, called RF-enhanced, alters the imputation likelihood mean calculation with random forest prediction without changing the variance while the second proposed technique, called RF-moment, alters both likelihood mean and variance. Both frameworks consider the prior spatial distribution in the same way as parametric

imputation. The proposed techniques improve the imputation accuracy in certain circumstances. Numerous examples are presented to provide guidance on technique selection.

Random forest regression does not always perform better in predicting missing values. Yet, both proposed imputation techniques still perform quite well and have a promising result for imputation development.

Acknowledgments

First of all, I would like to praise the Almighty God for all His blessings and graces throughout my research work until the end of the path.

I would like to thank Prof. Clayton Deutsch for his incredible teaching ability, his patience, his great humour and support that made me believe the impossible can be done when we try.
Thanks to Lembaga Pengelola Dana Pendidikan (LPDP), Centre for Computational Geostatistics (CCG) and all the sponsors from the industry for the financial support. And also to all my colleagues inside and outside of the research group for being a great mentor or/and discussion partner.

Thanks to all acquintances and friends that are too many to mention in this short page for helping me with everything I had to figure out from the first time I arrived in Canada three years ago with such a horrible English skill until now I have a decent self-confidence to continue my life forward.

The last but not the least, I would love to thank my parents for all their prayers that go along with every step of my journey. And also to my only brother in this life, Tedo, even though you never seem to care about me, I know deep inside you are the one who always supports every decision I make. Love you kak.

Table of Contents

List of Figures

Chapter 1
Introduction

Geostatistics is a statistical methodology that has been used to help people understand the resources and reserves of a mineral deposit. Geostatistics uses all the available information to predict rock properties at unsampled locations and assess uncertainty in the predictions. In general, mining practitioners should carefully manage the risks. The feasibility of exploitation over an area must be well studied before deciding to develop a mine. Resources are estimated and calculated with different exploration methods.

The accuracy of the predicted model is usually measured by the closeness of the exploration model block value to the value at the time of production. The output from the processing plant will likely be different from what is predicted; in some cases, metallurgists face difficulties to process the ore. Understanding geology, quantifying the spatial distribution of rock properties and managing metallurgical performance are collectively referred to as “geometallurgy”.

1.1. Background and Motivation

Geometallurgy is a scientific practice designed to integrate all relevant disciplines to maximize the economic value of a mining operation considering metallurgical responses like recovery, throughput and reagent/power consumption. This is motivated by many global factors such as tighter environmental regulation and commodity price fluctuations. Geometallurgy has evolved from its early simplicity of ‘geology + metallurgy’ conception. It is also recognized as an approach that can predict the risks related to resource development.

The main difference between geostatistical modeling of geology and geometallurgy variables is that the latter is a multivariate problem often with complex relationships between variables. Mineral resources in general are multivariate systems with many factors required to characterize their overall complexity. Yet, ore is typically described by using grade only, despite the fact that many geology factors such as grain sizes, mineralogy and rock texture may lead to different processing results.

Geometallurgy domains are qualitative attributes amenable to spatial block modeling (Dominy & O'Connor, 2016). Yet, geometallurgy does not replace the importance of geology and grade modeling and process design. Conventional geology or ore type/domains may not necessarily be appropriate for the various metallurgical processes being considered.

Mining engineers would plan and forecast with specific expectations, and metallurgists would know the type of ore being received at the plant. An integrated system would have geometallurgical data as part of the database. There are additional requirements to support a successful geometallurgy application: (1) in situ ore characterization from exploration drill data; (2) determination of ore types relative to specific applications; (3) setting up processing strategies based on ore characterization.

Underestimating the variation of ore properties is a prevalent problem (Dominy & O'Connor, 2016). It happens because specific ore types that need special treatment are not adequately sampled from core sampling in exploration. For instances, clay material occurrence in heap leaching may create some ponds on a heap leach cell that will prevent chemical percolation that extracts the mineral from the rock. To make mine projects viable in uncertain times, it is crucial to anticipate and manage surprises throughout the life of mine. Geometallurgy contributes to the sustainable extraction of the resource by informing optimal allocation of resources before production starts.

Optimization in geometallurgy may involve finding the optimal processing choice based on available information. Modern geometallurgy aims to understand grade, metallurgical and mining variability based on factors such as geochemistry, mineralogy, lithology, and alteration collected from spatially distributed samples. Other goals of geometallurgy include integrated mine planning, better modeling based on more information, improved understanding of the mineral deposit and enhanced management of the mining process leading to greater value.

There are four different perspectives on how to define domains and ore types: geology, mining, metallurgy, and geometallurgy (Jackson & Young, 2016). (1) Geology: Domains are defined to explain the geology domains that affect the resource definition such as rock types, lithology, and mineralogy; (2) Mining: Domains are used to optimize the resource that can be obtained by mine design and mine operation, and most of the times the domains will be simplified for operation efficiency; (3) Metallurgy: Domains are described by the ore behavior in a specific processing treatment; (4) Geometallurgy: Domains are derived from the fundamental geology domains with the intention to enable metallurgical workflow. Therefore, a hierarchy of domains is implied from geology to geometallurgy with each level having a different purpose.

Geometallurgy testing is often performed separately from geology grade testing. The differences are based on: (1) geometallurgy tests often require a larger mass to minimize sampling error and improve test reproducibility (Bax, et al., 2016) as compared to geological grade testing. The amount of sample needed is different and unique for different types of tests. Geometallurgy testing in exploration stage may be quite challenging as small diameter core drilling may not be sufficient. (2) Some geometallurgy tests will take more time to be complete than geological grade samples. For these reasons, geometallurgy samples are more expensive. This partially explains why geometallurgy tests are rare early in project appraisal. (3) Geometallurgy tests may also need

the mixing of ore that is considered representative to understand the treatment of ore at the time of mining. For instance, when the ore mixing behavior is 80% rocky ore and 20% clayey ore then the sample also has to have this composition. This may be unknown at the exploration drilling stage.

Coward and Dowd (2015) summarize the current general approach to geometallurgy modeling as: (1) identify the geology variables expected to explain crucial metallurgical performance variables; (2) sample and test these variables; (3) analyze the result and understand the relationship between geology variables and geometallurgy variables; and (4) develop techniques to calculate the value of these variables and incorporate them into geological block models. A detailed mineralogical description may help the geologist and metallurgist define a geometallurgy model in the deposit.

Estimating metal recovery and other plant performance variables is difficult because they are influenced by many multivariate factors, as mentioned above. This multivariate problem may be simplified by using constant recovery factors and plant efficiencies based on past experiences. These simplifications may be suitable for the prefeasibility stage of mineral exploration, but when it comes to reserve estimation stage, multivariate modeling should be utilized to improve plant performance predictions. A high resolution model of geometallurgical variables would allow more accurate mine planning and economic forecasting.

Managing multiple variables is sometimes troublesome because some of them are non-additive. Non-additive variables can be kriged (Deutsch, 2013) to understand the spatial features of the variable. Kriging generates a smoothed model that underestimates the variation of properties. Kriging may cause a bias for variables that do not average linearly. Non-additive variables should be simulated and combined using a relevant mixture rule calibrated from real data.

Machine learning has been well recognized as a merged idea from many disciplines to make computer learn, modify, and adapt so that it gets more accurate with the amount of data it uses. There are several ways to understand whether or not the machine is learning and they will lead to the classification of different machine algorithm types (Marsland, 2015): (1) supervised learning: a training set with response values is provided and algorithm is generalized based on this training set, (2) unsupervised learning: response values are not provided and algorithm tries to identify the similarities between the predictor values, (3) reinforcement training: when the algorithm is told when the response is incorrect but is not told how to correct it therefore the algorithm will try out many possibilities until it works out, and (4) evolutionary learning: biological adaptation is seen as a learning process. The most common type is the supervised learning where it includes regression and classification technique such as artificial neural network and random forest regression.

Neural networks are a wide class of flexible nonlinear regression and discriminant models, data reduction models, and nonlinear dynamical systems. (Sarle, 1994). Neural networks are modeled after biological neurons and are commonly used in predicting values. They work by working on a multilayer system where input layer consists of several nodes that work by using binary system that is controlled by hidden layer consists of several controlling nodes that control the combination of nodes in the input layer (Odom & Sharda, 1990). Then these combinations use binary system of 0 and 1 to predict a response neuron in the output layer. Random forest regression will be discussed in Chapter 2.

1.2. Variables Denotation

Geometallurgy variables are sometimes referred to as process variables, processing variables, and metallurgy variables, but in this thesis, the term geometallurgy variables will be used. Geology variables are more conventional rock properties. Some examples of geology variables are grade, mineralogy, lithology, and alteration. On the other hand, geometallurgy variables are rock properties that describe the performance in processing operations. Some examples of geometallurgy variables are throughput, metal recoveries, reagent process consumption, and tailings properties.

A geometallurgy variable is denoted with y and will be considered as a response variable that is dependent on predictor variables. Geology variables are denoted with x and will be considered as predictor variables that are also considered independent variables.

1.3. Problem Statement and Limitations

There are several techniques for building multivariate models and dealing with missing data. This thesis will help to establish the more appropriate of two evident workflows: (1) imputing the missing geometallurgy variables at the locations of the geology data then proceed with multivariate modeling or; (2) model the geologic variables, then apply response surface modeling to predict the geometallurgical response. A better way to predict missing values will also be proposed for the first workflow.

Machine learning techniques such as random forests may help with the response surface modeling workflow. The variogram in the imputation workflow can help explain the spatial features of the geometallurgy variable. An integrated workflow that builds multivariate aspects of

prediction using random forest and spatial aspects from variogram model will be proposed and tested.

The data in this thesis will be used under several assumptions: (1) there is no error in data measurements, and (2) all the data belong to one stationary domain. Access to the full sampling protocol is not available from an academic setting. Multiple domains would be treated one at a time.

1.4. Thesis Outline

Chapter 2 will review concepts and techniques related to geometallurgy modeling and discuss the research challenge. Workflows for model building techniques will be represented in Chapter 3. Comparisons between techniques and observed advantages and disadvantages will be discussed. Chapter 4 will show the proposed technique for improved modeling. Chapter 5 focuses on the practical application and implementation of the techniques to data. Validation and checking will be addressed. The main contributions and future work are discussed in Chapter 6.

Chapter 2
Background

Direct estimation of geometallurgy variables would not consider the more sampled grade variables. Also, the smoothing of estimation will cause the estimates to converge to the mean values that are non-informative. Estimated grade models give a unique result and are based on the data only as compared to simulation that will give non-unique results and show stochastic high and low areas that may be partly the result of a random number generator and not local measurements.

Geology grade variables average linearly and are easier to predict than geometallurgy variables. One approach is to model the geology variables then forecast the geometallurgy variables with a transfer function. Uncertainty of the prediction results is affected by geology uncertainty and this brings simulation with multiple realizations as a way to understand the uncertainty of geometallurgy variables.

2.1. Geometallurgy and Geometallurgy Models

Having all ore types predicted will help to calculate the output from mine process with less uncertainty. A geometallurgy model built in the early stages of mining, will improve predictions of revenue and capital expenditure. A geometallurgy model could help the mine planner optimize the sequence of extraction to maximize value and minimize cost. Lulea University of Technology in Sweden has demonstrated that a geometallurgy program can give up to 25% shorter payback time compared to cases when no geometallurgy information is available (Lishchuk, 2016).

Two methodologies to build a geometallurgy model are considered. The first approach is to impute all the missing values using existing imputation methods such as parametric or Gaussian Mixture Model (GMM) and then proceed with multivariate modeling. The second approach is to model all geologic variables in the area of interest and apply response surface modeling to predict the response geometallurgy variable based on the geology block model.

2.2. Multivariate Geostatistical Modeling

Geostatistics considers multiple variables to be modeled. The data must be multivariate Gaussian for most simulation techniques. Therefore transformation such as stepwise conditional transform (SCT) (Leuangthong & Deutsch, 2003; Rosenblatt, 1952), projection pursuit multivariate transform (PPMT) (Barnett, et al., 2014; Friedman, 1987), minimum maximum autocorrelation factors (MAF) (Desbarats & Dimitrakopoulos, 2000; Switzer & Green, 1984), 1984), and principal component analysis (PCA) (Davis & Greenes, 1983; Hotelling, 1933) are considered. Most of these transformation methods require all variables to be available at all data locations.

There are usually missing values due to technical issues such as missing core or a decision to save cost and time. The percentage of missing values in a database is varied. Predicting missing values before block modeling may help the geostatistical modeling process.

Multiple realizations of data are generated when imputing missing multivariate data. Each realization is then used through the multivariate transformation workflow to generate a geostatistical realization. The result of imputing missing values leads to greater accuracy and less uncertainty than removing incomplete data (Barnett & Deutsch, 2013).

Imputation considers spatial correlation and whatever information is available from collocated variables. An unbiased available information and a representative variance should increase the accuracy of imputation. An early imputation application to geologic data sampled conditional distributions with probability field simulation to generate realizations (Barnett & Deutsch, 2012). Simulating missing data requires the data to be preprocessed and transformed to have a Gaussian distribution. This step is important because conditional distributions are fully defined by a mean and variance; multivariate relationships are fully parameterized by correlation coefficients.

At each location for each variable, the original parametric imputation starts by calculating a prior mean ($\bar{y}_P$) and prior variance (σ^2_P) using geometallurgy variable data and the spatial correlation between data locations and the location data being predicted. The next step is calculating a likelihood mean ($\bar{y}_L$) and likelihood variance (σ^2_L) using collocated geology variables with weights calculated based on the multivariate correlation between geometallurgy and geology variable and also correlation between geology variables.

The next step is calculating updated mean ($\bar{y}_U$) and updated distribution (σ^2_U) using the result from second and third step with following equations (Ren, 2007):

$$\bar{y}_u = \frac{\bar{y}_L\, \sigma_P^2 + \bar{y}_P\, \sigma_L^2}{\sigma_P^2 - \sigma_P^2\, \sigma_L^2 + \sigma_L^2} \tag{1}$$

$$\sigma_U^2 = \frac{\sigma_L^2\, \sigma_P^2}{\sigma_P^2 - \sigma_P^2\, \sigma_L^2 + \sigma_L^2} \tag{2}$$

Monte Carlo Simulation is the final step of imputation and it uses a random probability value *p* and the standard normal CDF, *G*. Both of them are used to calculate simulated realization of the missing value, y_s with the following equation:

$$y_s = \sigma_U . G^{-1}(p) + \bar{y}_u \tag{3}$$

There are other imputation methods such as non-parametric imputation which was established for the imputation of complex multivariate variables by Barnett and Deutsch (2012) and the other method is based on Gaussian Mixture Model (GMM) that was proposed by Silva and Deutsch (2015). An alternative to predict missing values is to use a response surface.

2.3. Response Surface Methodology (RSM)

Response Surface Methodology (RSM) is a method to predict a missing geometallurgy variable values using its relationship to geology variables. RSM does not normally consider the spatial location of the data. This might be appropriate when dealing with a variable that has poor continuity or a high nugget effect. RSM considers some form of regression model. An important step is to determine the optimum combination of independent variables for optimal prediction.

The modern approach to RSM is to split the dataset into a training dataset and test dataset. The training dataset is a dataset where the geometallurgy variable and geology variables values are both known. The test dataset is a dataset where the geology variables values are known but geometallurgy variable values are kept back for testing. The RSM will build a function based on the relation of the geometallurgy variable to its collocated geology variables from the training data. This function would then be applied to predict missing values in the test dataset.

It is important to select variables that improve predictions of the response variable. This can be detected by analyzing the sensitivity of each variable and the coefficient of each variable when predicting the geometallurgy variable. Geology variables that have better correlation to the geometallurgy variable will give more impact to the result. This analysis is important because in most response surface problems, there are several geology variables that can be left out to improve the result.

There are many RSM techniques with different properties. Guidance on the appropriate technique to be used for different circumstances will be discussed in this thesis.

2.3.1. Linear Least Squares

The simplest RSM technique is based on a linear function. The geometallurgy variable is fit to a first order equation (Watson, 1967):

$$y^* = a_0 + \sum_{n=1}^{N} a_n x_n \tag{4}$$

Where y^* is the geometallurgy variable prediction result, n is the number of geology variables, a_0 is a constant term, a_n represents the coefficients of the linear parameters, x_n represents the geology variables. The equation does not have a residual value because the result is an estimated value of the geometallurgy variable.

For more complex relationships the polynomial function could be expanded to include quadratic terms according to the following equation:

$$y^* = a_0 + \sum_{n=1}^{N} a_n x_n + \sum_{n=1}^{N} b_n x_n^2 + \sum_{n=1}^{N} \sum_{m=1}^{M} c_{nm}\, x_n x_m \;\; \forall\, n \neq m \tag{5}$$

Where y^* is the response, x_n are the independent geology variables, the second part of the equation with the b coefficients are the second-order quadratic model for evaluating curvature, the third part describes the interaction between the different independent variables. The a coefficient in the linear model and the a, b, and c coefficients in the quadratic model are fitted to minimize the squared difference between the predicted and actual y values using the training data.

Linear least-squares cannot straightforwardly be used for categorical variables. Yet another limitation is its sensitivity to outliers. Nevertheless, linear least squares regression is widely used

because of its simplicity. There are types of data that are better described by non-linear functions. Real data often show non-linear relationships.

2.3.2. Alternating Conditional Expectations (ACE)

ACE is a non-parametric regression technique that can be very effective to understand and predict with complex multivariate data. The ACE algorithm is implemented by considering conditional expectations (Breiman & Friedman, 1985). The ACE algorithm determines optimal transformation functions between the geology variables and geometallurgy variable. ACE functions can be stable and reliable or overfit and unreliable. Barnett & Deutsch (2013) propose a modified algorithm that provides insight into the uncertainty of the ACE functions

ACE allows variables with any distribution to be considered. ACE models a function of the geometallurgy variable as the summation of smoothed functions for the geology variables:

$$\theta(y) = \sum_{i=1}^{n} \phi_i(x_i) \tag{6}$$

Where $\theta(y)$ is a function of the geometallurgy variable and ϕ_i are functions of the geology variables. No linear, quadratic, or logarithmic form needs to be assumed for these functions.

The order of the geology variables could change the fit. ACE can be sensitive to outliers for both response and geology variables. ACE functions can provide a substantial improvement over parametric regression models, particularly in the presence of non-linear features.

2.3.3. Random Forest

Ensemble learning starts with bagging of classification trees as described by Breiman (1996) where each tree is independently constructed using a bootstrap sample of the dataset. Breiman (2001) proposed random forests which add an additional layer of randomness to bagging.

Random Forest is a machine learning models for predictive analytics, widely used in practice. The random forest can be thought of as a bootstrap version of regression tree analysis where many trees are built based on subsets of the data (Breiman, 2001). They are a type of additive model that makes predictions by combining decisions from a sequence of base models.

Random Forest works by investigating different thresholds in different geology variables to find out which split leads to the greatest difference in the geometallurgy variable. Then it will investigate a second split and so on until there is no significant difference to be explained with further splits.

2.4. Research Challenge

Defining the proper technique for imputing missing values is challenging due to many possibilities of missing data categories: (1) missing completely at random (MCAR) when the probability of being missing is the same for all cases, (2) missing at random (MAR) when the probability of being missing is the same only within the same predictor variable group, and (3) missing not at random (MNAR) where the probability of being missing varies for unknown reason (Rubin, 1976).

Non-linear variables such as geometallurgical variables will have a different behavior from linear variables while defining the relation to the geological variables is also challenging. The research in this thesis will consider using mixture model to capture non-linearity in the data.

Random forests may capture the relationship with collocated variables better than parametric imputation. Altering prior distribution or imputation technique with random forest result may improve the quality of imputation result. Calculating a suitable variance so that the distribution

can be merged with the spatial information will be another challenge. Exploring random-forest-imputation techniques will be presented in this thesis.

Chapter 3
Model Building Framework

Understanding the spatial and multivariate characteristics of the geometallurgy variable is required. If the geometallurgy variable has clear spatial continuity, spatial modeling would be a reasonable way to build the model. On the other hand, when the spatial continuity is low relative to the data spacing, directly predicting the geometallurgy variable from the available geology variables may be more viable.

The fraction of missing data also affects the suitability of each prediction method. Imputation needs a variogram model for the geometallurgy variable; a lack of data will make variogram fitting very uncertain. RSM techniques requires understanding the relationships between variables that may be unstable with limited data.

All prediction methods are affected by the conditions mentioned above. The recommended approach in different circumstances will be based on the accuracy of the predictions and reproduction of spatial features. The two different frameworks are pictured in Figure 3.1 and are discussed below.

3.1. Imputation Framework

Imputation in recent geostatistical application is based on Bayesian Updating (BU) to build conditional distributions to simulate realizations of data (Ren, 2007). Conventional BU assumes the data all have normal distributions. Therefore, normal score transformation of all variables is a mandatory first step. Normal score transform is independently applied to each variable.

Calculating experimental variograms and fitting variogram models is done to describe the spatial characteristics of variables. Missing values are sampled using Monte Carlo simulation from an appropriate conditional distribution. Missing value imputation generates multiple realizations. These realizations reflect geometallurgy variable uncertainty.

After all missing values have been imputed, y (geometallurgical) and x (geological grade) variables are simulated at the same time where x data values are static and the y values are changed for each realization. This will generate multiple geometallurgy models that reproduce all available data.

Imputation does not necessarily capture complex multivariate relationships between variables. Another concern about imputation is when there are too few data values available for stable variogram inference. These conditions should be considered before applying imputation.

3.2. Response Surface Methodology (RSM) Framework

RSM determines the transformation functions that will optimize the prediction of a dependent variable. RSM is based on known values from a training dataset where the relationship can be fitted. RSM is useful to estimate variables without consideration for spatial continuity. Yet, the relationship to the geology variables should be reasonable. Moreover, outliers must be managed since they can have a large impact on the estimated values by affecting the relationship between variables. When RSM function is built from a training dataset with outlier, there will be a tendency to reproduce the outlier but the function that does not consider spatial characteristic may reproduce it at a wrong location that has similar independent variable value. RSM does not, in general, capture uncertainty.

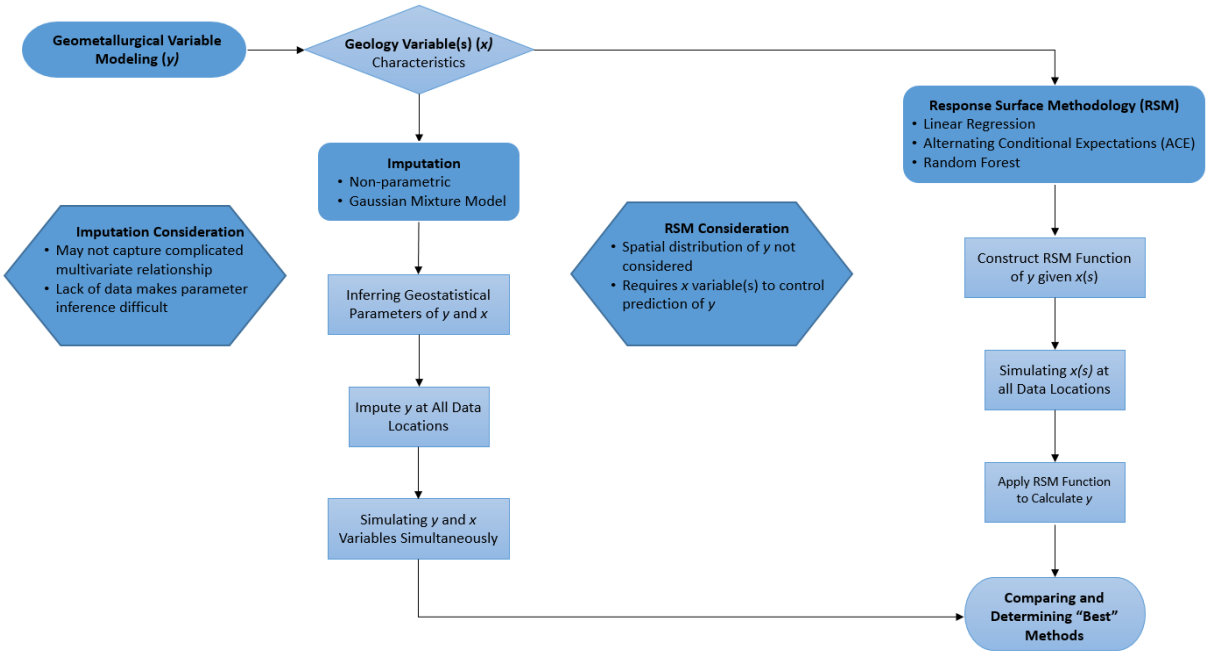


Figure 3.1: Geometallurgical modeling workflow diagram

3.3. Methodology for Comparison

Establishing the framework that performs the best is important. Performance can be measured by the coefficient of determination or R^2 defined as:

$$R^2 = 1 - \frac{E[[y^* - y]^2]}{E[[y - m]^2]} \tag{7}$$

Where the top part on the right is the mean squared error between the predicted value (y^*) and the real value (y) also called the residual sum of squares. And the bottom part on the right is the variance of the real values.

R^2 should fall between 0 and 1. If R^2 is 1, it means all predicted values fall precisely on the regression line and are equal to the real data. When it goes lower, it means the prediction correlation is getting lower and when R^2 equals to 0 means that the prediction is not related to the real data.

3.4. Geometallurgy Modeling Framework Application

The two different frameworks will be applied to synthetic data where all the samples are equally sampled, no error, and having Gaussian distribution. The two methods will be compared.

The synthetic dataset has two variables with a variogram range of 15 and correlation between them equal to 0.6. One geometallurgy variable (y) and one geology variable (x) are generated on a 50x50 2D grid with square spacing of 1 unit yielding 2500 data values. The data are then randomly split into 3 datasets which are: (1) Dataset 1 which has 100 locations with known x and y; (2) Dataset 2 which has 200 locations that are different from dataset 1 with known x but unknown y; (3) Dataset 3 which has the remaining 2200 locations to validate the prediction result. Details of the data are shown in Figure 3.2.

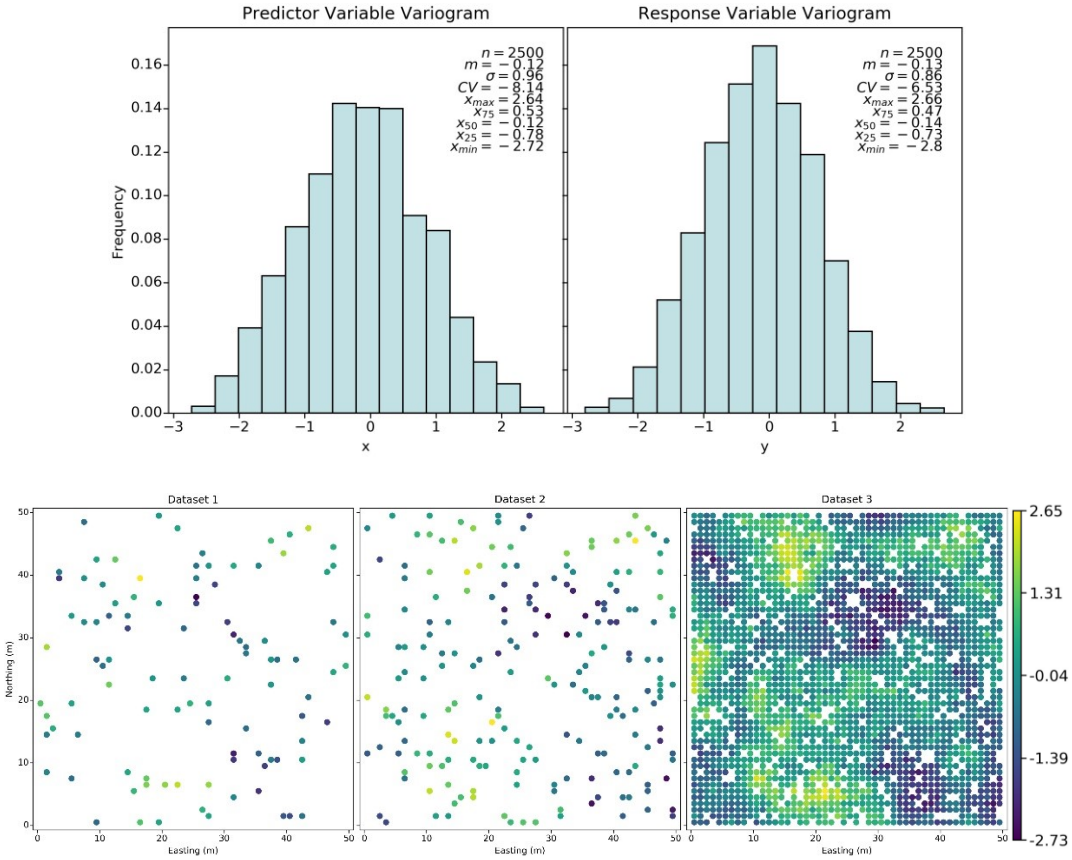
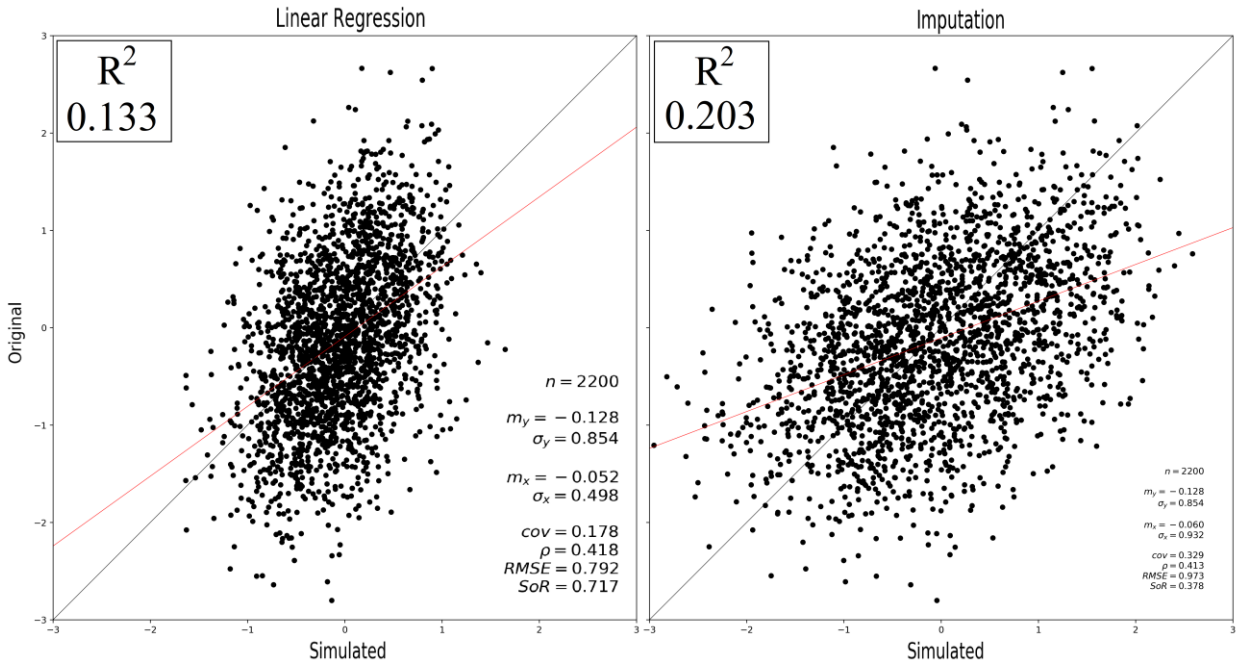


Figure 3.2. Characteristic of dataset used in this chapter: histograms of the two variables (top); and location maps of dataset 1, 2, and 3 (bottom)

RSM is only done using linear least squares method with following steps: (1) Regression formula is calculated using dataset 1 of y given x; (2) 100 realizations of x on the full grid are simulated using all 300 x values from dataset 1 and dataset 2; (3) RSM function is applied to the 100 realizations at 2200 locations to estimate y values at each location in each data files.

Regarding the imputation workflow, the steps include: (1) Missing y values at 200 locations in dataset 2 are imputed yielding 100 realizations of y data; (2) 100 realizations of y data are concatenated with dataset 1 resulting in 100 data files where 100 values are fixed and 200 values are changed for each data file; and (3) 100 realizations of y at 2200 locations are simulated using all 300 values in each dataset.

The results show that imputation performs better than regression. The average of 100 R^2 values for linear regression is 0.133 while it is 0.203 for imputation. The results are not very good due to lack of training data values but still can be used to compare the two methods. The comparison between methods is shown in Figure 3.3. The imputation result has more variance than linear regression; 0.932 and 0.498 respectively. That happens because the imputation result is simulated instead of being regressed.



(a) (b)

Figure 3.3: Realization of one dataset cross-plot validation of (a) linear regression and (b) imputation method

Multiple realizations would have different results. Using the average of 100 realizations may show the result more clearly because it averages some random variations as shown in Figure 3.4. The average of 100 realizations gives the R^2 of 0.210 for linear regression and 0.263 for imputation.

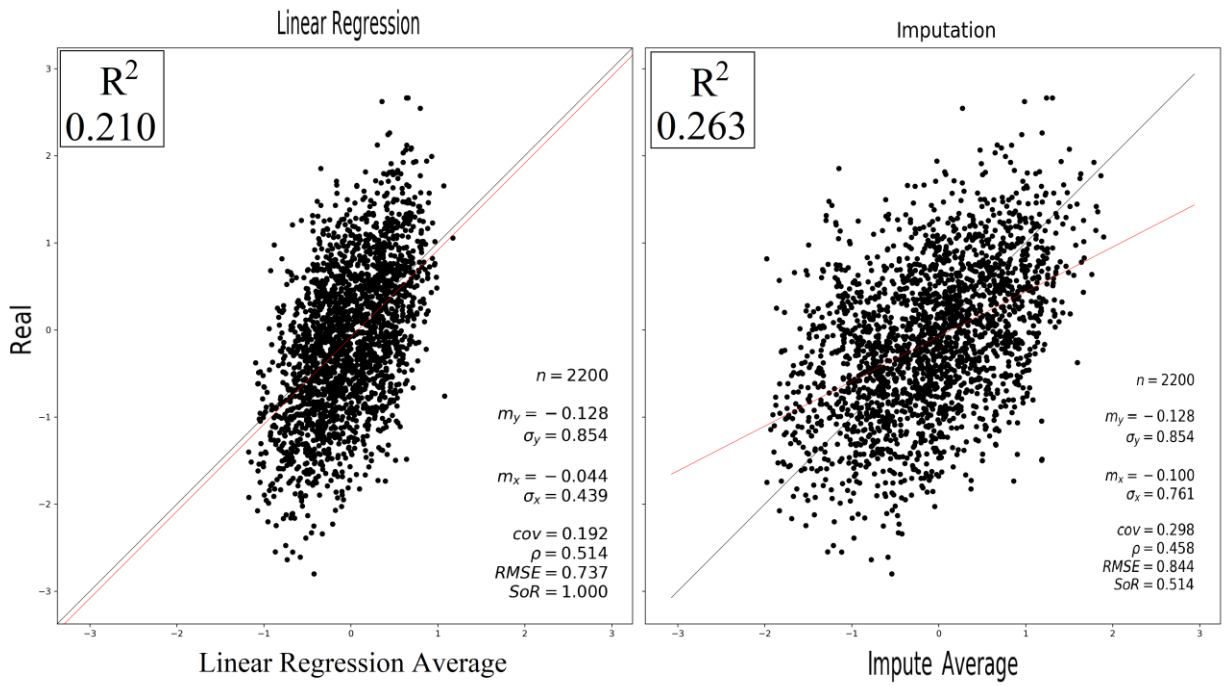


Figure 3.4: Realizations average cross-plot validation of (a) linear regression and (b) imputation method

Two cross-plots on Figure 3.4 show that the imputation framework is somewhat better than the linear regression framework. The conclusion is not definitive because the result may change for different datasets. As expected, the averages from both frameworks are closely related, see Figure 3.5.

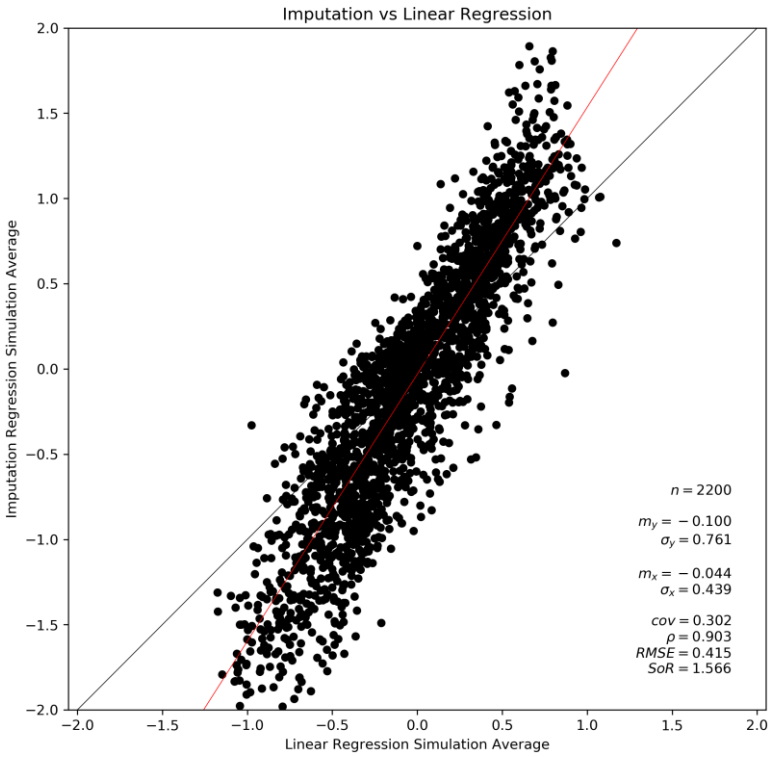


Figure 3.5: Cross-plot of simulation average between linear regression and imputation

3.5. Discussion

Linear regression and imputation frameworks perform almost at the same level. This happens because the synthetic dataset has reasonably good spatial features and correlation between variables. It shows that the two modeling frameworks are valid to build geometallurgical models. The advantage of the imputation framework is that the simulated realizations of the geometallurgy variable reproduce the specified variogram for the variable. The RSM method provides a too-smooth result.

There are some advantages and disadvantages of each framework and using the less appropriate framework on a certain dataset may lead to increased error. A new framework that can be used in most situations will be discussed in the next chapter.

Chapter 4
Proposed Random Forest Imputation

Variables with known clear spatial correlation may be better predicted using imputation. Response surface modeling may outperform imputation when the spatial structure is not well understood and there is a reasonable relationship to measured geological variables. The proposed prediction technique is believed to provide robust predictions for a wide variety of dataset characteristics.

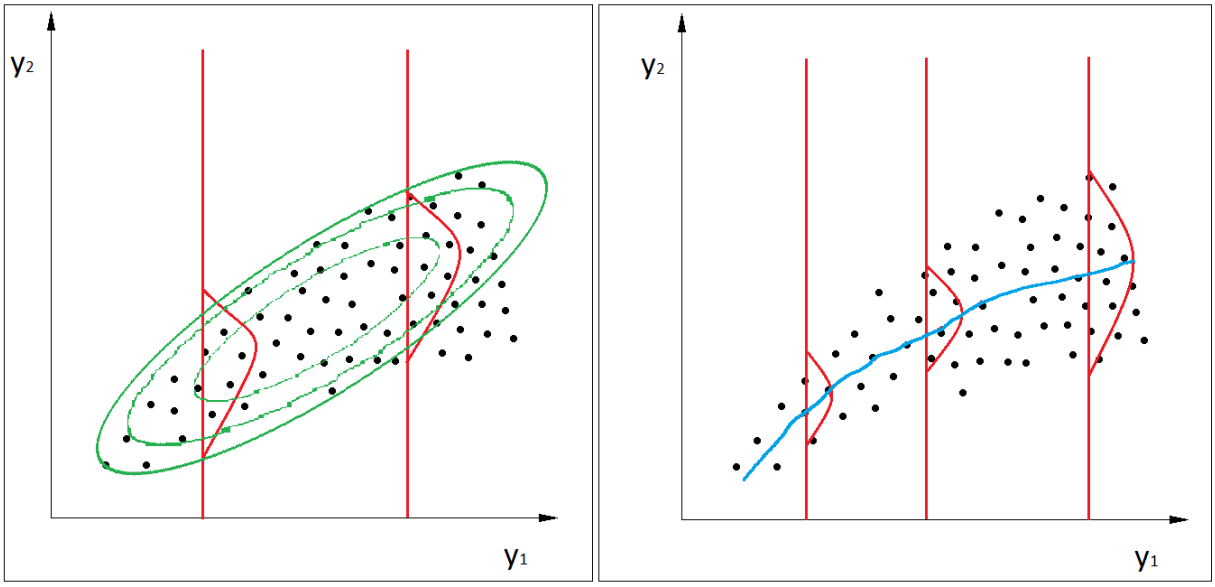
4.1. Proposed Framework

The random forest response surface regression technique performs well to capture multivariate relationships between variables, but response surface techniques fail to capture spatial characteristics of the variable being predicted. On the other hand, imputation can capture both multivariate and spatial features, but multivariate relations are not captured as well as the random forest. Injecting aspects of the random forest technique in the imputation method will be developed here.

There are two aspects of imputation discussed in Chapter 2 that are (1) the prior distribution that accounts the spatial information, and (2) the likelihood distribution that accounts for information from the collocated variables. Note that reference to "prior" and "likelihood" distributions is consistent with published papers on the subject, but not consistent with conventional statistical notation. The idea developed here is to alter the likelihood distribution with values calculated from the random forest method. After normal score transform of the original

variable, the likelihood distribution is often considered to be Gaussian defined by a mean and variance.

Parametric imputation calculates the parameters of the likelihood distribution using linear regression that has the same equation for every location as shown on the left side of Figure 4.1. The black dots show the linear correlation of two different variables and the green line shows how Gaussian the variables are. The red lines on the left figure show the variance and mean calculation of the data at different conditioning values. This is correct for Gaussian linear data, but may not be suited to geometallurgy variables with complex non-linear behavior. The likelihood distribution is supposed to be different at each location due to its non-linearity. On the other hand, the random forest technique can adapt to complex features in the data, as shown with blue line in Figure 4.1. This flexibility could be used to adapt the distribution calculation to every data location.



(a) (b)

Figure 4.1: Comparison between calculating likelihood distribution using (a) linear regression and (b) random forest

The first two steps of the proposed framework are the same as the parametric imputation method described as parametric Bayesian Updating (BU) in Barnett & Deutsch (2013) which are (1) normal score transformation of all variables, and (2) calculating the prior distribution of each variables. The normal score transformation is required for the assumption of the multivariate Gaussian distribution where the conditional distributions Gaussian and are fully defined by a mean and variance. The mean and variance are calculated using simple kriging with the spatial correlation between the unsampled location and the data locations.

The third step is different. The proposed framework calculates the likelihood mean and variance using the random forest regression technique. The updated distribution will be calculated as before with Equations (1) and (2) shown in the Chapter 2. Monte Carlo Simulation samples the updated distribution to simulate realizations of the missing value.

4.2. Altering Likelihood Mean

The likelihood mean, $\bar{y}_L(\mathbf{u})$, at location $\mathbf{u}$ will be substituted with the expected result of random forest regression given the collocated geology variables. The random forest technique randomly splits training dataset with one geology variables criteria at a time until a minimum number of data fall in each branch. The number of splitting (called trees) affects the accuracy of the regression.

The likelihood mean of this proposed framework is calculated in normal score units. This is done to facilitate merging with the normal score prior distribution. Different random seed values can be used to draw multiple realizations. The random forest can adapt to complex relationship with geology variables, which can be an improvement to the current imputation method.

4.3. Altering Likelihood Variance

The variance is the expected value of the squared deviation from the mean. From the definition, variance, $\sigma^2_L(\mathbf{u})$, at location $\mathbf{u}$ can be expressed as:

$$\sigma_L^2(\boldsymbol{u}) = E[y^2(\boldsymbol{u})] - E[y(\boldsymbol{u})]^2 \tag{8}$$

$E[y(\boldsymbol{u})]$ is the likelihood mean. The expected squared value $E[y^2(\mathbf{u})]$, could also be calculated using random forest as follows:

$$\mathrm{E}[y^2(\mathbf{u})] = E_{x,y}\left(y^2 - (\frac{1}{n}\sum_{n=1}^{N} h(x^2{}_n)\right)^2 \tag{9}$$

Considering squared values of x_n with n=1,2,3,…,N that have been normal score transformed is proposed to calculate the variance value and $h(x^2{}_n)$ is the number of trees defined by numerous predictors. The difference of the proposed likelihood variance and the one from parametric imputation is that the former has variance values that depend on the data values. This can be an advantage for complex non-linear data.

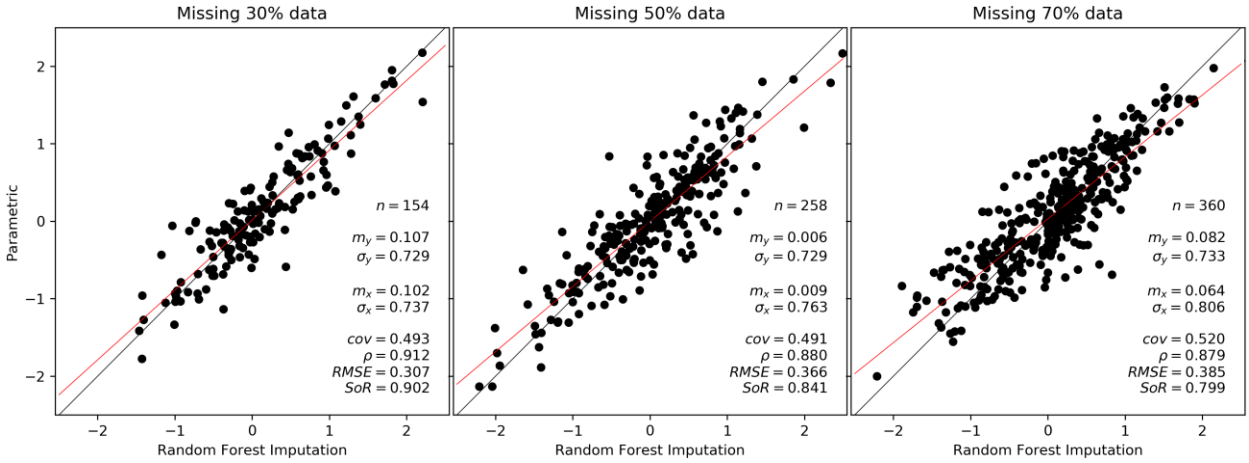
Numerical experiments show that the likelihood variance is unstable when calculated this way. In fact, the likelihood variance calculated from x^2 may even be negative. A global likelihood variance calculation will be considered with the mean coming from the random forest imputation framework. The global likelihood variance is calculated using the mean of the random forest regression result. The proposed likelihood variance is as follows:

$$\sigma_L^2(\boldsymbol{u}) = \overline{E[\,y^2(\boldsymbol{u})\,]} - \overline{E[[\,y(\boldsymbol{u})\,]^2]} \tag{10}$$

Where $\overline{E[\,y^2(\boldsymbol{u})]}$ is the global statistical mean of random forest squared prediction and $\overline{E[[\,y(\boldsymbol{u})\,]^2]}$ is the squared value of the global statistical mean of normal random forest. The equation will only yield one value for every data location, which is the likelihood variance that is data value independent.

4.4. Performance of Proposed Workflow

The likelihood mean calculated using linear least squares regression and the random forest are highly correlated to each other because they are both valid techniques to predict missing values. Their correlation is close to 0.75 on the synthetic Gaussian dataset as shown in Figure 4.2. This difference is significant and the results of the two approaches would be different. The performance of random forest compared with linear regression will be discussed in the next chapter.



(a) (b) (c)

Figure 4.2: Cross-plots between likelihood mean of parametric imputation and random forest imputation on three different datasets of missing (a) 30 percent; (b) 50 percent; and (c) 70 percent

Likelihood variance from proposed workflow tends to be 35% lower due to random forest incapability to reproduce data variation and tend to be more stable than parametric imputation as seen in Table 4.1. Therefore, result from proposed frameworks will be more narrow and cannot cover a wide spread data. The difference may be higher when it is applied to non-linear variable.

Table 4.1. Comparison of likelihood variance between parametric imputation and random forest imputation

Missing Percentage	Likelihood Variance Parametric Imputation	Likelihood Variance Random Forest Imputation	Difference
30%	0.712	0.412	42%
50%	0.680	0.440	35%
70%	0.728	0.473	35%

In this thesis, there are two frameworks to understand how well the proposed likelihood mean calculation and likelihood variance calculation work. The first framework called RF-enhanced uses the proposed likelihood mean with the variance from linear regression. The second called RF-moment uses both proposed likelihood mean and likelihood variance from the random forest. These two frameworks are applied to Gaussian dataset of 1 geology variable and 1 geometallurgy variable with known variograms and relationships. The dataset contains 2500 equally sampled data locations and some values that will be randomly left out from 10% to 90% of the data yields 9 datasets and each of them is simulated 100 times. Then, the prediction result is compared to the real data values and the R^2 calculated for every missing data percentage.

The result from the two frameworks are shown in Figure 4.3. The comparison uses one out of 100 realizations and shows how the proposed frameworks perform at the same level as parametric imputation. Among the two proposed frameworks, RF-moment consistently performs better than RF-enhanced. From this example, the proposed frameworks could lead to an improvement in imputation.

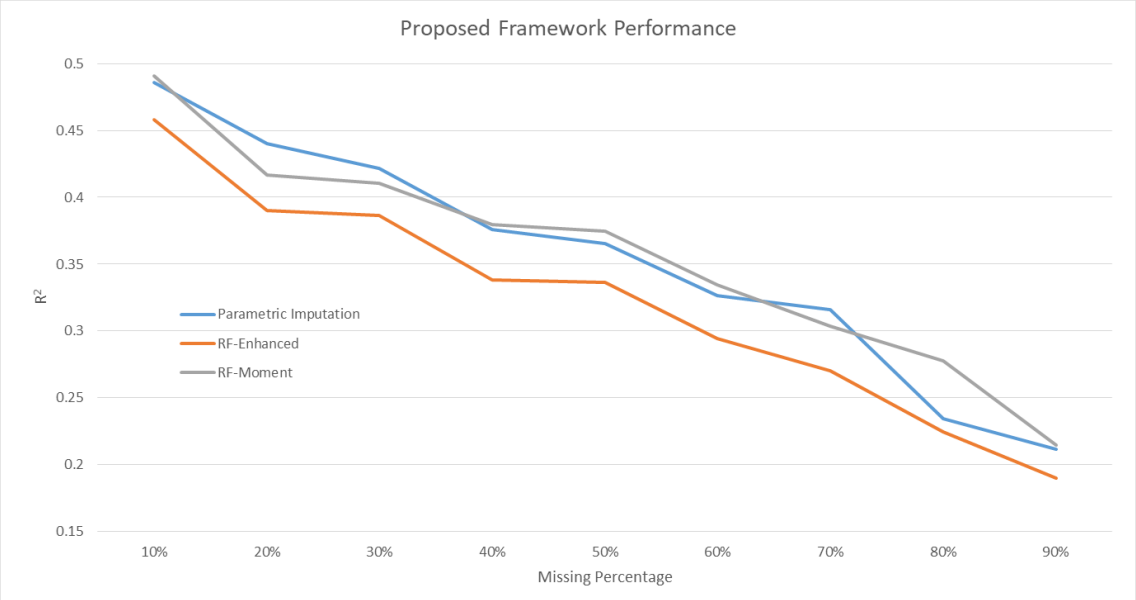


Figure 4.3: Accuracy level comparison between two proposed frameworks and parametric imputation

Only using one realization may lead to randomness influencing the results. Averaging will capture all the realizations uncertainty but will only give final result. The averaging of the results are shown in Figure 4.4. Parametric imputation performs better than two proposed frameworks. This happened because when the variability of random number given to linear regression is averaged, the result will be a kriging result whereas kriging is proven to be the best linear unbiased prediction method (Cressie, 1989). On the other hand, with less variability of random forest and its tendency to not overestimate the prediction result, averaging the simulation results does not help to increase the accuracy.

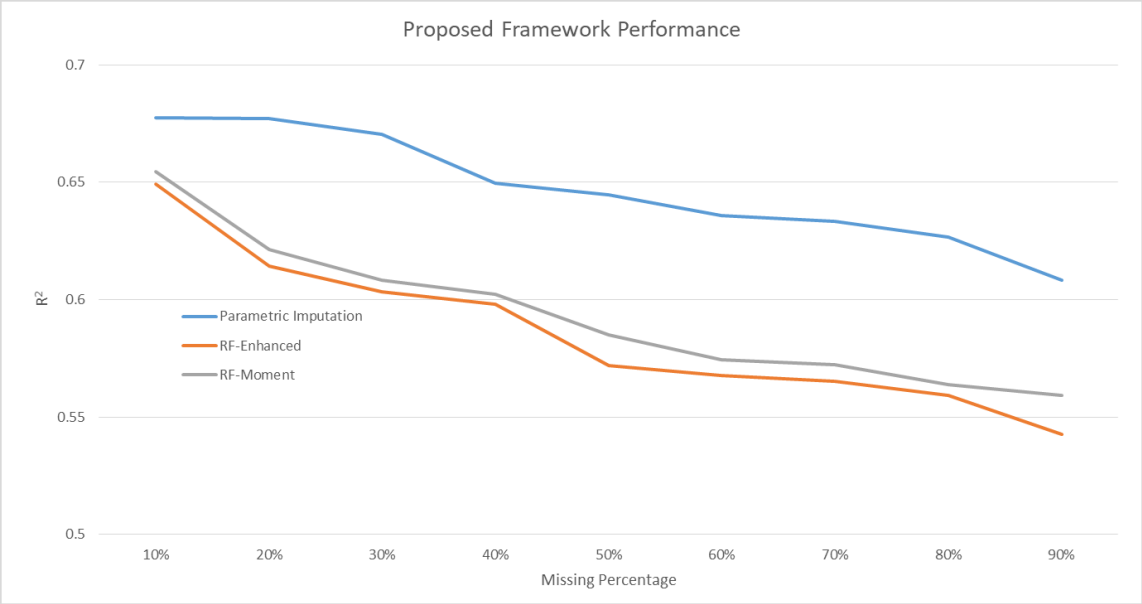


Figure 4.4: R^2 comparison between realizations average result of two proposed frameworks and parametric imputation

As mentioned above, random forest is more stable than linear regression therefore likelihood mean calculation result from many realizations is not as varied as the one that is calculated using linear regression in the imputation method. That leads to stable result of 100 realizations as seen in Figure 4.3 and Figure 4.4 where there is no significant difference between one realization and the average of realizations.

4.5. Discussion

The likelihood mean may be calculated by a random forest in the imputation method. The Gaussian synthetic dataset gives reasonable results with the proposed frameworks. The preliminary conclusion of the proposed frameworks from the simulation is that they are valid to predict missing values where there is only one estimated result needed. Validation is important to evaluate whether the proposed frameworks are overfit.

These methods will be tested on the other datasets to establish their place. More geological variables and non-linear variables should perform better. The next chapter compares the proposed frameworks to other prediction method including RSM. Defining the best out of three different RSM techniques mentioned in Chapter Two will also be discussed in the next chapter.

Chapter 5
Implementation and Case Study

Comparing RSM and imputation may seem unfair as the former does not consider spatial characteristics; however, the spatial characteristic comparison can also be a disadvantage to the prediction particularly if the spatial characteristics are poorly known. The comparison with different datasets will help define the most suitable technique for given cases. There are three techniques discussed in this thesis; choosing the best technique for different situations is important.

The RSM techniques discussed in this thesis are Linear Least Squares (Watson, 1967), ACE (Breiman and Friedman, 1985), and Random Forest (Breiman, 2001). The timeline indicates how the techniques evolved and perhaps improved over the years. Yet, each technique may have its advantages and disadvantages in different situations. All of them could be used to predict missing values on datasets where a certain percentage of values are left out for testing.

5.1. Comparison of RSM Techniques

5.1.1. Synthetic Dataset

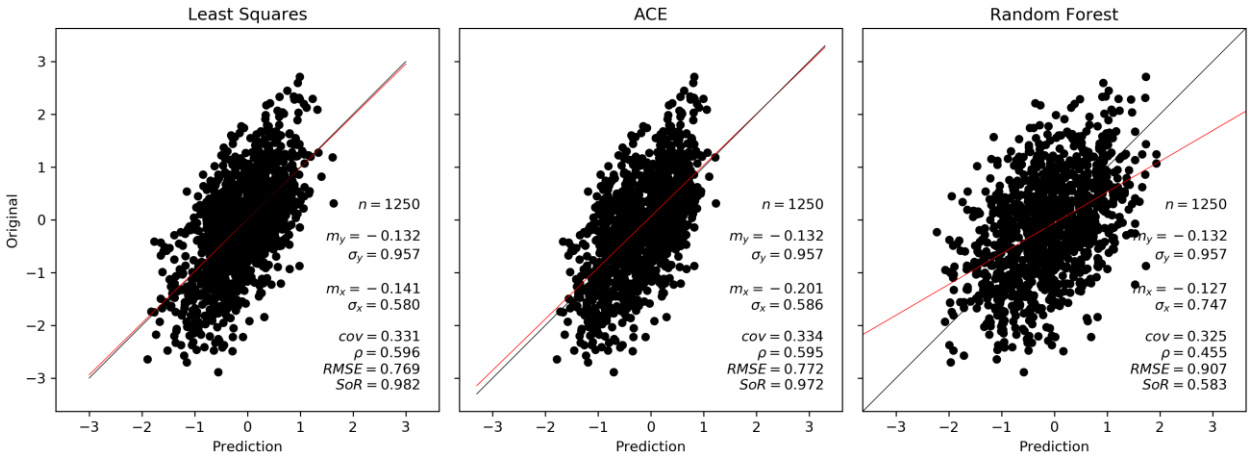
There are three different synthetic datasets used with one geometallurgy variable and a different number of geological variables. Additional collocated geological variables could increase the accuracy of prediction depending on the dataset characteristics.

The comparison focuses on the missing value prediction and the steps of prediction are: (1) Generate data with y as the geometallurgy variable name and x_n as the geological variables where $n = 1, 2,\ldots,N$. (2) 10% of data values are randomly left out 100 times yielding 100 new datasets with 10% missing data. (3) 100 new datasets are split into a training dataset with known x_n and

known y and a test dataset with known x_n and unknown y. (4) Step (2) and (3) are repeated with increasing missing values for 10% intervals, yielding 900 training datasets and 900 test datasets in total. (5) The training datasets are used to generate RSM function for the three techniques. (6) RSM function is applied to predict missing y in each test dataset. (7) Estimated and original y values are compared. (8) Prediction accuracy of each technique are calculated using the coefficient of determination (R^2) and the average R^2 of 100 datasets is compared.

5.1.1.1. Synthetic Gaussian Dataset with One Geology Variable

The first dataset is the same dataset used in Chapter 3 and the characteristics are shown in Figure 3.2. The cross-plot between prediction result and original y value of one dataset with 50% missing data percentage is shown in Figure 5.1. Prediction accuracy of least squares, ACE, and Random forest for this dataset is 0.355, 0.354, and 0.207 respectively. Least squares and ACE perform at almost the same level while random forest is worse. Nonetheless, random forest can reproduce the mean and variance better



(a) (b) (c)

Figure 5.1. Cross-plots of original and prediction result on synthetic Gaussian dataset with one geology variable using (a) least squares; (b) ACE; and (c) random forest

Least squares and ACE perform well on all 100 datasets with one geology variable as shown in Figure 5.2. From the missing data values percentage, all three techniques seem stable with increased missing percentage. It can be said that for predicting missing values on a dataset that has one geology and one geometallurgy variable with 0.6 correlation coefficient regardless the percentage of missing values, least square and ACE are suitable techniques to use with no significant difference between them two. This is likely due to the simplicity and Gaussian distribution of the data where a linear prediction is theoretically correct. Also, the R^2 does not decrease because there are enough training data in all cases to provide a reasonable response surface fit.

Random forest does not perform well in this case. This can be caused by the number of geology variables since more geology variables will make random forest able to sample more variables as candidates at each split. In this thesis, number of times the out of bag data are permuted per tree for assessing variable importance is set to 1. Larger than 1 will give slightly more stable estimation even though will not be very effective time-wise.

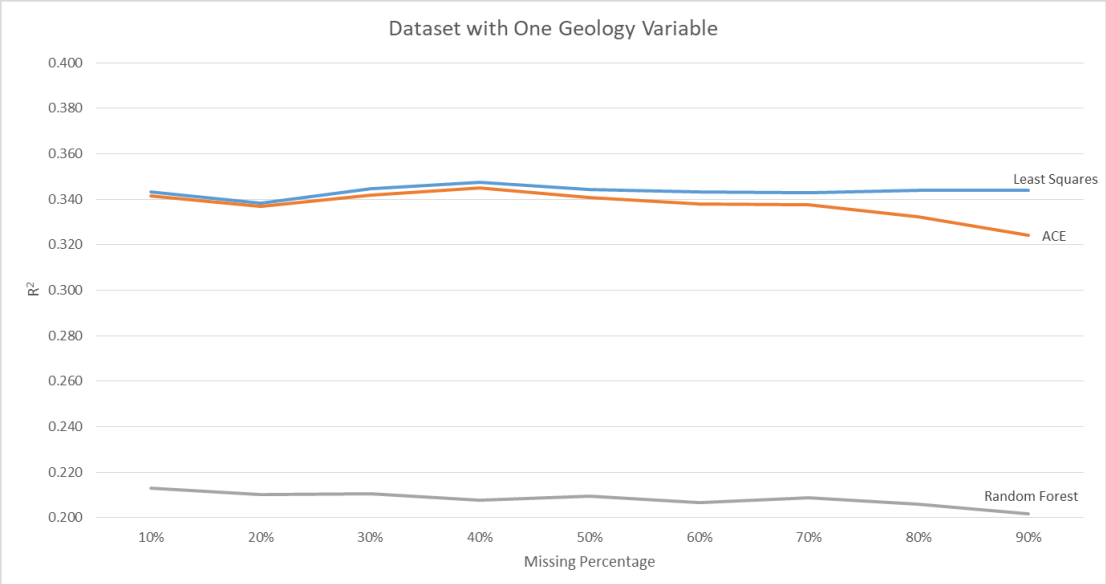


Figure 5.2. Average R^2 from 100 simulations comparison between the three techniques on synthetic Gaussian dataset with one geology variable

5.1.1.2. Synthetic Gaussian Dataset with Three Geology Variables

The second Gaussian dataset has three geology variables and one geometallurgy variable with correlation matrix shown in Figure 5.3. The correlation coefficient between the geology variables is relatively low to show that the two different predictor variables increases the prediction accuracy.

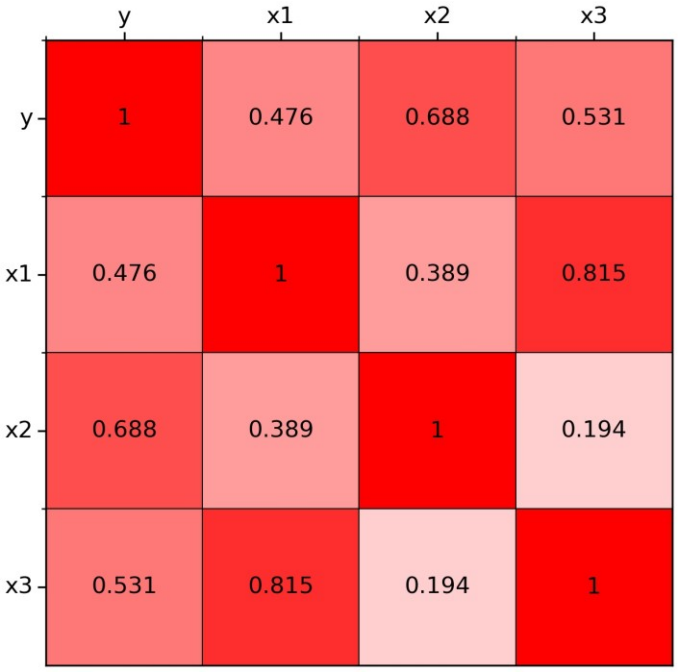
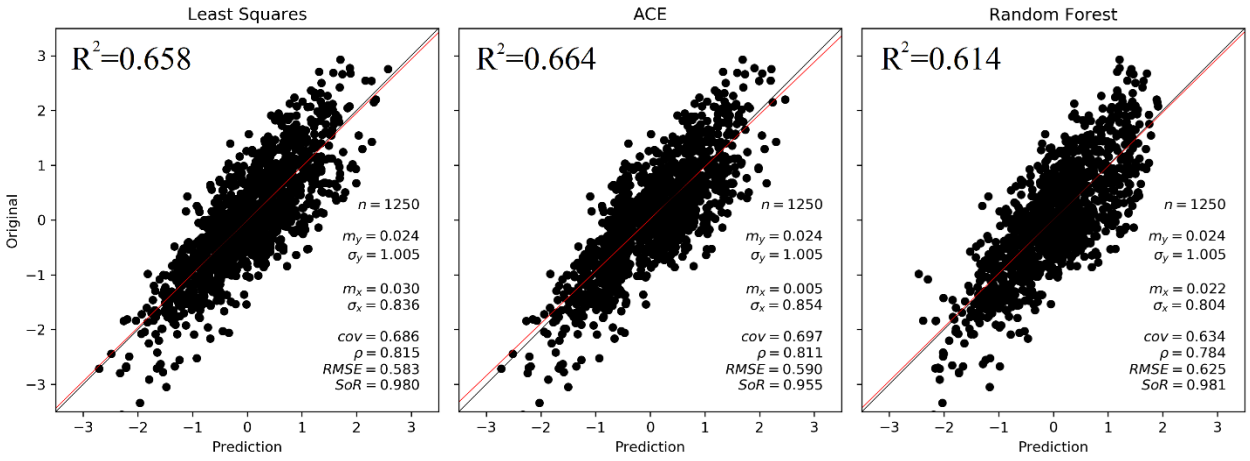


Figure 5.3. Correlation matrix of variables in the second synthetic Gaussian dataset

Figure 5.4 shows the cross-plot between prediction result and original *y* value of one dataset with 50% missing data percentage. R^2 of least squares, ACE, and random forest technique for this dataset is 0.658, 0.664, and 0.614 respectively. Adding a geology variable improves the prediction R^2. Random forest still performs the worst between the three techniques but the gap is smaller. ACE and least squares perform at the same level with no significant difference between them.



(a) (b) (c)

Figure 5.4. Cross-plots of original and prediction result on synthetic Gaussian dataset with three geology variables using (a) least squares; (b) ACE; and (c) random forest

The results with different missing data value percentage is summarized in Figure 5.5. As before, the missing value percentage does not affect the accuracy of prediction because there is enough training data in all cases. All calculations show that least squares and ACE perform better than random forest. A linear approach is a suitable technique for a Gaussian dataset.

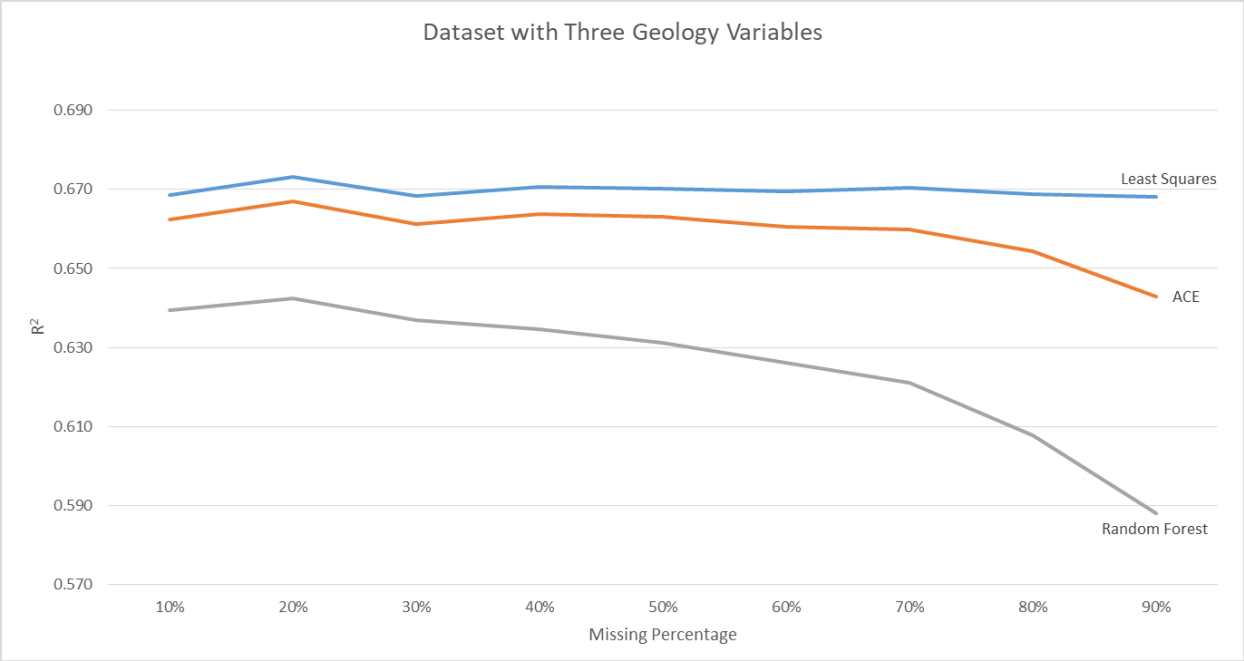


Figure 5.5. Average R^2 from 100 simulations comparison between the three techniques on synthetic Gaussian dataset with three geology variables

5.1.1.3. Synthetic Gaussian Dataset with Five Geology Variables

The third Gaussian dataset has five geology variables and one geometallurgy variable with correlation matrix shown in Figure 5.6. Five geology variables are used to understand what happens when more redundant data are considered. The geometallurgy variable is reasonably correlated with all geology variables with the minimum correlation of 0.334 and maximum of 0.623. The geology variables have various correlations ranging from 0.236 to 0.905.

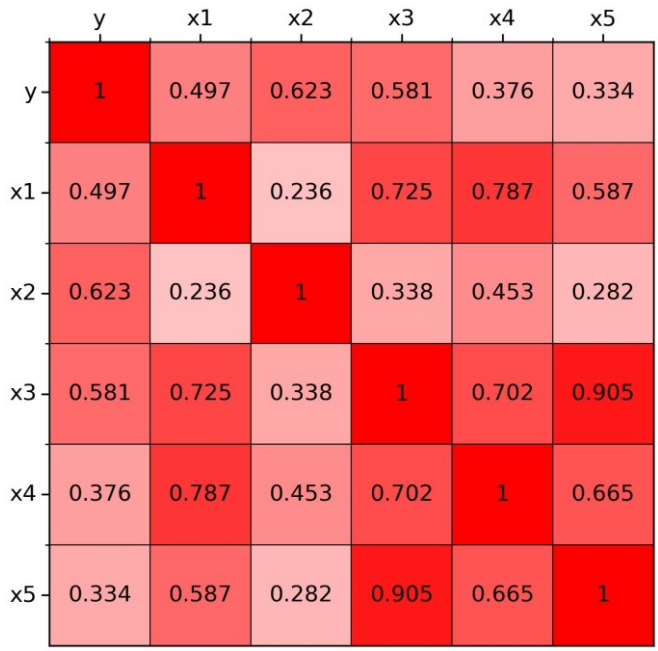
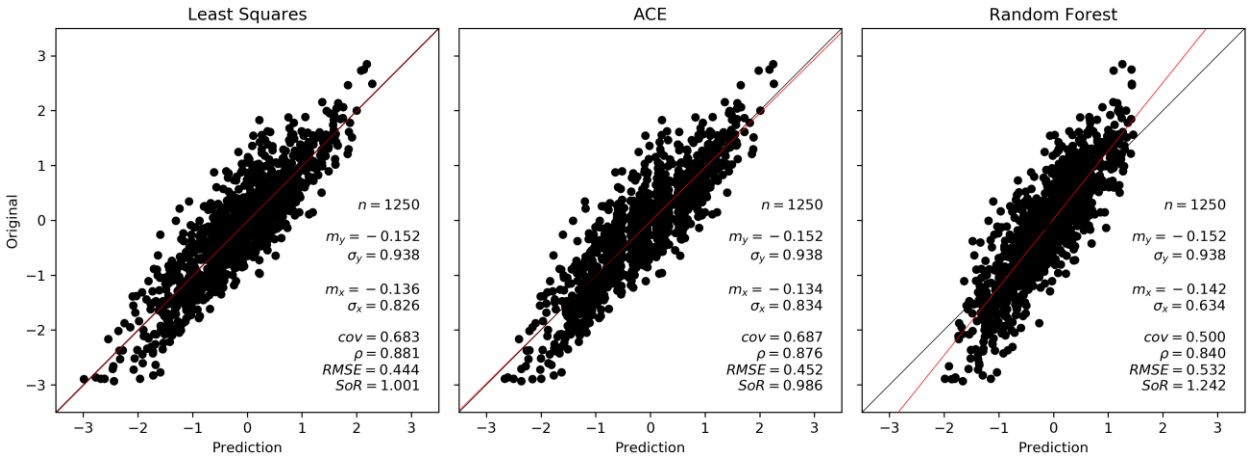


Figure 5.6. Correlation matrix of variables in the third synthetic Gaussian dataset

Cross-plots in Figure 5.7 show how the prediction result of least squares, ACE, and Random forest technique on 50% missing values dataset getting improve with more geology variables to R^2 values of 0.776, 0.768, and 0.705 respectively. As above, the true underlying linear Gaussian nature of the data leads to the simpler techniques performing better.



(a) (b) (c)

Figure 5.7. Cross-plots of original and prediction result on synthetic Gaussian dataset with five geology variables using (a) least squares; (b) ACE; and (c) random forest

The results for different missing values percentage is shown in Figure 5.8. Random forest appears to perform worse with less training data. The R^2 of existing techniques is pretty high even for 90% missing data, which implies there are sufficient training data to fit the response surface.

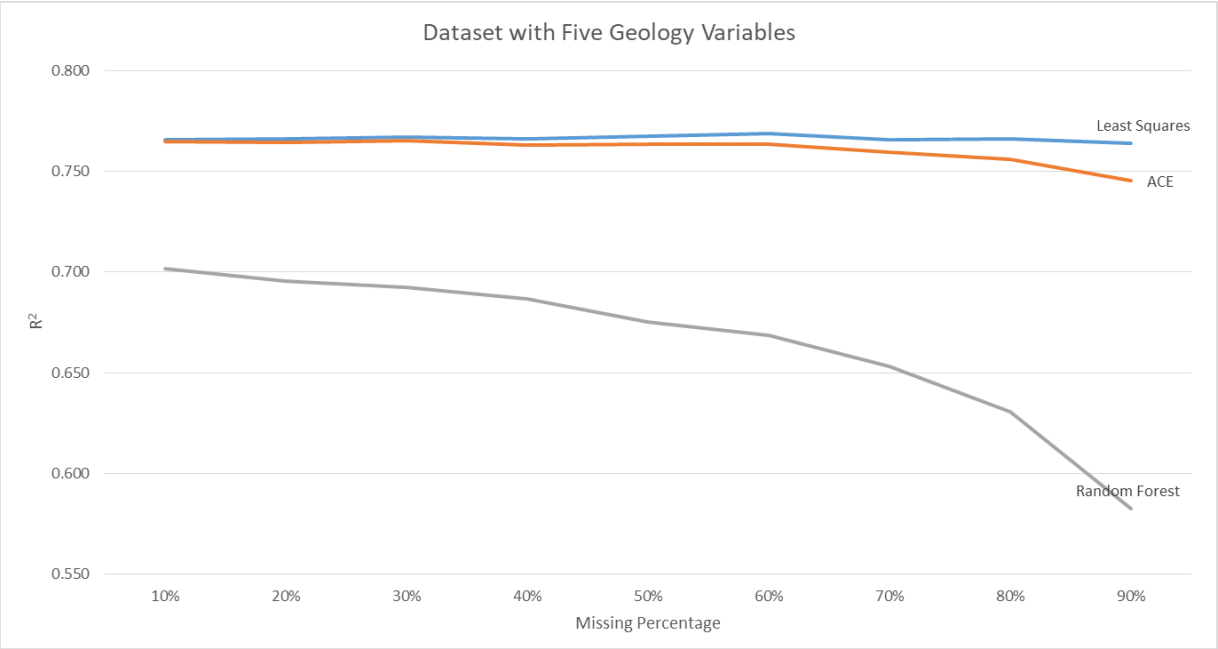


Figure 5.8. Average R^2 from 100 simulations comparison between the three techniques on synthetic Gaussian dataset with five geology variables

All three examples support the same conclusions. The result of least squares technique is appropriate for predicting linear Gaussian datasets.

5.1.2. Non-Linear Dataset

A non-linear dataset is generated from synthetic dataset used above by randomly removing some values away so the variables no longer have a linear Gaussian distribution. This dataset represents non-linearity that would be expected in most geometallurgy datasets and utilized to understand how well prediction techniques perform for such situation.

Non-linear data with a single geology variable is shown in Figure 5.9. A threshold following the red line in the figure is considered. The threshold reduces the number of values from 2500 to 799 and changed the histogram of *x* and *y* to have skewed distribution as seen in Figure 5.10. Yet, the correlation coefficient between two variables remains unchanged, at around 0.6.

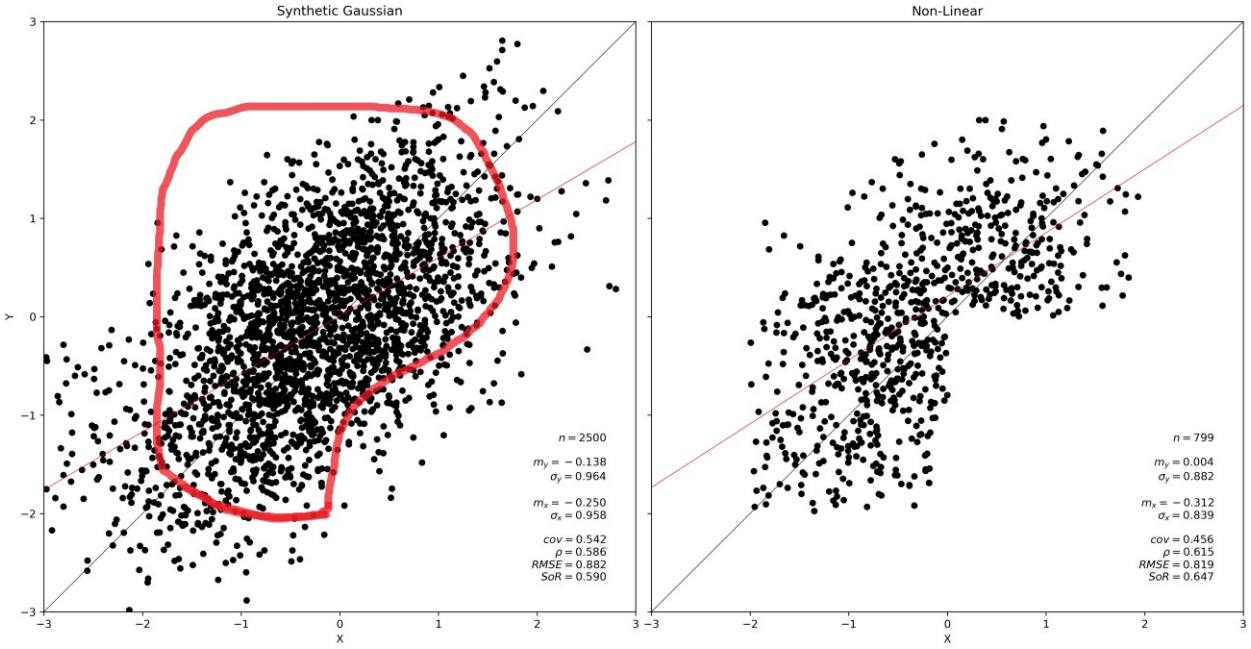


Figure 5.9. Non-linear synthetic Gaussian dataset

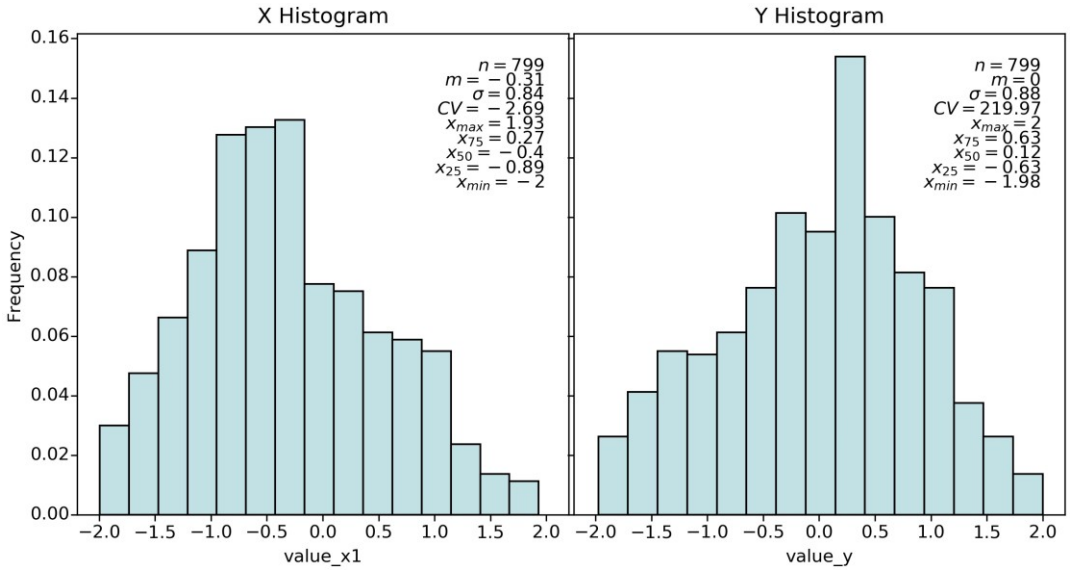
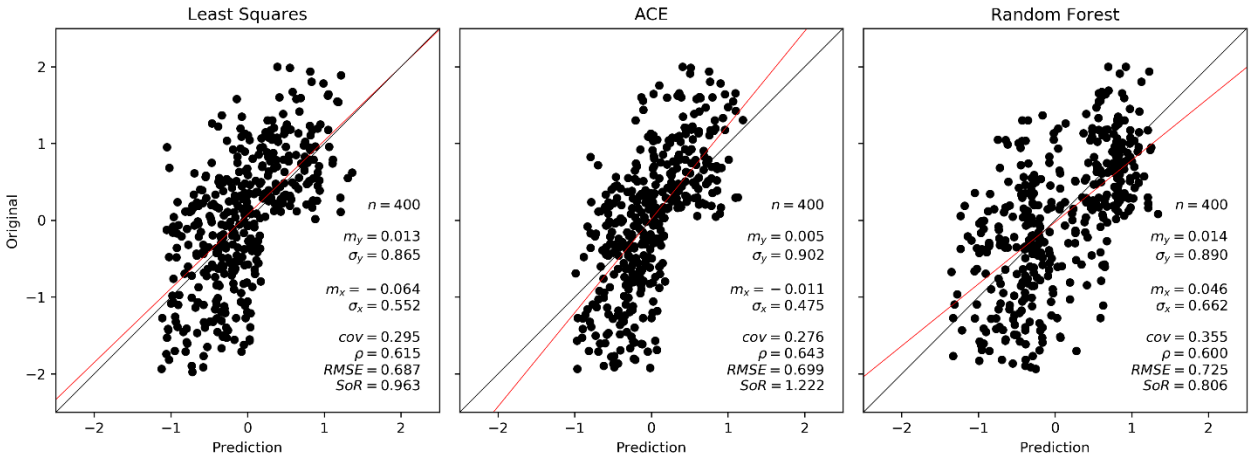


Figure 5.10. Histograms of variables in non-linear dataset

The three RSM techniques are applied to the modified data following the same steps. Cross-plots in Figure 5.11 display how the prediction result of least squares, ACE, and Random Forest with R^2 of 0.378, 0.413, and 0.360 respectively. Random forest comes as the worst technique to predict the missing values on this dataset, because even though the variables have non-linear relationship, each of them is still averaged linearly. Therefore, other RSM techniques that also can predict linear variable can outperform random forest. Nonetheless, random forest R^2 has the highest jump as compared to the same dataset in section 5.1.1.



(a) (b) (c)

Figure 5.11. Cross-plots of original and prediction result on non-linear synthetic Gaussian dataset using (a) least squares; (b) ACE; and (c) random forest

The percentage of missing data affects the ACE result as seen in Figure 5.12. ACE is the best method when missing 10% values but the performance degrades above 40% missing. Random forest does not work well as implemented. ACE result drops significantly as compared to the result from other techniques because lack of data makes ACE algorithm, which maximizes variables correlation, misinterprets the correlation between variables that leads to the bad prediction accuracy. The experiments show that for all Gaussian data, linear or non-linear, least squares is a suitable prediction technique to use. Real data will be used to compare all the techniques below.

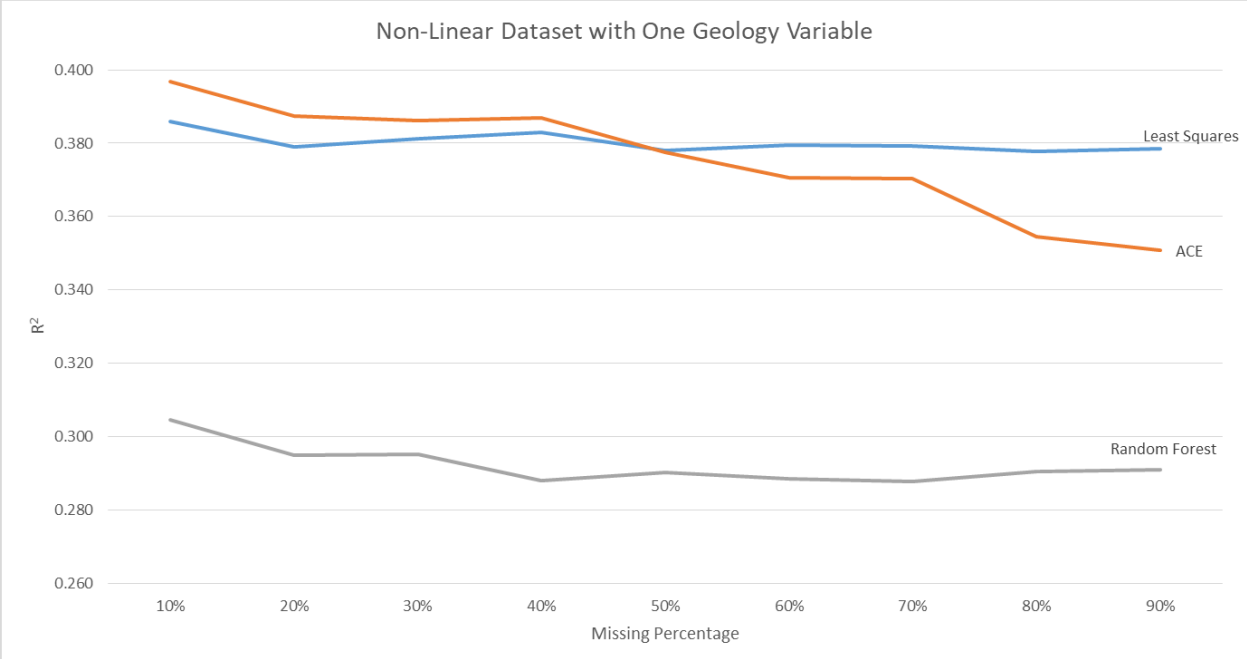


Figure 5.12. Average R^2 from 100 simulations comparison between the three techniques on non-linear synthetic Gaussian dataset

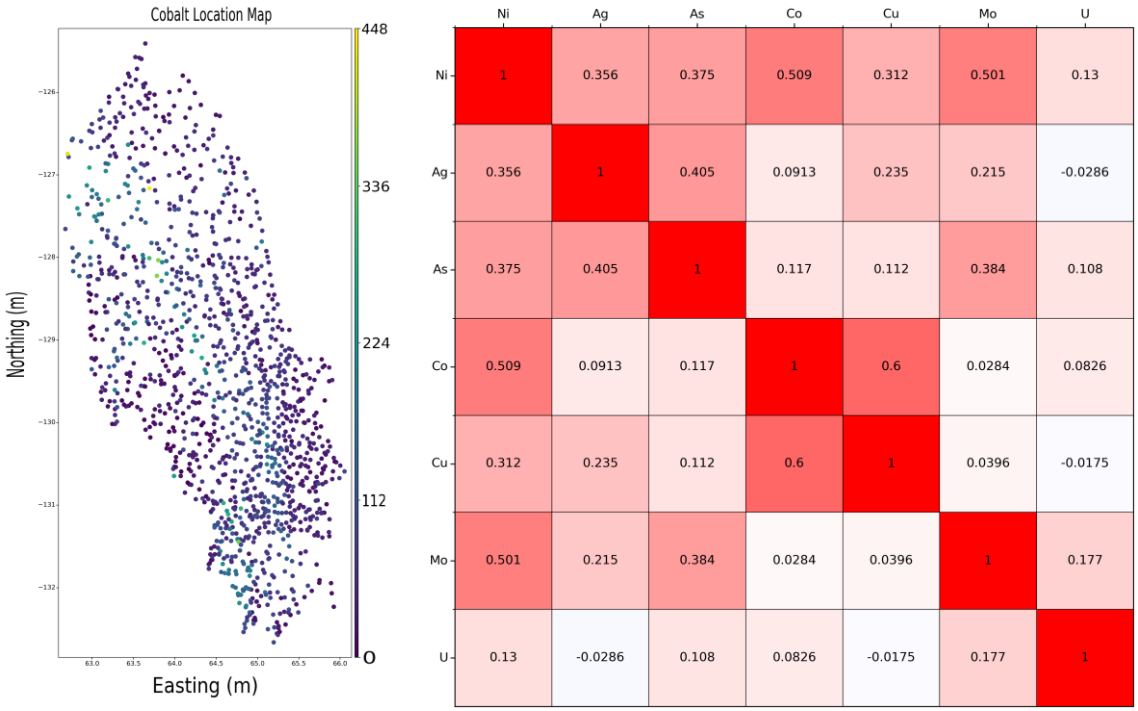
5.1.3. Real Dataset

The three RSM techniques will now be demonstrated with real datasets. There will be two different datasets used for this section. Both have one geometallurgy variable whereas one dataset has six geology variables and the other one has one geology variable. Both datasets are complete so the accuracy level can be judged by data left out. A different percentage will be left out ranging from 10% to 90% with the interval of 10%. The prediction steps are similar to the steps for predicting missing values in the synthetic datasets discussed above.

Sensitivity analysis was done on the dataset with 6 geology variables to understand the importance of the predictor variables. When one particular geology variable has a very low sensitivity coefficient, not using it may not change the prediction accuracy and the variable could be considered for removal (Kumara & Deutsch, 2018).

5.1.3.1. Heavy Metal Composite (HMC) Data

This dataset was used by Prades & Deutsch (2017) and followed by Pinto & Deutsch (2018). The dataset consists of 1308 samples with irregular sample spacing as shown in Figure 5.13 on the left. There are 63 unique metal variables in this dataset but only 7 variables with various correlation between each other as shown in Figure 5.13 on the right will be used in this thesis. Nickel (Ni) will be treated as the geometallurgy variable and 6 other variables including Silver (Ag), Arsenic (As), Cobalt (Co), Copper (Cu), Molybdenum (Mo), and Uranium (U) are considered as the geology variables. These six geology variables have linear bivariate relationship with the geometallurgy variable as shown in Figure 5.13 on the bottom.



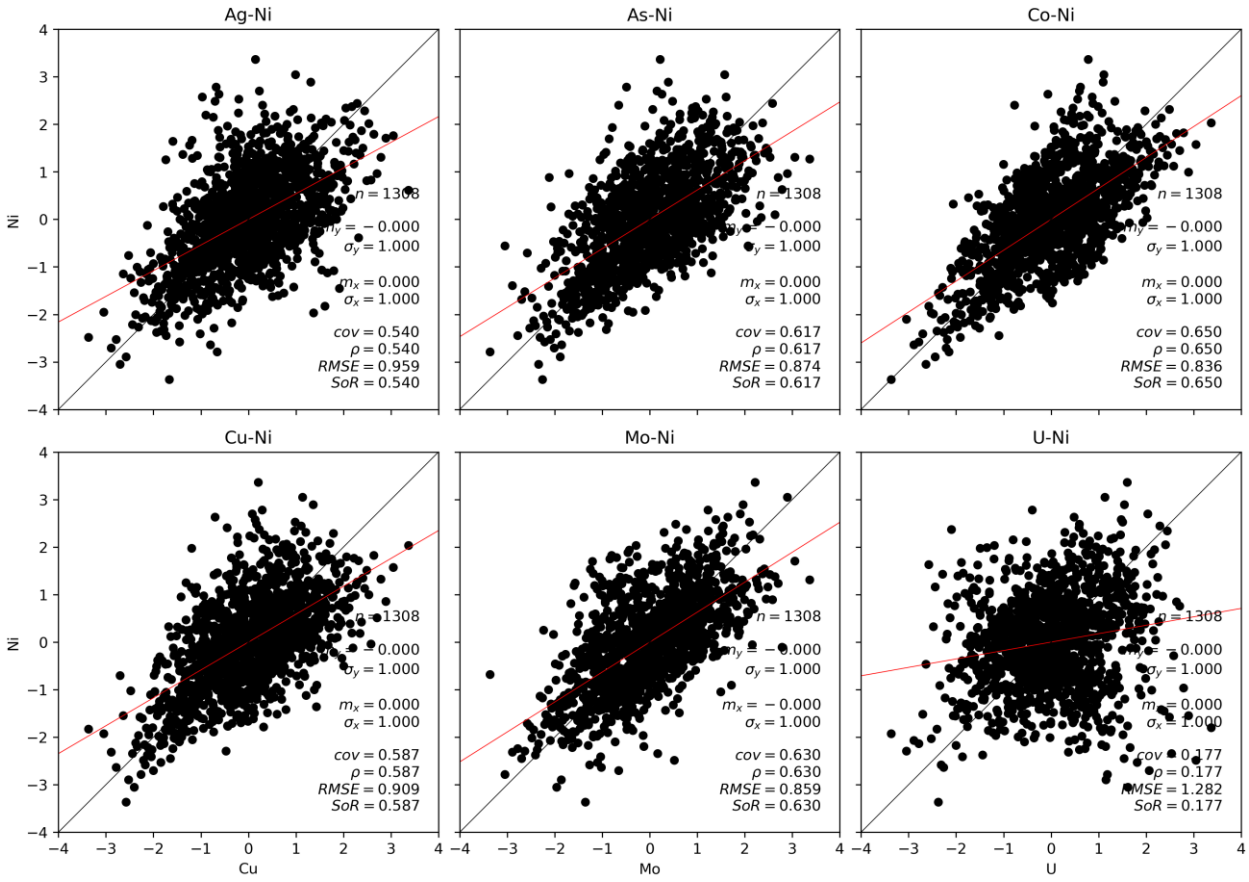
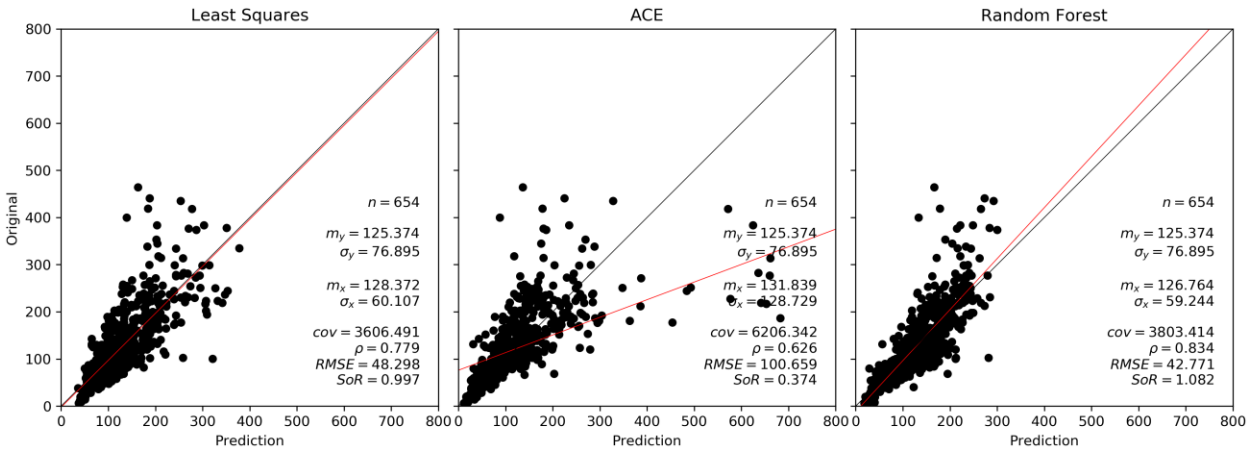


Figure 5.13. Location map, correlation matrix, and bivariate relationships between geology variables and geometallurgy variable of HMC dataset

All variables have positively-skewed distributions with some extreme values and different scales. The geology variables are ordered alphabetically. The three RSM techniques are applied to 900 different data without normal score transformation.

The cross-plots from 50% missing values are shown in Figure 5.14. All predictions have less variance than the original values. The coefficient of determination for the results are 0.607, 0.392, and 0.694 respectively from left to right. ACE is the technique that reproduces some outliers but most of the time they are not at the right location and this makes ACE look the worst. On the other hand, random forest and least squares are overly smooth with almost no outlier reproduced. Random forest has the least variance prediction result, but appears the best.



(a) (b) (c)

Figure 5.14. Cross-plots of original and prediction result on hmc dataset using (a) least squares; (b) ACE; and (c) random forest

The average of multiple realizations for each technique is presented in Figure 5.15. Missing data affects all three RSM techniques where the R^2 values go down as the missing percentage increases. Random forest is the best at predicting missing values in this dataset by about 10% better R^2. Least squares comes in second place with the least affected accuracy level among the three techniques while ACE performs the worst. ACE predicting extreme values comes as a disadvantage for mean squared error and R^2. Random forest is the most suitable technique for this dataset.

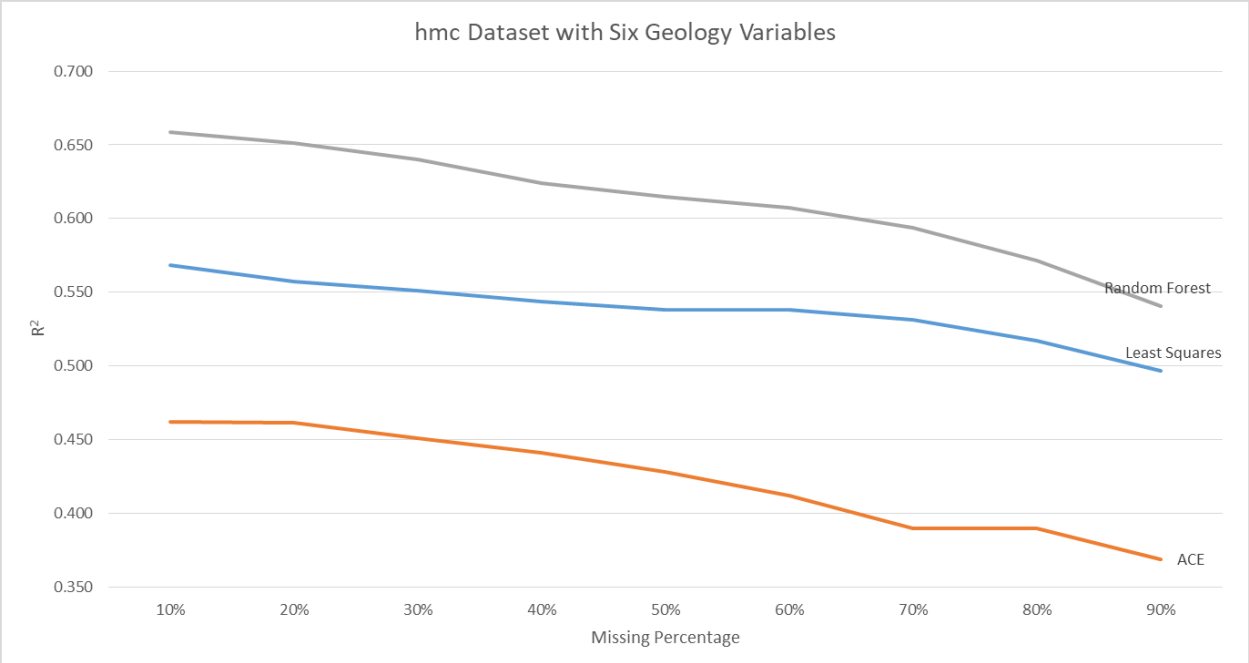


Figure 5.15. Average R^2 from 100 simulations comparison between the three techniques on hmc dataset

5.1.3.2. Porphyry Data

The second dataset was used by Deutsch (2018) and has 2 variables which are Copper (Cu) and Gold (Au) at 2634 locations as shown in Figure 5.16 on the top left. The samples spread out from south-west to north-east with the high grade located near the south-west end. The sample spacing is very regular. Both variables are related with correlation coefficient of 0.692 with linear bivariate correlation as shown in Figure 5.16 on the top right. The histograms of the two variables are shown in Figure 5.16 on the bottom. Both of them have some extreme values.

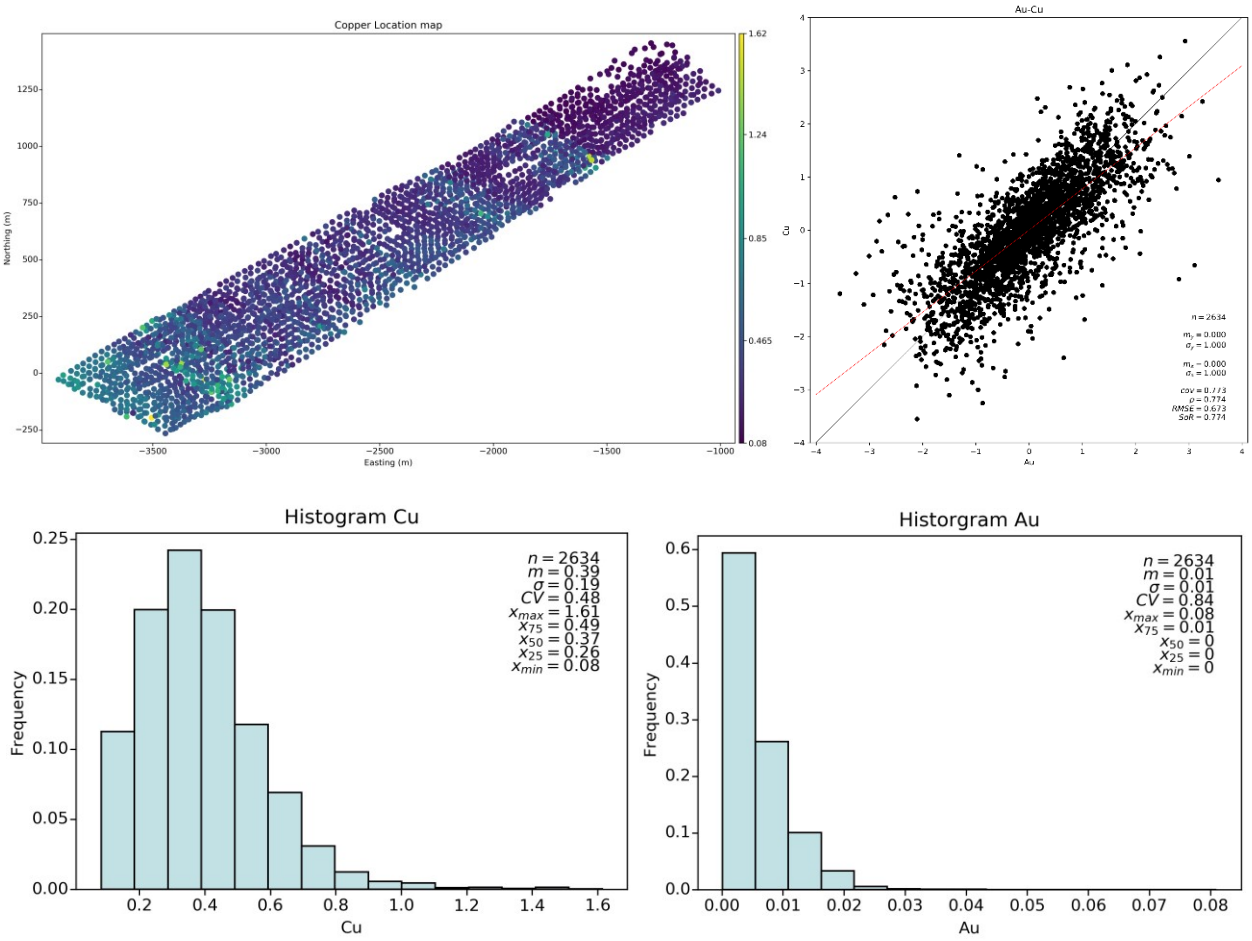
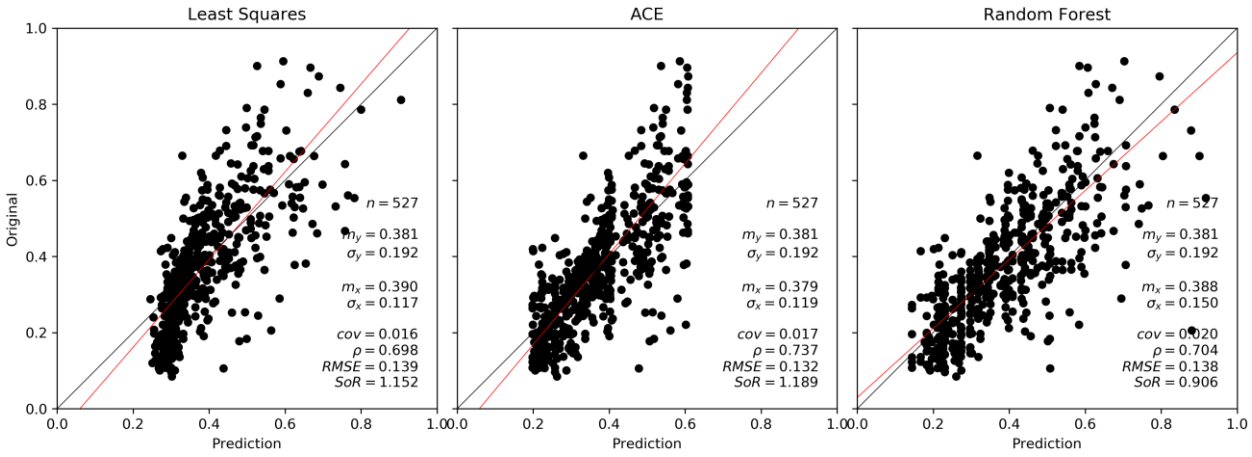


Figure 5.16. Location map, bivariate relationship between variables and histograms of all variables for porphyry dataset

Figure 5.17 shows the cross-plots with 20% missing values of the three techniques. The result seems different from previous real dataset where random forest and least squares reproduce more outliers than ACE. The coefficient of determination for the results are 0.487, 0.543, and 0.495 respectively from left to right. Without producing excessive outliers and a decent variables relationship, ACE comes out having the best accuracy among the three techniques for this dataset. Random forest result has the highest variance that means random forest prediction on this dataset is not overly smooth like the previous dataset.



(a) (b) (c)

Figure 5.17. Cross-plots of original and prediction result on porphyry dataset using (a) least squares; (b) ACE; and (c) random forest

Simulation averages are shown in Figure 5.18. From all simulations, ACE always has the best accuracy while random forest and least square come in the second and third place. It can be said that ACE is the most suitable RSM technique with this dataset. When there is less geology variables to predict the geometallurgy variable, the missing values percentage has less influence on the result. ACE has artificial limitation of prediction result whereas all the prediction results do not exceed the third quartile of the original data which is efficient in controlling the overestimation due to data outliers.

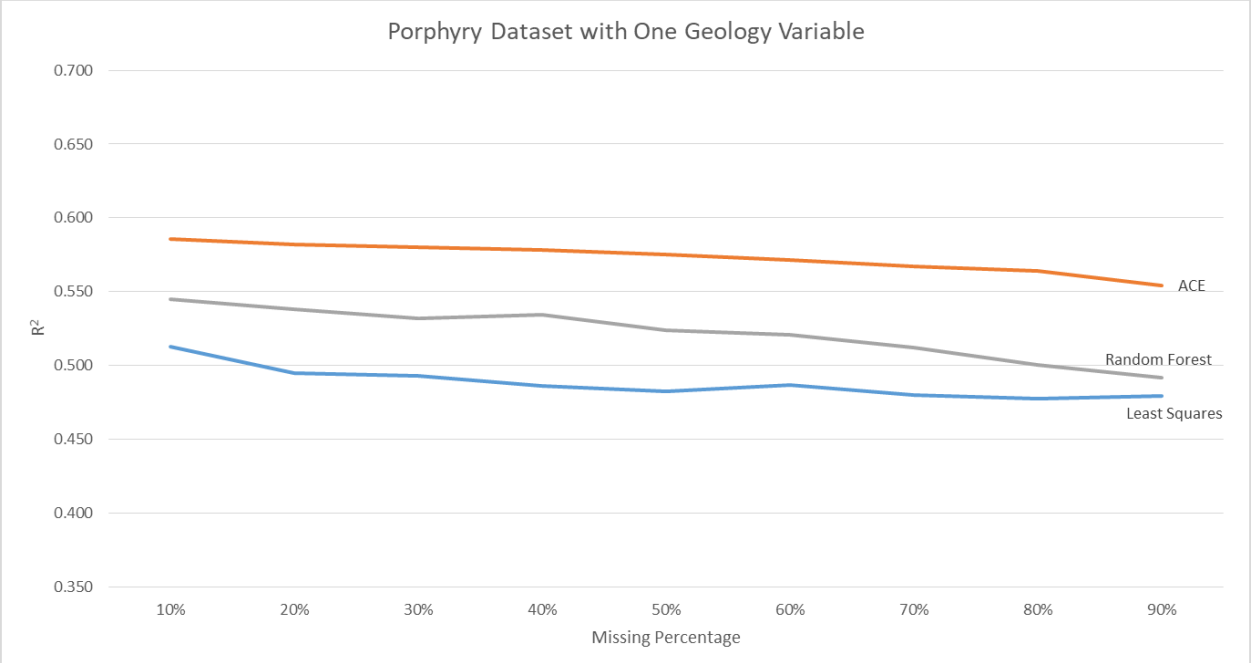


Figure 5.18. Average R^2 from 100 simulations comparison between the three techniques on porphyry dataset

Determining the best RSM technique result appears very specific for each dataset characteristic. The study has not considered the spatial information of the dataset yet. In the next section, the best RSM technique for each dataset is compared with prediction techniques that consider the spatial information of the data which are parametric imputation, random forest enhanced, and random forest moment.

5.2. Parametric Imputation and Case Study

The same five datasets as used above plus a sixth non-linear case will be considered. The imputation has the advantage of capturing data uncertainty by doing many simulations for each data realization while RSM only gives one prediction result. To make all of them comparable, the result of 100 imputation realizations are averaged so there is only one result of each case. The total

cases will be the same 900 as the previous section for each dataset resulting from 9 different missing values percentages and 100 realizations for each missing values percentage.

All imputation methods will consider normal score transformation for non-Gaussian variables. Back transform to the original value is performed before calculating the coefficient of determination of the result. Variogram calculation and variogram modeling consider Gaussian transformed data.

5.2.1. Synthetic Gaussian Dataset with One Geology Variable

Variograms of two variables in the first dataset are shown in Figure 5.19. Both variables have well defined spatial continuity with one spherical variogram structure and a small nugget effect. These variograms are modeled using the complete dataset. The variogram model will be different for each missing *x* case and they are modeled using autofit feature of *varmodel* software (Deutsch, 2015).

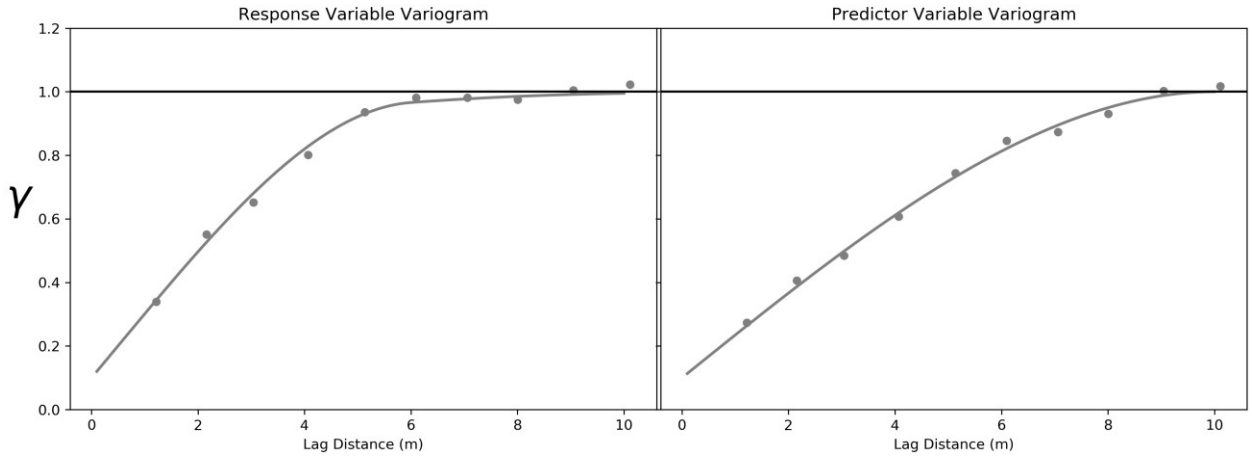
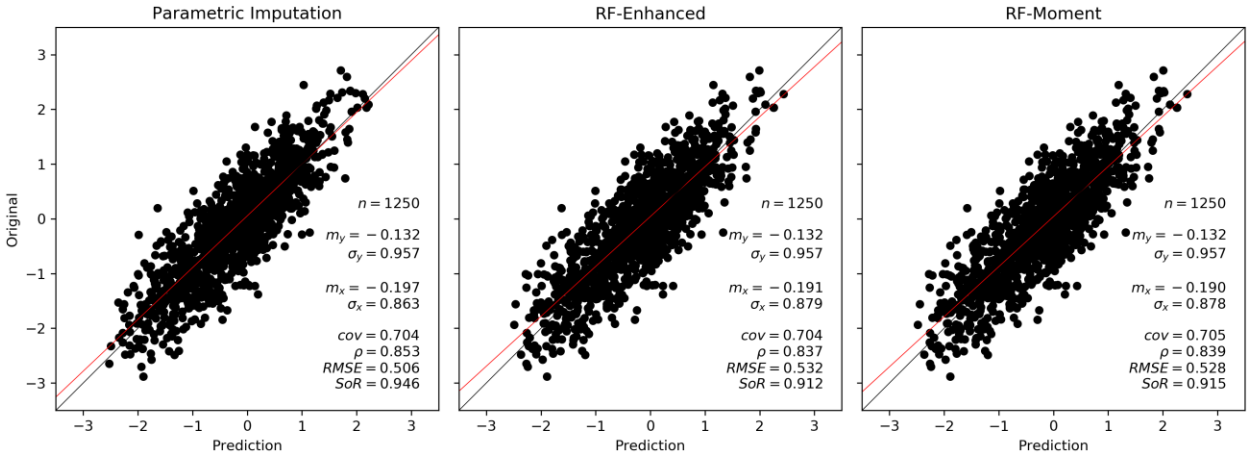


Figure 5.19. Variograms of (a) response variable, and (b) predictor variable from complete dataset on synthetic Gaussian dataset with one geology variable

The cross-plots from one of the result with 50% missing values are shown in Figure 5.20. The coefficient of determination for the results are 0.728, 0.701, and 0.704 respectively from left to right. The synthetic dataset has a variable with decent spatial continuity that makes the accuracy significantly higher than the RSM techniques. Both proposed frameworks perform at the same level as parametric imputation.



(a) (b) (c)

Figure 5.20. Cross-plots of original and prediction result on synthetic Gaussian dataset with one geology variable from (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

Figure 5.21 shows the average from 100 realizations of 100 different datasets for each missing values percentage. Parametric imputation performs better than two proposed frameworks. This happens because least squares technique that is used in parametric imputation performs better than random forest that is used in the proposed techniques as seen in Figure 5.2. Substituting likelihood distribution with less accurate result will degrade the quality of prediction.

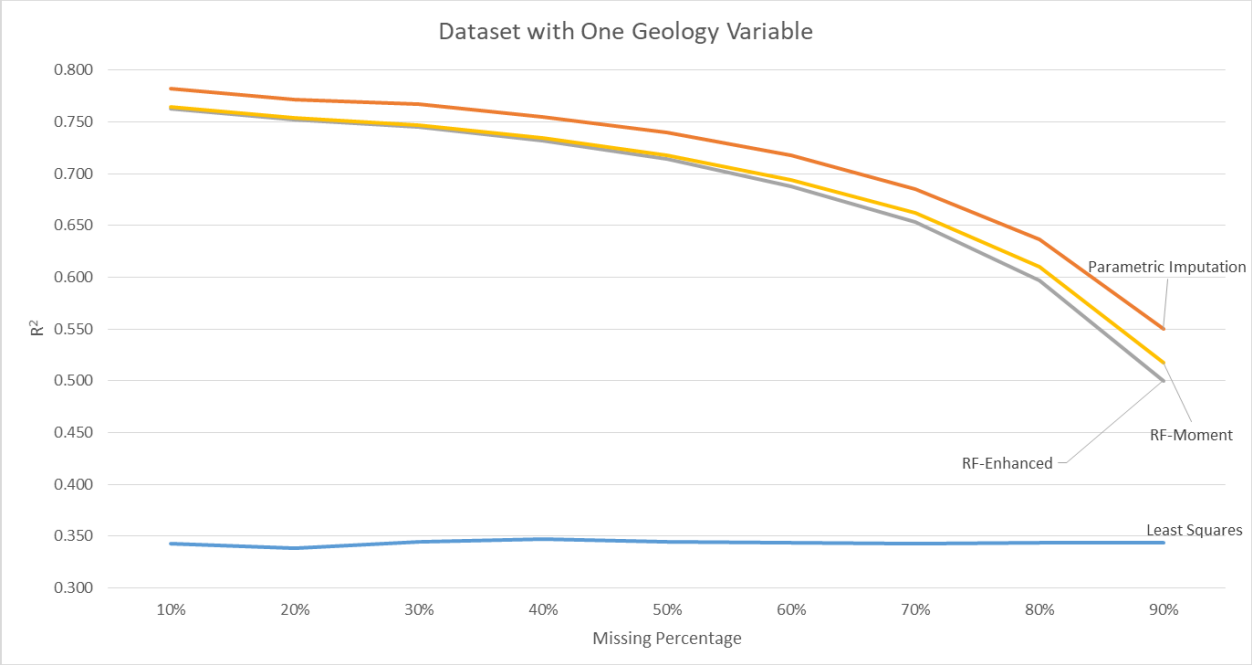


Figure 5.21. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on synthetic Gaussian dataset with one geology variable

For this dataset, using spatial characteristics for prediction improves the result. The three imputation techniques have almost the same accuracy. The accuracy may be further improved if the variogram is manually modeled.

5.2.2. Synthetic Gaussian Dataset with Three Geology Variables

The variograms from all four variables in the second dataset are shown in Figure 5.22. All of them appear to have the same variogram range of around 10. They have nugget effects ranging from 0.05 for x_2 to 0.3 for x_3 as the highest. The four variogram models consider one spherical structure.

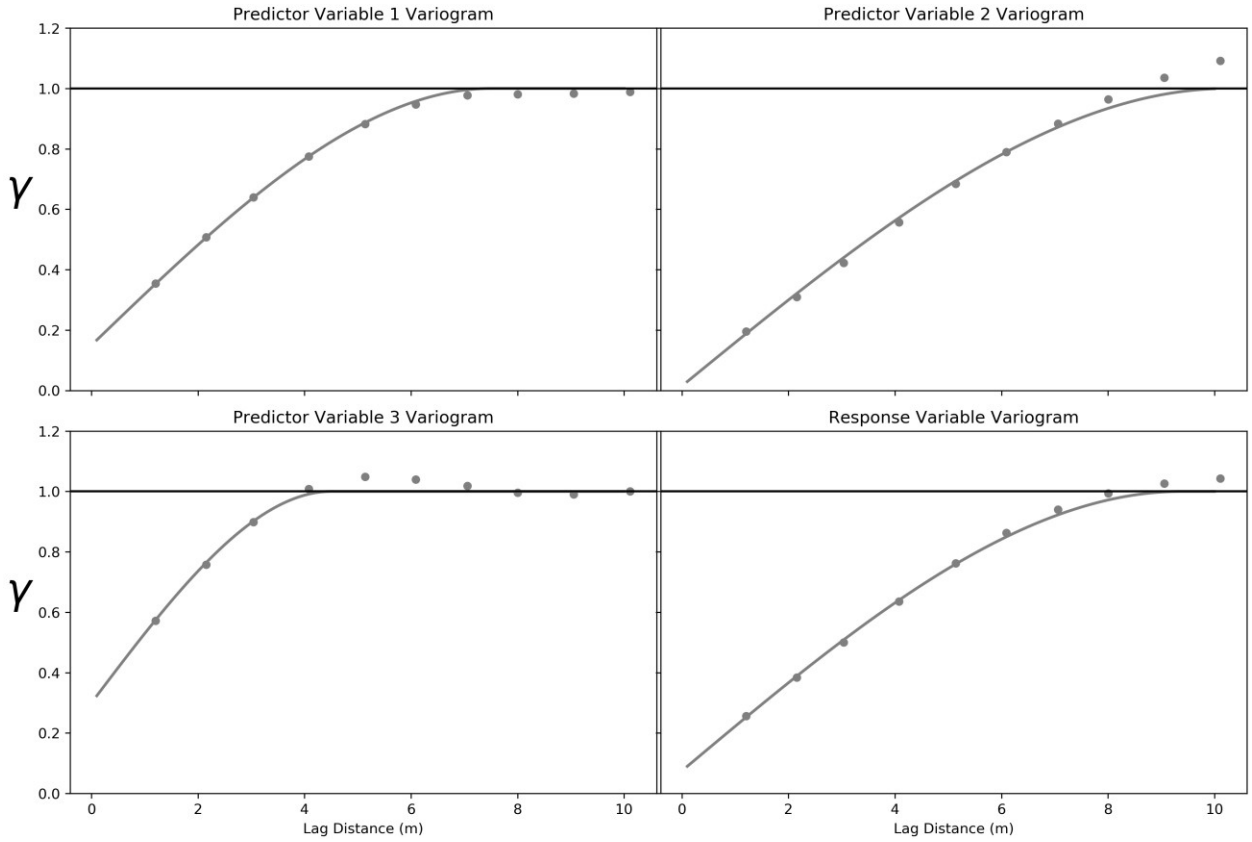
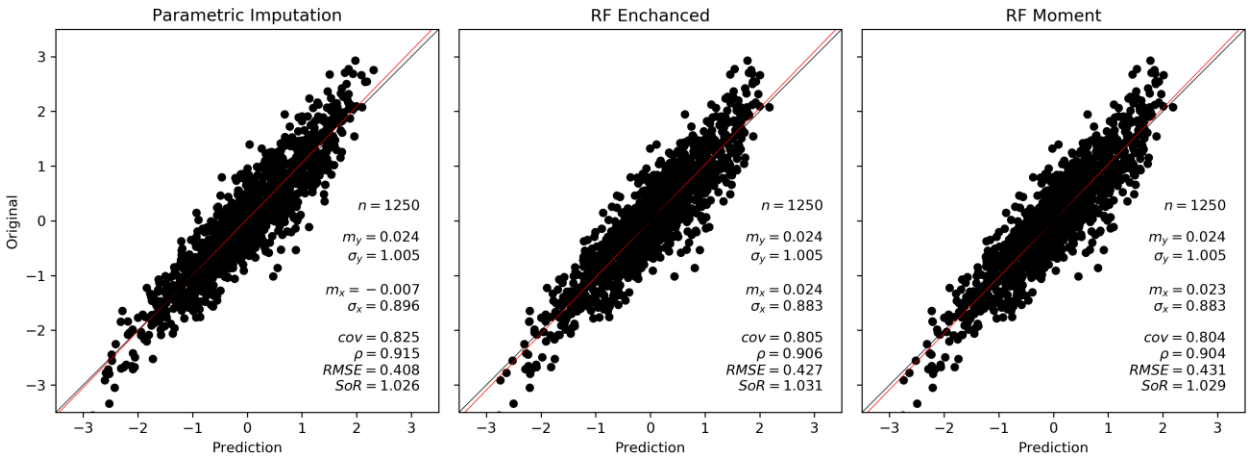


Figure 5.22. Variograms of all variables from complete dataset on synthetic Gaussian dataset with one response variable and three predictor variables

The cross-plots between original values and predicted values with 50% missing values using parametric imputation, RF-enhanced, and RF-moment are shown in Figure 5.23 with R^2 values of 0.838, 0.820, and 0.817, respectively. Although parametric imputation has slightly better R^2, the proposed frameworks can reproduce the mean closer to the original value. Care should be taken in practice since unbiasedness is an important property of geostatistical models.



(a) (b) (c)

Figure 5.23. Cross-plots of original and prediction result on synthetic Gaussian dataset with three geology variables using (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

All simulations and realizations are summarized in Figure 5.24. Increasing the number of geology variables makes the gap between RSM and imputation closer. Parametric imputation comes as the best imputation technique for 90% missing values as compared to the proposed frameworks. RF-moment performs at almost the same level as RF-enhanced. Yet, at 90% missing values, the gap between them becomes bigger and RF-moments comes as the worst parametric imputation technique.

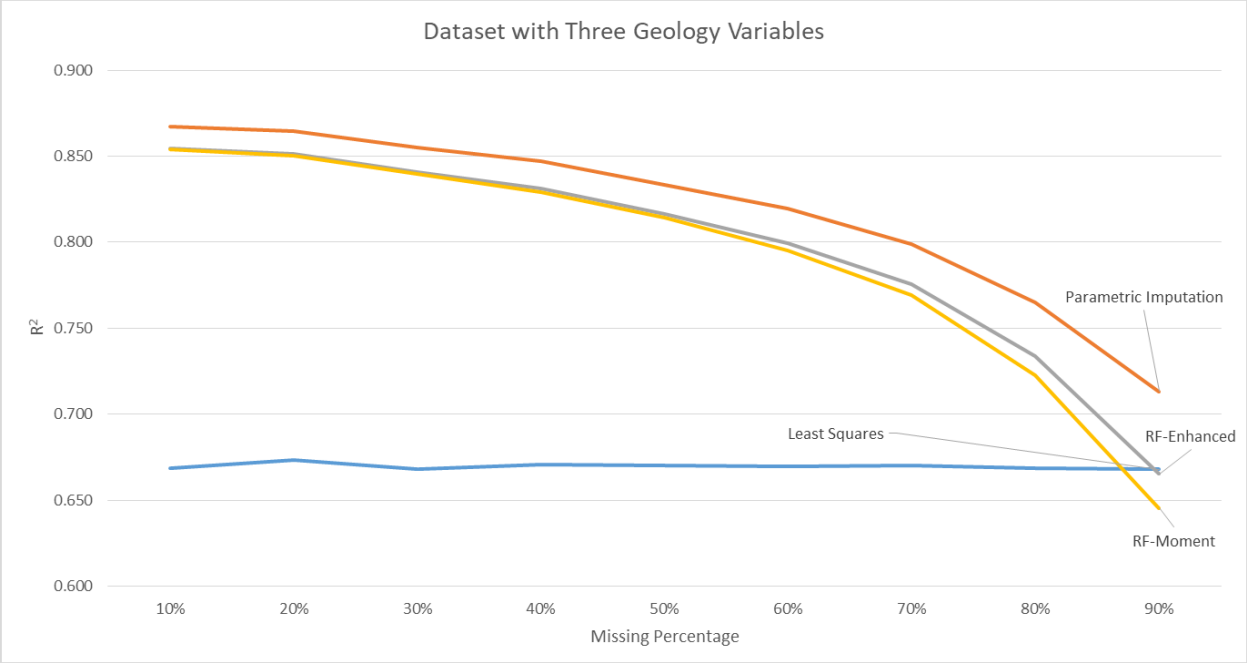


Figure 5.24. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on synthetic Gaussian dataset with three geology variables

5.2.3. Synthetic Gaussian Dataset with Five Geology Variables

Figure 5.25 represents the variograms of all variables in the third dataset. They are correlated to each other on some extent, which is why they have similar variograms. The variogram range is around 10 meters with a single spherical structure variogram model. All the variograms are omni-directional.

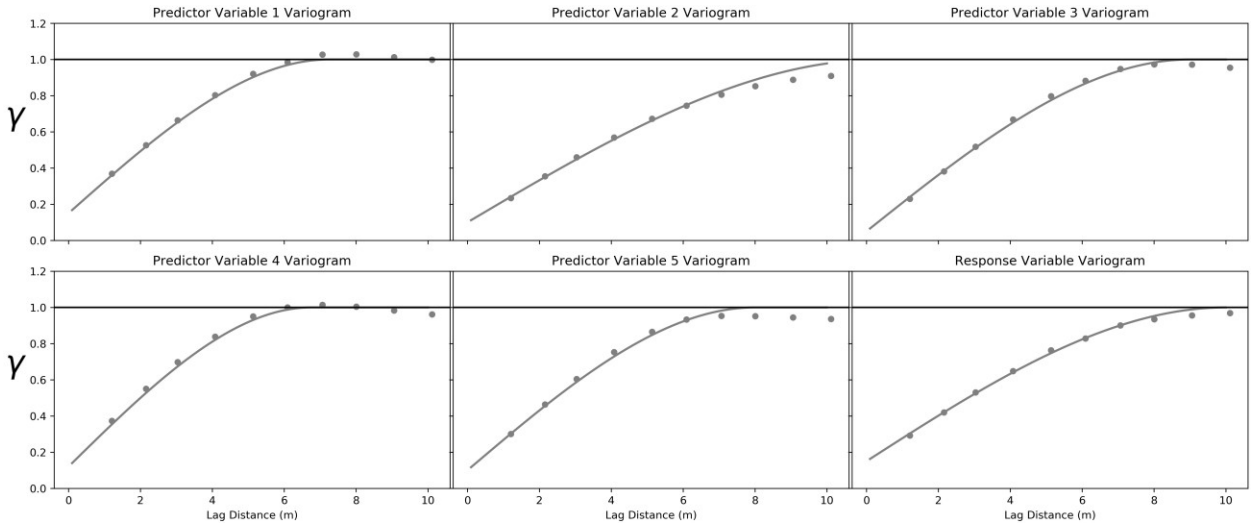
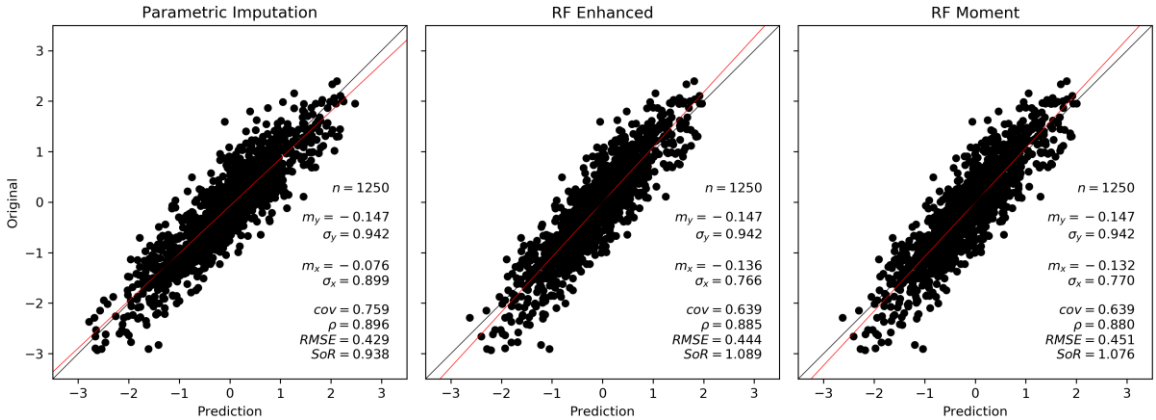


Figure 5.25. Variograms of all variables from complete dataset on synthetic Gaussian dataset with five geology variables

Figure 5.26 shows the cross-plots between the predicted and the original values from the three imputation techniques on missing 50% of the values. The R^2 of them are 0.801, 0.783, and 0.774, respectively from left to right. Proposed frameworks tend to be better at reproducing the mean of missing values despite has slightly lower R^2 due to random forest out-of-bagging algorithm works better with more prediction variables. Moreover, they have lower variances as compared to parametric imputation.



(a) (b) (c)

Figure 5.26. Cross-plots of original and prediction result on synthetic Gaussian dataset with five geology variables using (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

The R^2 of all prediction results are shown in Figure 5.27. With five collocated variables, the least square technique from RSM method performs better than the proposed frameworks starting from 50% missing values. And least squares is the most stable technique without any noticeable drop as the missing proportion increases. Random forest does not perform better than imputation despite having five geology variables and this is the reason behind proposed frameworks failure to outperform parametric imputation. Overall, both proposed frameworks perform quite well and RF-enhanced always outperforms RF-moment. Parametric imputation still becomes the most recommended technique because of how stable the prediction results are as compared to other techniques as the missing percentage increases.

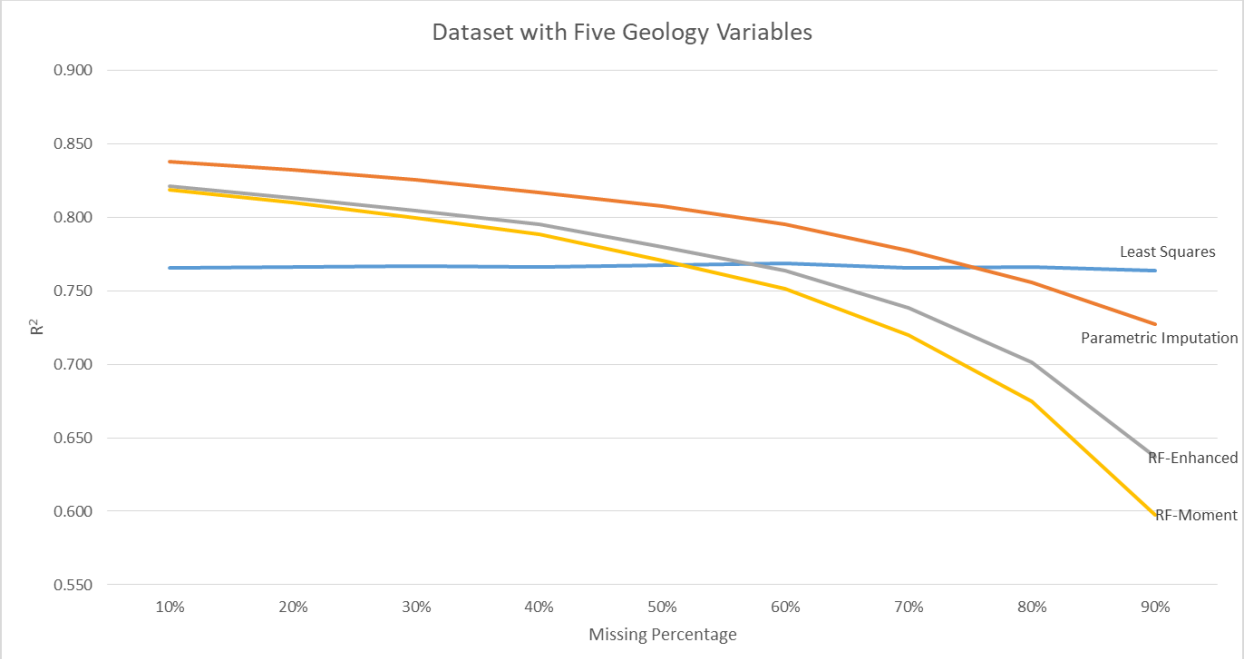
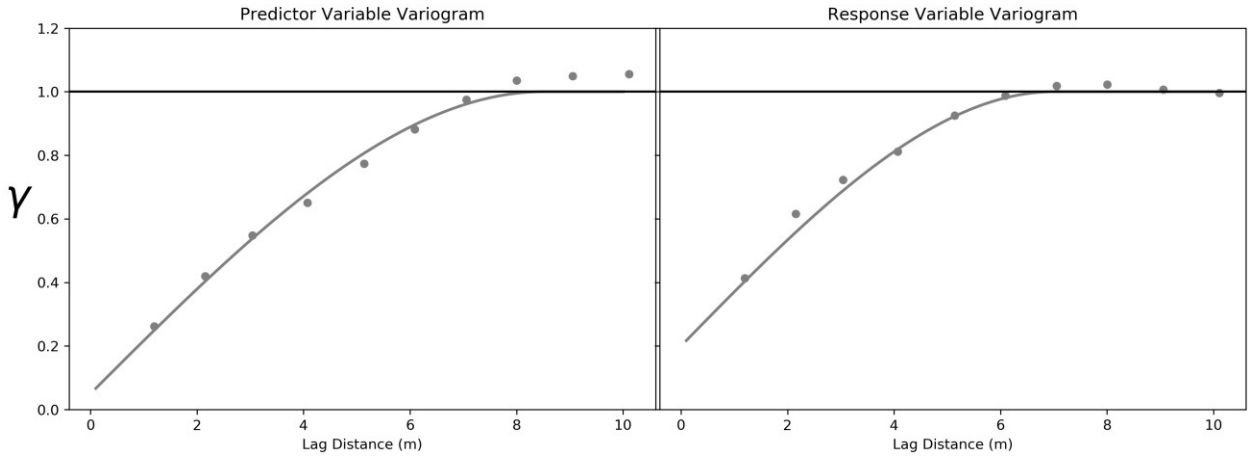


Figure 5.27. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on synthetic Gaussian dataset with five geology variables

5.2.4. Non-linear Dataset

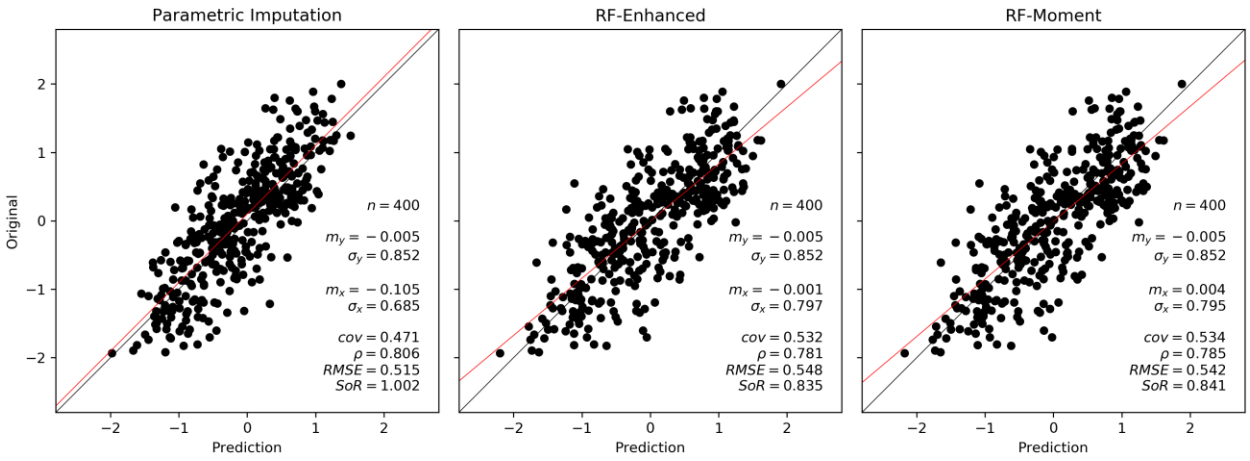
Throwing away some values to make the variables non-linear makes some changes to their variograms. The x_1 nugget effect drops down from 0.10 to 0.05 and the variogram range also decreases from 10 to 8.5. On the other hand, the y nugget effect increases from 0.15 to 0.20 and the variogram range decreases from 8 to 7. The variograms shown in Figure 5.28 are omni-directional.



(a) (b)

Figure 5.28. Variograms of (a) predictor variable, and (b) response variable from complete dataset on non-linear dataset

The proposed frameworks can reproduce the data distribution really well including the high values. Their mean values are too high while parametric imputation results have a slightly lower mean than the original. R^2 values of each cross-plot in Figure 5.29 from left to right are 0.649, 0.610, and 0.617, respectively. The proposed frameworks predict the extreme values fairly well. The result from both proposed frameworks also have almost the same mean as compared to the original value while parametric imputation slightly underestimates the missing values



(a) (b) (c)

Figure 5.29. Cross-plots of original and prediction result on non-linear dataset using (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

The comparison is shown in Figure 5.30. The two proposed frameworks perform at almost the same level for the entire simulation. Parametric imputation comes out better at reproducing R^2 values while proposed techniques are better at reproducing mean. These results are better than the result predicted using RSM technique. Parametric imputation is less stable as the missing values percentage increases. Parametric imputation is the most recommended technique to predict missing non-linear variable values.

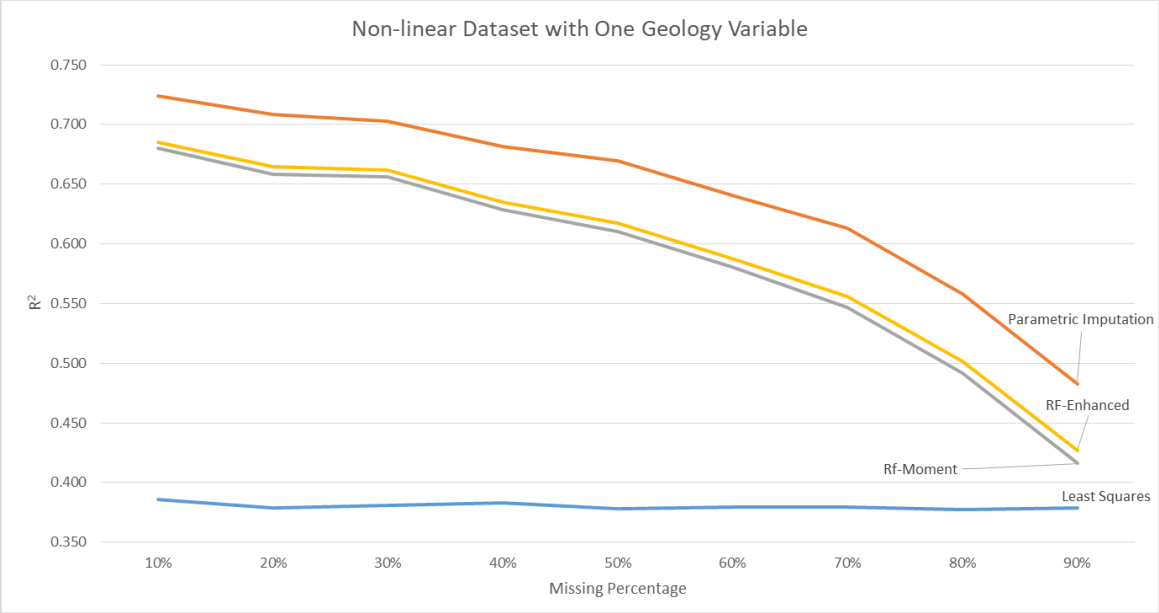


Figure 5.30. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on non-linear dataset

5.2.5. Heavy Metal Composite (HMC) Data

The variograms of normal score transformed variables of HMC dataset are shown in Figure 5.31. The real dataset have various variogram features such as different number of variogram structures, nugget effect, and variogram type. But, they have almost similar variogram range which is around 1.20. They are all omni-directional variograms.

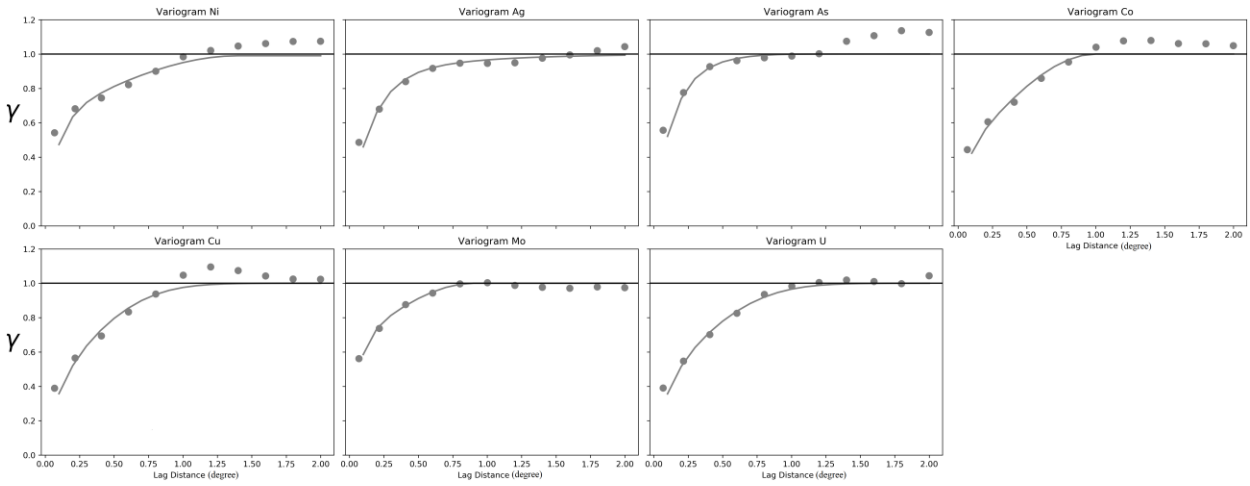
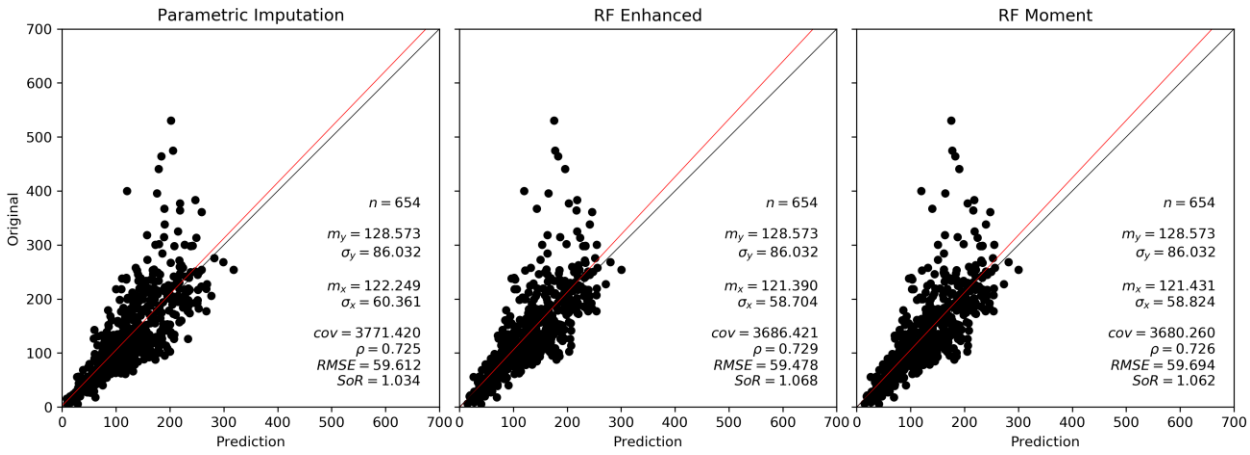


Figure 5.31. Variograms of all normal score transformed variables from complete hmc dataset

On this dataset, the results from the proposed frameworks have a slightly lower mean. The cross-plots shown in Figure 5.32 are from the dataset with 50% missing values. R^2 values of each cross-plot from left to right are 0.504, 0.515, and 0.512, respectively. Both proposed frameworks have higher R^2 than the parametric imputation result.



(a) (b) (c)

Figure 5.32. Cross-plots of original and prediction result on hmc dataset from (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

The comparison of all techniques for this dataset is shown in Figure 5.33. Random forest performs better than parametric imputation. For this dataset, RF-enhanced is the most recommended technique. Not only because it has the highest accuracy, but it can capture uncertainty using realizations.

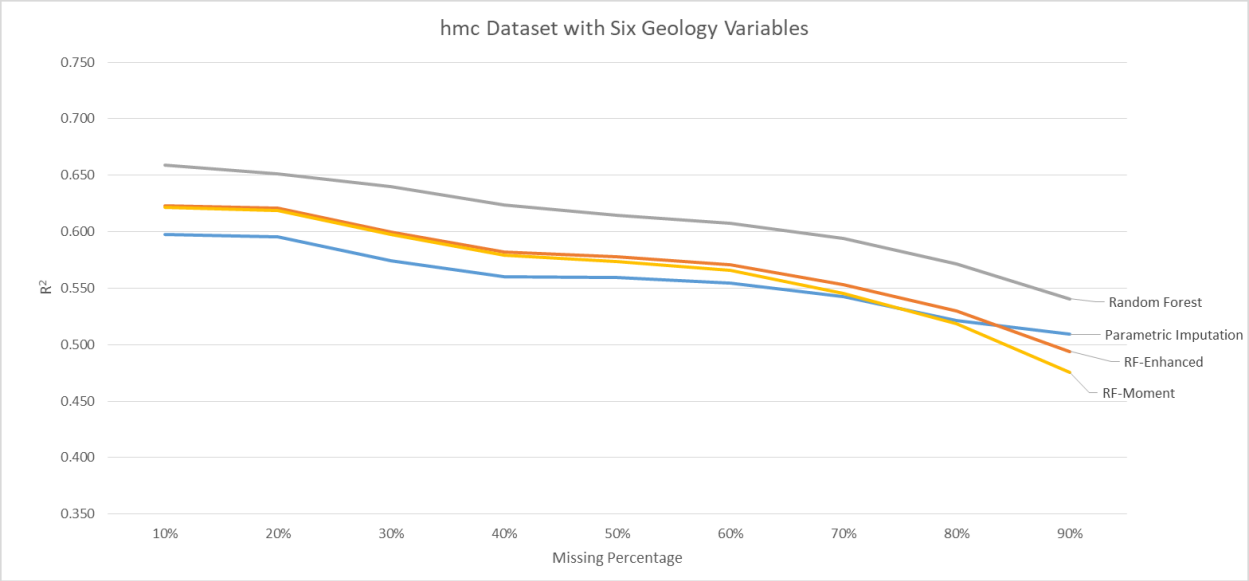


Figure 5.33. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on hmc dataset

5.2.6. Porphyry Data

The dataset has two variables with variograms as shown in Figure 5.34. The two variables have different variogram range where the geometallurgy variable (Cu) is more continuous than the geology variable (Au) with variogram range of 400 and 500, respectively. They also have different number of variogram structures where Cu can be modeled with only one spherical structure and Au needs two structures to be modeled.

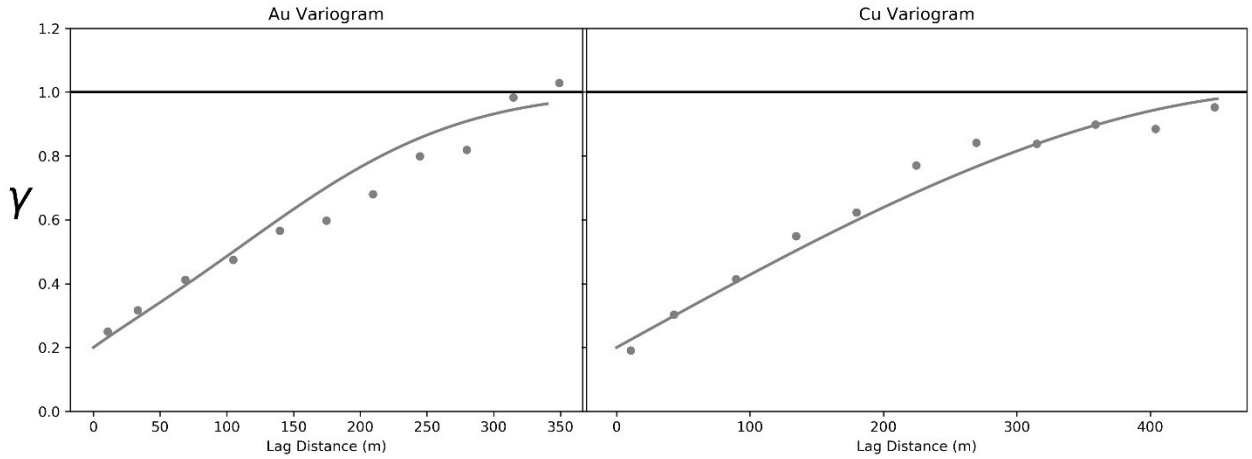
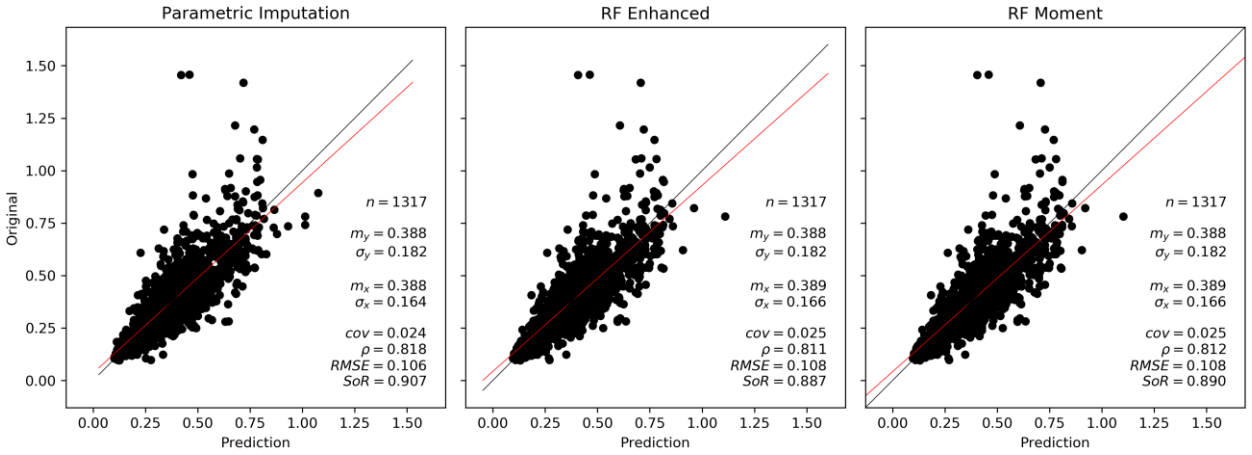


Figure 5.34. Variograms of all normal score transformed variables from complete porphyry dataset

The cross-plots between original and predicted values from 50% missing data values dataset are shown in Figure 5.35. The R^2 of them are 0.656 for parametric imputation, 0.646 for RF-enhanced, and 0.647 for RF-moment. The results are similar.



(a) (b) (c)

Figure 5.35. Cross-plots of original and prediction result on porphyry dataset from (a) parametric imputation; (b) RF-enhanced; and (c) RF-moment

Figure 5.36 shows the R^2 comparison from all prediction results. Parametric imputation and proposed frameworks perform at comparable level. Random forest itself cannot outperform

parametric imputation so that the proposed frameworks cannot perform better than parametric imputation. Using spatial features for predicting missing values for this dataset gives more accuracy.

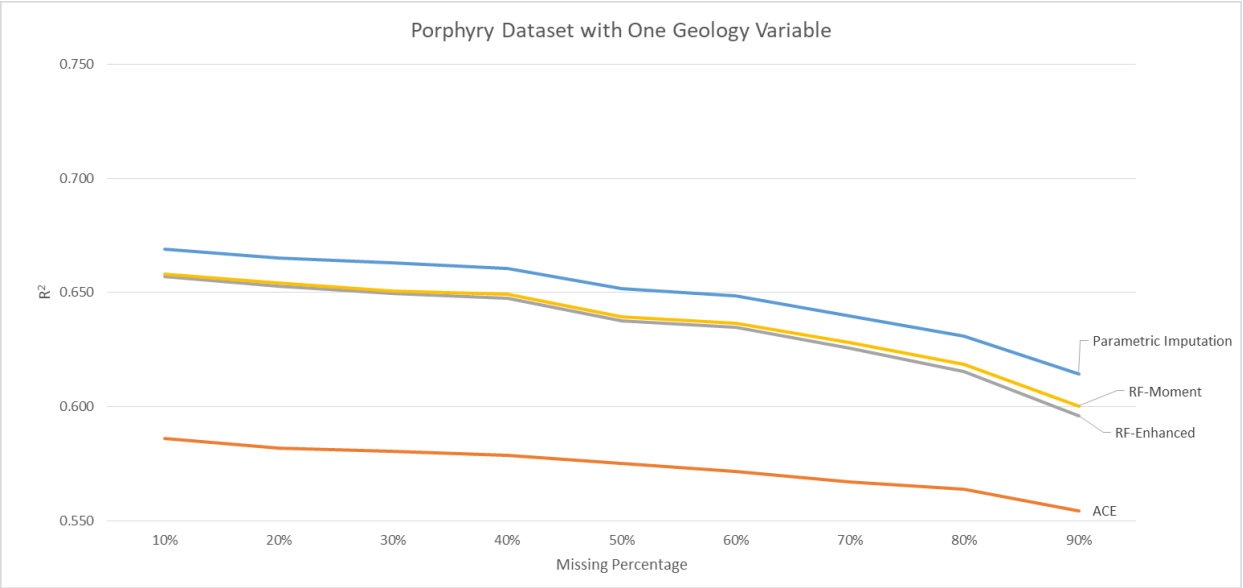


Figure 5.36. Average R^2 from 100 simulations comparison between the three imputation techniques and one RSM technique on porphyry dataset

5.3. Discussion

Each prediction technique has a place to predict missing values with certain characteristics. When the geometallurgy variable has decent spatial continuity, imputation will have better R^2 than RSM technique. On the other side, RSM technique R^2 will get better when the number of geology variables increases. Imputation techniques perform as well as RSM techniques when predicting geometallurgy variables with decent spatial continuity and many collocated variables. Least squares regression is the best RSM technique to predict synthetic Gaussian linear dataset.

The random forest is the best to explain real dataset with multivariate relationship and ACE is the best when the dataset only has one collocated variable. That happened because random forest

is the best to utilize all the collocated variables even with low correlation coefficient and ACE performs the best because it can find a good transformation.

Furthermore, ACE will find a non-linear transformation that makes it a candidate RSM technique. ACE can stabilize the error variance and normalization of error distribution. But, it has the limitation when the data becomes too sparse and least square prediction outperforms ACE at that point. Random forest cannot perform well on a dataset with very few collocated variable even though it can well capture non-linearity of the variable. Spatial features help imputation techniques to perform adequately even with non-linear variable.

Both proposed frameworks perform at almost the same level for all case studies in this thesis while RF-enhanced performs generally better than RF-moment. To narrow it down, RF-enhanced will be the only framework proposed in this thesis and will be called as Random Forest Imputation. Altering likelihood variance following formula (10) can be unstable. When the dataset becomes bigger, this problem will get worse, which explains why RF-moment performs slightly worse on a dataset with many collocated variables.

Chapter 6
Conclusions and Future Work

This thesis focuses on common challenges encountered when working with multivariate geometallurgical data. The main goal is to provide guidance and additional methods to improve the accuracy of geometallurgical modeling. A problem with geometallurgical data is its non-linearity and missing data values. The influence of various dataset characteristics and percentage of missing values are investigated. A novel prediction technique is proposed.

6.1. Contributions

The imputation and RSM frameworks for model building are comparable with as shown in Chapter 3. The difference in accuracy depends on the dataset characteristics.

The least squares technique is the best method to predict missing variable values in a linear dataset with no noise. Yet, when the variable has decent spatial continuity, imputation techniques will perform better than RSM techniques. Least squares prediction accuracy can be quite good as the number of geology variables increases.

The least squares technique works well with a single non-linear variable. The random forest can theoretically explain the non-linearity better; however the synthetic non-linear dataset in this thesis did not confirm this. The existence of additional geology variables is the key feature for random forest to perform better. Nonetheless, all imputation techniques outperform the RSM approach and parametric imputation comes out as the best technique.

For real dataset, the number of collocated variables determines the best RSM technique to predict missing values. Random forest can perform well when the dataset has many geology variables while ACE is a well-established at a certain accuracy level disregarding the number of geology variables. When the real dataset spatial characteristics can be well defined, all imputation technique outperform all RSM techniques. As the number of collocated variable increases, random forest outperforms parametric imputation.

The proposed RF-enhanced imputation technique appears to have promise for imputation. Development of random forest regression technique will improve the RF-enhanced result. This could be addressed in future research.

6.2. Future Work

This thesis does not consider categorical variables. The other limitation is not having non-linear dataset with more than one geology variable. Future work could overcome these limitations.

There are several prediction techniques that are not brought up to this thesis such as Gradient Boosted Model (GBM) by Hastie et al. (2009) and GMM imputation by Silva and Deutsch (2016). They should be compared to the existing method and also to random forest imputation. Considering the GBM method to alter the likelihood distribution in imputation could be another step.

A research in random forest regression technique will help improving the result of RF-enhanced. In this thesis, *rpy2* Python library was used. Changing some parameters or using other random forest program such as *scipy* Python library may give different result.

Developing software related to random forest imputation will be a good contribution to establishing these novel imputation techniques. Random forest imputations in this thesis are done

manually using python software. Writing random forest imputation code as a standalone software will be a convenience for some people to use random forest imputation on predicting missing values.

References

Barnett, R. & Deutsch, C. V., 2012. *Data Replacement in a Complex Multivariate Context,* Edmonton, Alberta: CCG Annual Report 14.

Barnett, R. M. & Deutsch, C. V., 2013. *Assessing the Uncertainty and Value of ACE Transformations,* Edmonton, Canada: Centre for Computational Geostatistics, University of Alberta.

Barnett, R. M. & Deutsch, C. V., 2013. *Imputation of Geologic Data.* Edmonton, CCG.

Barnett, R. M., Manchuk, J. G. & Deutsch, C. V., 2014. Projection Pursuit Multivariate Transformation. *mathematical geosciences,* pp. 46:337-359.

Bax, A. R. et al., 2016. *An Integrated Liberation-leach Model and Ore Characterisation Procedure for Gold Ores.* Perth, Australia, The Australasian Institute of Mining and Metallurgy, pp. 315-319.

Breiman, L., 1996. Bagging Predictors. In: *Machine Learning Edition 24.* Boston: Kluwer Academic Publishers, pp. 123-140.

Breiman, L., 2001. Random Forests. *Machine Learning,* pp. 5-32.

Breiman, L. & Friedman, J. H., 1985. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association.*

Coward, S. & Dowd, P. A., 2015. Geometallurgical Models for the Quantification of Uncertainty in Mining Project Value Chains. *Proceedings of the 37th APCOM Conference,* Volume Soc. Mining, Metallurgy and Exploration (SME) ISBN 978-0-87335-417-2, pp. 360-369.

Cressie, N., 1989. The Origins of Kriging. In: *Mathematical Geology.* s.l.:Kluwer Academic Publishers-Plenum, pp. 239-252.

Davis, B. M. & Greenes, K. A., 1983. Estimating Using Spatially Distributed Multivariate Data. *Mathematical geology,* pp. 15(2):287-300.

Desbarats, A. J. & Dimitrakopoulos, R., 2000. Geostatistical Simulation of Regionalized Poresize Distributions Using Min/Max Autocorrelations Factors. *Mathematical Geology,* pp. 32:919-942.

Deutsch, C. V., 2018. *Session III of Magíster en Modelamiento Geoestadístico de Depósitos Minerales.* Chile: Universidade Adolfo Ibanez.

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

Deutsch, J. L., 2015. *Variogram Program Refresh,* Edmonton, Canada: Centre for Computational Geostatistics.

Dominy, S. C. & O'Connor, L., 2016. *Geometallurgy - Beyond Conception.* Perth, Australia, The Australasian Institute of Mining and Metallurgy, pp. 3-10.

Friedman, J. H., 1987. Exploratory Projection Pursuit. *Journal of the American Statistical Association,* pp. 82:249-266.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning.* Second ed. Berlin: Springer Series in Statistics.

Hotelling, H., 1933. Analysis of a Complex Statistical Variables Into Principal Components. *Journal of Educational Psychology,* pp. 24:417-441.

Kumara, P. & Deutsch, C. V., 2018. *Sensitivity Analysis in Response Surface Methodology.* Edmonton, Canada, Centre for Computational Geostatistics, University of Alberta.

Leuangthong, O. & Deutsch, C. V., 2003. Stepwise Conditional Transofrmation for Simulation of Multiple Variables. *Mathematical Geology,* pp. 35(2):155-173.

Lishchuk, V., 2016. *Geometallurgical Programs - Critical Evaluation of Applied Methods and Techniques.* Lulea, Sweden: Lulea University of Technology.

Marsland, S., 2015. *Machine Learning: An Algorithmic Perpective.* 2nd ed. Boca Raton: CRC Press.

Odom, M. D. & Sharda, R., 1990. *A Neural Network Model for Bankruptcy Prediction.* San Diego, California, USA, IJCNN International Joint Conference on Neural Networks.

Pinto, F. & Deutsch, C. V., 2018. *Improved Multivariate Clustering Classification with Geochemistry Data,* Edmonton, Canada: Centre for Computational Geostatistics, University of Alberta.

Prades, C. & Deutsch, C. V., 2017. *Data Transformation for Cluster Anaylsis,* Edmonton, Canada: Centre for Computational Geostatistics, University of Alberta.

Ren, W., 2007. *Exact Downscaling in reservoir Modeling.* Ph.D. Thesis ed. Edmonton, Alberta: University Of Alberta.

Rosenblatt, M., 1952. Remarks on a Multivariate Transformation. *Annals of Mathematical Statistics,* pp. 23(3):470-472.

Rubin, D. B., 1976. Inference and Missing Data. *Biometrika,* 63(3), pp. 581-592.

Sarle, W. S., 1994. *Neural Networks and Statistical Models.* Cary, North Carolina, USA, SAS Users Group.

Silva, D. S. & Deutsch, C. V., 2015. *Multivariate Data Imputation Using Gaussian Mixture Models.* Edmonton, CCG.

Switzer, P. & Green, A. A., 1984. *Min/Max Autocorrelation Factors for Multivariate Spatial Imaging,* s.l.: Technical Report, Stanford University.

Watson, G. S., 1967. Linear Least Squares Regression. *Ann. Math Statist.,* 38(6), pp. 1679-1699.