AES/RE/13-03 Decision Making on the Ocean Floor
A Simulation Framework to Assess Uncertainty
and Variability in Deep Sea Mining

22/03/2013 Tom Wambeke

Title : Decision Making on the Ocean Floor -
A Simulation Framework to Assess Uncertainty and
Variability in Deep Sea Mining

Author : Tom Wambeke

Date : March 2013

Supervisors : dr. Ing. MPhil. Jörg Benndorf (TUDelft)
dr. Mario Alvarez Grima MSc (IHC Merwede)
Prof. dr. Michael A. Hicks (TUDelft)
Prof. dr. Ir. André Vervoort (KULeuven)

TA Report number : AES/RE/13-03

Postal Address : Section for Resource Engineering
Department of Geoscience & Engineering
Delft University of Technology
P.O. Box 5028
The Netherlands

Telephone : (31) 15 2781328 (secretary)

Telefax : (31) 15 2781189

Copyright ©2013 Section for Resource Engineering

*All rights reserved.
No parts of this publication may be reproduced,
Stored in a retrieval system, or transmitted,
In any form or by any means, electronic,
Mechanical, photocopying, recording, or otherwise,
Without the prior written permission of the
Section for Resource Engineering*

Decision Making on the Ocean Floor

A Simulation Framework to Assess Uncertainty and Variability in Deep Sea Mining

MASTER OF SCIENCE THESIS

For the degree of Master of Science in Resource Engineering
at Delft University of Technology

Tom Wambeke

March 22, 2013

The work in this thesis was supported by IHC Merwede. Their cooperation is hereby gratefully acknowledged.

Copyright © Applied Earth Sciences (AES)
All rights reserved.

Abstract

As land resources decrease, commodity prices increase and technology evolves, deep sea mining is becoming a viable and sustainable alternative to meet the increasing demand for minerals. Successful deep sea mining operations are built on a sound identification of the resource, proper selection of equipment, a thoughtful production plan and good project management. These four key activities can be further optimized by analyzing how the spatial variability and uncertainty of the ore body properties impact the final mining operation.

To address this problem, IHC Merwede and TUDelft are in the process of developing a risk quantification framework, which makes use of state of the art geostatistical simulation methods and transfer functions to quantify geological uncertainty and translate it into decision or project risk.

The contribution of this master thesis is twofold. First, it offers a new methodological procedure to simulate realizations of the spatial deposit variability, in order to characterize risk and uncertainty. The Local Average Subdivision method (LAS) is introduced into mining geostatistics and new extensions are presented to (a) simulate under a nested spherical structure, (b) deal with not normally distributed data, and (c) incorporate point measurements of geotechnical and ore grade properties. Due to its computational efficiency, the presented method is suitable for simulating largely extended sea floor deposits on a scale of the Smallest Mineable Unit (SMU).

Secondly, the research illustrates, through several example case studies, the significant benefits of a risk based approach. The geological uncertainty is propagated through the entire mining operation, and as such translated into financial or operational performance indicators. In particular, the focus is on assessing financial project risk and on the design optimization of the deep sea extraction equipment by analyzing cutting forces, specific energies and power requirements.

Table of Contents

Acknowledgments	xiii
1 Introduction	1
1-1 Relevance of Deep Sea Mining	1
1-2 Added Value of Geostatistical Simulations	4
1-3 Purpose and Research Objectives of the Thesis	6
1-4 Outline of the Thesis	7
2 Current State of the Art in Modelling Geological Uncertainty	9
2-1 A Probabilistic Approach to Describe Spatial Variation: Random Function Model	9
2-1-1 Necessity of Probabilistic Modeling	9
2-1-2 Random Variable	10
2-1-3 Random Function	11
2-2 Characterization of Spatial Variation and Some Inevitable Assumptions	12
2-2-1 Spatial Covariance Function	12
2-2-2 Stationarity Assumptions	12
2-2-3 Assumption of Ergodicity	13
2-2-4 Spectral Density Function	14
2-3 Geostatistical Estimation	14
2-3-1 The Kriging Paradigm: Weighted Linear Combination, Unbiasedness Condition and Minimized Error Variance	14
2-3-2 Ordinary Kriging Equations	15
2-4 The Necessity of Geostatistical Simulations	16
2-4-1 Application Targets for the Generated Realization(s)	16
2-4-2 The Undesirable Effects of Geostatistical Estimations	17
2-4-3 Requirements for Simulated Realizations	18
2-4-4 The Flaw of Averages	20

2-5	Geostatistical Simulation Techniques	20
2-5-1	Selection Criteria for a Suitable Simulation Method	20
2-5-2	Generating Conditional Realizations: Two Possible Approaches	21
2-5-3	Brief Overview of some Existing Simulation Techniques	22
2-5-4	Conclusion and Final Selection	26
3	Conditional Simulation via Local Average Subdivision	29
3-1	The Principles Behind the Simulation Methodology	29
3-1-1	Local Average Theory and Change of Support	30
3-1-2	One Dimensional Local Average Subdivision	31
3-1-3	Two dimensional Local Average Subdivision	32
3-1-4	Conditioning of the Random Field	34
3-2	Critical Assumption and Limitations	35
3-2-1	Assumptions of the Random Function Model	36
3-2-2	Assumptions Induced by the LAS Simulation Algorithm	36
3-2-3	Limitations induced by the Conditioning Operation	37
3-3	Necessary Adjustments and Extensions	39
4	A Practical Process for the Simulation of Deep Sea Deposits: The Theoretical Principles	41
4-1	Roadmap of a simulation-based geostatistical analysis	41
4-1-1	Main Operations	43
4-1-2	Data and Area Statistics	44
4-1-3	Validation Procedures	44
4-2	Analyzing Results: Data and Area statistics	45
4-2-1	Data or Global Statistics	45
4-2-2	Area or Spatial Statistics	46
4-3	The Spine of the Developed Practical Process: The Main Operations	48
4-3-1	Declustering	48
4-3-2	Normal Score Transformation	49
4-3-3	Characterization of Spatial Variation	51
4-3-4	Support Adjustment	53
4-3-5	Ore Body Simulation	56
4-3-6	Back Transformation	56
4-3-7	Transfer Functions	56
4-4	Implementation and Software Language	56

5	A Practical Process for the Simulation of Deep Sea Deposits: Walker Lake Case Study	59
5-1	Description of Exploration Data and Completely Known Environment	60
5-1-1	Origin of the Data Set, Design of the Simulation Grid and Performed Data Adjustments	60
5-1-2	Exploration Data and Known Environment	61
5-2	Application of the Practical Process; An Example Regarding the Necessary Calibration, Validation and Adjustment Steps	62
5-2-1	Declustering and Exploratory Data Analysis	62
5-2-2	Normal Score Transformation	65
5-2-3	Model Covariance Function	66
5-2-4	Data Support Adjustment	67
5-2-5	Ore Body Simulation	69
5-2-6	Back Transformation	72
5-3	Assessment of the Performance; Validation of the Simulation Results in a Completely Known Environment	73
5-3-1	Cumulative Distribution Function, Histogram and Experimental Semi-Variance Function	76
5-3-2	Evaluation of the Estimation and Simulation Error	77
5-3-3	Assessment of Calculated Confidence Intervals	78
5-3-4	Representation of Geological Uncertainty	80
6	Case Studies Illustrating the Potential Benefits of a Simulation-based Approach	83
6-1	Case Study 1: Economic Evaluation of a Gold Deposit	84
6-1-1	Introduction	84
6-1-2	Exploration Assistance; Areas for Additional Drilling?	85
6-1-3	Grade-Tonnage Curve and Recovered Metal	86
6-1-4	Optimization of Ore/Waste discrimination	87
6-1-5	Cash Flow Analysis	89
6-1-6	Net Present Value of the Mining Operation	91
6-2	Case Study 2: Decision Support for Installed Power Supply	91
7	Conclusions and Recommendations	95
7-1	Conclusions	95
7-2	Recommendations for Future Work	97
A	Characterization of Spatial Variation and the Corresponding Assumptions	99
A-1	Covariance Function	99
A-2	Stationarity Assumptions and Ergodicity	100
A-3	Spectral Density Function	102
A-3-1	Spectral Representation of One Dimensional Process	102
A-4	Spectral Representation of the Covariance Function of a one-dimensional Process	105

B	Derivation of Ordinary Kriging Equations	107
B-1	Weighted Linear Combination and the Unbiasedness Condition	107
B-2	The Random Function Model and Error Variance	109
B-3	Ordinary Kriging Equations	111
C	Mathematical Principles of the Local Average Subdivision Method	113
C-1	Local Average Theory and Variance Reduction	113
C-2	Variance and Covariance of Local averages	116
C-3	One Dimensional Local Average Subdivision	116
C-4	Conditioning of the Random Field	121
C-5	Weights and Gauss Points	124
D	Detailed Discussion on the Data and Area Statistics	125
D-1	Data or Global Statistics	126
D-2	Area or Spatial Statistics	129
D-3	Summary	136
E	Declustering of Clustered Preferential Sampling	139
E-1	Polygon declustering	141
E-2	Cell declustering	142
E-3	Quality control of the declustering operation	144
F	Input Data and Design of Simulation Grid	147
F-1	Original Data Set	147
F-2	Design of Simulation Grid and Performed Data Adjustment	148
F-3	Underlying Sampling Strategy	151
G	Walker Lake Exploration Data	155
G-1	Walker Lake Exploration Data: Gridded + Clustered	155
G-2	Walker Lake Exploration Data: Gridded	166
	Bibliography	173
	Glossary	177
	List of Acronyms	177
	List of Symbols	178

List of Figures

1-1	Offshore Production System	3
1-2	Conceptual visualization of the added value of the combination between geostatistical simulations and transfer functions.	5
2-1	Estimation versus Simulation	18
2-2	Concept of the flaw of averages	19
2-3	Overview of Simulation Methods	22
3-1	Resolution improvement during a LAS simulation of a local average random process	30
3-2	Cell subdivision in a two dimensional local average subdivision procedure	32
3-3	Problem definition and possible solution regarding conditioning limitations	38
4-1	Roadmap of the developed practical process for the simulation of deep sea deposits	42
4-2	Histogram and cumulative frequency distribution of Walker Lake exploration data	45
4-3	Experimental semi-variance and (auto)covariance function	47
4-4	Cell Declustering	49
4-5	Quality control in declustering operation	50
4-6	Normal Score Transformation	51
4-7	Omnidirectional characterization of spatial variation in normal score domain	52
4-8	Effect of changing support	53
4-9	The affine correction plotted on a $q - q$ curve.	55
5-1	Color coded plot of Walker Lake exploration data and completely known field	60
5-2	Quality control of declustering operation	62
5-3	Histogram and cumulative distribution function of Walker Lake exploration data	63
5-4	Experimental semi-variance function of Walker Lake exploration data	64
5-5	Normal score transformation of Walker Lake exploration data	65

5-6	Investigation of the spatial correlation of the Walker Lake exploration data in the normal score domain	66
5-7	Data support adjustment through an affine correction	68
5-8	Histogram and experimental semi-variance function of support adjusted normal score values	69
5-9	Comparison of simulated and estimated random field, projected in the normal score domain	70
5-10	Validation of the unconditional simulation results in the normal score domain . .	71
5-11	Validation of the conditional simulation results in the normal score domain . . .	71
5-12	Validation of the conditional simulation results in the original concentration space	73
5-13	Comparison of simulated and estimated random field	74
5-14	Resolution Adjustment of the Simulated Deposit Models	75
5-15	Validation of cumulative distribution function	76
5-16	Validation of histogram	77
5-17	Validation of experimental semi-variance function	78
5-18	Evaluation of the error characteristics	79
5-19	Assessment of the calculated 50% confidence intervals	79
5-20	Assessment of the calculated 95% confidence intervals	80
5-21	Representation of geological uncertainty	81
6-1	Simulation results belonging to the investigated gold deposit	84
6-2	Identify areas for additional drilling	86
6-3	The grade-tonnage and recovered metal curve	87
6-4	Optimized ore/waste selection in mine planning	89
6-5	Future cash flow analysis of the investigated gold mining project	90
6-6	Net Present Value of the investigated gold mining project	91
6-7	Cumulative discounted cash flow of investigated gold mining project	92
6-8	Decision support regarding installed cutting power	93
A-1	Two-sided spectral density function $S(\omega)$	103
A-2	Two-sided spectral density function $G(\omega)$	104
B-1	Data configuration in an ordinary kriging example	108
C-1	Effect of local averaging on variance	114
C-2	Top-down approach to generate one dimensional local average process.	117
C-3	Indexing of one-dimensional LAS cells for stage i (top) and stage $i+1$ (bottom) .	119
D-1	Frequency distribution of Walker Lake Exploration Data	127
D-2	Cumulative Distribution Function of Walker Lake Exploration Data	128
D-3	Basic spatial description of Walker Lake exploration data	129
D-4	Proportional effect of Walker Lake exploration data	131

D-5	h-scattergram of gridded normal scores located 40m apart.	132
D-6	Experimental semi-variance and (auto)covariance function	133
D-7	Experimental (auto)correlation function or correlogram	134
D-8	Lag mean and lag variance of Walker Lake exploration data	135
D-9	Pairwise relative experimental semi-variance function	136
E-1	Preferential sampling	140
E-2	QQ Plot	141
E-3	Polygon Declustering	142
E-4	Cell Declustering	143
E-5	Quality control in declustering operation	145
E-6	Quality Control of Declustering Operation	146
F-1	Walker Lake data adjustment	150
F-2	Color coded plot of Walker Lake exploration data	152

List of Tables

2-1	Selection of a simulation method using a weighted decision table	27
4-1	Change of support, example comparing change in global statistics	54
5-1	Global statistics of gridded, complete and declustered Walker Lake exploration data	63
5-2	Calculation of the variance adjustment factor	67
6-1	Tonnage, average grade and recovered metal for a cut-off grade of 2 <i>g/ton</i> . . .	88
C-1	Weights and Gauss points to evaluate the Gaussian quadrature approximation . .	124
F-1	Walker Lake data adjustment	149
G-1	Walker Lake Exploration Data	155
G-2	Walker Lake Exploration Data gridded	166

Acknowledgments

This research would not have been possible without the support of MTI Holland (member of the IHC Merwede group), and in particular Dr. Mario Alvarez Grima under whose guidance, I chose the topic and further developed it into the framework that exists today.

I would like to gratefully acknowledge the supervision of my advisor, Dr. Ing. Jörg Benndorf MPhil, who has been abundantly helpful and has assisted me in numerous ways. I specially thank him for his infinite patience. The discussions I had with him were invaluable.

I would like to express my sincere thanks to Prof. Dr. Gordon A. Fenton who introduced me into the Local Average Subdivision simulation method and provided with some existing fortran code. The detailed explanation on the software structure was really useful and provided a solid basis for the further expansion.

I am also thankful for the many interesting discussions I had, during the beginning of my master thesis, with Prof. Dr. Michael Hicks and Prof. Dr. Ir. Andre Vervoort. Their useful advice is gratefully acknowledged.

Finally, the author would also like to thank his girlfriend, Marijke Laporte, for her unending patience, care and encouragement during the course of the research and the preparation of this document.

Delft, University of Technology
March 22, 2013

Tom Wambeke

“Imagination is more important than knowledge: for knowledge is limited to what we know and understand while imagination embraces the entire world and all that ever will be known and understood.”

— *Albert Einstein*

Chapter 1

Introduction

1-1 Relevance of Deep Sea Mining

Over the past decade, the global mineral markets have experienced an enormous pressure due to the increasing demand for mineral resources. The supply struggled to keep pace causing the longest boom in prices since the Second World War. This unbalance induced the fear of mineral scarcity and brought mining and metals back on the policy agenda.

The continuous industrialization and urbanization of large developing countries let analysts believe that the world economy has entered an unusually resource intense phase of expansion. The sustained, massive investments in infrastructure, rapid urbanization, and the relocation of a substantial fraction of global manufacturing capacity resulted in an extraordinarily high demand for base metals in China. Also several other large emerging economies will have similar needs in the coming years (e.g. the other BRIC countries: Brazil, Russia and India). However, not only the demand for base metals will keep growing rapidly, many high-tech applications and most notably ‘green’ technologies require growing amounts of more exotic specialty and minor metals. These elements have become an indispensable part in the production of mobile phones, flat screens, hybrid cars, wind turbines or are used in a long list of other applications in aerospace, automobile, defense, optical, semiconductor, and advanced chemical industries, as well as in many emission reduction technologies.

While demand kept growing rapidly, the supply lagged behind, seriously causing a volatile rally in prices. The policymakers and the media began to worry about the threat of mineral scarcity, but the debates mainly focused on the concerns that the ever growing world population and global economy might quickly reach the limits of the planet.

In contrast, careful analysis shows that the cyclic nature of the mining industry, long lead times for setting up new mines and a sore lack of investments are at least as much to blame for the sputtering supply as the exhaustion of available resources. In fact, there is plenty more to be found on our planet, although the ‘low hanging fruit’ has been picked. The exhaustion of many existing mines, shrinking size, increasing remoteness, greater depth, and lower ore grade of new mineable deposits pose significant challenges to expand the global supply. Mining and

refining becomes technically more difficult and increasingly energy intense, seriously driving up the production costs (Rademaker and Koororshy, 2010).

The changing economical and geopolitical climate triggered exploration for unusual resources, thereby increasing the likelihood that seafloor mineral occurrences become economic reserves. Offshore diamond, phosphate, iron sand and gold deposits are already being mined on a regular basis using innovative dredging technology. In addition to these placer mineral deposits, also marine minerals in deep waters are currently being assessed as future potential resources. Such deposits are located near the spreading plate boundaries, which can be found in the ocean basins. These spreading boundaries are deemed to be an active source of mineralization in the form of metalliferous sediments, Seafloor Massive Sulphides (SMS), manganese nodules, manganese crust and gas hydrates (Murton et al., 2001). Currently, companies involved in Deep Sea Mining (DSM) are developing state of the art, groundbreaking and innovative solutions to mine the ocean floor in a cost effective and efficient way (OceanfLORE, 2012; Nautilus Minerals Inc., 2012).

One of these companies, Nautilus Minerals, is pioneering in the exploration and development of high grade copper and gold SMS deposits in the Bismarck Sea. The indicated and inferred mineral resources attributed to the companies first development project, Solwara 1, are 1030 *kt* and 1540 *kt* respectively. This SMS deposit is located approximately 1600 *m* below sea level and has an average copper grade of about 7 %, well above the majority of onshore mining projects which fall below 2%.

When the presence of precious metals (gold and silver) is taken into account, in order to evaluate the equivalent copper grade, the difference is even more significant, i.e. 12% compared to the onshore majority that falls below 3% (Lipton, 2012).

To extract such SMS deposits from the ocean floor, Nautilus designed an offshore production system that can be relocated and reused on subsequent sites (Figure 1-1). The proposed system performs following key operations: seafloor excavation and gathering, ore (slurry) recovery to the surface, slurry dewatering and ore discharge to transportation barges.

During production, three Seafloor Production Tools (SPT) are responsible for the cutting and gathering of ore from the sea floor. First, the auxiliary cutter assists in leveling and grinding remnant volcanic chimneys to create flat operational areas. Thereafter, the bulk cutter is deployed on the prepared leveled benches, where it transfers cutting power to the rock mass in an efficient way to ensure high productivity. Finally, the collecting machine gathers the excavated rock fragments with its slewing recovery head and pumps the resulting slurry to the Riser and Lifting System (RALS) via a flexible transfer pipe.

The RALS includes a Subsea Slurry and Lifting Pump (SSLP), which is connected with a vertical riser assembly suspended below the Production Support Vessel (PSV). Once on deck, the slurry is dewatered and discharged to a cargo barge, which transports the dewatered ore to shore for stockpiling. Subsequently, the ore will be loaded into bulk carriers for the transportation to a treatment plant, where the further processing takes place. Eventually the concentrate is delivered to the market (Smith, 2011).

Figure 1-1 shows an artist impression of the Offshore Production System (OPS) including the Seafloor Production Tools, the Riser and Lifting System and the Production Support Vessel (Nautilus Minerals Inc., 2012; Soil Machine Dynamics Ltd., 2012).

Figure 1-1: Offshore Production System; top left: equipment performing key operations; top right: auxiliary cutter, middle right: bulk cutter, bottom right: collecting machine; bottom: production support vessel (Courtesy of Nautilus Minerals Inc. and Soil Machine Dynamics).

1-2 Added Value of Geostatistical Simulations

As land resources diminish and technology evolves, DSM is becoming a viable and sustainable alternative to meet the increasing demand for minerals. While most effort is put in new technological developments in the field of underwater mining, it is equally vital to understand the unique management processes and risks of constructing and operating an offshore mine. The importance of establishing thorough geotechnical and mineralogical knowledge and understanding of the resource can not be overstated. During deep sea exploration programs, mineral composition and metal content are often of primary interest, because a proper reserve estimation can turn a deposit into a bankable project. On the other hand, for the development of an underwater mine, a proper geotechnical characterization is also crucial to design and select the proper mining tools, which have to meet all the desired functional requirements (Searle and Smit, 2011).

Unfortunately, due to the high operating, maintenance and investment costs of the diving support vessel and the ROV (Remotely Operated Vehicle), a sample coverage equivalent to a common land based exploration program would be extremely expensive and highly unlikely. Also the remote location and the technical difficulties to collect samples, further decrease the reliability of the gathered information (van de Ketterij, 2010).

Nevertheless, the minimal amount of information is interpreted by scientists and engineers and used to build 2D/3D deposit models. Generally, a substantial part of the resulting deposit models contain spatial estimates of geotechnical and resource parameters. Geostatistics and spatial statistics are the scientific fields that deal explicitly with data distributed in space and aim at modelling the spatial relationships between the individual data points. The available models, however, rarely focus on the calculation of reliable measures of uncertainty. At most, the conventional modelling techniques are able to give an indication on how well the prediction algorithm is performing (Caers, 2011). In the context of deep sea mining, the uncertainty on the spatial inferred information is increasing and thus an improved understanding on the reliability of the results is indispensable.

Another argument, that supports the application of an uncertainty quantification framework, is the desire to translate the geological uncertainty into project risk. A successful mining project is built on the sound identification of the resource, proper equipment selection, a thoughtful production plan and good management. These four key activities can be further optimized by analyzing how the spatial variability and uncertainty of the ore body properties impact the final mining operation.

To address this problem, the R&D institute of IHC Merwede (MTI Holland) decided to initiate a geostatistical research program (Alvarez Grima et al., 2011). The proposed research aims to develop a general simulation framework, in which the spatial predictions are calculated together with their corresponding uncertainties. This master thesis is a reflection of the first steps towards that direction.

Instead of obtaining a single prediction value for a certain geotechnical or resource parameter at a specific location or region of interest, the focus now lies on obtaining a probability distribution. The resulting distribution can be regarded as a representation of the geological uncertainty.

In order to identify the (financial) project risk or to optimize the design of deep sea extraction equipment, it is necessary to propagate the characterized geological uncertainty through the complete value chain. The obtained spatial distributions thus have to be inserted into physical

Figure 1-2: Conceptual visualization of the added value of the combination between geostatistical simulations and transfer functions. The main advantage of the proposed simulation framework lies in the fact that the spatial variability and uncertainty is propagated through the whole equipment design and mine planning process. As such, the whole procedure results in a risk robust decision, increasing the value of the project.

models, which translate the geological parameters into financial or operational performance indicators. These models, the so called utility or transfer functions, can be used to calculate for example the cutting force, power requirement, bearing capacity, cash flows, grade tonnage curve or Net Present Value (NPV).

To further aid the decision making process, the likelihoods and the corresponding (economic) consequences of different scenarios can be compared and evaluated. Often the entire analysis can be summarized, via a decision or forecast model, into a single monetary value which estimates the expected profit, loss or cost associated with each investigated scenario. Typically the scenario that maximizes the monetary value of the mining project is preferred and implemented. Such a scenario analysis does not necessarily need to be performed explicitly. For certain decision problems, optimization algorithms exist, which automatically explore the different scenarios and produce the ‘best’ decision.

Figure 1-2 presents a conceptual example of such an analysis. Data, obtained by drilling boreholes during the exploration phase, is used to calculate the spatial probability distributions of the Uniaxial Compressive Strength (UCS), Brazilian Tensile Strength (BTS) and friction angle in the region(s) of interest, in this case a single mining block. The uncertainty in the requisite geotechnical parameters is taken into account during the calculation of the cutting force and power requirement.

Such information can then be used to calculate, for example, the probability that the cutting forces in the selected mining block are exceeding a certain specified threshold value. A similar threshold analysis can be done for the power requirement. The obtained threshold probabilities are subsequently used to decide on the size of the deployable equipment and the installed engine type. If the calculated probability of exceeding a threshold is too high, engineers can even decide to modify the initial mining method or to completely avoid the risky areas.

1-3 Purpose and Research Objectives of the Thesis

The main purpose of the research is to initiate the development of the previously outlined modeling framework. Three main components need to be considered:

- **Spatial stochastic model:** The spatial stochastic model needs to be able to capture and describe the spatial variability and the geological uncertainty. A suitable simulation technique needs to be used to generate equally probable realizations of the in-situ orebody variability. Contrary to a single estimated smooth orebody model, the simulated realizations provide the tools necessary to quantify uncertainty and risk (Dimitrakopoulos et al., 2002).
- **Physical model (transfer/utility function):** The physical model calculates the operational or financial performance indicators of the corresponding inserted realization. The entire set of simulated realization can subsequently be converted into a distribution of the response variable, which defines its space of uncertainty (Benndorf, 2012a).
- **Decision or forecast model:** The decision or forecast model can be used to translate the operational or financial performance indicators into risk robust decisions. During this phase of the analysis, the goal is to optimize the mine planning or equipment selection under the modeled geological uncertainty.

The performed research, presented in this thesis, mainly focused on the development of a new simulation methodology that is able to generate equally probable deposit models. The physical and decision models, on the other hand, are just briefly discussed during some case studies to illustrate the added value of geostatistical simulations in deep sea mining.

To structure and guide the development of the new simulation methodology, following research objectives were formulated:

1. Evaluate the available geostatistical simulation techniques and assess their applicability in the context of deep sea mining. Select a method that meets most of the formulated requirements (The requirements for a suitable simulation method are discussed in the following chapter).
2. Adapt the selected simulation method, so it can deal with some specific conditions, inherent to the nature of the modeling problem. Generally, such conditions concern non-normally distributed data, preferentially collected in high valued clusters, on a support that differs from the one that is used during decision making.
3. Develop a framework/methodology to assess the geological uncertainty and variability in deep sea mining. The developed methodology needs to result in some transparent guidelines which ensure that the generated deposit realizations reproduce the available data, the descriptive statistics and the characterized variability.
4. Illustrate how the methodology needs to be used in practice and validate the simulation results in an exhaustively known environment. Such a validation creates confidence that the developed methodology works fine.
5. Present a case study that illustrates how the geological uncertainty impacts the mining operation and influences the corresponding decisions.

1-4 Outline of the Thesis

The outline of the master thesis reflects the formulated research objectives.

After discussing the relevance and revival of deep sea mining, **Chapter 1** introduces the idea of a stochastic modeling framework that is able to translate geological uncertainty into project risk. Further, this chapter presents the research objectives and the thesis structure.

Chapter 2 provides a brief theoretical background on topics, related to the current state of the art in the modeling of geological uncertainty. The chapter first presents the theoretical principles behind the random function model. Thereafter, the chapter discusses the stationarity and ergodicity assumptions that are frequently, and often implicitly, used during geostatistical estimation or simulation. Subsequently, the justifications for using a simulation technique are highlighted. Then, the requirements for a suitable simulation technique are listed and elucidated. The chapter finally concludes with a brief comparison of the possible simulation techniques.

Chapter 3 contains a more detailed explanation of the selected simulation method. Further, the chapter clearly states the inherent assumptions and discusses some of the limitations regarding the selected Local Average Subdivision technique (LAS). The chapter concludes with a critical discussion on the necessary adjustments to overcome some of the irreconcilable limitations.

Chapter 4 discusses solutions to overcome some of the undesirable constraints, that are inherently connected with the selected LAS simulation method. Moreover, the proposed solution package is integrated into a general simulation framework that is systematic, robust, and easy to follow. The transparent methodology also contains a few validation procedures to check whether the simulated realizations reproduce the available data, the descriptive statistics and the modeled variability.

Chapter 5 presents a case study that illustrates how the developed methodology needs to be used. The investigated exploration data originates from an exhaustive benchmark data set, which can be used to validate the simulation results and assess the quality of the developed methodology.

Chapter 6 contains two case studies, illustrating the advantages of the application of a simulation based approach. Both case studies show that the geological uncertainty can have a significant impact on the project risk, which is otherwise underestimated or completely ignored during the conventional estimation analysis.

Chapter 7 is the last chapter of the thesis and presents the final recommendations and conclusions.

Current State of the Art in Modelling Geological Uncertainty

2-1 A Probabilistic Approach to Describe Spatial Variation: Random Function Model

2-1-1 Necessity of Probabilistic Modeling

The total cost of offshore exploration cruises is significant, due to the remoteness of the deep sea and the technical difficulties to extract small bodies of material from the ocean floor. Such high costs generally result in a small number of measurements, separated from one another by relatively large distances. Yet the engineers, who have to make decisions regarding equipment design and mine planning, would like to characterize the ground properties in the intervening space. In other words, they would like to predict in a spatial sense from their more or less sparse data.

Important to realize is that a model, is but a simplified representation of the largely unknown reality. Although the reality is unique, in practice, it is impossible to come up with a single conclusive ‘picture’ of the area under investigation. Generally, many possible realizations can be generated, which all explain the observed ground properties. To be acceptable and valid, a model should include all relevant information, should be tractable and should be tuned to the decision question at hand (Webster and Oliver, 2007).

The described prediction problem can be solved through the use of either a probabilistic or a deterministic model.

The outcome of a deterministic model is precisely captured through known relationships and measured states, without including any random variation. This type of model results in a unique single representation of the reality. Theoretically, it is possible to formulate a physical model, describing the geochemical, biological and erosional processes governing the observed spatial variation in the ground properties. Such a deterministic model would associate to any unsampled location a single estimate without documenting the potential error

or uncertainty. Such a procedure might be justified, if the estimate is based on many data or on a sound understanding of the interactions between all the physical processes responsible for the spatial variation. The limited scientific understanding, together with sparse data, make the deterministic modeling approach highly unreliable (Webster and Oliver, 2007; Isaaks and Srivastava, 1989)

In the context of deep sea mining, where the stakes are high due to required investment capital, it would be better to assess the range of values the unknown could take and calculate the probability of occurrence of each outcome (Goovaerts, 1997). The outcome of such a probabilistic approach reflects our imperfect knowledge of the unsampled value and more generally the distribution of the property within the area. Instead of obtaining a single estimated value for the unknown, the approach results in a probability distribution.

To conclude this subsection, two reasons for selecting the probabilistic modeling approach are highlighted.

The first reason is motivated by the current inadequate scientific understanding of certain phenomena, governing the formation of deep sea deposits. The complex and chaotic interactions of geological processes further complicate the development of physical models which are able to describe the spatial variation.

The second reason to apply a probabilistic modeling approach is provided by the desire to calculate risk indication values, i.e. probabilities of certain specified negative events. Prior to the calculation of such indication values, interest lies in obtaining, for each unsampled location, a set of possible outcome values and their corresponding probability of occurrence. This formulated concept is the basic idea behind the theory of the *random function* or *regionalized variable*, also known as geostatistics due to its applications in mining and petroleum engineering (Matheron, 1971).

2-1-2 Random Variable

A random variable, as opposed to other mathematical variables, does not have a single fixed value. Rather it can take on a set of possible different values, each with an associated probability. A realization of a random variable should be considered as the outcome of a random event, which is determined by a probabilistic mechanism (Isaaks and Srivastava, 1989). This type of variable ideally suits the requirement to include uncertainty and risk into the inference procedure.

Since the originating processes are so complex and can not be disentangled, the outcome of the ground properties can be regarded as random. The uncertainty in the state of the properties can be described by two different types of random variables: the continuous or the discrete variable (Goovaerts, 1997). An additional type, the indicator variable is ideally suited to calculate the probability of certain events.

A continuous random variable is used to model an attribute which has a continuous range of possible outcome values, exhibiting a natural order. Such a random variable, $Z(\mathbf{x}_\eta)$, is fully characterized by its cumulative distribution function (cdf), which expresses the probability that the variable Z at location $\mathbf{x}_\eta$ is no greater than any given threshold, b (Webster and Oliver, 2007; Goovaerts, 1997):

$$F(\mathbf{x}_\eta; b) = \text{Prob}\{Z(\mathbf{x}_\eta) \leq b\} \quad \forall b. \quad (2-1)$$

Depending on the property under investigation, the reliability can be expressed as the probability that the attribute does or does not exceed a specified threshold value. For example, from a financial perspective, it would be favorable if the metal grade in a certain block exceeds a specified cut-off grade. Operational downtime, on the other hand, can be prevented by ensuring that the rock strength of the same block is below a critical threshold. Both probabilities are easily calculated as the expectation of the indicator variable over m simulations or realizations (a realization of a random variable is denoted by a lowercase letter):

$$F(\mathbf{x}_\eta; b) = \text{Prob}\{Z(\mathbf{x}_\eta) \leq b\} = E[I(\mathbf{x}_\eta; \mathbf{b})], \quad (2-2)$$

where the indicator variable for continuous processes is defined as:

$$I(\mathbf{x}_\eta; b) = \begin{cases} 1 & \text{if } z(\mathbf{x}_\eta) \leq b, \\ 0 & \text{otherwise.} \end{cases} \quad (2-3)$$

The current interest lies in the spatial variability of the geotechnical and mineralogical properties, hence the discrete variable will not be discussed.

The basic paradigm of the probabilistic modelling approach is to handle an unknown attribute as a random variable. In this way, the trouble of assessing the uncertainty and risk is now reduced to an exercise in modelling the probability distribution of a random variable at a particular location. By accounting for the dependence between random variables at different locations, the random function model allows to account for structures in the spatial variation (Goovaerts, 1997; Webster and Oliver, 2007).

2-1-3 Random Function

The random function or random process model defines a set of dependent random variables $Z(\mathbf{x})$, one for each location $\mathbf{x}$ in the study area $\mathcal{A}$, $\{Z(\mathbf{x}), \forall \mathbf{x} \in \mathcal{A}\}$. The objective of the random function model is to describe how the ensemble of possible values evolves with position (Isaaks and Srivastava, 1989).

Now, let $\mathcal{E}$ denotes an arbitrary set of k locations $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$. Then the random function Z on $\mathcal{E}$ is defined as a vector of k random variables $\{Z(\mathbf{x}_1), Z(\mathbf{x}_2), \dots, Z(\mathbf{x}_k)\}$. In general, the set $\mathcal{E}$ is considered to be finite, made up of points in the two or three dimensional space. The random function on $\mathcal{E}$ is said to be fully specified if for any k elements the joint cdf is given (Kolmogorov, 1950; Yaglom, 1962):

$$F(\mathbf{x}_1, \dots, \mathbf{x}_k; b_1, \dots, b_k) = \text{Prob}\{Z(\mathbf{x}_1) \leq b_1, \dots, Z(\mathbf{x}_k) \leq b_k\}. \quad (2-4)$$

This function expresses the probability that all the k elements have a value that does not exceed their corresponding threshold limit. Such a multivariate cdf is ideally suited to describe the joint uncertainty about the k inferred values. Unfortunately, it is almost impossible to use Eq. (2-4) in practice, not only because of the existing mathematical challenges (multidimensional integral of probability density function), but also in that it is difficult to infer its parameters from real data.

To simplify the characterization of the problem, the random process is often described by a Gaussian joint probability density function (pdf) (Diggle and Ribeiro Jr., 2007). The great advantage of a multivariate Gaussian distribution is that the complete distribution can be

characterized by just a mean vector and a covariance matrix. The multivariate Gaussian pdf expresses the likelihood that the set of k dependent random variables $\{Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_k)\}$ take on the values $\mathbf{b} = \{b_1, \dots, b_k\}$ (Anderson, 1958):

$$f(\mathbf{x}_1, \dots, \mathbf{x}_k; b_1, \dots, b_k) = \frac{1}{(2\pi)^{k/2} |C|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{b} - \boldsymbol{\mu})^T C^{-1}(\mathbf{b} - \boldsymbol{\mu})\right\}, \quad (2-5)$$

where $\boldsymbol{\mu}$ is a vector containing the mean values for all $Z(\mathbf{x}_i)$, C is the covariance matrix and $|C|$ is its determinant. The elements of the covariance matrix C_{ij} give an indication of the dependence between $Z(\mathbf{x}_i)$ at location $\mathbf{x}_i$ and $Z(\mathbf{x}_j)$ at location $\mathbf{x}_j$.

Yet, this simplification will not be sufficient. In order to make the random function model applicable, some additional assumptions regarding stationarity and ergodicity have to be made. Without these assumptions, it is impossible to estimate, from measurement data, the parameters governing the behavior of the Gaussian process. The overall spatial behavior of the random process is prompted by a covariance function, describing the variation of the ensemble in a simple and convenient way.

2-2 Characterization of Spatial Variation and Some Inevitable Assumptions

2-2-1 Spatial Covariance Function

The covariance function needs to capture and describe the spatial dependence between two regionalized variables. Presumably, two regionalized variables in close proximity are similar, whereas those that are at more distant locations are less similar. Their covariance can be inferred as (Webster and Oliver, 2007; Banerjee et al., 2004; Fenton, 1990):

$$\begin{aligned} C(\mathbf{x}_1, \mathbf{x}_2) &= E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\} \cdot \{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}] \\ &= \int_{\theta} \int_{\theta} b_1 b_2 f(\mathbf{x}_1, \mathbf{x}_2; b_1, b_2) db_1 db_2 - \mu(\mathbf{x}_1) \cdot \mu(\mathbf{x}_2), \end{aligned} \quad (2-6)$$

where the integral of the joint pdf is evaluated over the entire sample space θ of the random variable Z and where $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ represent the mean values of the property Z at location $\mathbf{x}_1$ and $\mathbf{x}_2$.

Notice that this definition can not be used for statistical inference, since only one realization per location is available. This impasse is overcome by introducing the stationarity assumptions, which allows us to treat the values at different positions as though they are different realizations of the same property (Journel and Huijbregts, 1978). More information can be found in Appendix A-1.

2-2-2 Stationarity Assumptions

When a spatial stochastic process is said to be stationary, it means that the process has certain attributes that are the same everywhere. In following discussion, it is assumed that the spatial process has a mean, say $\mu(\mathbf{x}) = E[Z(\mathbf{x})]$ associated with it, and that the variance of $Z(\mathbf{x})$ exists for all $\mathbf{x} \in \mathcal{A}$.

Generally three types of stationarity are widely accepted in geostatistics: strict stationarity, weak stationarity and intrinsic stationarity.

The most restrictive assumption, *strict stationarity*, states that the random variables corresponding to two sets of locations should have the same joint distribution if the second set of locations is a translated version of the first one (Journel and Huijbregts, 1978).

A less restrictive condition is given by *weak stationarity*, which only dictates that the first and second order moments of the joint distribution have to be constant. Weak stationarity thus assumes that:

- the individual realizations fluctuate around a constant mean,
- the *a priori* or point variance is finite and constant,
- the covariance between two random variables is only depended on the displacement vector $\mathbf{h}$ and independent of the absolute locations.

This weak stationarity assumption thus postulates that the covariance relationship between the values of the process at any two locations can be summarized by a covariance function $C(\mathbf{h})$, which depends only on the separation vector $\mathbf{h}$. Note that strict stationarity implies weak stationarity. The converse is not true in general, but does hold for a Gaussian process because its higher order moments are known constants (Banerjee et al., 2004).

The third type of stationarity, *intrinsic stationarity*, was introduced by Matheron (1971) to overcome the problem of a mean that changes across a region, and a variance that increases without bound as the region of interest expands. In these circumstances, intrinsic stationarity places restrictions on the first and second order moments of the differences between a regionalized variable at any pair of locations. Matheron took the view that, although the mean might not be constant over the whole region, it will be in smaller subregions, so that the expected difference of a pair would approach zero for short lag vectors.

Further he replaced the covariance function with a function that expresses the expected variance of the difference as a function of the lag vector. This so called variogram is an equivalent measure of the spatial relation between two regionalized variables separated by a vector $\mathbf{h}$:

$$\begin{aligned}\gamma(\mathbf{h}) &= \frac{1}{2} \text{E} \left[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2 \right] \\ &= \frac{1}{2} \text{Var} [Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})].\end{aligned}\tag{2-7}$$

Only under the condition of ergodicity, it is possible to recover the covariance function from the semi-variogram. For additional information, the reader is referred to Appendix A-2.

2-2-3 Assumption of Ergodicity

The additional ergodicity assumption implies that, when the regional boundaries expand towards infinity, the moments inferred from the spatial observations (one realization per location) converge towards those of complete unknown ensemble (Webster and Oliver, 2007; Chiles and Delfiner, 1999). Ergodicity essentially says that the mean and covariance structure

of a random process can be inferred from a single realized field of infinite extent (Fenton and Griffiths, 2008). As a result, the covariance between two points in space should vanish as the separation distance becomes infinitely large.

It is important to realize that the assumption of ergodicity is implicitly connected with the use of the random function model and that the property can not be proved or disapproved. From a practical point of view the condition relates to the following questions:

- *Is the investigated area large enough to properly infer the complete variogram or covariance structure?* After all, the correlation length needs to be considerably lower than the extent of the investigated domain.
- *Is there enough data available to robustly calculate the necessary expectations?* After all, the statistical moments inferred from the scarce data need to approach those of the exhaustive sampled reality.

Please notice, that the assumptions of stationarity and ergodicity are not properties of the ore body attribute itself, but are merely required to follow a certain modelling strategy. The user has to judge whether the random function model and its related assumptions correspond to the analyzed data and the characteristics of the ore body itself (Benndorf, 2012b).

2-2-4 Spectral Density Function

The covariance and semi-variogram function are by far the most widely used descriptive tools in geostatistics to characterize the spatial relation of geological attributes. However, there is another equivalent approach which takes advantage of the spectral representation of a random process. Information, concerning the second moment of a stationary Gaussian random process, is eventually summarized in the so called spectral density function. The spectral density function describes how much each frequency component contributes to the overall variance of the random process. More information can be found in Appendix A-3.

2-3 Geostatistical Estimation

2-3-1 The Kriging Paradigm: Weighted Linear Combination, Unbiasedness Condition and Minimized Error Variance

Mainly due to reasons of economy, the number of samples gathered during a feasibility study is limited. During this early phase, managers and engineers nevertheless would like to have a complete characterization of the deposit. In order to comply with these stringent requirements, the properties of the deposit at unsampled locations need to be estimated from the available data.

In the context of the probabilistic modeling approach, each measurement at a sampling location can be considered as a realization of a regionalized variable. Hereby, it is sensible to estimate an unknown true value at a particular location $\mathbf{x}_\eta$, as the most probable realization of an unknown random variable $Z(\mathbf{x}_\eta)$ (Webster and Oliver, 2007). The modeled random

variable, denoted by $\hat{Z}(\mathbf{x}_\eta)$, can be obtained by following linear combination:

$$\hat{Z}(\mathbf{x}_\eta) = \sum_{\alpha=1}^{n_k} \lambda_\alpha z(\mathbf{x}_\alpha), \quad (2-8)$$

where each λ_α assigns a weight to the corresponding datum $z(\mathbf{x}_\alpha)$ sampled at location $\mathbf{x}_\alpha$. Subsequently, the weights need to be calculated in such a way, that they account for the correlation between the data and the unknown value, and they are able to deal with the redundancy of the available information (Caers, 2011). Under the assumption that the mean of the random process is unknown, *ordinary kriging* is the Best Linear Unbiased Estimate (BLUE) suitable for this job. The term ‘best’ is inherently connected with the following requirements:

- The **unbiasedness condition** states that the expected error, calculated over all the prediction points, should be zero:

$$E[\hat{Z}(\mathbf{x}_\eta) - Z(\mathbf{x}_\eta)] = 0. \quad (2-9)$$

When the mean of the random process is unknown, the unbiasedness condition further implies that the ordinary kriging weights, λ_α , need to sum up to one. Proof is given in Appendix B-1.

- The second constraint imposes that the **error variance**, calculated over all the prediction points, needs to be **minimized**:

$$\sigma_R^2 = E[\{\hat{Z}(\mathbf{x}_\eta) - Z(\mathbf{x}_\eta)\}^2] \rightarrow \text{minimize} \quad (2-10)$$

A detailed discussion on the further elaboration of the error variance can be found in Appendix B-2.

The minimization of the fully formulated error variance, while taking account of the unbiasedness condition, finally results in the ordinary kriging equations. The solution of the resulting system of equations will yield a unique set of weights, which safeguards both formulated requirements. The mathematical details are omitted here, but can be reviewed in Appendix B-3.

2-3-2 Ordinary Kriging Equations

Ordinary kriging takes advantage of the modeled covariance between the data locations $\mathbf{x}_i$ and the target point $\mathbf{x}_\eta$, $C_{i\eta}$. Moreover, in order to reduce the data redundancy, the covariance between individual data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, is also taken into account, C_{ij} . To find the weights, λ_α , which are necessary to calculate the estimate in Eq. (2-8), ordinary kriging essentially solves following linear system of equations:

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n_k} & 1 \\ C_{21} & C_{22} & \cdots & C_{2n_k} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n_k 1} & C_{n_k 2} & \cdots & C_{n_k n_k} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n_k} \\ \psi \end{bmatrix} = \begin{bmatrix} C_{1\eta} \\ C_{2\eta} \\ \vdots \\ C_{n_k \eta} \\ 1 \end{bmatrix}. \quad (2-11)$$

The ordinary kriging weights are finally obtained by inverting Eq. (2-11).

The bottom row of Eq. (2-11) ensures that the unbiasedness condition is not violated (Webster and Oliver, 2007). The additional Lagrange parameter ψ needed to be inserted in order to convert the constrained minimized problem into an unconstrained one. Further information can be found in Appendix B-3.

2-4 The Necessity of Geostatistical Simulations

The main goal of this section is to illustrate the necessity and the additional advantages of geostatistical simulations. First, the text reconsiders some very important application targets for the generated orebody realization(s). Subsequently, some inevitable side effect of kriging are exposed. In the face of the formulated application targets, these effects seem to be very undesirable. Next, some simulation objectives, regarding statistics that are deemed to be the most consequential for the problem at hand, are formulated. Finally, the concept behind the flaw of averages is introduced, confirming the dire necessity of geostatistical simulations. The entire section actually illustrates how geostatistical simulation kills multiple birds with one stone.

2-4-1 Application Targets for the Generated Realization(s)

A crisp description of the application targets for the generated realization(s) is presented here, in order to show further ahead that the side effects of kriging are deemed to be problematic.

The first target is a correct **representation of the spatial variability** of the characteristics of the recoverable reserves at various stages in the mining project (Journel and Huijbregts, 1978). The successful design and planning of following operations largely depends on a correct prediction of the spatial dispersion characteristics:

- selection of deep sea mining equipment and the estimation of the potential mining recovery rate,
- ore/waste discrimination and the estimation of material volumes that needs to be pumped to the surface,
- development of stockpiling and blending strategies,
- performance and flexibility of processing plant and the estimation of metal recovery.

This short list shows that the variability has a significant impact on downstream processing steps and fundamentally drives realized value (Vann et al., 2012). The consequences of an incorrect characterization of the spatial variation should not be underestimated (Dimitrakopoulos et al., 2002).

The second target is to detect **patterns of extreme attribute values**, such as zones of high rock strength or zones rich in metal. Depending on the modeled attribute, such zones are particularly important since they entail an upside potential or a downside risk. Both opposite situations can have a significant impact on the entire mining operation.

The third and last target is a **realistic characterization of the geological uncertainty**, which is recognized to be a major source of technical risk. A correction quantification and propagation of geological uncertainty results in the sheltering of strategic investments and creates an operation that performs closer to its potential (Dimitrakopoulos et al., 2002; Vallee, 2000). Following statement stresses the importance of this third target:

... in the first year of operation after start-up, 60% of mines surveyed had an average rate of production less than 70% of the designed capacity. Although shortfalls in production are also due to problems in scale-up from pilot plant to commercial plant, the quantity and grade of ore are a major contributor to potential shortfalls ... (Vallee, 2000)

2-4-2 The Undesirable Effects of Geostatistical Estimations

Although the beginning of this chapter took off with the ambition to model the probabilistic mechanism of a largely unknown orebody model, the effort eventually resulted in a way to calculate only the ‘best’ estimate at a single location. The objective of such a calculation is to provide, at a specific location, an estimator which is as close as possible to the true unknown grade. The quality of each estimation is measured separately, i.e. independently of neighboring estimates, through its unbiasedness and error variance (Journel and Huijbregts, 1978).

A deposit model, consisting of an entire collection of such ‘best’ estimates, however might not be the best as a whole. After all, there is no reason to believe that these estimators reproduce the spatial variability of the true grades. In the case of kriging, for instance, the minimization of the estimation variance causes an insurmountable smoothing of the true dispersions (Journel and Huijbregts, 1978; Goovaerts, 1997; Vann et al., 2012). Yet, this variability has a significant impact on downstream processing steps and fundamentally drives realized value (Vann et al., 2012).

As shown in Figure 2-1, estimation algorithms tend to smooth out local details of the spatial variation of the orebody attribute. Typically small values are overestimated, whereas large values are underestimated. This conditional bias is a serious shortcoming and immediately violates the first two application targets. Another drawback, related to this smoothing effect, is that the degree of spatial variation depends on the proximity and density of data locations (Goovaerts, 1997). The smoothing tends to be minimal near data locations and increases when the prediction points are located farther away. Therefore, an estimated orebody model appears to be more variable in densely sampled areas. This results in artifact structures, which are misleading and completely disturb the correct characterization of the spatial continuity (David, 1977).

Probably the most detrimental consequence of this smoothing effect is the loss of variance, compared to the one that is present in the initial data. Furthermore, the modeled covariance function, obtained from the input data, is not reproduced by the obtained kriging estimates (Figure 2-1). The larger is the kriging variance on average, the more variance is lost (Webster and Oliver, 2007). The estimated deposit model is thus also biased, with respect to the dispersions of the true grades.

Finally, it is also worthwhile to mention that the kriging variance, corresponding to a particular prediction point, is not really a representative measure of uncertainty. After all,

Figure 2-1: Estimation versus Simulation; Estimation algorithms tend to smooth out local details of spatial variability. The simulation approach, on the other hand, generates a more realistic representation of the in-situ variability. The dispersion of the data histogram and the sill value of the modeled variogram are reproduced. ((Benndorf, 2005)).

a kriging variance is calculated based solely on the spatial configuration of the sampling locations and the modeled covariance function. The actual measured values in the near proximity of a prediction point are not considered (Diggle and Ribeiro Jr., 2007). The use of the kriging variance as an uncertainty indicator can be called into question. Moreover, the non-linear nature of several transfer functions limit the use of confidence intervals and thus the ‘modeled’ uncertainty can not be further propagated during the analysis of the mining operation (Dimitrakopoulos, 1998).

Both remarks show that the geostatistical estimation techniques are not suitable nor practical to characterize the geological uncertainty in a realistic way (violation of third application target). This, in combination with the smoothing effect and its undesirable consequences, lead to the conclusion that a simulation based approach might be more suitable to meet previous formulated application targets.

2-4-3 Requirements for Simulated Realizations

Geostatistical simulation techniques are designed to generate realizations with the same first two experimentally found moments as the true grades. Therefore they do not suffer from the previously exposed undesirable effects.

Each orebody realization originates from the same probabilistic mechanism, i.e. described in the first two section of this chapter. Once certain model parameters are inferred from the available data, several possible simulation techniques can be used to generate equally probable orebody models from this probabilistic mechanism. Following prerequisites are common to most simulation methods (Goovaerts, 1997; Benndorf, 2005):

- Measured values are honored at their sampling locations.

Figure 2-2: Concept of the flaw of averages; The upper part of the figure indicates the situation where an expected value $E[Z(x)]$ of an unknown attribute $Z(x)$ is inserted into a non-linear transfer function. This illustrates approach is deemed to give incorrect results. The lower part, on the other hand, indicates the correct situation where the entire probability distribution is sampled and inserted into the non-linear transfer function. This results in a set of output values from which subsequently an expected value can be calculated. The collection of output values also immediately gives an indication about the uncertainty. Notice that the same procedure can be performed with sampled orebody realizations (adopted from (Martinez, 2009)).

- The histogram of the simulated values closely reproduces the representative histogram of the sample data.
- Simulated realization reproduce the spatial continuity of the data, i.e. the modeled covariance function is reproduced in each generated orebody model.

Figure 2-1 demonstrates that the simulation approach results in a more realistic representation of the spatial variability, compared to the estimation approach. This time, the dispersion measures of the simulated values are similar to those of the available data. Also the shape of the histogram and the modeled covariance function are reproduced in the simulated realizations.

Before proceeding, it is important to be fully aware of the difference between estimation and simulation. Generally the objectives of both modeling strategies are not compatible. An estimated realization will on average be closer to the reality, whereas a simulated realization is a better reproduction of the spatial variability. Estimations are preferable to infer and locate reserves, while simulations are preferred to study the dispersion of the characteristics of these reserves (Journel and Huijbregts, 1978).

It is important to mention that a collection of simulated orebody realizations can be converted (averaged) into a single estimated deposit model, making a separate estimated realization unnecessary. The simulation based approach thus entails great flexibility and is actually the only correct way when dealing with non-linear transfer functions. This statement is further elucidated in the following subsection.

2-4-4 The Flaw of Averages

The problems associated with geostatistical estimation are not only caused by its incapability of representing the correct spatial variability, but also by the fact that only a single estimated orebody model is generated. Such a single model does not provide the necessary support for risk-based decision making. Moreover, the use of a single estimated orebody model might lead to incorrect decisions (Martinez, 2009; Savage, 2003).

The reason behind this problem can be found in the so called Jensen's inequality. In finance, this inequality states that because the value of a project, v , is a random variable and the option value of the project OV is a convex function of the project value, then $OV(E[v]) \neq E[OV(v)]$. Professor Savage, from Stanford University, calls this phenomena the *flaw of averages*.

Martinez (2009) has published a very intuitive scheme which shows that average inputs do not always yield average outputs when dealing with uncertainty and non-linear transfer function (Figure 2-2). The concept is definitely applicable to the current environment and shows that serious problems can arise when a local probability distribution or global collection of deposit models is substituted respectively for a single estimated number or deposit model.

To summarize, the concept of the flaw of averages further stimulates the use of simulated orebody models.

2-5 Geostatistical Simulation Techniques

2-5-1 Selection Criteria for a Suitable Simulation Method

The discussion in the previous section clearly illustrated the urgent need and the associated advantages of geostatistical simulations. At this point, it is important to select the right simulation technique that corresponds best with the needs induced by the problem at hand.

The most important prerequisites were already formulated before, and postulate that each generated realization should reproduce the statistics, which are deemed to be consequential for a correct representation of the in-situ variability (Ref. subsection 2-4-3). The final selected method will certainly have to meet all three formulated prerequisites.

Additionally, some less stringent criteria are described, which might facilitate a quick decision regarding the final selection:

- *easy to use*: Does the method requires a complicated calibration procedure? In order to convince the industry to apply the previously discussed framework, it should be easy to use.
- *easy to implement/adjust*: How complex is the software structure of the selected algorithm? A clear and simple structure makes it much easier to add/adjust new features. Sufficient flexibility makes it possible to address all kind of questions related to the spatial variability and uncertainty.
- *theoretical complexity*: How complex are the theoretical principles behind the selected simulation algorithm? In order to raise some public support for the developed framework, the concept behind the simulation algorithm should be appealing and intuitive. Too complicated techniques will not be convincing and the implementation of an entire risk and uncertainty based strategy might be put on hold.

- *resolution/support*: What is the resolution of the simulated field? How does it relate to the sample volumes used during decision making? Is re-blocking necessary? Preferably, the resolution of the simulated field should be the same as the one used during decision making. Such an approach makes it possible to explicitly validate that the statistics are consistent with the field resolution.
- *data management*: How much memory is required to perform a simulation? Which calculated information is stored and reused later?
- *Computational Efficiency*: How long does it take to generate a reasonable amount of realizations? Is the method designed to make smart use of previously calculated information, thereby significantly improving the computation speed?
- *size of simulation*: How many blocks or grid nodes can be generated within an acceptable time frame? How much memory is required to accomplish this?
- *scientific contribution*: Does a detailed investigation and a further implementation of the selected method lead to a new scientific contribution?

These formulated questions will be further taken into account during the discussion on the available simulation techniques. Especially the first three criteria are decisive. The compliance of each with each listed criteria is indicated as: excellent (4), good (3), mediocre (2) or bad (1). Finally, a weighted decision table will be used to evaluate, rate and compare the alternative simulation methods on these formulated selection criteria.

2-5-2 Generating Conditional Realizations: Two Possible Approaches

Before presenting a brief overview of some of the most used geostatistical simulation techniques, it is useful to point out the existence of two separate classes. The subdivision is based on the way how the method generates a conditional realization.

A conditional realization is a realization that reproduces the experimentally measured values at the actual data locations. In mining applications, conditioning is important to ensure that the simulated deposits and real one have the same clusters of high and low values at the same positions. In this way, characteristics which are not explicitly modeled by the random function are introduced into the realizations, creating a certain robustness (Journel and Huijbregts, 1978). It is obvious, that in order to comply with the first prerequisite (honoring the data values), all final obtained realizations have to be conditioned on the experimental data.

Generally, two possible approaches exist to perform the conditioning operation during a geostatistical simulation. On one hand, there is a group of simulation methods that directly generate conditional realizations, which reproduce the measurements at the actual data locations. This group of methods is called the ‘direct single step conditional’ simulators.

On the other hand, it is also possible to design simulation methods consisting out of two major parts: the unconditional simulation phase and the conditioning phase. The first phase generates random fields which still comply with the modeled second order moments, but who do not honor the data values. The second phase subsequently conditions the obtained random fields using following formula (Journel and Huijbregts, 1978; Fenton and Griffiths, 2008):

$$Z_c(\mathbf{x}) = Z_u(\mathbf{x}) + [Z_k(\mathbf{x}) - Z_s(\mathbf{x})], \quad (2-12)$$

where $Z_c(\mathbf{x})$ represents the conditional field. The three other components are defined as follows: a) $Z_u(\mathbf{x})$ is an unconditional simulation b) $Z_k(\mathbf{x})$ denotes the kriging estimate based

Block	DBSim		LAS		
	SGS	GSGS	DFT CMD	FFT TB	MA
Point					
	Direct Single Step Simulation		Random Field Simulation		

Figure 2-3: Overview of Simulation Methods; The figure shows whether a method is designed to simulate either block or point value and whether the method relies on an additional conditioning operation (MA: moving average, CMD: covariance matrix decomposition, DFT: discrete Fourier transform, FFT: fast Fourier transform, TB: turning bands, SGS: sequential Gaussian simulation, GSGS: generalized sequential Gaussian simulation, DBSim: direct block simulation, LAS: local average subdivision).

on known measured values at the sampling locations c) $Z_s(\mathbf{x})$ denotes the kriging estimate based on the unconditional simulated values at the sampling locations.

The conditioning operation calls for solving a kriging system at each location. However, for each location, the covariance matrix needs to be inverted only once and can subsequently be used to condition multiple realizations. Random field simulations (unconditional) are generally faster than direct single step conditional simulators. Unfortunately, the gain in speed is often negated during the additional required conditioning operation.

2-5-3 Brief Overview of some Existing Simulation Techniques

For this study, some of the most used techniques for simulating continuous variables are reviewed. An overview of the discussed methods is displayed in Figure 2-3. The figure immediately shows to which class each method belongs and whether the method is designed to simulate either node or block values.

Moving Average (MA) methods The *moving average* technique of simulating random processes is based on the expression of the process as a weighted average of underlying white noise process. Formally, for each location of interest, a realization is obtained through the calculation of an integral of the product between an incremental white noise process and a weighting function. The weighted function is determined by the second order statistics, inferred from available data (theoretical complexity = 2).

In practice, the integral is approximated as a summation, its extent and resolution directly affecting the accuracy of the resulting field. Because the sum needs to be calculated separately at each point, the MA technique can be very computationally intensive. For reasonable sized fields in two or three dimensions, it can be orders of magnitudes slower than some of the other methods to be discussed (Fenton, 1990) (computational efficiency = 2, size of simulation = 3).

Another problem, which impedes the widespread use of the method, has to do with the fact that the weighting function is difficult to find for an arbitrary covariance structure in higher dimensions. Typically, inverse Fourier transformation methods are used to numerically calculate the weighting function (easy to use = 2, easy to implement = 2).

However, it is worthwhile to mention that Journel and Huijbregts (1978) and Mantoglou and

Wilson (1981) have constructed some weighting functions corresponding to several common one dimensional covariance functions. This makes the MA method thus possibly useful to generate one dimensional random fields. In higher dimensions, the method rapidly becomes cumbersome.

Covariance Matrix Decomposition (CMD) method The *covariance matrix decomposition* technique is a very simple and straightforward method of producing a random field with prescribed covariance structure. The method can be described as follows: first, create a grid of prediction points and calculate the lag distance between all possible pairs of points. Then, construct the covariance matrix such that each entry represents the spatial correlation between the two corresponding points. A lower triangular matrix can subsequently be obtained, typically through the use of a Cholesky decomposition. Thereafter, a vector with zero mean, unit variance Gaussian random variables is created and multiplied with the lower triangular matrix. The resulting vector finally contains a collection of realized values, one for each prediction point (easy to use = 4, easy to implement = 4, theoretical complexity = 4). Although appealing in its simplicity (and accuracy), this method is only useful for very small fields. For fields with a medium size, the Cholesky decomposition of the covariance matrix would be time consuming and prone to considerable round-off error (Fenton and Griffiths, 2008) (computational efficiency = 1, size of simulation = 1) .

Discrete Fourier Transform (DFT) method The Fourier transform methods are based on the generation of unconditional simulations in the spectral domain using the Fourier integral. The first and second order moments of the random process are typically described by the physically meaningful spectral density function, which makes this class of simulation methods intuitively attractive (Yaglom, 1962).

In the case of the *discrete Fourier transform*, the Fourier integral is approximated by a sum, which needs to be evaluated separately at each location of interest. To start the simulation, it is necessary to convert a spatially discretized covariance field, via a discrete Fourier transformation, into a collection of discrete spectral densities. These densities are subsequently being used to calculate the variances of the random amplitudes. With these variances and knowing that the amplitudes have a zero mean, it is possible to generate for each amplitude spectrum a realization. These realization are finally being multiplied with their corresponding evaluated cosines and inserted into the above mentioned sum. This eventually results in a realization of an orebody attribute at a single location (easy to use = 1, easy to implement = 1, theoretical complexity = 1).

The DFT method is potentially accurate, but might be computationally slow for reasonable field sizes and typical spectral density functions. Also the discretization of the covariance function and the spectral density function are rather complicated (Fenton, 1990) (computational efficiency = 3, size of simulation = 2).

Fast Fourier Transform (FFT) method If both the space and frequency are discretized into a series of equispaced points, then the *fast Fourier transform* method can be used to approximate the above mentioned Fourier integral (Cooley and Turkey, 1965). The FFT simulation method is much more computationally efficient and therefore also a lot faster than the DFT method (computational efficiency = 4).

The gain in efficiency, however, does not come without any price. In order to apply the FFT method, two additional assumptions need to be made. First, the random process is assumed to be mean zero, real and discrete. Secondly, the process is also considered to be periodic, with the same number of spatial and frequency discretization points.

As far as the one dimensional simulation is concerned, the process proceeds as follows: first, the variances of the independent normally distributed amplitudes are calculated using a discrete approximation to the Wiener-Khinchine relationships. These relationships are based on the one-sided spectral representation of the covariance function. Then, some amplitude values are randomly drawn from the modeled distributions. Next, symmetry relationships are used to produce the remaining Fourier coefficients. After all the generated amplitudes are inserted, the Fourier transform can be finally evaluated. The entire described procedure eventually results into a single point realization. Fenton and Griffiths (2008) state that the simulation procedure for two and three dimensions is almost identical to the one followed in the one dimensional case. The only difference is that the Fourier coefficients need to be generated randomly over the half plane or volume rather than the half line (easy to use = 1, easy to implement = 1).

A alternative approach can be used to generate realizations in higher dimensions. Borgman et al. (1984) present a similar relatively fast FFT method to simulate geological variables in the three dimensional frequency domain. The method simulates parallel two dimensional layers, while accounting for the cross-spectral density. When the simulation is finished, the fast Fourier transform is applied to project the simulated values back in their original domain. Perhaps, the greatest disadvantage of the entire FFT method is that, regardless of the target covariance, the actual covariance of the process will always be symmetric about the midpoint of the simulated field (Fenton, 1990). This deficiency can be overcome by generating a field twice as long as required in each coordinate direction and keeping only the first quadrant of the field (size of simulation = 2).

Another problem relates to the close relationship between the spatial and spectral discretization. Considerable care must be taken when initially defining the discretization of the spatial field, since this directly impacts the discretization of the covariance function which is again related to the approximation of the target spectral density function. The size and the lay-out of the simulation grid can thus not be chosen independently. Often, the user would like to have the freedom to easily consider other geometries or spectral density functions. Without careful thought and research, the FFT approach can easily yield highly erroneous results (resolution/support = 1).

Turning Bands (TB) Method The *turning bands* method, as originally introduced by Matheron (1973), approximates a simulation of a random process in the two or higher dimensional space by using a sequence of uniformly distributed one dimensional lines. On each line an independent simulation is performed based on an intermediate one dimensional covariance function. This intermediate covariance function can be derived analytically from an integral equation involving the two or three dimensional covariance function (Mantoglou and Wilson, 1981) (easy to use = 1).

The TB method actually proceeds as follows: first, create a collection of lines crossing the domain and generate a one-dimensional process on each line. Then, orthogonally project the field point of interest on all available lines and collect the corresponding process value. Finally, add these values up and divide by the number of lines to obtain a single simulated orebody

attribute (Journel and Huijbregts, 1978) (theoretical complexity = 2, easy to implement = 1).

Actually, the TB method is not a fundamental generator, since an existing simulation algorithm needs to be used to generate one dimensional process lines (FFT, MA, LAS, ...). Depending on the selected simulation technique, it might even be necessary to replace the intermediate covariance function by its equivalent spectral density function, which further complicates the method.

One of the main problems related to the use of the TB method is the streaked appearance of the realizations. The elimination of the streaks comes at the price of construction more one dimensional lines, which requires additional computation time (computational efficiency = 2, size of simulation = 1). Since the streaks are present in the field itself, this type of error is generally more serious than errors in the variance or covariance field (Fenton and Griffiths, 2008). This streaking effect, in combination with the necessity of finding an equivalent one dimensional covariance function, makes the method less attractive.

Local Average Subdivision (LAS) method The *local average subdivision* method is a fast and generally accurate method for producing realizations of a discrete local average random process. The motivation behind the method arose out of the need to properly account for the specific sample volumes on which engineering properties are measured. The method is designed in such a way that it is able to produce a field of local averages, whose statistics are consistent with the field resolution. As such, the method is ideally suited to deal with problems that require average properties over a set of elements. Changing the element size automatically results in an adjustment of the statistics (Fenton and Griffiths, 2008) (resolution/support = 4).

The construction of a LAS process essentially follows a top-down approach, in which each parent cell is subdivided into four equal sized child cells. The values of the child cell are generated in such a way *a)* that upwards averaging is preserved, *b)* that they show the correct variance according to the local average theory, and *c)* that they are properly correlated with one another. Each child value is randomly drawn from a distribution that is conditioned on the parent cells and the three other child cells. The procedure is actually an Autoregression Moving Average (ARMA) model in which the past is represented by the coarser resolution stages (theoretical complexity = 3, computational efficiency = 4).

Perhaps, the major advantage of the LAS method is that it produces fields of local average cells, whose statistics are consistent with the field resolution. This makes the method particularly useful to simulate block realization, since the spatial variability and dispersion of the average values will be inferred correctly. Another remarkable feature is that the resolution can be increased or decreased quite easily in specific subregions. This flexible block size could be very beneficial compared to other block simulation method. Finally, it is also interesting to mention that the, from all reviewed simulation methods, the LAS method is probably the easiest to use. The implementation on the other hand can be rather difficult, due to some complex circumstances that appear when simulating near boundaries or corners easy to use = 4, easy to implement = 2).

Sequential Gaussian Simulation (SGS) method The *sequential Gaussian simulation* approach is one of the most straightforward methods for simulating a random process. Each value is simulated sequentially according to its conditional normally distributed cdf, which

must be determined at each location to be simulated. The conditioning data comprise all the original data and all previously simulated values within the neighborhood of the point being simulated (Webster and Oliver, 2007).

The sequential simulation approach proceeds as follows: first, the coordinates of all the simulation points need to be specified. Then, the sequence in which points will be visited during the simulation needs to be determined. Subsequently, simple kriging is used to calculate the conditional mean and the conditional variance at each node. These computed values make it possible to construct the conditional cdf. Finally, a value is drawn at random and added to the data set. The whole procedure is repeated at the next node in the sequential path (Goovaerts, 1997) (easy to use = 3, easy to implement = 4, theoretical complexity = 4). The algorithm is fairly simple and very easy to use. However, due to the necessity of solving kriging systems at each location, the method is computationally intensive (Benndorf, 2005) (computational efficiency = 2, data management = 3). Fortunately some recent developments considerable improved the computation speed (Dimitrakopoulos and Luo, 2004).

Generalized Sequential Gaussian Simulation (GSGS) method A considerable improvement is introduced by the *generalized sequential Gaussian simulation* method. This technique considers the simultaneous realizations in groups of close nodes, instead of the node by node sequential simulation process of the SGS method (Dimitrakopoulos and Luo, 2004). The sequential simulation of group of nodes takes advantage of the shared neighborhood for the closest nodes, which leads to an improvement of the computation efficiency (computational efficiency = 4, theoretical complexity = 4, data management = 3). The method showed to be successful in generating conditional simulation fast and accurate. Because of time efficiency, a well-implemented GSGS algorithm has to potential of meeting the requirements of the industry (Benndorf, 2005) (size of simulation = 4).

Direct Block Simulation (DBSim) Godoy (2003) even further extended the GSGS method to optimize the data management and to directly simulate on a block scale (data management = 4, resolution/support = 4). When the interest lies in the simulation of large grid, millions of nodes need to be retained as conditioning information, which of course requires an awful lot of memory. Moreover, the computational efficiency is considerably reduced due to the increasing search times for constructing the neighborhood.

To address this problem, a block support simulation algorithm was developed, using some features of the GSGS method. Once the internal nodes of a certain group are simulated, they are reblocked and the individual points values are discarded. Only the calculated block value is subsequently added to the conditioning information. This operation requires the integration of additional conditioning data on a block support (computational efficiency = 4, size of simulation = 4).

2-5-4 Conclusion and Final Selection

In order to select the optimal simulation technique, a weighed decision table was constructed (Table 2-1). The ‘complex’ decision is first decomposed into some basic selection criteria, which were already discussed before. Then, a weight reflecting the importance, is assigned to each criteria. Subsequently, each selection criteria receives a score between one and four to

evaluate the alternatives. Finally, the total score for each simulation option is calculated. This entire procedure shows that the LAS method receives the highest score and will thus be investigated through the remained of the research.

Table 2-1: Selection of a simulation method using a weighted decision table; The compliance of each with each listed criteria is indicated as: excellent (4), good (3), mediocre (2) or bad (1).

	weights	MA	CMD	DFT	FFT	TB	LAS	SGS	GSGS	DBSim
1 easy to use	30	2	4	1	1	1	4	3	3	3
2 easy to implement	20	2	4	1	1	1	2	4	3	3
3 theoretical complexity	15	2	4	1	1	2	3	4	4	4
4 resolution/support	10	2	2	1	1	2	4	3	3	4
5 data management	7,5	1	1	2	2	2	4	3	3	4
6 computational efficiency	7,5	2	1	3	4	2	4	2	4	4
7 size of simulation	5	3	1	2	2	1	4	2	4	4
8 scientific contribution	5	1	1	2	2	1	4	1	2	2
Score	100	48	76	33	35	35	86	78	81	85

Conditional Simulation via Local Average Subdivision

The previous chapter stressed the importance of geostatistical simulation techniques in spatial variability problems. A concise analysis showed that the LAS method met most of the listed requirements and is most suited for the tasks at hand.

This chapter contains a more detailed explanation of the LAS method, bearing in mind the final aim to characterize the uncertainty and spatial variability at the scale of the Smallest Mineable Unit (SMU). The detailed explanation is split up into three distinct sections.

- The first section explains the most relevant principles behind the selected simulation methodology.
- The second section explicitly states the associated critical assumption and the consequent limitations.
- The third section finally discusses the necessary adjustments and extensions that needed to be developed to make the method better applicable in the context of deep sea mining.

All information regarding the LAS method is based on Fenton and Griffiths (2008); Griffiths and Fenton (2007); Fenton (1990); Fenton and Vanmarcke (1990)

3-1 The Principles Behind the Simulation Methodology

The selected simulation technique essentially consists of two major parts; the simulation phase and the conditioning phase.

First, during the simulation phase, an unconditional random field is generated with a resolution in agreement with the desired SMU scale. The core of this simulation exists out of a sequence of calculation stages, improving the coarse resolution of an initially generated random field. At each calculation stage, the current image is further discretized by increasing

Figure 3-1: Resolution improvement during a LAS simulation of a local average random process. The coarse resolution of an initially generated random field is further improved through a sequence of calculation stages. At each calculation stage, the current image is further discretized by increasing the total number of cells, while reducing their sizes and adjusting their associated values accordingly.

the total number of cells, while reducing their sizes and adjusting their associated values accordingly (Figure 3-1). The final value of each simulated pixel represents the local average of the modeled ground property. The described focusing operation eventually results in a picture (realization) of the random process, whose statistics are consistent with the desired field resolution.

Subsequently, during the second phase, a conditioning operation is performed, which ensures that each simulated random field honors the known measurement data. Although, the general concept of the conditioning operation has already been explained before, it still requires some additional attention.

The final design of the simulation methodology will prove to be very useful in mining applications, since the focus in this industry lies on the distribution of average grades in SMUs. It is obvious that the resolution, which is related to the desired cell size, can be tuned to the SMU of interest.

3-1-1 Local Average Theory and Change of Support

In mining geostatistics, the simulated or measured material volume with its particular size, shape and orientation is known as the support of the sample. Before explaining the details of the LAS simulation technique, it is important to realize that the sampling size has a considerable influence on the distribution of the obtained values. This so-called support effect definitely needs to be incorporated in the proposed simulation model.

Apparently, attribute values measured on rock chip sized samples can be highly variable, whereas those obtained on the basis of large truckloads would demonstrate a more consistent behavior. After all, the mixing of high and low grade values in large volumes result in a less erratic sampling set (Isaaks and Srivastava, 1989). The dispersion variance of simulated block values (B) thus needs to be remarkably smaller than the variance of point samples (P). This conclusion is also captured by Krige's relation (Webster and Oliver, 2007; Journel and Huijbregts, 1978):

$$\sigma_{P \in A}^2 = \sigma_{P \in B}^2 + \sigma_{B \in A}^2. \quad (3-1)$$

In other words, the dispersion variance of points in the area of interest (A) equals the sum of the dispersion variance of points in a block and the dispersion variance of blocks in the area. Eq. (3-1) indicates that the dispersion variance of blocks is always smaller than the dispersion

variance of points (Webster and Oliver, 2007).

Remark that, for not normally distributed data, the change of support is not only reflected by a reduction in variance, but also by an increase in symmetry. This additional consequence can be explained by the Central Limit Theorem, which states that the sample means tend towards a normal distribution regardless of the initial distribution of the samples (Isaaks and Srivastava, 1989).

Since the LAS simulation technique is build upon a normally distributed random function model, the reduction in dispersion variance only is sufficient to characterize the effect of a change in support. Later on, an indirect correction technique will be introduced to incorporate also non Gaussian measurement values.

The practical importance of previous conclusions should not be underestimated. Especially in deep sea mining applications, the financial impact of disregarding the change of support effect can be substantive. In most cases, an ore/waste classification is made on relatively large blocks, which will differ in size from the sampling support. This results in a difference between the total volume of the mining units exceeding a given economic cut-off grade, and the total volume of points fulfilling the same condition. The quantity of recovered metal generally decreases with an increase in support. From an economic point of view, it is really vital to characterize the discrepancy between the support of the samples and the intended support of the simulations.

Appendix C-1 mathematically captures this change of support effect for a local averaging operation of a one dimensional random process. The derivation finally results in the so called variance reduction formula, which expresses the variance of a one dimensional local average in function of the cell size and the covariance model. Appendix C-2 subsequently introduces a numerical approximation that can be used to calculate the variance of a two dimensional local average in an efficient and accurate way.

A detailed study of the resulting formulas reveals, that local averaging both smooths the process and reduces its variance. The amount of variance reduction turns out to be proportional to the high frequency content present in the initial point signal. In other words, the variance reduction increases with an increased independence in the random field (Griffiths and Fenton, 2007). The mean of a local average, however, remains constant with a change of support.

To summarize, the two main effects of local averaging are the reduction of the variance in the signal and the damping of the high frequency components. The entire benefit of the LAS method lies in the fact that this important change of support effect is explicitly modeled. This is in strong contrast with some of the ‘conventional’ methods, which implicitly model the change of support effort through a reblocking of point simulations.

3-1-2 One Dimensional Local Average Subdivision

Appendix C-3 introduces the concepts and mathematical formulas to generate a one dimensional random process. Although final interest lies in the generation of a two dimensional random field, the elaboration of the one dimensional simulation methodology is not entirely useless. A quick review of the mentioned section might place the subsequent discussion into perspective and improves the understanding of the simulation principles. Therefore, the reader is recommended to quickly read through the one dimensional theory.

Figure 3-2: Cell subdivision in a two dimensional local average subdivision procedure; The procedure exists of a sequence of calculation stage, during which each cell is further subdivided into four equal sized cells. The large squares are the parent cells, whereas the small cells represent the child cells. The connection between the colored squares indicate the three different types of block-to-block covariances that need to be calculated (after Fenton and Vanmarcke (1990)).

3-1-3 Two dimensional Local Average Subdivision

The acquired understanding of the local average theory (and the one dimensional LAS methodology) can now be used to extend the previously introduced simulation principles (Figure 3-1) to a proper formulated technique. In two dimensions, the procedure essentially exists of a sequence of calculation stages, during which each cell is further being subdivided into four equal sized cells. This subdivision procedure continues until the desired field resolution is obtained.

Figure 3-2 illustrates a general case, where a certain parent cell, Z_5^i , is further subdivided into four associated child cells Z_j^{i+1} , $j = 1, 2, 3, 4$. Although eventually, the resolution of all the parent cells needs to be improved, only one subdivision is showed in order to retain a certain simplicity.

Since the single element notation would result in a cluttered collection of equations, the decision was made to use a vector-matrix notation. As such, the four child variables of parent Z_5^i are stored in one column vector, $\mathbf{Z}^{i+1} = \{Z_1^{i+1}, Z_2^{i+1}, Z_3^{i+1}, Z_4^{i+1}\}$, while the nine conditioning parent values are stored in another one, $\mathbf{Z}^i = \{Z_1^i, \dots, Z_9^i\}$.

The unknown local averages of the four child cells are subsequently modeled as normally distributed random variables with a mean and variance selected to satisfy the following four criteria:

1. All four local variables average to the parent value, such that the global average remains constant throughout the sequence of subdivisions.
2. All four local variables show a correct variance according to the local average theory.
3. All four local variables are properly correlated with one another.

4. All four local variables are properly correlated with the neighboring child values across the parent boundaries.

At this point, the goal is to obtain one realization from each of the four modeled probabilistic mechanisms. The required local averages are eventually generated as the sum of a mean vector and a random component vector. The entire subdivision procedure is essentially an ARMA model, in which the ‘past’ is represented by the previous coarser resolution stages. The model is formulated as:

$$\mathbf{Z}^{i+1} = \mathbf{A}^T \mathbf{Z}^i + \mathbf{L}\mathbf{U}. \quad (3-2)$$

The first term on the right hand side, $\mathbf{A}^T \mathbf{Z}$, defines the best unbiased linear estimates for the four child cells. Each element in the resulting 4×1 mean vector is calculated as a linear combination of the direct and adjacent generated parent values. To reduce computation time and memory usage, while preserving an acceptable level of accuracy, only nine conditioning parent values are extracted from a 3×3 neighborhood. This implies that the matrix $\mathbf{A}$ must contain four columns, each filled with a specific set of nine weights. The four sets of weights, each used to calculate the mean of the corresponding cell, are obtained from the following system of simple kriging equations:

$$\mathbf{A} = \mathbf{R}^{-1} \mathbf{S}, \quad (3-3)$$

where $\mathbf{R}$ and $\mathbf{S}$ are both matrices filled with covariances calculated over two specified regions, which do not necessarily need to be of equal size. These covariances, between all the different two dimensional domains of the child and parent cells, are defined as complex four double integrals. Appendix C-2 however discusses an efficient and accurate way to approximate numerically the necessary covariances by Gaussian quadrature summations (Fenton and Griffiths, 2008).

The matrix $\mathbf{R}$ contains the covariances between all possible pairs of parent cells, located within the 3×3 neighborhood (Figure 3-2, orange color). Evidently, the elements on the diagonal of this 9×9 covariance matrix just express the unconditional variance of a local average with a size equal to that of the parent cell. The matrix $\mathbf{R}$ is defined as:

$$\mathbf{R} = \mathbf{E} [\mathbf{Z}^i \mathbf{Z}^{iT}]. \quad (3-4)$$

Each element in the matrix $\mathbf{S}$, on the other hand, expresses the relation between one of the nine parent cells and one of the four child cells. The k^{th} column of this 9×4 matrix thus contains the nine covariances between a particular child cell and all its parent cells in the 3×3 neighborhood (Figure 3-2, purple color). The matrix $\mathbf{S}$ is defined as:

$$\mathbf{S} = \mathbf{E} [\mathbf{Z}^i \mathbf{Z}^{i+1T}]. \quad (3-5)$$

Once both matrices $\mathbf{R}$ and $\mathbf{S}$ are constructed, it becomes fairly straightforward to calculate the necessary simple kriging weights and evaluate the first component of Eq. (3-2).

The second term on the right hand side of Eq. (3-2), $\mathbf{L}\mathbf{U}$, defines a generated realization from a four dimensional multivariate zero mean Gaussian distribution. The second order moments of this distribution are determined by the conditional covariances between all possible pairs of child cells. The necessary covariance matrix is defined as:

$$\mathbf{L}\mathbf{L}^T = \mathbf{B} - \mathbf{S}^T \mathbf{A}, \quad (3-6)$$

where the matrix $\mathbf{B}$ contains the unconditional covariances between all possible pairs of child cells (Figure 3-2, green color). Evidently, the elements on the diagonal of this 4×4 covariance matrix just express the unconditional variance of the child cells. The matrix $\mathbf{B}$ is defined as:

$$\mathbf{B} = \mathbf{E} \left[\mathbf{Z}^{i+1} \mathbf{Z}^{i+1T} \right]. \quad (3-7)$$

Eq. (3-6) and Eq. (3-2) hint that the random vector can be easily generated using the CMD simulation method. A Cholesky decomposition of the conditional covariance matrix results in the lower triangular matrix, $\mathbf{L}$, which is subsequently being used in a linear combination with four independently drawn elements from $N(0, 1)$. The resulting 4×1 vector, $\mathbf{LU}$, contains the random components which have to be added to the mean vector, $\mathbf{A}^T \mathbf{Z}^i$, to obtain a realization for each of the four child cells.

Notice that the upwards averaging over the different calculation stages is preserved, since the matrix on the right hand side of Eq. (3-6) is only of rank 3. This implies that $L_{44} = 0$ and thus that $Z_4^{i+1} = 4Z_5^i - Z_1^{i+1} - Z_2^{i+1} - Z_3^{i+1}$. Due to this condition, one of the child cells will be explicitly determined once the other three are known (Fenton, 1990).

Before proceeding, it is good to stress the computational efficiency of the LAS simulation method. In contrast to other kriging related simulation techniques, the LAS method only requires a solution for a handful of simple kriging systems. Due to the cleverly designed child-parent configuration and the fact that the covariances are independent of absolute positions, **the matrices $\mathbf{A}^T$ and $\mathbf{L}$ need to be constructed only once during a single subdivision stage**. The obtained matrices are subsequently being used multiple times, to evaluate Eq. (3-2) for each parent that needs to be subdivided. Once a certain stage is completed, the domain is further being subdivided, the matrices $\mathbf{A}^T$ and $\mathbf{L}$ are constructed, and the entire process is repeated.

The main improvement in efficiency results from the fact that most of the computation effort goes into the matrix multiplications of Eq. (3-2), which is not an intensive operation and can be performed very fast.

3-1-4 Conditioning of the Random Field

This subsection seeks to explain how each simulated variability field, $Z_u(\mathbf{x})$, eventually has to be conditioned on measurement data. To accomplish this, the simulated collection of local average cells needs to be partitioned into two categories: (1) $\mathbf{x}_\alpha$, $\alpha = 1, \dots, n_k$ being those cells for which measurement data is available, and at which the random field takes on the deterministic values $z(\mathbf{x}_\alpha)$; and (2) $\mathbf{x}_\eta$, $\eta = 1, \dots, N - n_k$ being those locations at which random realizations need to be generated from the modeled conditional distributions.

What actually needs to happen is an elevation, some sort of local translation, of the simulated zero mean variability field towards the known information. This procedure is called the conditioning of the random field, and the necessary conceptual formula was already introduced before (Ref. Eq. (2-12)). It is fairly straightforward to prove that this conditioning procedure returns for each known cell the correct measured value (see Appendix C-4).

In order to obtain a conditional simulated value at one of the $N - n_k$ unsampled cells, Eq. (2-12) needs to be evaluated. The first component of this conditioning formula is typically generated by an unconditional simulation technique, in this case the LAS method. The difference between the second and third component, on the other hand, can be obtained at

once, using following efficient formula (Fenton and Griffiths, 2008; Journel and Huijbregts, 1978):

$$Z_k(\mathbf{x}_\eta) - Z_s(\mathbf{x}_\eta) = \sum_{\alpha=1}^{n_k} \beta_\alpha [z(\mathbf{x}_\alpha) - Z_u(\mathbf{x}_\alpha)]. \quad (3-8)$$

Please refer to Appendix C-4 for more information regarding the conditioning operation and the derivation of the necessary equations. The weights, β_α , necessary to model this linear combination of differences, are obtained from the following set of simple kriging equations:

$$\beta = \mathbf{C}^{-1} \mathbf{b}, \quad (3-9)$$

where β is the vector containing the n_k weighting coefficients β_α , and where $\mathbf{C}$ is the $n_k \times n_k$ matrix containing the covariances between the sampled local average cells.

The matrix $\mathbf{C}$ is thus constructed with the following elements:

$$C_{\alpha\beta} = \text{Cov} [Z(\mathbf{x}_\alpha), Z(\mathbf{x}_\beta)], \quad (3-10)$$

for $\alpha, \beta = 1, \dots, n_k$. The vector $\mathbf{b}$, on the other hand contains the n_k covariances between the current local average cell of interest, $\mathbf{x}_\eta$, and the available sample cells: $\mathbf{x}_\eta$, and the available measurements from the sampled cells:

$$b_\alpha = \text{Cov} [Z(\mathbf{x}_\eta), Z(\mathbf{x}_\alpha)], \quad (3-11)$$

for $\alpha = 1, \dots, n_k$. The covariances of Eq. (3-10) and Eq. (3-11) can be calculated efficiently with the Gaussian quadrature numerical approximation, described in Appendix C-2.

Since the matrix $\mathbf{C}$ is dependent on the covariances between known cells, its inverse actually needs to be calculated only once. The resulting inverted matrix can then repeatedly being used to evaluate Eq. (3-9) and Eq. (3-8) at each simulated cell that needs to be conditioned. Moreover, since both matrices $\mathbf{C}$ and $\mathbf{b}$ are not related with any measured or simulated values, they can be reused over several simulation loops, improving the computation speed of the entire conditioning operation.

The code files, governing the conditioning of the random field, were adjusted to perform the operation in a more efficient and optimal way. Clearly stated, the entire set of simulated values, corresponding to a particular cell, is conditioned at once. The advantage of this procedure is that the memory requirements are significantly reduced, since it is not necessary anymore to store the vector $\mathbf{b}$ for all the different cells (or to recalculate it each time again).

3-2 Critical Assumption and Limitations

This section lists and discusses some of the critical assumptions and limitations that are inherently connected with the application of the selected simulation methodology. Notice that, the selected LAS method is merely a simulation technique to generate equally probable realizations from a specific random function model.

First, the most important assumptions, regarding the application of a random function model, are highlighted. The previously introduced stationarity and ergodicity assumptions are recalled, since they still apply and are independent of the selected simulation technique. Thereafter, a short discussion is devoted to the assumptions that are induced by the design of

the numerical simulation algorithm. These calculation assumptions are intended to facilitate a computational efficient generation of a random field realization. Therefore, the concerning assumptions can also be more or less considered as numerical approximations.

Finally, the section concludes with an identification of some critical limitations regarding the conditioning of the simulated random field. Although the unconditional LAS random field simulation is ideally suited to deal with the discrepancy between the support of the samples and the intended support of the simulations, the conditioning operation unfortunately nullifies much of the gained advantages.

3-2-1 Assumptions of the Random Function Model

Bear in mind that, in order to infer the model parameters necessary to characterize the spatial variation, the stationarity and ergodicity assumptions were introduced (Section 2-2). Since the LAS method is based on a random function model, it is important to realize that both assumptions still apply.

- *Stationarity assumption:* The first and second order moments of the random field are constant and are invariant to the translation by an arbitrary displacement vector.
- *Ergodicity assumption:* The mean and covariance structure of a random process can be reliably inferred from the few samples collected in a single realized field of finite extent.

Both assumptions were extensively discussed in the previous chapter.

Before proceeding, it should be stressed that the random function model is generally depicted by a multivariate **Gaussian distribution** (Eq. (2-5)). Theoretically, it is possible to perform a LAS simulation for a non-Gaussian random field. However, reasons of computational efficiency and software complexity dictates otherwise.

3-2-2 Assumptions Induced by the LAS Simulation Algorithm

The design of the numerical LAS algorithm further introduces a set of assumptions. The conditioning child-parent configuration is essentially a numerical approximation and is subject to following suppositions:

- *Correlation Assumption 1:* One of the four criteria, for the conditional distribution of a local average, stated that a generated child value needs to be properly correlated with other child values across the parent boundaries. This requirement can not be explicitly modeled, unlike the other three, since the neighboring child values are not always known during the construction of a new subdivision stage. Therefore, the LAS method implicitly assumes that the correlation across parent boundaries can be accomplished through a proper conditioning on the adjacent parent cells. Notice that, on the other hand, the correlation within the parent boundaries is explicitly and correctly modeled, without the need of any additional assumption.
- *Correlation Assumption 2:* In order to increase the computational efficiency and to reduce the memory requirements, the range of parent cells, on which to condition each

distribution, needs to be limited to a restricted neighborhood. The current implemented simulation algorithm resorts to a square 3×3 neighborhood, including one direct and eight adjacent parent cells (Figure 3-2).

- *Correlation Assumption 3*: As mentioned before, the covariances between the different two dimensional domains of child and/or parent cells are approximated by an efficient and accurate Gaussian quadrature summation. This approximation is further discussed in Appendix C-2.
- *Correlation Assumption 4*: The covariance vector, $\mathbf{b}$, which needs to be constructed for each cell during the conditioning operation (Eq. (3-11)), can be computed in an even more efficient way. This vector typically contains the covariances between a single simulation cells and all the sampled cells. Instead of calculating n_k four dimensional summations, the required block-to-block covariances can be further approximated by just a single multiplication of the point model and the variance reduction factor (Webster and Oliver, 2007). Currently, such an approximation is optional and only applicable for the construction of the covariance vector. The reduction in computation time is enormous, while the lost accuracy remains acceptable. Some test showed that the differences per cell stayed within one percent.

Moreover, the conditioning child-parent configuration also imposes some practical obstructions to the use of the LAS method. Due to the typical subdivision strategy, one is not entirely free to choose the final lattice dimensions arbitrarily. Generally, the values of an initial $k_1 \times k_2$ lattice of large cells are simulated using the previous discussed CMD method. The resolution of this very coarse picture is subsequently being refined through m LAS subdivision stages. The final field dimensions can be expressed as:

$$N_1 = k_1 \cdot 2^m, \quad (3-12)$$

$$N_2 = k_2 \cdot 2^m, \quad (3-13)$$

where N_1 and N_2 are respectively the number of cells in the x and y direction. The design of the initial field lattice and the selection of the number of subdivision thus need to be made with the eventual desired field configuration in mind.

3-2-3 Limitations induced by the Conditioning Operation

As mentioned before, the LAS method is ideally suited to deal with the discrepancy between the support of the samples and the intended support of the simulations. Unfortunately, the required conditioning operation has not been developed to meet the same goals, and is therefore responsible for certain limitations that might jeopardize the use of the proposed simulation framework in the context of deep sea mining.

The major problem is related to the support of the measurement data. The previously discussed conditioning operation assumes that the known measurement values have a similar support as the simulated local averages. In reality, the volumes of the samples are much smaller than the intended support of the simulations. Ignoring this observation would entail a serious violation of Krige's relation with all its consequences (Eq. (3-1)).

Figure 3-3: Problem definition and possible solution regarding conditioning limitations; The complex nature of the conditioning problem is indicated by the difference between the number of sampling points on which to condition and the number of unconditionally simulated cells located over these sampling points (left). The combination of simulating on a finer resolution, reblocking within the parent boundaries and performing a support adjustment on the input data results in a computational efficient and rather straightforward solution to overcome the problem regarding the support discrepancy (right).

One could argue, that the introduction of an additional point-to-block kriging system, would resolve the problem regarding the support discrepancy. The left hand side of Figure 3-3, however, illustrates the complex nature of the proposed solution. One of the first things to realize, regarding the evaluation of Eq. (3-8) and Eq. (2-12), is that the number of sampling points on which to condition might differ substantially from the number of unconditionally simulated blocks (overlying the same sampling points). The possible occurrence of such a situation has some serious impact on the design of the conditioning algorithm. Following list gives an overview of the necessary adjustments, that need to be made, in order to incorporate a point-block covariance computation.

- Instead of a single covariance matrix, two different matrices need to be constructed. One contains the covariances between all possible pairs of sampling points, the other contains the covariances between local average cells located over the sampling points. The illustrated example indicates that both matrices do not need to be of equal size. Both matrices, again, need to be calculated only once and can be reused multiple times.
- Instead of a single covariance vector, two different covariance vectors need to be constructed for each specific cell. One contains the covariances between the cell of interest and all the sampling points, the other contains the covariances between the cell of interest and other cells located over the sampling points. The illustrated example indicates that both vectors do not need to be of equal size. Notice that only the construction of the first vector requires the calculation of point-block covariances.
- Consequently, it is necessary to calculate two sets of simple kriging weights, for each simulated cell that needs to be conditioned. One set of weights is used to calculate the BLUE based on the measurement data, the other is used to calculate the BLUE based on the unconditionally simulated cells located over the sampling points. Finally their difference is evaluated and used to compute Eq. (2-12), which results in a conditioned value for the specific cell of interest.

On second thoughts, the implementation of this solution strategy would result in a conditioning algorithm that is considerably slower. The approximation of the covariances between equal sized blocks can not be used that frequently anymore (fourth correlation assumption). Instead, the computation of the point-block covariances require each time again the full evaluation of a Gaussian quadrature summation. This, in addition with the second parallel conditioning sequence, considerable increase the computation time.

Furthermore, the most undesirable effect has not been discussed yet. It is important to realize that, once a certain cell contains one or more sampling points, the final simulation will always return the same value, i.e. a block kriged estimate based on measurement data. This is definitely unwanted, since one of the main reasons to apply geostatistical simulations was to characterize the uncertainty on a block scale.

The next section proposes a computational efficient and rather straightforward solution to overcome the problem regarding the support discrepancy.

3-3 Necessary Adjustments and Extensions

The previous discussion identified two major limitations that might further obstruct the practical application of the LAS simulation methodology.

First, many deposit variables in mining applications tend to show an asymmetric distribution. This is bad news, since the Gaussian random function model imposes a normal distribution constraint on the input data. This constraint, however, can be easily eliminated through the implementation of a normal score transformation. The main advantage of this selected transformation technique lies in the fact that a single method can be used to deal with multiple types of distributions. This introduces some flexibility into the simulation framework.

Second, the implemented Gauss Markov covariance model is not capable of describing the spatial continuity features which generally occurs in (mining) geostatistics. Instead, a nested spherical model was implemented to correctly capture the observed variability, especially at short distances.

Third, generally deposit variables are measured on material volumes that are very small compared to the intended support which is used during the decision making. Nevertheless the conditioning algorithm expects that the input data is collected on a support that is similar to that of the final simulated cells. This discrepancy and the related problems were extensively discussed before.

The previous outlined problem can be overcome by a combination of simulating on a finer resolution, reblocking within the parent boundaries and performing a support adjustment on the input data. The right hand side of Figure 3-3 illustrates the proposed computational efficient and rather straightforward solution.

The proposed solution package will be further integrated in a simulation framework that is systematic, robust and easy to follow. The resulting practical process is specifically designed for the simulation of deep sea deposits and contains a clear documentation on several calibration and validation steps, thereby safeguarding the quality of the entire geostatistical simulation analysis. The next chapter is entirely devoted to these improvements embedded in the developed practical process.

A Practical Process for the Simulation of Deep Sea Deposits: The Theoretical Principles

To further increase the applicability of the selected simulation technique, it becomes necessary to provide solutions to overcome the identified undesirable restraints, i.e. (a) the necessity of normally distributed data, (b) the limitations of the implemented Gauss Markov model, and (c) the requirement that the data support needs to be similar to the pixel size of the simulated images.

Moreover, the proposed solution package will be integrated in a general simulation procedure that is systematic, robust and easy to follow. The generally obscured calibration, validation and adjustment steps, which are vital to the success of a simulation-based geostatistical analysis, are now explicitly documented in order to safeguard the quality of the entire analysis. Such a procedure reduces the likelihood of costly mistakes and insures the validity of the simulation results (Nowak and Verly, 2004).

4-1 Roadmap of a simulation-based geostatistical analysis

The previously described LAS simulation technique constitutes the core of the developed practical process. The other required operations are ordered and embedded in a visual roadmap (Figure 4-1). The displayed roadmap indicates that the entire practical process is mainly constructed from three types of components.

1. **The Main Operation** (7): all numbered and located in a vertical sequential path.
2. **The ‘Data and Area Statistics’ Boxes** (6): all including the same set of tools used to summarize the results of the operations in a few global and spatial statistics.
3. **The Validation Procedures** (2): all designed to control and safeguard the representativeness of the simulation results.

Figure 4-1: Roadmap of the developed practical process for the simulation of deep sea deposits; The practical process consists of seven main operations, each numbered and located in a vertical sequential path. The practical process also repeatedly relies on a particular set of tools gathered in the 'Data and Area Statistics' boxes. This set of tools is able to summarize the results of a selected operation into a few global and spatial summary statistics. The validation procedures, indicated as dark gray diamonds, extensively use these calculated summary statistics to approve the simulation results. Notice that the support and projection space changes during the analysis; both are printed in the light blue ovals. The internal part of the simulation framework, which operates in the normal score domain, is indicated by thick lined boxes. The displayed framework contains all the necessary calibration, simulation, validation and adjustment steps vital to the success of each simulation-based geostatistical analysis.

Please notice that projection space and data support changes during the analysis. Both properties are displayed in the light blue ovals.

4-1-1 Main Operations

The seven main operations, displayed in Figure 4-1, constitute the spine of the developed simulation framework. Following lists gives a concise overview.

- *Declustering*; A declustering algorithm is implemented to remove the bias due to preferential sampling and to make the data more representative for the area under investigation.
Input: original scale point values; *Output*: original scale point values + declustering weights.
- *Normal Score Transformation*; The normal distribution constraint on the original data is removed through the application of a normal score transformation. The operation transforms the original point data into normal scores, using the declustered cumulative distribution.
Input: original scale point values; *Output*: normal score point values.
- *Model Covariance*; During the third operation, a covariance model is calibrated in order to capture the spatial variability of the deposit.
Input: normal score point values; *Output*: analytical covariance model.
- *Data Support Adjustment*; An affine correction is performed to overcome the second constraint, which required that the support of the data was similar to the pixel size of the simulated images. The point normal scores are transformed into block normal scores.
Input: normal score point values; *Output*: normal score block values.
- *Ore Body Simulation*; The ore body simulation constitutes the core of the geostatistical analysis framework. The combination of random field simulation and conditioning operation results in the generation of equally probable representations of the in-situ ore body variability.
Input: normal score block values (conditioning data); *Output*: simulated normal score block values.
- *Back Transformation*; The back transformation is responsible for the representation of the generated ore body models in the original space. The simulated normal score values are translated into the required block concentrations.
Notice that also a back transformation box is connected with the data support adjustment operation (Figure 4-1). The output of this back transformation is merely indented to support the validation procedure.
Input: simulated normal score block values; *Output*: simulated original scale block values (operation in sequential path). **OR**
Input: normal score block values; *Output*: original scale block values (box connected with data support adjustment operation).

- *Transfer Functions*; Economical and technical transfer functions are used to translate the simulated geological and geotechnical models into financial and operational performance indicators. *Input*: original scale simulated block values; *Output*: performance indicators.

4-1-2 Data and Area Statistics

The inference process, embedded in each DAS box, aims at estimating the same relevant summary statistics for each inserted collection of data values.

Each box thus contains the same set of tools to compare the effects and the results of the relevant operations. As the name suggests, the ‘Data and Area Statistics’ can be subdivided into two categories: a group of global and a group of spatial statistics.

The group of global statistics focuses on the characterization of the general data properties (mean, standard deviation, inter quantile range, histogram and cumulative distribution). The conventional formulas are adapted in such a way, that they are able to correct the bias due to preferential sampling.

The spatial or area statistics, on the other hand, aim at the estimation of the spatial correlation structures inherently present within the area under investigation. The semi-variance, the covariance and the correlation function belong to the set of tools frequently used to achieve this goal. Also the calculation of the lag mean, lag variance and some moving window statistics turns out to be particularly useful.

The theoretical principles behind the calculation of the mentioned global and spatial statistics are referred to, later on in this chapter and in Appendix D.

4-1-3 Validation Procedures

The goal of the embedded validation procedures is to safeguard the quality and representativeness of the simulation results. Without some kind of validation procedure, even the most sophisticated technique will likely render unreliable results.

The proposed practical process contains two main validation steps, displayed in Figure 4-1 as dark gray diamonds.

1. The first validation procedure is called the inner validation check and is responsible for the quality control within the most internal part of the simulation framework. This internal part, which operates solely in the normal score domain, is indicated by the thick lined boxes (Figure 4-1).
The simulation results are approved if the deviations between the calculated ‘Data and Area Statistics’ of the simulated normal score block values and those of the conditioning normal score block values stay within acceptable limits. If the deviations, on the other hand, turn out to be too large, certain calibration parameters in the third or fourth operation have to be adjusted and the simulations have to be ran again.
2. The second validation procedure is called the blanket validation check and is responsible for the approval of the complete set of geostatistical simulation results. This blanket validation procedure checks if the final simulation results (after back transformation)

Figure 4-2: Histogram and cumulative frequency distribution of Walker Lake exploration data; The proportion of data values falling within each class is expressed by the heights of the bars (left). The cumulative frequency indicates the proportion of the exploration data that does not exceeds the corresponding concentration value (right).

reproduce the global and spatial statistics of the conditioning block data.

If the deviations between the series of calculated statistics are too large, certain calibration parameters have to be adjusted and the simulations have to be ran again. The calibration of the model is thus an iterative process that continues until the simulation statistics are acceptable.

The designed encapsulated validation procedure makes it easier to detect where in the process the calibration parameters have to be adjusted.

4-2 Analyzing Results: Data and Area statistics

This section contains a brief explanation on the calculation of several relevant global and spatial statistics. The location dependent declustering weights are used in each DAS box to convert certain sample statistics into representative population parameters. A more extensive discussion, illustrating the mathematical principles behind the adjusted formulas, can be found in Appendix D.

4-2-1 Data or Global Statistics

Arithmetic average and variance The mean is calculated as an equal-weighted linear combination of all available data values. To correct for possible bias, each constant weight in the conventional formula needs to be replaced by a variable location dependent declustering weight.

By analogy, the variance formula can also be adjusted. More information about the calculation of these declustered statistics can be consulted in Appendix D-1.

The histogram or frequency distribution The distribution of continuous values is generally depicted by a histogram, with the range of data values discretized into a specific number of classes, each of equal width. In order to construct a representative histogram for a biased data set, a correction is applied which removes the negative effects of preferential sampling.

The frequency of each class is recalculated as the sum of the relevant declustering weights. The proportion of data points falling within each class is finally expressed by the height of the drawn bars. The left hand side of Figure 4-2 shows an example.

cumulative distribution function A graph of cumulative frequencies is another convenient tool to capture and summarize certain interesting characteristics of the data distribution (Figure 4-2). Cumulative frequencies are very useful to study the proportion of a population, in this case a deposit, that has an attribute value above or below a critical threshold value. The resulting cumulative distribution function (cdf) does not only provide an answer to the threshold question, but also gives an insight into the spread of the underlying population values. Moreover, the well-known quantiles are ideally suited to report confidence intervals. In the case of preferential sampling, it is vital that the declustering weights are taken into account to calculate correct and representative cumulative frequencies. More information about the technical details can be found in Appendix D-1.

4-2-2 Area or Spatial Statistics

None of the previously introduced descriptive global estimates are able to capture spatial features. Hence, this subsection is devoted to some techniques which are frequently used to investigate the spatial continuity and variability of exploration data.

Generally, an h-scattergram is used to explore the spatial relations hidden in the measurement data. Such a scattergram displays all possible data pairs which are separated by a specific displacement vector $\mathbf{h}$. When the selected data points get more dissimilar, the displayed cloud gets fatter and more diffuse. The inflation of the cloud with increasing separation distance reflects a decreasing correlation.

In order to summarize the visual conclusions into quantitative information, some techniques need to be introduced which are able to capture the fatness of the cloud by a single correlation related number. Such a computation can be repeated for each class of displacement vectors and finally results in a graph that visualizes the spatial relation as a function of distance and direction.

Semi-variance The experimental semi-variance expresses how much the data points drift away from the bisector in an h-scattergram (Appendix D-2). The semi-variance for a particular displacement vector $\mathbf{h}$ can be calculated as:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{x}_\alpha) - z(\mathbf{x}_{\alpha+\mathbf{h}})]^2, \quad (4-1)$$

where $N(\mathbf{h})$ represents the number of data points plotted in the corresponding h-scattergram. An ordered set of calculated semi-variances, displayed as a function of their corresponding lag distance, is called an experimental semi-variogram. In general, the semi-variance increases with decreasing dissimilarity. The left hand side of Figure 4-3 shows two examples of such semi-variograms.

Figure 4-3: Spatial description of Walker Lake data set. Experimental semi-variance function or semi-variogram showing the calculated semi-variances as a function of the lag distance (left). Experimental (auto)covariance function showing the calculated covariances as a function of the lag distance (right). A green colored symbol (+) indicates a calculated statistic which was obtained from regularly spaced gridded data points. The red symbols (x), on the other hand, correspond with calculated statistics of the complete exploration data set, also including clustered data points.

Covariance An alternative index to express the fatness of the cloud or the spatial continuity is the covariance. The covariance between $N(\mathbf{h})$ data pairs separated by a vector $\mathbf{h}$ is computed as follows:

$$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} z(\mathbf{x}_{\alpha}) \cdot z(\mathbf{x}_{\alpha+\mathbf{h}}) - m_{-\mathbf{h}} \cdot m_{+\mathbf{h}}, \quad (4-2)$$

where $m_{-\mathbf{h}}$ and $m_{+\mathbf{h}}$ are the means of the tail and head values. An ordered set of computed covariances is called an experimental (auto)covariance function. In general, the covariance coefficient decreases with decreasing similarity. Eventually the covariance coefficient reaches a value of zero, indicating that after a certain distance, the data becomes uncorrelated. The right hand side of Figure 4-3 shows an example of two series of calculated experimental covariance functions.

Correlation Finally, also a unit free measure of similarity is available. This measure is derived from the well known correlation coefficient and is actually a standardized form of the covariance function. The resulting function expresses the relation between the correlation coefficient of an h-scattergram and the displacement vector $\mathbf{h}$. A correlation coefficient of one corresponds with perfect correlation, while a value of zero denotes that the set of values is spatially uncorrelated. An ordered set of computed correlation coefficients for different lags is called the experimental (auto)correlation function or correlogram and is calculated using following formula:

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{\sqrt{\sigma_{-\mathbf{h}}^2 \cdot \sigma_{+\mathbf{h}}^2}}, \quad (4-3)$$

where $\sigma_{-\mathbf{h}}^2$ and $\sigma_{+\mathbf{h}}^2$ are the variance of the tail and head values.

4-3 The Spine of the Developed Practical Process: The Main Operations

The following discussion further elaborates on the backbone of the developed practical process (Figure 4-1). A concise description of the tasks of each main operation was already presented before. This section, on the other hand, is devoted to a more detailed explanation on the technical principles.

4-3-1 Declustering

The first operation aims to remove the undesirable effects caused by preferential sampling. Data locations, that are neither regularly nor randomly distributed over the study area, likely originate from a biased exploration campaign. It is not uncommon that highly mineralized zones are sampled more intensively than the remainder of the investigated area.

The phenomena, though, is not necessarily connected with a thoughtless exploration strategy. The sampling scheme, however, is rarely designed with the objective of reducing its negative influence on the statistical analysis. In certain circumstances, the surveyors even deliberately choose for clustered data locations (Goovaerts, 1997).

Great attention should be paid to preferential sampling, since it may skew the results of any data analysis. A possible correction procedure is to only retain the regularly spaced or gridded data points.

Whenever data sparsity does not allow one to ignore the clustered data points, other procedures have to be implemented. Intuitively, data points in densely sampled areas should receive less weight than those in sparsely sampled areas. The cell declustering technique is able to perform such an adjustment.

During the cell declustering approach, the entire study area $\mathcal{A}$ is divided into rectangular cells. Each sample receives a weight that is inversely proportional to the number of samples that are located within the corresponding cell. The approach calls for counting the number B of cells containing at least one sample and the number n_b of datum locations falling within each cell b . Finally, each datum location, $\mathbf{x}_\alpha$ receives a declustering weight λ_α that equals $1/(B \cdot n_b)$ (Figure 4-4). Such a declustering weight generally gives less importance to clustered data points, since the cell in which they are located also contain several other samples.

The cell declustering algorithm is fairly easy to implement into computer software. The parameter calibration of this approach might provide some difficulties, but is generally considered to be straightforward if the initial sampling phase is based on a pseudo-regular grid (Goovaerts, 1997; Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978). A more detailed account on preferential sampling and declustering algorithms is presented in Appendix E.

Since the cdf plays a key role in the remainder of the geostatistical analysis, it is important to estimate it right. In order to avoid errors that go undetected, a quality control methodology was implemented in the declustering operation. Figure 4-5 gives a clear overview of the sequence of actions that need to be completed. The gray shaded diamonds indicate the positions and concepts of the two inserted quality checks.

The first quality check supports the decision of running the declustering algorithm. The second quality check, on the other hand, is designed to validate the declustering operation.

Figure 4-4: Cell Declustering; The declustering weights are inversely proportional to the product of the number of cells containing at least one sample and the number of samples within the cell. n indicates the number of samples within each cell, while w indicates the declustering weight of each sample located in that cell. Clustered samples generally receive lower weights because the cells in which they are located will also contain several other samples. To obtain optimal declustering weights following parameters have to be calibrated correctly: cell width, cell height, origin and orientation of the grid.

Once the declustering operation is performed, the computed weights are stored into the central memory of developed software. In this way, they are accessible from all the DAS boxes where representative global and spatial statistics are computed.

4-3-2 Normal Score Transformation

The normal score transformation is designed to remove the normal distribution constraint on the original data. Since many variables in earth sciences tend to show an asymmetric distribution, such a transformation is definitely not just a luxury. Moreover, the application of a normal score transformation introduces some flexibility regarding the distribution of the input data into the simulation procedure.

In contrast with other techniques, this type of transformation also does not suffer from the restriction that the input data needs to be strictly positive. The main advantage of the normal score transformation lies in the fact that a single technique can be used to deal with any kind of distribution.

The normal score transformation is easy to understand and consists of a short sequence of relatively straightforward steps, which can be executed numerically or graphically.

As a starting point, two tables or graphs need to be constructed. The first table contains the ordered set of original measurement values together with their corresponding calculated cumulative frequencies. A visual representation of such a table is displayed in the left hand side of Figure 4-6.

The second table, on the other hand, contains a numerical approximation of the standard normal cdf. The cumulative frequencies of the listed normal scores are computed numerically as integrals of the standard normal distribution. The right hand side of Figure 4-6 gives a visual representation of the tabulated relationship.

Once the necessary preparations are completed, the actual normal score transformation can

Figure 4-5: Quality control in declustering operation; The gray shaded diamonds indicate the positions of two quality checks inserted into the declustering operation. The first quality check supports the decision of running the declustering algorithm. The second quality check evaluates the performance of the declustering operation. If the results of the second check are not satisfactory, the declustering model needs to be recalibrated.

Figure 4-6: Normal Score Transformation; The cumulative frequency indicates the proportion (y-axis) of the data that does not exceeds the corresponding value (x-axis). The red curve (x) represents the cdf of the complete equal-weighted data set, while the green (+) and orange (o) curves represent respectively the cdf of the gridded equal-weighted and declustered complete data set. The crosses indicate the specific p -quantiles, ranging from 0.05, with steps of 0.05 to 0.95 (left). The cumulative frequency indicates the proportion $\Phi(y)$ (y-axis) of the standard normal distribution that does not exceeds the corresponding normal score y (x-axis). The stars represents specific p -quantiles, ranging from 0.05, with steps of 0.05 to 0.95 (right). The actual normal score transformation procedure is depicted by the black arrows.

start. All the necessary steps are visualized in Figure 4-6 by the black arrows. The procedure starts with locating a given measurement value on the x-axis of the first constructed graph. Subsequently, a vertical line is drawn to project the measurement value onto the cdf curve, which yields the corresponding cumulative frequency. This cumulative frequency is then matched with the proper quantile of the standard normal distribution (horizontal line). A vertical projection finally yield the required normal score.

The discussed projections each have a straightforward numerical counterpart, which mainly exists of locating the boundary values in a table and subsequently perform an interpolation.

4-3-3 Characterization of Spatial Variation

During the third operation, a theoretical model is fitted to the computed experimental semi-variances or covariances in order to capture the spatial variability of the area under investigation.

The necessary experimental values are calculated in the DAS box, which is connected with the normal score transformation (Figure 4-1). To ensure that the covariance can be defined for any possible lag distance and to smooth out sample fluctuations, it is necessary to fit a continuous analytical model. The challenge is to select the proper parameters, so that the theoretical model describes the observed spatial variability/continuity in the best possible way.

The Gauss Markov model, which was implemented initially, is not capable of describing the spatial continuity features, which generally occur in (mining) geostatistics. The existing LAS simulation code had to be adjusted in order to incorporate a nested spherical model (isotropic). Such a model is better suited to capture the behavior of the experimental semi-variogram at short distances (nugget component and the initial linear course).

Figure 4-7: Omnidirectional Characterization of spatial variation in normal score domain; Omnidirectional experimental semi-variances and fitted theoretical isotropic model (left). Omnidirectional experimental covariances and fitted theoretical isotropic model (right). The red (x) and the green (+) curve represent respectively the spatial characterization of the equal-weighted complete and gridded normal score data. The blue solid line represents the fitted theoretical model (Nested spherical covariance structure). The model parameters are also indicated in both graphs.

Typically, a graphical interactive procedure is used to calibrate the four model parameters, $\{a_0, a_1, \sigma_0, \sigma_1\}$, of the following analytical semi-variance model:

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{if } h = 0 \\ \sigma_0 \cdot \left(\frac{3}{2} \frac{h}{a_0} - \frac{1}{2} \frac{h^3}{a_0^3}\right) + \sigma_1 \cdot \left(\frac{3}{2} \frac{h}{a_1} - \frac{1}{2} \frac{h^3}{a_1^3}\right) & \text{if } 0 > h > a_0 \\ \sigma_0 + \sigma_1 \cdot \left(\frac{3}{2} \frac{h}{a_1} - \frac{1}{2} \frac{h^3}{a_1^3}\right) & \text{if } a_0 \geq h > a_1 \\ \sigma_0 + \sigma_1 & \text{if } h \geq a_1. \end{cases} \quad (4-4)$$

The necessary theoretical covariance function is subsequently recovered from the calibrated semi-variance function, see e.g. Eq. (A-7).

Figure 4-7 shows an example of a successful isotropic calibration procedure. At the same time, the figure illustrates how some initial parameter values are obtained.

- *Nugget*, σ_0 : The nugget variance is usually determined by extrapolating the behavior of the first few semi-variance values to the vertical axis. This component arises from measurement errors and spatial variation at distances smaller than the shortest sampling interval.
- *Sill*, $\sigma_0 + \sigma_1$: The sill, on the other hand, focuses on the characterization of the spatial variation at larger distances. The sill is more or less defined as the asymptotic value of the experimental semi-variance function. Once the sill is obtained, the second variance component, σ_1 , can be calculated as the difference between the sill and the nugget. This parameter can also be estimated by extrapolating the first few experimental covariances to the vertical axis (right hand side Figure 4-7).
- *Range*, a_1 : The range is defined as the distance from which two points in space become uncorrelated. Any existing correlation fades away as soon as the semi-variance reaches its maximum value or as the covariance approaches zero.

Figure 4-8: Effect of changing support; The left hand side shows a histogram of the 73728 points values which were collected on a regular $1m$ by $1m$ grid in the Walker Lake area. The middle graph show a histogram of the 288 re-blocked block values with a support of $16 \times 16 m^2$. The right graph shows the calculated cumulative frequencies of the point (x) and block ($\square$) values. The symbols (x or $\square$) indicate the specific p -quantiles, ranging from 0.05, with steps of 0.05, to 0.95.

- *Short range, a_0 :* The nugget component is actually modeled as an extremely short ranged spherical model with a sill σ_0 and a range a_0 . Test simulations have shown that this particular implementation strategy is superior to the theoretical discontinuous nugget component. The implemented numerical approximations are apparently not capable to deal with a nugget discontinuity. The short range of the nugget structure should be set to a fraction of the normal range ($1/20$).

These initial values are subsequently inserted into the theoretical covariance or semi-variance structure, which is then superimposed on its corresponding set of calculated experimental values. The parameter values can be further optimized through an iterative process of adjusting, plotting and comparing. Often, only a few iterations are necessary to end up with a fairly good calibrated covariance structure.

4-3-4 Support Adjustment

The support adjustment algorithm is embedded into the practical simulation process to overcome the third constraint, imposed by the previously discussed simulation technique. This constraint states that the conditioning data should have a support that is similar to the pixel size of the simulated images. To overcome this restriction, the point normal scores are transformed into block normal scores using an affine correction.

Following example illustrates the practical implications of the change of support concept. The global statistics of an exhaustive collection of point values are compared with those of the corresponding $16 \times 16 m^2$ block values (Figure 4-8b and Table 4-1).

As the support of the data increases, one notices that the frequency of extreme low and high values decreases. This is due to the fact that extreme values are diluted by less extreme values as we average them over larger areas. The dilution of extreme values generally has the effect of reducing the variance of the data, making the distribution more symmetric. The standard deviation, the coefficient of variation, the interquartile range and the difference between the mean and median all decrease as the support of the data increases (Isaaks and Srivastava,

Table 4-1: Change of support, example comparing change in global statistics; The example is based on an exhaustive collection of 73728 point values, which were measured on a $1m$ by $1m$ grid. Subsequently, the exhaustive set of data points was reblocked into 288 block values with a support of $16 \times 16 m^2$. A comparison of the calculated statistics enhances the comprehension of the effects of a change in support.

Data Support	Point	Block ($16 \times 16 m^2$)
No. data points	73728	288
Mean [ppm]	283.3	283.3
Median [ppm]	227.3	247.0
Standard deviation [ppm]	251.7	206.1
Variance [ppm ²]	63351	42478
Coefficient of variation [/]	0.888	0.727
Min. [ppm]	0	0
0.05 Quantile [ppm]	0	9.9
0.25 Quantile [ppm]	72.7	132.5
0.75 Quantile [ppm]	436.2	404.6
0.95 Quantile [ppm]	767.8	651.0
Max. [ppm]	1631.2	1137.5

1989). Though the support has a direct effect on the spread of the data values and the shape of its distribution, it does not affect the mean.

Comparing the calculated cumulative frequencies immediately reveals the practical importance of the support effect. If the decisions are made on a block support, while the analysis is performed on a point support, serious errors can occur. In such a situation, the proportion of the area that exceeds an economic cut-off grade might be severely over- or underestimated (overestimated if the cut-off grade is higher than 300 ppm, underestimated otherwise). Also the variability of the ore grade in blocks will be vastly overestimated if point values are used during the analysis.

This basic example clearly illustrates the problem of the discrepancy between the support of the samples and the intended support of the estimated values. If both supports turn out to be different, it is necessary to apply a correction. The applied correction technique needs to be designed in such a way that it reduces the variance, while keeping the mean constant.

The rate, at which the spread of the distribution decreases and the symmetry increases, is dependent on the spatial continuity of the area under investigation. For homogeneous areas, where the values are rather pervasive, the reduction of spread and asymmetry will occur more rapidly than for heterogeneous areas in which there is a strong zonation (Isaaks and Srivastava, 1989). This link between the effect of a change in support and the spatial continuity makes the covariance model the most suitable tool to estimate the reduction in variance.

Unfortunately the covariance model alone does not capture all the relevant spatial characteristics of a data set. Another important consideration is the degree of connectivity between extreme values. This feature is known as entropy and determines how rapidly the distribution will symmetrize as the support increases. Earth sciences data sets tend to have low entropy and thus will symmetrize slower.

The required mathematical procedure is merely a transformation that maps quantiles of one

Figure 4-9: The affine correction plotted on a $q - q$ curve; The blue line ($\square$) shows Eq. (4-5) on a $q - q$ plot with mean zero and variance one. The orange line ($\circ$) shows the same equation but with variance adjustment factor of one. The symbols ($\circ$ or $\square$) indicate the specific p -quantiles, ranging from 0.05, with steps of 0.05, to 0.95.

distribution to those of another using their $q - q$ plot. Unfortunately, for the support effect problem, the intended distribution of block values is not available and therefore it would not be possible to calculate the corresponding quantiles. The quantiles of the block distribution however are essential for the construction of the $q - q$ plot or transformation curve. To overcome this problem, certain assumptions have to be made regarding the shape of the $q - q$ plot.

Since the support adjustment algorithm is placed behind the normal score transformation, the symmetry issue and the related entropy characterization is not relevant anymore. The starting point of the support adjustment step is a standard normal distribution, which as we all know is “perfectly” symmetric. Therefore, it is only necessary to implement a technique that can reduce the variance of the distribution without changing its mean value. This goal can be easily accomplished by squeezing all the values closer to the mean. In mathematical terms, this operation can be expressed as:

$$q_b = \sqrt{f} \cdot (q_p - m) + m, \quad (4-5)$$

where m represents the mean of both distributions, q_p are the point quantiles and q_b are the block quantiles. The factor f is called the variance adjustment factor (Eq. (C-4), Appendix C-1) and is defined in such a way that, if the variance of the point distribution is σ^2 , then the variance of the block distribution is calculated as $f \cdot \sigma^2$. Since the input of this step consists of point normal score values, Eq. (4-5) can be further simplified ($m = 0$).

Figure 4-9 clearly illustrates the effect of an affine correction performed on normal score values. The straight line is a reflection of the desire to preserve the shape of the original distribution and implicitly assumes that the symmetry of the distribution does not increase with increasing support. The blue line with square markers shows how the affine correction alters the given distribution. The orange line with round markers represents the base case corresponding to a variance adjustment factor of one. The graph clearly illustrates that, especially the more extreme values are squeezed closer towards the mean in an absolute sense.

4-3-5 Ore Body Simulation

The ore body simulation forms the core of the developed practical process. The principles behind the very efficient LAS technique were extensively discussed before and will not be repeated here. This subsection just reviews how this particular component is linked with the other main operations.

First, the normal score point covariance model needs to be inserted into the LAS simulation algorithm. This algorithm performs several unconditional random field simulations. The resulting normal score variability fields are subsequently conditioned on the support adjusted normal score input data. This entire operation results in a collection of conditional simulated equally probable normal score deposit models.

4-3-6 Back Transformation

The back transformation is responsible for the representation of the generated ore body models in the original data space. This final process translates the simulated block normal score values into the required block values. The transformation is merely the inverse of the previously discussed normal score transformation and is thus also based on the declustered cumulative frequency distribution.

4-3-7 Transfer Functions

The transfer function is problem specific and is thus not an inherent part of the developed simulation framework. The flexible and well structured Fortran code, however, is developed in such a way to allow easy implementation of any transfer function. Some example transfer functions will be discussed later on in this thesis.

4-4 Implementation and Software Language

This section shortly discusses the implementation strategy and software language of the developed practical process for the simulation of Deep Sea Deposits. The entire framework is developed as a hybrid FORTRAN-MATLAB software suite.

The seven main operations all consist of specific mathematical actions. The corresponding calculations are all programmed in FORTRAN.

The existing FORTRAN code for the unconditional LAS simulation and the subsequent conditioning operation was adjusted in order to cope with a nested spherical model and normal score values.

The other six remaining operations were programmed from scratch by the author and integrated in existing (adjusted) FORTRAN code.

The six DAS boxes are developed in MATLAB to facilitate the visualization of the calculated global and spatial statistics. The separate MATLAB packages are able to read the output of the main operations (text files), calculate the required statistics, visualize the results and export the constructed figures and tables.

The validation procedures are not automated and just rely on the visual inspection of the generated figures by the user. A detailed discussion on this topic follows later on in this thesis.

A Practical Process for the Simulation of Deep Sea Deposits: Walker Lake Case Study

Previous chapter presented the developed practical process for the simulation of deep sea deposits. The proposed procedure was designed to safeguard the validity and representativity of the simulation results and to reduce the likelihood of costly mistakes. Moreover, the constructed roadmap and accompanying documentation facilitate a clear communication regarding the calibration, validation and adjustment steps vital to the success of a simulation-based geostatistical analysis.

In order to further increase the reliability and the value of the developed framework, a test simulation is performed in an almost completely known environment. The targets of such an analysis are two-fold. On one hand, the test simulation is useful to illustrate the application of the practical process. On the other hand, the completely known environment is ideally suited to validate the performance of the developed practical process and to demonstrate that the framework meets the previously formulated objectives.

Both mentioned topics are discussed in the course of this chapter. The further explanation can essentially be subdivided into three main part.

- *Part 1: Walker Lake data set*; The first section contains a brief explanation regarding the origin of the investigated database. A short description of the exploration data and the completely known environment are given.
- *Part 2: The practical process*; The second section serves as an example regarding the necessary calibration, validation and adjustment steps embedded in the developed practical process.
- *Part 3: Assessment of the performance*; The third and final sections describes the validation of the simulation results in the completely known environment.

Figure 5-1: Color coded plot of Walker Lake exploration data and completely known field; The primary exploration points were positioned on a 20 m by 20 m grid. This initial sampling grid resulted in 195 regularly spaced data points (a). The subsequent exploration cruises focused on the characterization and the delineation of the highly mineralized zones. This strategy added 275 clustered data points to the initial exploration data (b). The known field is represented by a collection of 73728 data points located on a 1 m × 1 m grid (c).

5-1 Description of Exploration Data and Completely Known Environment

5-1-1 Origin of the Data Set, Design of the Simulation Grid and Performed Data Adjustments

In order to safeguard the reproducibility of following geostatistical analysis, a brief explanation on the origin of the input data is given. The investigated collection of spatial measurements is actually a modified version of the generally available and well known Walker Lake data set. This data set has been introduced in 1989 by Isaaks and Srivastava and has ever since been used in multiple academic case studies (Isaaks and Srivastava, 1989). The Walker Lake database distinguishes itself from many others in that it also contains the almost completely known reality. This known environment makes the database particularly useful to investigate the performance of newly developed simulation and estimation techniques.

The spatial extent of the original data is taken into account during the design of the optimal simulation mesh. Due to the applied LAS methodology and the corresponding sequence of cell quartering, the dimensions of the simulation grid have to comply with certain restrictions (see Section 3-2).

Suitable grid dimensions are obtained by choosing for 5 subdivisions and selecting an initial 8 × 9 lattice containing 72 local averages, each with an area of 32 × 32 m². This specific set of parameters finally results in a mesh that is 256 m wide and 288 m high. The cleverly chosen mesh configuration subsequently enables the reblocking, within the respective parent boundaries, of the 73,728 1 × 1 m² cells into 4608 4 × 4 m² blocks, 1152 8 × 8 m² blocks or even 288 16 × 16 m² blocks.

The original data is subsequently projected onto a local Cartesian coordinate system, whose origin coincides with that of the simulation mesh. Generally, the sampling locations just need to be translated so that all the selected points nicely fit within the designed physical

simulation boundaries. Notice that the modifications merely change the absolute locations of the data points and thus have no impact at all on further calculations and possible conclusion. The coordinates of all the data points, contained in the original Walker Lake database, are shifted 1.5 *m* westwards and 5.5 *m* southwards. Any measurement points, located outside the simulation boundaries, are eventually removed from the database. The results are shown in Figure 5-1. A detailed account on the design of the simulation grid and the performed coordinate transformation can be found in Appendix F.

5-1-2 Exploration Data and Known Environment

The spatial configuration of the displayed exploration data immediately reveals the underlying sampling strategy. The exploration measurements are regarded to be a combination of regularly spaced and clustered data points.

Apparently, an initial exploration campaign focused on the collection of gridded measurements along the cell midpoints of a 20 *m* by 20 *m* sampling grid, which origin is located at (-2;-4.5). This first stage eventually resulted in 195 regularly spaced data points, which are displayed in Figure 5-1a and listed in Table G-2.

The subsequent exploration stages were intended to further investigate the profitable subareas. During these phases, the sampling tools were deployed to intensively sample and delineate the rich zones in the deposit. This approach added 275 clustered and preferentially sampled measurements to the initial exploration database. The resulting 470 data points are shown in Figure 5-1b and tabulated in Table G-1.

A preliminary examination of both plots does not reveal any clear indication of erroneously recorded spatial coordinates. Furthermore, there is also no reason to believe that certain measured concentration would be erroneous. The observed spatial pattern seems to be authentic.

The data density of the gridded exploration data makes it rather complicated to investigate the spatial continuity features and delineation of the rich zones with the naked eye. Figure 5-1b, on the other hand, already sheds some more light on the existing spatial structures. This plot discloses the existence of about three large and six smaller high mineralized zones. The small zones can be more or less delineated by circles with a diameter between 30 *m* and 50 *m*. The large zones, in contrast, are delineated by ellipses with a minor axis of about 40 *m* - 60 *m* and a major axis of 60 *m* - 120 *m*.

Previous observations and presumptions can now be compared with the completely known reality, displayed in Figure 5-1c. This known environment is essentially an exhaustively sampled field and exists out of 73,728 point measurements. These measurements were collected at the cell midpoints of a 1 *m* by 1 *m* grid with origin (0,0).

Before proceeding, it might be interesting to mention that the spatial extent of the large horizontal structure, ranging between 100 *m* - 200 *m* EW and 50 *m* - 100 *m* NS, was not entirely exposed during the exploration campaign.

Figure 5-2: Quality control of declustering operation; Q-Q plot of the distribution of the complete data versus that of the gridded data, before and after the declustering operation. Each blue cross corresponds with the position of a specific p-quantile pair $\{q_p^g, q_p^{g+c}\}$ (a) or $\{q_p^g, q_p^d\}$ (b). In total 18 pairs are calculated, ranging from a proportion p of 0.05 with steps of 0.05 to 0.95. Similar distributions should plot on the 45° line. The figure clearly illustrates that declustering improves the global estimates for the underlying population. Notice that also the quantiles of the completely known field are displayed.

5-2 Application of the Practical Process; An Example Regarding the Necessary Calibration, Validation and Adjustment Steps

5-2-1 Declustering and Exploratory Data Analysis

In order to properly perform a geostatistical simulation, it is important to understand the relevant statistical characteristics of the investigated deposit. Such an understanding is generally acquired during an Exploratory Data Analysis (EDA). To ensure that the calculated global and spatial statistics are representative for the area of interest, a declustering operation needs to be performed first.

The sampling documentation already suggested the need for such an operation. This necessity can be formally confirmed by the investigation of a Q-Q plot, which compares the probability distributions of the complete and gridded data by plotting their quantiles against each other. Figure 5-2a makes it immediately apparent that preferential sampling causes an overestimation of almost every population quantile. Of course, this conclusion is strongly dependent on the assumption that the gridded data set yields statistics that are representative for the underlying population. The displayed quantiles, obtained from the known environment, however substantiate this assumption.

The declustering operation subsequently aims to assign location dependent weights to each measurement in the complete data set (gridded + clustered), so that the calculated statistics become more representative for the studied area. Figure 5-2b illustrates that the implemented cell declustering algorithm is able to neutralize the negative effects of preferential sampling. In this particular situation, the calibration parameters, such as the cell size and the grid origin, naturally result from the previously discussed sampling pattern. More information, regarding the calibration of the declustering algorithm, can be found in Appendix E-2.

Eventually, the obtained declustering weights are used to transform the sample statistics of the

Figure 5-3: Histogram and cumulative distribution function of Walker Lake exploration data; Histogram of gridded, complete and declustered exploration data. The proportion of points falling within each class is expressed by the location of the symbols. Each class has a width of 50 ppm and the corresponding proportion is plotted above the midpoint of the class (a). The cumulative frequency indicates the proportion (y-axis) of the exploration data that does not exceed the corresponding concentration (x-axis). The dashed lines connect the cumulative frequencies of all available sorted data points. The symbols further indicate the specific p-quantiles ranging from 0.05 with steps of 0.05 to 0.95 (b). Both plots also contain the corresponding statistics calculated from the completely known field.

complete data into reliable and representative population parameters. A second procedure, which only retains the regularly spaced data points, is executed in parallel as an additional quality check.

Figure 5-3 clearly illustrates that serious mistakes can be made if the bias, due to preferential sampling, is not removed.

The histogram, based on the equal weighted complete data (x), underestimates the occurrence of low mineralized zones (0 ppm - 350 ppm), whereas it seriously overestimates the occurrence of high mineralized zones (600 ppm - 1000 ppm). The proportion of the deposit with a medium concentration, however, is estimated more or less correctly. The red curve also gives the incorrect impression that the investigated population can be evenly divided among a certain series of classes. This observation is inherently connected with the relatively high estimated

Table 5-1: Global statistics of gridded, complete and declustered Walker Lake exploration data. Besides the three sets of computed global statistics, the table also contains the statistics calculated from the completely known field. The results indicate that generally the declustering algorithm works fine (*gridded and clustered data).

Walker Lake Exploration Data	gridded	complete*	declustered	known field
No. data points	195	470	470	73,728
Equal Weighted?	yes	yes	no	yes
Mean [ppm]	275.3	435.4	288.5	292.5
Median (50% Quantile) [ppm]	209.1	424.0	234.3	227.3
Variance [ppm ²]	62,356	89,752	64,568	65,476
Standard deviation [ppm]	249.7	299.6	254.1	255.9
50 % IQR	360.6	456.9	377.6	363.6
90% IQR	802.4	915.6	773.3	767.8

Figure 5-4: Experimental semi-variance function of Walker Lake exploration data; The experimental semi-variance function shows the calculated semi-variances as a function of their corresponding lag distance. The graph displays the semi-variograms calculated for gridded, complete and declustered data set. Also the semi-variogram for the completely known field is shown as a reference.

measures of spread (Table 5-1).

The green (gridded; +) and orange (declustered; o) curve, on the other hand, show that actual population is likely to be positively skewed, and that most of the concentrations in the deposit can be classified as low or medium. The proportion of points, belonging to the first class (0 ppm - 50 ppm), even turns out to be respectively 23.6% and 22.5%, compared to biased underestimated proportion of 11.9%.

Figure 5-3a also shows that both histograms, calculated from either the gridded or the declustered data, represent the known field rather well.

Instead of studying the individual proportions of certain classes, it is also instructive to have a look at the cumulative frequencies. After all, a representative cdf gives a rough indication about the percentage of the entire material volume that exceeds a critical (economic) threshold value. The cumulative distribution is also particularly useful to examine the spread and the shape of a certain distribution.

Figure 5-3b presents an interesting overview of the effects of declustering and preferential sampling on the calculated population quantiles. The red curve (x), calculated from the equal weighted complete data, indicates that about 30% of the deposit has a concentration that can be classified as high. The green (gridded; +) and orange (declustered; o) curve, on the other hand, show that the actual proportion is as low as 12%. The economic consequences, concerning the disregarding of preferentially sampled clusters, are obvious.

Again, both the gridded as well as the declustered cumulative distribution constitute an utmost acceptable approximation of the reality. Both graphs, displayed in Figure 5-3, demonstrate that the declustering algorithm works fine.

The global statistics presented in Table 5-1 confirm this observation. Generally, the results obtained from the equal weighted complete data overestimate almost every population parameter, no matter whether it concerns a measure of spread or central tendency. In contrast, the gridded and the declustered data set are very well capable of capturing the statistical behavior of the known field.

The last point of discussion, before proceeding to the normal score transformation, is related to the analysis of the spatial correlation. Figure 5-4 shows some calculated experimental

Figure 5-5: Normal score transformation of Walker Lake exploration data; The cumulative frequency indicates the proportion $\Phi(y)$ (y-axis) of the standard normal distribution that does not exceeds the corresponding normal score (x-axis). The stars represent specific p -quantiles, ranging from 0.05 with steps of 0.05 to 0.95 (a). Histogram of the gridded, declustered and known normal scores. The proportion of points falling within each class is expressed by the location of the symbols. Each class has a width of 0.2 and the corresponding proportion is plotted above the midpoint of the class (b). Direct transformation curves for the gridded, complete, declustered and known field data. The symbols indicate the same afore mentioned specific p - quantiles (c).

semi-variance functions of the Walker Lake exploration data.

One of the first things to notice is the severe overestimation of the semi-variance calculated from the complete data. This should not come as a surprise, since the point variance listed in Table 5-1 was about 30% higher than the one calculated from the gridded and declustered data. Apparently, the semi-variogram calculated from the gridded data is representative for the known field. The declustered semi-variogram, in turn, is able to capture the spatial continuity fairly well. It is remarkable though, that the declustered semi-variances appear to be more erratic.

Once a reliable experimental semi-variance function is obtained, a theoretical model can be calibrated. This specific model, however, won't be applied in any algorithm, but just turns out to be useful during the later validation of the simulation results. The blue solid line indicates the fitted theoretical model, built up from a nested structure. The primary structure models the spatial variability on a scale less than the smallest sampling interval. The so-called nugget is set to 8000 ppm^2 (remember that the nugget structure is actually model as a spherical model with an extremely short range: 2.5 m , see 4-3). The variability on the global scale is subsequently modeled by a superimposed spherical function with a sill of 58,000 ppm^2 and a range of 50 m . As shown in Figure 5-4, the resulting model is perfectly capable of describing the observed spatial correlation.

5-2-2 Normal Score Transformation

The previous EDA demonstrated that the input data is definitely not normally distributed. Since the simulation procedure relies on a Gaussian random function model, it is recommended to perform a normal score transformation.

Due to the inherent nature of the applied transformation, the resulting normal scores will always be characterized by the same standard normal distribution. Calculated quantiles are ought to describe a certain distinctive cumulative distribution, which is displayed in

Figure 5-6: Investigation of the spatial correlation of the Walker Lake exploration data in the normal score domain; Experimental semi-variance function or semi-variogram showing the calculated semi-variances as a function of the lag distance (a). Experimental (auto) covariance function showing the calculated covariances as a function of the lag distance (b). The results are calculated for the gridded, declustered and known field normal scores. Both graphs also display the fitted theoretical model (see text).

Figure 5-5a.

This theoretical consequence, however, does not make the analysis of global statistics redundant. It still remains useful to construct, for example the histogram, if only to check whether the declustering is properly taken into account. Figure 5-5b shows that histogram calculated for the declustered normal scores is almost indistinguishable from that of the gridded normal scores and the known field.

Finally, Figure 5-5c illustrates that depending on the selected transformation curve, certain mineralization pockets receive different normal scores. This observation stresses the importance of a correct characterization of the transformation curve. The graph demonstrates that both the gridded as well as the declustered approach will result in a reliable normal score transformation.

5-2-3 Model Covariance Function

Figure 5-6 displays both the experimental semi-variance and covariance function calculated from the gridded, declustered and known field normal scores.

Both graphs illustrate that the spatial correlation computed from the gridded data is representative for the known field. The declustered correlation structure, on the other hand, does not yield entirely reliable results for short lag distances (until 25 m). For larger lag distance, the declustering approach works fine.

The calculated correlation numbers clearly expose the presence of a nugget component, which amounts about 0.18 [—]. The global correlation structure, on the other hand, is captured through the superposition of a spherical model with a sill of 0.85 [—] and a range of 50 m. Figure 5-6 demonstrates that the fitted simulation model is able to describe the spatial correlation present in the input data.

Table 5-2: Calculation of the variance adjustment factor; Structure 0 represents the model that describes the short range variability or the nugget component. Structure 1, on the other hand, is superimposed to model the global spatial correlation. The range of both models remain constant with a change in support (Assumption). Structure 0 has a range of 2.5 *m*, whereas structure 1 has a range of 50 *m*.

Support	Structure 0		Structure 1		f_{tot}
	C_0	f_0	C_1	f_1	
Point	0.180	1.000	0.850	1.000	1.000
Block 1 x 1 m^2	0.131	0.726	0.838	0.986	0.940
Block 16 x 16 m^2	0.015	0.081	0.662	0.778	0.656

5-2-4 Data Support Adjustment

Before starting the ore body simulation, one last adjustment operation needs to be performed. The change in support implies an adjustment of the point statistics.

The entire data support adjustment essentially consists of three distinct actions. First, the variance reduction coefficients are calculated and the necessary variance adjustment factor is obtained. Subsequently, an affine correction is performed which transforms the point normal scores into the required block normal scores. Finally, a block correlation model is constructed, together with a representative block histogram.

In order to correctly calculate the variance adjustment factor (f_{tot}), the previously calibrated double nested spherical model needs to be dismantled into the individual constituents. The sill value of each individual structure needs to be rescaled according to the change in support. The individual variance reduction coefficients (f_0 and f_1) are calculated numerically with a Gaussian quadrature approximation (see Appendix C-2). Some computation results for a 1 x 1 m^2 block and a 16 x 16 m^2 are listed in Table 5-2.

The displayed results confirm that the variance reduction increases with an increased independence in the random field. Such an increased independence is typically reflected by a smaller value for the range. As expected, the extent of the support has also a considerable impact of the variance reduction.

Once the individual variance reduction coefficients are calculated, both rescaled models can be assembled again. The sill value of the resulting model is finally used to calculate the necessary variance adjustment factor. For a 1 x 1 m^2 block, this adjustment value amounts:

$$f_{tot} = \frac{f_0 \cdot c_0 + f_1 \cdot C_1}{C_0 + C_1} = \frac{0.726 \cdot 0.18 + 0.987 \cdot 0.85}{0.18 + 0.85} = 0.940. \quad (5-1)$$

Now that the variance adjustment factor is obtained, it becomes rather straightforward to perform the affine correction. This particular correction requires the multiplication of each individual point normal score value with the square root of the variance adjustment factor. Figure 5-7 demonstrates the effect of a change of support on a selection of calculated quantiles. Both graphs indicate that the distribution (in the normal score domain) of 1 x 1 m^2 blocks does not differ substantially from the one of the point values. The spread of the resulting block normal scores is only slightly smaller, i.e. a variance of 0.94 compared to the point variance of 1. In contrast, the figures also illustrate that an increase in support to 16 x 16 m^2 blocks has a more immediate impact on the spread of the distribution.

Figure 5-7: Data support adjustment through an affine correction; Q-Q plot displaying the calculated quantiles of the point normal scores versus those of the obtained block normal scores (a). Cumulative distribution function of the obtained block normal score values (b). The symbols in both graphs indicate specific p -quantiles, ranging from 0.05 with steps of 0.05 to 0.95. The performed affine correction is based on the variance reduction model described in Table 5-2.

The straight line in the Q-Q plot indicates that the shape of the original distribution is preserved. The affine correction essentially squeezes especially the more extreme values closer towards the mean. The impact of the support adjustment is thus only limited to a symmetric reduction of the spread.

Due to the selected block size of $1 \times 1 \text{ m}^2$, the variance reduction effect is not very pronounced at first sight. Although the impact on the histogram and cdf seems to be rather insignificant, the effect on the semi-variogram is more apparent (Figure 5-7 and Figure 5-8).

This is because it can be rather difficult to notice a variance reduction of only 6% on a histogram. A closer look at Figure 5-8a, however, reveals that the frequency of the middle four classes is always higher in the block normal score histogram compared to point equivalents. The frequencies in the tail classes, on the other hand, are smaller in block histograms. Both observations indicate that the spread of the block normal scores is somewhat smaller.

The same conclusion could have been obtained after a detailed study of Figure 5-7b. This graph shows that for example the 90% IQR of the $1 \times 1 \text{ m}^2$ block normal scores is just a few percent smaller than the one of the point values.

The statistical difference, induced by both types of support, are best observable in Figure 5-8b. This figure illustrates the experimental semi-variance functions calculated for the gridded and declustered point and block normal score values. The 6% variance reduction, related to the point-block transition, is immediately visible. To check the performance of the entire support adjustment operation, one can check if the computed theoretical block model is able to describe the course of the corresponding experimental semi-variances. The variance parameters of the nested spherical block model are listed in Table 5-7, together with those of the point model. Both necessary range parameters are the same as the ones of the point model ($a_0 = 2.5 \text{ m}$ and $a_1 = 50 \text{ m}$). Eventually, this computed block model is used further on to validate the simulation results in the normal score domain.

Figure 5-8: Histogram and experimental semi-variance function of support adjusted normal score values; Histogram of gridded and declustered point and block normal scores. The proportion of values falling within each class is expressed by the location of the symbols. Each class has a width of 0.2 (a). Experimental semi-variance function showing the calculated semi-variances as a function of the lag distance. The results are calculated for the gridded and declustered point and block normal scores. The fitted theoretical point and block models are also shown (b).

5-2-5 Ore Body Simulation

The obtained support adjusted conditioning data and covariance model are subsequently inserted into the simulation algorithm to generate an ensemble of equally probable ore body models. In order to properly capture all relevant characteristics of the calibrated random function model, 100 deposit realizations are simulated. The developed software is able to complete this task within a timespan of about 220 seconds. An example of one of the resulting conditional simulated random fields is shown in Figure 5-9b.

The directives, regarding the application of the developed practical process, suggest that the simulation results need to be validated in the normal score domain, before proceeding to the back transformation step. Although the formulated guidelines only refer to a validation procedure for the conditioned images, following text also briefly discusses some interesting features regarding the statistics of the unconditional simulated random fields.

An example of such an unconditional simulated random field is displayed in Figure 5-9a. The validation results, presented in Figure 5-10, clearly illustrate that the unconditional LAS simulation technique is not entirely capable of reproducing the second order moments of the calibrated random function model at large correlation lengths. The sill value of the different realizations fluctuates between 0.6 [-] and 1.4 [-]. A possible explanation for this rather poor reproduction is given later on in this chapter (see Section 5-3).

The generated unconditional random fields, however, do not take full advantage of the available information, for this would imply that the known blocks are honored at their corresponding locations. Figure 5-9b shows the conditioned version of Figure 5-9a. The effect of the conditioning operation reaches further than just the reproduction of measurement data. A detailed comparison between Figure 5-10 and Figure 5-11 learns that the implemented conditioning operation also further improves the reproduction of the modeled distribution function and the spatial continuity. The differences between the individual simulated histograms, cdfs and semi-variance functions become remarkably smaller after the conditioning step.

Figure 5-9: Comparison of simulated and estimated random field; Color coded plot of the unconditionally simulated $1 \times 1 \text{ m}^2$ block normal score values (a). Color coded plot of simulated $1 \times 1 \text{ m}^2$ block normal score values, after conditioning on the inserted known block values (b). Color coded plot of the estimated $1 \times 1 \text{ m}^2$ block normal score values (c). All three figure are based on the same calibrated random function model.

The greater spread in the unconditional simulated semi-variance lines (Figure 5-10) is probably caused by the introduction of the first two correlation assumptions (see Section 3-2). These assumptions state that the correlation between the child cells, across the parent boundaries, can be approximated by a proper conditioning on the adjacent parent cells in a restricted 3×3 neighborhood. This would indeed explain why each unconditional simulated semi-variance function matches rather well with the calibrated model at the short lag distance, while a substantial deviation might occur at larger distances.

An additional reason for the observed phenomena lies in the fact that the global mean of each unconditional simulation is slightly different. This global mean is determined during the realization of each initial coarse resolution 8×9 lattice and is maintained during the subsequent subdivision stages. The apparent horizontal shift between the different calculated histograms and cdfs are a direct consequence of this changing global mean.

The conditioning operation subsequently adjusts this global mean and reduces its variability, hereby eliminating some of the spread between the simulation curves. Somehow, the conditioning algorithm also manages to correct the strong deviations of the calculated semi-variances at large lag distances. This correction is presumably induced by the fact that the conditioning algorithm takes account of all available information in the entire field.

Figure 5-11 clearly illustrates that the conditioning operation contributes, to a certain extent, to the overall performance of the simulation algorithm. The simulation results almost perfectly describe the histogram and cdf of the support adjusted input data. Also the reproduction of the spatial continuity, visualized by the set of simulated semi-variance functions, will not get any better.

The discussed results support the decision to approve the internal simulation results. This final normal score validation procedure indicates that the internal part of the simulation framework meets the previously formulated objectives.

This conclusion can be further substantiated by a brief comparison between the simulation and estimation results. Figure 5-9c shows a color coded plot of estimated $1 \times 1 \text{ m}^2$ block normal score values, which were calculated with a block kriging algorithm. A comparison with

Figure 5-10: Validation of the unconditional simulation results in the normal score domain; Histogram of gridded and declustered block normal scores, superimposed on the unconditional simulation results. Each class has a width of 0.2 and the corresponding proportion is plotted above the midpoint of the class (a). Cumulative distribution function of the gridded and declustered block normal scores, superimposed on the unconditional simulation results (b). Experimental semi-variance function showing the calculated semi-variances as a function of the lag distance. The estimated and simulated fields were sampled at a 5 m by 5m grid. The calibrated spatial correlation model is also shown (c).

Figure 5-11: Validation of the conditional simulation results in the normal score domain; Histogram of gridded and declustered block normal scores, superimposed on the conditional simulation results. Each class has a width of 0.2 and the corresponding proportion is plotted above the midpoint of the class (a). Cumulative distribution function of the gridded and declustered block normal scores, superimposed on the conditional simulation results (b). Experimental semi-variance function showing the calculated semi-variances as a function of the lag distance. The estimated and simulated fields were sampled at a 5 m by 5m grid. The calibrated spatial correlation model is also shown (c).

Figure 5-9b immediately confirms that kriging tends to smooth out local details of spatial variability. Remember that a correct reproduction of the spatial variability was one of the key reasons to choose for a simulation-based approach.

Moreover, Figure 5-11c illustrates that the disadvantages of kriging are not only limited to an incorrect representation of the short range variability. The computations indicate, that in this particular case, kriging severely underestimates the overall spatial variability by about 30 %. This unrealistic representation of the spatial variability is directly related to the previously discussed smoothing effect.

This smoothing effect is responsible for the overestimation of low values and the underestimation of high values, and immediately impact the histogram and cdf, displayed respectively in Figure 5-10a and Figure 5-10b. The histogram, calculated from the kriging results, definitely has a lower variance than the ones obtained from the simulation results and the conditioning data. The proportions of the tail classes is underrated, whereas the middle classes contain too much material. A similar conclusion results from an investigation of the cdf. The steeper cumulative distribution reflects the reduction of variance and indicates that the possible occurrence of extreme values is not properly modeled.

The previous validation steps demonstrate the performance of the method up to this step and guarantee reliable results.

To finalize the previous discussion, it is important to realize that performed extensive validation procedure is just intended to ensure a correct reproduction of the calibrated random function model in the normal score domain. The validation procedure is not entirely designed to issue warnings against very poor modeled spatial correlation structures. Although there is no guarantee, it is nevertheless possible that certain procedures, embedded in the practical process, will give a clear indication whenever the fitted covariance model is inappropriate.

5-2-6 Back Transformation

The necessary back transformation is merely the inverse of the previously discussed normal score transformation. This final operation eventually converts the simulated block normal score values into the required average block concentrations.

However, remember that also a back transformation operation is connected with the data support adjustment block (Figure 4-1). This additional back transformation is intended to project the support adjusted normal score conditioning data back onto the original concentration scale. The output is necessary to calculate the validation curves of the histogram, the cdf and the semi-variance function (orange and green curves in Figure 5-12). Another useful tool, to aid the blanket validation procedure, results from a support adjustment of the previously fitted point semi-variance function (Figure 5-4). Since the selected range parameters are the same as the ones of the fitted normal score model, the variance reduction coefficients of the nested structures can also be found in Table 5-2. The resulting theoretical block model, expressing the spatial continuity in the concentration domain, is shown in Figure 5-12c. The nugget component is modeled as a spherical model with a sill of 5800 ppm^2 and a range of 2.5 m . The second structure, on the other hand, is described by a spherical model with a sill of 57,000 ppm^2 and a range of 50 m .

Subsequently, the simulation results are superimposed on the calculated validation curves and the differences are evaluated. Figure 5-13 illustrates that the back transformed simulation

Figure 5-12: Validation of the conditional simulation results in the original concentration space; Histogram of gridded and declustered block concentrations, superimposed on the back transformed simulation results. Each class has a width of 50 ppm and the corresponding proportion is plotted above the midpoint of the class (a). Cumulative distribution function of the gridded and declustered block concentrations, superimposed on the back transformed simulation results (b). Experimental semi-variance function showing the calculated semi-variances as a function of the lag distance. The estimated and simulated field were sampled at the nodes of a 5 m by 5 m grid. The calibrated and support adjusted spatial correlation model is also shown (c).

results are perfectly capable of reproducing the validation statistics. Again, the observed range of the simulated semi-variance functions more or less reflects the uncertainty in the calculated experimental sill value. It is interesting to notice, that after the back transformation, the simulated semi-variance functions even reproduce the observed hole, which appears around 250 m (Figure 5-12c). The discussed simulation results are of sufficient quality to approve the entire simulation-based geostatistical analysis.

To finalize the discussion on the practical process, the simulation results are evaluated against some estimation results, obtained via block kriging. Figure 5-13 shows a comparison between both approaches. It is immediately apparent that the transition between rich and poor zones is modeled in a rather different way. The simulation results are seemingly capable of reproducing a realistic picture of the spatial variability, whereas the estimation results, on the other hand, tend to be oversmoothed. Moreover, Figure 5-12a and Figure 5-20b both illustrate that the implemented geostatistical simulation technique is able to overcome the conditional bias problem, which is inherently connected with the conventional kriging techniques. Last but not least, the simulated deposit models are also ideally suited to quantify the geological uncertainty. Furthermore, the entire collection of realizations can be propagated through a selection of transfer functions to optimize decision under the presence of this type of technical risk.

5-3 Assessment of the Performance; Validation of the Simulation Results in a Completely Known Environment

The previous section illustrated that the newly developed simulation methodology is perfectly capable of reproducing the second order moments of the calibrated random function model. To further establish confidence, the simulation results are subsequently evaluated against the completely known environment. A comparison with their estimation counterpart eventually

Figure 5-13: Comparison of simulated and estimated random field; Color coded plot of the conditional simulated $1 \times 1 \text{ m}^2$ block concentrations (top). Color coded plot of estimated $1 \times 1 \text{ m}^2$ block concentrations (bottom). A comparison between both figures reveals that the spatial variability is reproduced in a rather different way.

Figure 5-14: Resolution Adjustment of the Simulated Deposit Models; Color coded plot of simulated concentrations in $4 \times 4 \text{ m}^2$ blocks. The results are obtained via a reblocking of the conditional simulated $1 \times 1 \text{ m}^2$ blocks (a). Color coded plot of simulated concentrations in $8 \times 8 \text{ m}^2$ blocks. The results are obtained via a reblocking of the simulated $4 \times 4 \text{ m}^2$ blocks (b). Color coded plot of simulated concentrations in $16 \times 16 \text{ m}^2$ blocks. The results are obtained via a reblocking of the simulated $8 \times 8 \text{ m}^2$ blocks (c).

benchmarks the performance and demonstrates the obtained improvements.

Furthermore, the simulated deposit models are evaluated on three different block sizes to ensure that the important concept of ‘change in support’ is properly modeled. The investigation focuses on the simulation and estimation results calculated for 4608 $4 \times 4 \text{ m}^2$, 1152 $8 \times 8 \text{ m}^2$, and 288 $16 \times 16 \text{ m}^2$ blocks, all covering the entire area of interest. In order to change the resolution of the investigated image, a different approach is required depending on the selected inference procedure.

- *Simulation:* The change in resolution is accomplished through a reblocking operation of the initially simulated 73,728 $1 \times 1 \text{ m}^2$. An example of such a procedure is shown in Figure 5-14. In this setting, reblocking refers to the calculation of the parent cell concentration, which is essentially an average of the respective child values.
- *Estimation:* In contrast, the estimated block concentrations have to be recalculated whenever the support changes. Reblocking kriging results would lead to a double smoothing effect, which is definitely undesirable.
- *Known environment:* To obtain the “known” block concentrations, the available exhaustive point measurements are averaged within the respective cell boundaries. Unfortunately, since only a single measurement is available in each $1 \times 1 \text{ m}^2$ block, it is not possible to calculate representative ‘known’ values for this particular size.

The next few sections present a detailed comparison between some global and spatial statistics calculated from the different inference results (simulations/estimation/reality). At the same time, the computed statistics also give an overview on the impact of a change in support. Subsequently, an evaluation of the estimation and simulation errors is made to illustrate that the simulated deposits have a spatial variability that is more less similar to the one of the reality. Thereafter, a short analysis demonstrates that the calculated confidence intervals

Figure 5-15: Validation of cumulative distribution function; Cumulative distribution of the estimated and known $4 \times 4 \text{ m}^2$ block concentrations, superimposed on the simulated values (a). Cumulative distribution of the estimated and known $8 \times 8 \text{ m}^2$ block concentrations, superimposed on the simulated values (b). Cumulative distribution of the estimated and known $16 \times 16 \text{ m}^2$ block concentrations, superimposed on the simulated values (c).

are reliable uncertainty indicators. Finally, some figures illustrate how the local geological uncertainty can be quantified.

5-3-1 Cumulative Distribution Function, Histogram and Experimental Semi-Variance Function

Figure 5-15 demonstrates that, concerning the reproduction of cdf, the performance of the simulation approach is generally better than the one of the estimation technique.

The graph on the left hand side shows that, upwards from the 40% quantile, all remaining quantiles of the estimated $4 \times 4 \text{ m}^2$ block concentrations are underestimated. This transition limit even further decreases to 25% and 5% for respectively $8 \times 8 \text{ m}^2$ and $16 \times 16 \text{ m}^2$ blocks. The simulated block concentrations, on the other hand all result in a more reliable calculation of the population quantiles.

It is remarkable to notice, that an increase in support, further enlarges the deviation between simulation and estimation results. The reliability and the quality of the simulation results more or less remain the same, whereas the kriging results become less accurate.

Moreover, Figure 5-15 indicates that the symmetry in the distribution of the block concentrations increases with an increase in support. This effect was to be expected and is reflected by the deformation of the cdf. The curves, connecting the first half of the calculated quantiles, generally becomes more straight with an increase in block size.

A quick investigation of Figure 5-16, obviously, leads to the same conclusion. Please notice, that the increasing spread between the displayed histograms, is due a reduction in the number of blocks used to compute the frequency of occurrence for each class. To resolve the problem, the bin width can be enlarged, but this would impede the comparison between the histograms. The histogram on the left hand side shows that kriging tends to underestimate the frequency of occurrence of $4 \times 4 \text{ m}^2$ blocks with a high concentration. Besides, the frequencies regarding the occurrence of blocks with a low concentration, is rather overestimated. Figure 5-16b and Figure 5-16c more or less confirm this observation, although it is less clear to the reduction in the number of blocks. The increasing spread unfortunately makes it impossible to formulate substantiated conclusions regarding the effect of a change in support.

Figure 5-16: Validation of histogram; Histogram of estimated and known $4 \times 4 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (a). Histogram of estimated and known $8 \times 8 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (b). Histogram of estimated and known $16 \times 16 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (c). Each class has a width of 50 ppm and the corresponding proportion is plotted above the midpoint of the class. The increase in spread is due to the reduction in the sample size and indicates that the bin width is getting too small.

Theoretically, the variance of the previously discussed histograms should decrease with an increase in support. However, due to the changing symmetry and shape, it is rather difficult to notice. In contrast, Figure 5-17 clearly indicates that the overall variance of the obtained block concentrations decreases with an increase in support.

The graph on the left hand side shows that, on average, the variance of the simulated $4 \times 4 \text{ m}^2$ block concentrations is slightly larger than that of the known field. By reducing the range parameter of the nugget structure to about 1.5 m, it was possible compensate for this effect. However, since the experimental normal score semi-variance function did not entirely suggested such a value, it was decided to retain the current calibrated model. Also the simulation results are still more representative than the estimated block concentrations, so there is no real reason to worry.

A quick analysis of Figure 5-17a and Figure 5-17b indicates that the previously observed defect in spatial variability more or less disappears with an increase in support. Moreover, both figures indicate that the simulation approach really reproduce very well the known variability field.

Another interesting observation, is that the performance of block kriging is dependent on the block size of interest. As the support increases, the deviation between the known and the estimated Spatial variability tends to decrease. Apparently, kriging is not capable of correctly describing the observed drop in the three displayed experimental semi-variance functions.

5-3-2 Evaluation of the Estimation and Simulation Error

In order to further evaluate the performance of the developed simulation methodology, a brief evaluation of the estimation and simulation errors is made.

- The estimation error is calculated for each block as the difference between the known

Figure 5-17: Validation of experimental semi-variance function; Experimental semi-variance function of estimated and known $4 \times 4 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (a). Experimental semi-variance function of the estimated and known $8 \times 8 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (b). Experimental semi-variance function of the estimated and known $16 \times 16 \text{ m}^2$ block concentrations, superimposed on the calculated simulation values (c).

value and the estimated concentration.

$$E_{\text{estimation}} = \hat{Z}(\mathbf{x}) - z(\mathbf{x}) \quad (5-2)$$

The collection of estimation errors is shown by the red curves in Figure 5-18.

- The simulation error is calculated for each block as the difference between the simulated and estimated block concentrations.

$$E_{\text{simulation}} = Z_i(\mathbf{x}) - \hat{Z}(\mathbf{x}) \quad (5-3)$$

The collection of “simulation” errors is subsequently superimposed on the estimation errors to investigate the similarities between their behavior. Similarities are desirable, since this would express that characteristics of the simulated deposit models are equivalent to those of the reality. Figure 5-18 indicates that the obtained results are acceptable. The spread of the simulation errors, however, is generally slightly larger than that of the estimation error. The figure also clearly illustrates that an increase in support results in a more reliable estimation of the block values.

Again the observed increase in spread between the simulation lines is due to a reduction in the number of blocks, used to create the plots.

5-3-3 Assessment of Calculated Confidence Intervals

To check whether the applied simulation methodology results in a reliable characterization of the geological uncertainty, it is instructive to perform a brief investigation regarding the calculated 50% and 95% confidence intervals.

Figure 5-19 shows that, independently from the selected support, about 50% of the calculated confidence intervals contain the true block concentrations. The results, displayed in Figure 5-20, lead to a similar conclusion. Again, about 95% of the calculated confidence intervals contain the true block concentration.

Both observation indicate that the developed practical process result in a reliable estimation of the geological uncertainty.

Figure 5-18: Evaluation of the error characteristics; Histogram of the estimation errors of the $4 \times 4 \text{ m}^2$ blocks, superimposed on the calculated simulation errors (a). Histogram of the estimation errors of the $8 \times 8 \text{ m}^2$ blocks, superimposed on the calculated simulation errors (b). Histogram of the estimation errors of the $16 \times 16 \text{ m}^2$ blocks, superimposed on the calculated simulation errors (c). Each class has a width of 20 ppm and the corresponding proportion is plotted above the midpoint of the class. The increase in spread is due to the reproduction in the sample size and indicates that the bin width is getting to small.

Figure 5-19: Assessment of the calculated 50% confidence intervals; The green color indicates the collection of blocks for which the known concentrations are located in the corresponding 50% confidence intervals. 55.0% of the calculated 50% confidence intervals for the $4 \times 4 \text{ m}^2$ blocks contain the actual value (a). 52.9% of the calculated 50% confidence intervals for the $8 \times 8 \text{ m}^2$ blocks contain the actual value (b). 56.9% of the calculated 50% confidence interval for the $16 \times 16 \text{ m}^2$ blocks contain the actual value (c).

Figure 5-20: Assessment of the calculated 95% confidence intervals; The green color indicates the collection of blocks for which the known concentrations are located in the corresponding 95% confidence intervals. 95.7% of the calculated 95% confidence intervals for the $4 \times 4 \text{ m}^2$ blocks contain the actual value (a). 95.4% of the calculated 95% confidence intervals for the $8 \times 8 \text{ m}^2$ blocks contain the actual value (b). 94.4% of the calculated 95% confidence interval for the $16 \times 16 \text{ m}^2$ blocks contain the actual value (c).

5-3-4 Representation of Geological Uncertainty

Figure 5-21 eventually indicates that the simulation based approach is perfectly capable of representing the geological uncertainty. The displayed distributions characterize the local geological uncertainty regarding the concentration of four specifically selected $16 \times 16 \text{ m}^2$ blocks. It is interesting to notice that the uncertainty and the shape of the distribution changes with locations. Moreover, the position of the orange bar, which represents the simulation mean, is compared with the estimated and known concentration, respectively indicated by the red and purple bar.

Figure 5-21: Representation of geological uncertainty; The simulation based approach eventually results in a collection of equally probable deposit models. These realization are subsequently being used to characterize the local geological uncertainty and the spatial variability. The displayed distributions characterize the local geological uncertainty regarding the concentration of four selected $16 \times 16 \text{ m}^2$. The orange bar shows the simulation mean, whereas the estimated and known concentration are indicated respectively by red and purple bar.

Case Studies Illustrating the Potential Benefits of a Simulation-based Approach

Besides a correct statistical reproduction of the calibrated random function model, the simulation methodology also generates a reliable characterization of the geological uncertainty. Subsequently, this uncertainty is further propagated through the necessary transfer functions, finally leading to a decision which exploits the upside potential and is sheltered against the downside risk.

The results of the case studies, presented in this chapter, illustrate the potential benefits of the proposed simulation approach and the associated risk based decision strategy. The presented examples are of synthetic nature and can be more or less subdivided into two major application areas, each requiring a different set of transfer functions.

- *Economic transfer functions:* A collection of economic transfer functions is evaluated in order to make a financial assessment of a potential gold mining project. First, a general introduction of the simplified economic situation is presented and the cut-off grade is determined. Thereafter, following examples are studied:
 1. A probability analysis is performed to identify areas for additional drilling.
 2. A grade-tonnage curve is constructed to determine an initial estimate of the amount of material that needs to be processed.
 3. A cash flow analysis is conducted and the Net Present Value (NPV) of the entire mining operation is calculated.
- *Technical transfer functions:* A collection of technical transfer functions is evaluated to transform the spatial rock strength parameters into equipment requirements. The calculation of the cutting forces, specific energies, and power requirements eventually leads to an optimized selection of the installed engine. The performed risk assessment clearly results in an optimized design of the deep sea extraction equipment.

Figure 6-1: Simulation results belonging to the investigated gold deposit; Color coded plot of the simulated gold concentrations in $16 \times 16 \text{ m}^2$ blocks. The block concentrations are obtained via a rescaling of the previously simulated Walker Lake random field (a). Color coded plot of the computed standard deviations, characterizing the uncertainty in the $16 \times 16 \text{ m}^2$ blocks. The computation of each standard deviation is based on the corresponding 100 simulated block concentrations (b). Eventually, each block histogram, constructed from the corresponding 100 simulated values, can be used to calculate the probability of a specified event. In this case, the shown probability map expresses the likelihood that the true block grade exceeds a specified cut-off grade of 2 g/ton (c).

Both sets of examples demonstrate that the proposed simulation approach and the associated risk based decision making can add considerable value to the investigated mining project.

6-1 Case Study 1: Economic Evaluation of a Gold Deposit

6-1-1 Introduction

The investigated gold deposit is essentially a rescaled version of the analyzed Walker Lake deposit. All previously simulated, estimated and known concentrations are divided by 150. Moreover, the resulting data is expressed in the units of g/ton. This particular approach is pursued in order to make sure that the obtained results and the corresponding conclusions can be validated.

Figure 6-1a shows an example of one of the 100 simulated gold deposit models. Figure 6-1b, on the other hand, displays the standard deviations computed from the simulated block grades.

Before commencing the economic evaluation of the gold mining operation, a few fixed financial parameters need to be defined.

- *Mining Cost (C_m)*: The cost to excavate one ton from a deep sea deposit is assumed to amount about 25 \$. This financial parameter just covers the operational cost of these deep sea excavation equipment.
- *Processing Cost (C_p)*: The cost to process one ton from a deep sea deposit is assumed to amount about 75 \$. This financial parameter covers the costs related to the vertical

transport (pump material to surface via the vertical riser), the transport to the mill (material transport with barges to the shore) and the operation of the processing plant.

- *Recovery (r)*: The processing plant is designed to recover about 95% of the contained metal from the material feed.
- *Metal Price (p)*: The gold price is assumed to be constant and amounts about 53 \$/g.
- *Discount rate (R)*: The future cash flows are estimated and discounted to their present values with a monthly discount rate of 5%.
- *Rock Density (ρ)*: The density of the rock is assumed to be constant and amounts about 2.7 ton/m³.
- *Block Size*: The following analysis is based on a block size of 16 x 16 x 16 m³. Such a 4096 m³ block weighs about 11 kt.

Considering the listed parameters, a very intuitive and simple economic cut-off grade can be calculated (Godoy et al., 2001):

$$z_{cut-off} = \frac{C_m + C_p}{p \cdot r} \approx 2 \text{ g/ton} \quad (6-1)$$

This particular cut-off grade is defined as the minimum average grade one ton of material should contain before the economical break-even point is reached. In other words, the value of the processed metal should cover the sum of the corresponding mining and processing costs. Figure 6-1c shows a probability plot which expresses the likelihood that the true block grade exceeds the determined cut-off grade.

6-1-2 Exploration Assistance; Areas for Additional Drilling?

The first example discusses the results obtained via the simple transfer function ‘identify areas for additional drilling’ (Figure 6-2). The implemented transfer function is defined as follows:

- *Scenario 1, white blocks*: The white cells indicate blocks which have a greater than 80% change that the average block grade is above the economic cut-off grade.
- *Scenario 2, dark gray blocks*: The dark gray cells indicate blocks which have a greater than 80% change that the average block grade is below the economic cut-off grade.
- *Scenario 3, light gray blocks*: The light gray cells indicate the locations where additional drilling would most likely provide useful information.

The results are shown in Figure 6-2.

Figure 6-2: Identify areas for additional drilling (light gray). The white cells indicate blocks with a greater than 80% chance that the concentration is above the economic cut-off grade. The dark gray cells outline areas with an over 80% chance that the concentration is below the economic cut-off grade. The intermediate light gray cells represent the blocks where additional drilling would most likely provide useful information. The economic cut-off grade is considered to be 2 g/ton.

6-1-3 Grade-Tonnage Curve and Recovered Metal

The second example discusses the construction of both the grade-tonnage and recovered metal curve under the presence of geological uncertainty. Generally these figures are created to provide a preliminary answer on following questions (Figure 6-3):

- How much material needs to be processed, depending on the selected economic cut-off grade? In other words, how many kiloton of material is classified as ore?
- What is the average grade of the material that is classified as ore?
- On average, how much metal (gold) can be recovered from the material that is classified as ore?

Typically, each question results in a single estimated number, without any indication of the inherent uncertainty. The proposed simulation based approach, on the other hand, enables the formulation of a confidence interval or a probability statement. Computed probabilities could be used later on during a more advanced risk assessment.

The grade tonnage and recovered metal curve for a single obtained deposit model, no matter whether it is estimated, simulated or known, is computed as follows:

1. Define an economic cut-off grade.
2. Select all the blocks with an average grade exceeding the economic cut-off grade.
3. Count the number of selected blocks. Multiply the obtained number with the weight of a single block to compute the total amount of material that needs to be processed (*tonnage*).

Figure 6-3: The grade-tonnage and recovered metal curve; The tonnage curve shows the nature of the declining material volume that needs to be treated as the economic cut-off grade increases. The associated grade curve, however, illustrates the inclining average grade of the treated material. The tonnage can be read off from the left axis, while the average grade is displayed on the right axis (a). The recovered metal curve shows the declining amount of metal that can be recovered from the processed material (b). The curves are constructed for the estimated, the known and each of the 100 simulated $16 \times 16 \times 16 \text{ m}^3$ block deposit models.

4. Calculate the average grade of the selected blocks (*grade*).
5. Multiply the grades with the weight and sum over all selected blocks to compute the recovered metal (*recovered metal*).
6. Loop over steps 1-5 to obtain results for a collection of different cut-off grades.
7. Plot the results in a graph.

The computed curves, shown in Figure 6-3, clearly demonstrate that a simulation approach yields a more reliable computation of the tonnage curve (compare with known field). The results, obtained via block kriging, underestimate the potential amount of material, that can be processed in an economical feasible way. The simulation results turn out to be more realistic. Furthermore, they provide a certain indication regarding the present uncertainty. The average grade, on the other hand, is more or less predicted equally well by both methods. The results show that the conventional kriging techniques do not recognize the full potential of the gold mining deposit. The recovered metal, which has a direct impact on the profitability of the mining operation, is seriously underestimated (Figure 6-3b). Simulations, on the other hand, are better suited to predict the amount of gold which could potentially be sold on the market.

To conclude this subsection, Table 6-1 gives an overview of the expected tonnage, average grade and recovered metal, calculated for a cut-off grade of 2 g/ton .

6-1-4 Optimization of Ore/Waste discrimination

The third example illustrates, how the economic value of the mining operation can be significantly improved, by considering the inherent geological uncertainty during the planning

Table 6-1: Tonnage, average grade and recovered metal for a cut-off grade of 2 g/ton. The table presents the computation results corresponding to a completely known environment, an estimated field and a simulated field. The displayed 90% confidence interval gives an indication about the inherent uncertainty on the simulated expectation.

	Tonnage [kt]	Average Grade [g/ton]	Recovered Metal [kg]
Real	1349	3.18	4287
Kriging	1249	3.09	3863
Mean Simulations	1295	3.20	4140
0.05 quantile	1172	3.13	3793
0.95 quantile	1432	3.30	4521

phase. For each excavated block, the production engineer needs to decide whether the mined material can be classified as ore or as waste. Since the economic consequence of misclassification is not symmetric, it definitely pays off to implement a risk based selection strategy. Decisions thus should be made based on a comparison between the expected economic consequences of both classification scenarios. The proposed strategy is better capable of exploiting the full potential of each single block.

In order to keep the case study simple, one can assume that every block in the investigated deposit will be mined in a sequential way, starting from the lower left corner, ending up in the upper right (Figure 6-4a). A row of 16 blocks will be completely excavated during a single mining period.

Subsequently, each block needs to be classified either as ore or as waste:

- *Scenario 1:* Mined material is regarded as waste and transported to the waste dump located at the ocean floor. The costs of this scenario are only limited to mining related costs.
- *Scenario 2:* Mined material is classified as ore, excavated, pumped to the surface via a vertical riser, transported with barges to the shore, and treated in the processing plant. This scenario is more expensive and includes a mining related and a processing related cost. The financial parameter C_p also account for the transportation costs.

A mined block is typically placed in the second logistical loop if the price of the estimated recovered metal exceeds the sum of the mining, transportation and processing costs. Traditionally, the economic cut-off grade (2 g/ton) is compared with a single estimated block grade in order to classify the mined material (Figure 6-4b). However, such a classification strategy does not account for uncertainty and risk. Given the simulated probability distributions, the expected profit associated with each economic scenario of classification can be assessed and used to derive an economically optimal ore selection (Glacken, 1996). The expected profit of classifying a block as waste is given by:

$$E[Pr_{waste}] = -C_m - P_o \cdot [p \cdot r \cdot m^+ - C_p], \quad (6-2)$$

where P_o is the probability that the true grade of the block exceeds the cut-off grade, m^+ represents the mean grade in the event that the block is classified as ore, and $[p \cdot r \cdot m^+ - C_p]$ corresponds to the lost opportunity cost. The formula actually indicates that the mining

Figure 6-4: Optimized ore/waste selection in mine planning; Indication of planned extraction sequence. The orange blocks are planned to be excavated during the fifth mining period (a). Conventional ore/waste selection strategy based on the estimated $16 \times 16 \times 16 \text{ m}^3$ block grades (b). Improved decision making under the face of geological uncertainty. Risk-based selection based on simulated probability distribution (c). Green = ORE, Red = WASTE

costs need to be paid anyway and that there is a possible loss associated with the disposal of profitable material.

The expected profit of sending a block to the processing plant is:

$$E[Pr_{ore}] = -C_m + P_o \cdot [p \cdot r \cdot m^+ - C_p] + P_w \cdot [p \cdot r \cdot m^- - C_p], \quad (6-3)$$

where P_w is the probability that the true grade of the block is lower than the cut-off grade, m^- represents the mean grade in the event that the block can be regarded as waste, $[p \cdot r \cdot m^+ - C_p]$ is the possible profit in the case of a correct classification, and $[p \cdot r \cdot m^- - C_p]$ is a possible additional cost in the case of a misclassification.

Finally, a block is selected as ore if the expected profit for classifying this block as ore is greater than the expected profit of selection this block as waste. The results of the risk based selection are illustrated in Figure 6-4c. A comparison between both graphs reveals that the risk based selection strategy results in a greater volume of ore. A conventional selection strategy would lead to the processing of 113 blocks, which have a total weight of about 1250 kt. The risk based selection strategy, in contrast, directs 166 blocks as ore, and thus about 1836 kt of minded material is sent to the processing plant.

6-1-5 Cash Flow Analysis

Now, in order to fully grasp the impact of the improved selection strategy, the results need to be expressed in financial terms. Since the entire mining operations is planned to be executed in a sequence of separate mining periods, one may as well perform a future cash flow analysis. Remember that during each mining period, an entire row of 16 blocks is going to be excavated (Figure 6-4a).

Figure 6-5: Future cash flow analysis of the investigated gold mining project; The red and green curve indicate the estimated and real cash flow occurring planning scenario one. The gray and purple curves, on the other hand, indicate the simulated and real cash flow occurring under planning scenario two.

The cash flow of each single period is calculated as:

$$CF_t = \sum_{i \in \mathcal{B}_t} [-C_m - I(\mathbf{x}_i, ore) \cdot (p \cdot r \cdot \bar{z}_i - C_p)], \quad (6-4)$$

where, $\mathcal{B}_t$ denotes the collection of blocks scheduled to be mined during the t^{th} sequence, $I(\mathbf{x}_i, ore)$ is an indicator variable (Ref. Eq. (2-3)), and $\bar{z}_i$ is the estimated or simulated block grade.

Figure 6-5 shows the results of the performed cash flow analysis. To validate the results, both planned mining operations were evaluated synchronous in the exhaustively known environment. A brief overview is given below:

- **Selection Strategy 1:** Conventional ore/waste selection strategy based on estimated block grades (Figure 6-4b).
 - *Planned:* The red curve indicates the cash flow that is estimated to occur when the first selections strategy is pursued. The ore/waste discrimination and predicted profit returns are based upon estimated block grades (block kriging).
 - *Real:* The green curve indicates the cash flow that really will occur when the first selection strategy is pursued.
- **Selection Strategy 2:** Risk based selection based on probability distribution (Figure 6-4c).
 - *Planned:* The gray curves indicate the cash flows that are predicted to occur when the second selection strategy is pursued. The ore/waste discrimination and predicted profit returns are based upon simulated block grades.
 - *Real:* The purple curve indicates the cash flow that really will occur when the second risk-based selection strategy is pursued.

Figure 6-6: Net Present Value of the investigated gold mining project; The red and green curve indicate the estimated and real cash flow occurring planning scenario one. The gray and purple curves, on the other hand, indicate the simulated and real cash flow occurring under planning scenario two.

First of all, Figure 6-5 immediately illustrates that the economic situation of the gold mining operation can be seriously improved. The average periodically revenue increases by about 5 Mln. \$. Moreover, the graph also demonstrates that a simulation based approach results in an expression of the uncertainty. Except for the first three and the sixth mining period, the simulation results are capable of predicting the financial course of the mining operation. It is also interesting to notice that an estimation approach results in a very realistic financial evaluation of the planned project. Unfortunately, estimated ore body models do not constitute an optimal basis for decision making.

6-1-6 Net Present Value of the Mining Operation

To make the impact of a risk-based decision even more transparent, the previous computed cash flows are discounted to their present values. The sum of the resulting terms is called the NPV, and is computed as:

$$NPV = \sum_{t=1}^{18} \frac{CF_t}{(1+R)^t}, \quad (6-5)$$

where t indicates the mining period and R represent the discount rate.

Figure 6-6 shows that a risk-based selection strategy can increase the project value by about 25%. Besides a most expected NPV, the simulation approach also provides an indication of the uncertainty on the entire project value. This uncertainty is a direct consequence of the spatial variability and the limited amount of available information.

To conclude, Figure 6-7 displays the cumulative discounted cash flows calculated for both scenarios.

6-2 Case Study 2: Decision Support for Installed Power Supply

The second case study illustrates how a simulation-based geostatistical analysis can be used to support certain decisions regarding equipment selection and design. A collection of technical

Figure 6-7: Cumulative discounted cash flow of the investigated gold mining project; The red and green curve indicate the estimated and real cash flow occurring under planning scenario one. The gray and purple curves, on the other hand, indicate the simulated and real cash flow occurring under planning scenario two.

transfer functions is used to transform the spatial rock strength parameters into equipment requirements. The analysis was performed as follows:

1. A geostatistical simulation is performed to assess and characterize the spatial variability and geological uncertainty in the rock strength parameters (Brazilian Tensile Strength). The tensile strength of the investigated rock varies between 0.5 - 3.2 MPa .
2. The simulated deposit models are subsequently inserted into Evan's cutting formulate to calculate the forces required to excavate the material under investigation (Evans and Pomeroy, 1966). The required cutting forces vary between 0.70 and 4.41 kN .
3. A second transfer function computes based on the previously obtained cutting forces the specific energy. The resulting values expresses the amount of energy that is needed to cut one cubic meter of rock. The obtained specific energies vary between 417 and 2635 kJ .
4. Considering a desired production rate of 600 m^3/h , a third and final transfer function can be used to compute the required cutting power. The final calculated values vary between 69.43 and 437.71 $kWatt$.

Since the collection of technical transfer functions is evaluated over the entire set of simulated realizations, the uncertainty is automatically propagated through the calculations. This means that eventually each cubic meter of material is connected with a specific distribution of required cutting power. The width of each obtained distributions implicitly reflect the inherent uncertainty.

Now, in order to optimize equipment design and to identify areas for additional drilling, a comparison is made between the workability of excavation equipment with an installed power of 200, 300 or 350 $kWatt$. To facilitate a comparison, the previous simulated block distributions are divided into three categories Figure 6-8:

Figure 6-8: Decision support regarding installed cutting power; Excavation analysis for 200 *kWatt* cutting tool, 300 *kWatt* cutting tool, and 350 *kWatt* cutting tool. Gray areas have a more than 80% change that the required cutting power is lower than the maximum limit of the selected tool. The orange areas have a more than 80% chance that the required cutting power exceeds the limit of the equipment. The orange areas indicate the intermediate situation.

- *Event 1:* The gray cells indicate the regions, which have a greater than 80% change, that the required cutting power is below the limit of the selected mining tool. There is a high probability that the excavation equipment will be able to cut the rock.
- *Event 2:* The orange cells indicate the regions, which have a greater than 80% change, that the required cutting power exceeds the one installed in the mining equipment. In these regions, the cutting will face problems during operations. This could include, lower production rate, high wear, increased maintenance requirements or even breakdown.
- *Event 3:* The purple cells correspond to the intermediate scenario. Due to the local geological uncertainty, it is rather difficult to make reliable statements regarding the cuttability of the rock, in respect with the selected mining tool. The purple cells thus indicate the locations where additional drilling would most likely provide useful information.

Some results of the discussed analysis are shown in Figure 6-8. Only about 25% of the area can certainly be excavated efficiently with the 200 *kWatt* excavation equipment. This percentage can be further increased by selecting a more powerful mining tool. Besides a stronger mining tool, the figure also indicates that part of the problem could be solved by a more detailed exploration campaign. Additional information might likely change the classification of areas, now indicated as intermediate, into safe.

To even further optimize equipment design, a detailed risk assessment should be performed, comparing the likelihood and costs of breakdown, wear and additional maintenance with those of the extra investment costs associated with more powerful equipment.

Conclusions and Recommendations

The goal of this thesis was to develop a general risk quantification framework in the context of deep sea mining, making use of advanced geostatistical simulation methods and transfer functions to quantify geological uncertainty and translate it into decision and project risk. To achieve this goal, the research: (a) explored the current state of the art in the modeling of geological uncertainty and selected the most suitable simulation method; (b) elucidated the practical issues and limitations regarding the application of the selected Local Average Subdivision technique and identified the necessity of certain critical adjustments; (c) integrated the developed solution package in a larger transparent practical process, which through implemented validation procedures, safeguards the reproduction of the available data, the descriptive statistics and the spatial variability; (d) applied the developed practical process to a few case studies, to illustrate the potential benefits of risk based decision making .

7-1 Conclusions

The performed research resulted in following conclusions:

1. Deep sea mining is becoming a viable alternative to meet the increased demands for minerals. Whilst most effort is put in new technological developments, it is equally important to understand the unique management processes and risks of constructing and operating an ocean mine. The importance of establishing a thorough geotechnical and mineralogical knowledge and understanding of the resource cannot be overstated. Unfortunately, the high operating costs of the diving support vessel and the remotely operated vehicle render a dense sampling coverage, equivalent to a land based exploration program, highly unlikely. However, due to the high investment cost and the increased risk, it just becomes even more important to quantify the geological uncertainty and translate it into decision or project risk.

2. Although the reality is unique, in practice, it is impossible to come up with a single conclusive ‘picture’ of the area under investigation. The complex interactions of many physical processes, the limited scientific understanding, and the sparse data make a deterministic modeling approach highly unreliable. In the context of deep sea mining, where the stakes are high due to the required investment capital, it would be better to use a random function model, which assesses the range of values the unknown could take and describes the probability of occurrence of each outcome. In order to facilitate the calibration of such an applied random function model, the stationarity and ergodicity assumptions were introduced.

Generally, once the random function model is calibrated, conventional geostatistical estimation techniques are applied to derive a single deposit realization. The literature review, however, revealed that the associated algorithms were only designed to compute the ‘best’ estimate at a single location. A single deposit model, consisting of an entire collection of such ‘best’ estimates, might not be the best as a whole. Due to the unbiasedness condition and the constraint of a minimized error variance, the estimators are not capable of reproducing the spatial variability of the true grades. Yet, this variability has a significant impact on downstream processing steps and fundamentally drives realized value.

Geostatistical simulation techniques, in contrast, are specifically designed to generate realizations that reproduce the statistics, which are deemed to be consequential for a correct representation of the in-situ variability. A weighted decision table was constructed to support an objective selection of the optimal simulation technique. The computed total scores indicated that the Local Average Subdivision technique is the most suitable for the tasks at hand.

3. The entire benefit of the Local Average Subdivision method lies in the fact that the important scale of support effect is explicitly modeled, resulting in a correct characterization of the geological uncertainty and spatial variability at any scale of interest. Since decisions are generally made on a block scale, it is desirable to immediately generate the realizations on the scale of the Smallest Mineable Unit. This thesis recognized the potential of the Local Average Subdivision technique and introduced it into the field of mining geostatistics.

The computational efficiency of the selected method is a direct result of the clever design of the Local Average Subdivision method. The conditioning child-parent configuration is essentially a numerical approximation and is subject to certain critical assumptions. These assumptions and limitations were explicitly stated to avoid unawareness and wrong application.

4. Although the Local Average Subdivision method is ideally suited to deal with the discrepancy between the support of samples and the intended support of the simulation, the required conditioning step assumes that the known measurement values have a similar size as the simulated local averages. The normal distribution constraint, imposed by the implementation of a Gaussian random function model, further jeopardizes the practical application of the proposed simulation method, in the context of deep sea mining. Moreover, the popular nested spherical covariance model was not implemented in the existing Fortran code. The identified limitations eventually led to a formulation of necessary adjustments and extensions.

The proposed solution package was further integrated into a simulation framework that is systematic, robust and easy to follow. Accompanying validation guidelines were formulated to reduce the likelihood of costly mistakes and to ensure that the simulation results are representative for the area under investigation. The developed simulation framework can be used to generate an entire collection of equally probable representations of the spatial variability.

5. To further increase the value and the reliability of the developed framework, a test simulation was performed in almost completely known environment. On one hand, the test simulation served an example regarding the application of the developed practical process. On the other hand, the completely known environment made it possible to assess the performance of the designed simulation methodology. The analysis showed that the framework meets the formulated objectives.
6. Several case studies further demonstrated the benefit of a probabilistic evaluation approach in a deep sea environment. Integrating geological uncertainty into decision making substantially increases both, the economic performance of classification decisions in mining and the likelihood of optimal equipment design to guarantee a reliable and efficient operation.

7-2 Recommendations for Future Work

The following text points out some interesting topics, which might require some further research.

- *Transfer/utility function and decision/forecast model*: The introduction explained that a conclusive risk quantification framework exists out of three components. This master thesis mainly focused on the spatial stochastic model. Further research should also consider more advanced transfer functions and decision models.
- *Covariance Model*: The test simulation and the case studies both assumed a nested spherical covariance structure. The obtained results are not general and not applicable to other covariance structures. Future work could investigate the performance of the developed simulation methodology under covariance structures.
- *Anisotropy*: The performed study only dealt with an isotropic model. It is not entirely clear yet how anisotropic covariance models might affect the accuracy of the numerical Local Average Subdivision calculations.
- *Reproduction of spatial variability of unconditional simulated random field*: The fifth chapter revealed that the reproduction of the spatial variability for the unconditional simulated random field is rather poor. This is probably due to the limited 3 x 3 parent neighborhood which is used during each child-parent conditioning step. Increasing this neighborhood might solve the problem. This however requires some major adjustments in the software code.
- *Conditioning operation*: The required conditioning step assumes that the known measurement values have a similar size as the simulated local averages. The current

implemented simulation methodology performs a data support adjustment to make point measurements more representative for small blocks. It is however possible to perform a point-block conditioning operation, improving the accuracy of the final obtained results. Further work could be carried out in improving this conditioning step.

- *Estimation versus Simulation:* The test simulation and estimation in the exhaustively known environment revealed that the smoothing effect, related to the application of kriging algorithms, is dependent on the block size under investigation. As soon as the width of the blocks approach one-third of the range, it appears that the smoothing effect becomes irrelevant. Further research could extend this observation into some practical warnings and guidelines.
- *Case Study:* The case studies revealed that decision making under geological uncertainty can considerably enhance the value of the mining project. It is obvious that an improved understanding of the relevant processes can be very valuable.
- *From 2D to 3D:* The currently implemented simulation algorithms are only capable to generate two dimensional deposit realizations. Future work could extend the develop practical process to a three dimensional environment. Software is already available to simulate three dimensional unconditional deposit models. The other remaining steps, however, still need to be extended. The main challenge will lie in the conditioning of the three dimensional unconditional deposit models on data from drill holes and on the anisotropic three dimensional variogram modeling.

Appendix A

Characterization of Spatial Variation and the Corresponding Assumptions

A-1 Covariance Function

The covariance function needs to capture and describe the spatial dependence between two regionalized variables. Presumably, two regionalized variables in close proximity are similar, whereas those that are at more distant locations are less so. Most engineers and scientists will be familiar with the covariance of classical statistics, which describes the relation between paired observations. The covariance between two variables z_1 and z_2 is estimated using n pairs of observations $(z_{i,1}, z_{i,2})$ $i = 1, \dots, n$;

$$C(z_1, z_2) = \frac{1}{n} \sum_{i=1}^n \{z_{i,1} - \bar{z}_1\} \{z_{i,2} - \bar{z}_2\}, \quad (\text{A-1})$$

where $\bar{z}_1$ and $\bar{z}_2$ are the means of the variables z_1 and z_2 . This formula can now be adjusted as to describe the relation between a regionalized variable at two distinct locations. In our spatial setting, z_1 and z_2 become $Z(\mathbf{x}_1)$ and $Z(\mathbf{x}_2)$, i.e. they represent now values of the same property Z , at the locations $\mathbf{x}_1$ and $\mathbf{x}_2$. Their covariance can be described by (Webster and Oliver, 2007; Banerjee et al., 2004; Fenton, 1990):

$$\begin{aligned} C(\mathbf{x}_1, \mathbf{x}_2) &= E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\} \cdot \{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}] \\ &= \int_{\theta} \int_{\theta} b_1 b_2 f(\mathbf{x}_1, \mathbf{x}_2; b_1, b_2) db_1 db_2 - \mu(\mathbf{x}_1) \cdot \mu(\mathbf{x}_2) \end{aligned} \quad (\text{A-2})$$

where the integral of the joint pdf is evaluated over the entire sample space θ of the random variable Z and where $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ represent the mean values of the property Z at location $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$\mu(\mathbf{x}) = E[Z(\mathbf{x})] = \int_{\theta} b f(\mathbf{x}; b) db. \quad (\text{A-3})$$

Notice that this definition is not directly applicable in our spatial setting, since an observation at a specific location remains the same, no matter how many times the observation is repeated

(assuming that the measurement error is ignored). In other words, it is not possible to further elaborate Eq. (A-2), because there is only realization of Z at each point in space and thus the expected product of both differences can not be calculated.

This impasse is overcome by introducing the stationarity assumptions, which allows us to treat the values at different positions as though they are different realizations of the same property.

A-2 Stationarity Assumptions and Ergodicity

A spatial stochastic process is said to be stationary if the process has certain attributes that are the same everywhere. In following discussion, it is assumed that the spatial process has a mean, say $\mu(\mathbf{x}) = E[Z(\mathbf{x})]$ associated with it, and that the variance of $Z(\mathbf{x})$ exists for all $\mathbf{x} \in \mathcal{A}$. Generally three types of stationarity are widely accepted in geostatistics: strict stationarity, weak stationarity and intrinsic stationarity.

The most restrictive assumption, *strict stationarity*, states that the random variables corresponding to two sets of locations should have the same joint distribution if the second set of locations is a translated version of the first one. In other words, at any set of locations $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ and for any separation vector $\mathbf{h} \in \mathcal{E}$, the joint pdf of $\{Z(\mathbf{x}_1), \dots, Z(\mathbf{x}_k)\}$ should be the same as that of $\{Z(\mathbf{x}_1 + \mathbf{h}), \dots, Z(\mathbf{x}_k + \mathbf{h})\}$ (Fenton, 1990).

A less restrictive condition is given by *weak stationarity*, which only dictates that the first and second order moments of the joint distribution have to be constant. Weak stationarity assumes that the individual realizations fluctuate around a constant first moment or mean, $\mu = E[Z(\mathbf{x})]$ for all $\mathbf{x} \in \mathcal{E}$ (Webster and Oliver, 2007). This enables us to overcome the impasse described above by replacing $\mu(\mathbf{x}_1)$ and $\mu(\mathbf{x}_2)$ in Eq. (A-2) by a single value μ . The result, however, is still restricted to the two particular locations $\mathbf{x}_1$ and $\mathbf{x}_2$, which is not very useful, when interest lies in tracing the general characteristics of the spatial process. To apply the proposed random function model in a more general setting, it is necessary to make some further assumptions.

The first additional assumption concerns the behavior of the function when the two locations coincide. In such a situation, the above formulated covariance equation reduces to the variance formula, $\sigma^2 = E[(Z(\mathbf{x}) - \mu)^2]$, sometimes called the *a priori or point variance* of the process. This variance is assumed to be finite and constant for all $\mathbf{x} \in \mathcal{A}$.

The second premise concerns the description of the behavior of the spatial process when two locations do not coincide. In this situation, the behavior should ideally be characterized based on a separation vector $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$ instead of on absolute positions. This implies that equation (A-2) has to be rewritten:

$$\begin{aligned} C(\mathbf{h}) &= \text{Cov}[Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})] \\ &= E[\{Z(\mathbf{x}) - \mu\} \cdot \{Z(\mathbf{x} + \mathbf{h}) - \mu\}], \end{aligned} \quad (\text{A-4})$$

for all $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$ within $\mathcal{E}$. This weak stationarity assumption thus postulates that the covariance relationship between the values of the process at any two locations can be summarized by a covariance function $C(\mathbf{h})$, which depends only on the separation vector $\mathbf{h}$. Note that strict stationarity implies weak stationarity. The converse is not true in general, but does hold for a Gaussian process because its higher order moments are known constants (Banerjee et al., 2004).

The third type of stationarity, *intrinsic stationarity*, was introduced by Matheron (1971) to overcome the problem of a mean that changes across a region, and a variance that increases without bound as the region of interest expands. In these circumstances, one can not define the covariance. Intrinsic stationarity instead, places restrictions on the first and second order moments of the differences between a regionalized variable at any pair of locations. Matheron took the view that, although the mean might not be constant over the whole region, it will be in smaller subregions, so that the expected difference of a pair would approach zero for short lag vectors:

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] \approx 0 \quad \text{for small } \|\mathbf{h}\|. \quad (\text{A-5})$$

Further he replaced the covariance function with a function that expresses the expected variance of the difference as a function of the lag vector. This so called variogram is an equivalent measure of the spatial relation between two regionalized variables separated by a vector $\mathbf{h}$:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] \\ &= \text{Var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})]. \end{aligned} \quad (\text{A-6})$$

Eq. (A-6) only makes sense if the right-hand side depends on the separation vector $\mathbf{h}$, and not on the absolute location of both points. If both requirements Eq. (A-5) and Eq. (A-6) are fulfilled, the process is said to be intrinsically stationary (Clark, 1979). The function $\gamma(\mathbf{h})$ is then called the semi-variogram for the obvious reason of being half a variance of the difference. This value is nevertheless the variance per point when the points are considered in pairs. Also note that intrinsic stationarity, as well as weak stationarity, imposes restrictions on the first and second moments of the random process. It says nothing about the joint distribution of a collection of variables and thus provides no likelihood.

The remainder of this subsection shall illustrate the relationship between the covariance function (weak stationarity) and the semi-variogram function (intrinsic stationarity):

$$\begin{aligned} 2\gamma(\mathbf{x}) &= \text{Var}[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})] \\ &= \text{Var}[Z(\mathbf{x} + \mathbf{h})] + \text{Var}[Z(\mathbf{x})] - 2\text{Cov}[Z(\mathbf{x} + \mathbf{h}), Z(\mathbf{x})] \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})]. \end{aligned}$$

Thus,

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \quad (\text{A-7})$$

Eq. (A-7) indicates that the semi-variogram $\gamma(\mathbf{h})$ is easily recovered from the covariance function, but it should be stressed that the converse is in general not true (Diggle and Ribeiro Jr., 2007). The covariance function can only be obtained from the semi-variogram function if an additional assumption is made.

The additional *ergodicity* assumption states that $C(\mathbf{h}) \rightarrow 0$ as $\|\mathbf{h}\| \rightarrow \infty$. This makes sense, since it means that the covariance between the values of two points in space vanishes as the separation between the points becomes infinitely large. Theoretically the same assumption also requires that, when the regional bounds expand towards infinity, the moments of the single observable realization approach those of the complete ensemble. The covariance at lag $\mathbf{0}$ can be obtained by taking the limit of both sides of equation (A-7) for $\|\mathbf{h}\| \rightarrow \infty$. This results in:

$$\lim_{\|\mathbf{h}\| \rightarrow \infty} \gamma(\mathbf{h}) = C(\mathbf{0}). \quad (\text{A-8})$$

By inserting Eq. (A-8) into Eq. (A-7) and using a dummy variable $\mathbf{u}$ to avoid confusion, we then obtain:

$$C(\mathbf{h}) = \lim_{\|\mathbf{u}\| \rightarrow \infty} \gamma(\mathbf{u}) - \gamma(\mathbf{h}). \quad (\text{A-9})$$

The limit in Eq. (A-9) need not to exist, but if it does the process is weakly stationary. It is only under these circumstances that the covariance function can be recovered from a semi-variogram and thus weak stationarity implies intrinsic stationarity (Banerjee et al., 2004). The covariance and semi-variogram function are by far the most widely used descriptive tools in geostatistics to characterize the spatial relation of geological attributes. However, there is another equivalent approach which takes advantage of the spectral representation of a random process. Information concerning the second moment of a stationary Gaussian random process is eventually summarized in the so called spectral density function.

A-3 Spectral Density Function

The spectral density function describes how much each frequency component contributes to the overall variance of a random process. Although this function is expressed in the less familiar spectral domain, it still contains the same information as the covariance function. Both functions are merely transforms of one another. Following discussion is based on the very lucid explanation given in Fenton and Griffiths (2008). To avoid a lengthy explanation, the discussion will mainly focus on the spectral representation of a one-dimensional random process and its corresponding spectral density function. The extension to the two-dimensional setting is fairly straightforward and tedious and will be omitted. Only the final results for the two dimensional setting will be presented.

A-3-1 Spectral Representation of One Dimensional Process

Priestley (1981) has shown that a one-dimensional stationary Gaussian process $Z(x)$ can be considered as a sum of sinusoids, with mutually independent amplitudes and phase angles. Such a representation is possible, provided that the covariance function is continuous at the origin, the random amplitude C_k is Rayleigh distributed and the random phase angle Φ_k is uniformly distributed on the interval $[0, \pi]$. Under these conditions, one can express the Gaussian process as:

$$Z(x) = \mu + \sum_{k=-N}^N C_k \cos(\omega_k x + \Phi_k). \quad (\text{A-10})$$

Now consider the expectation and variance of the k_{th} component of the Gaussian process, $Z_k(x) = C_k \cos(\omega_k x + \Phi_k)$, while ignoring the process mean μ (Fenton and Griffiths, 2008):

$$\mathbb{E}[Z_k(x)] = \mathbb{E}[C_k \cos(\omega_k x + \Phi_k)] = \mathbb{E}[C_k] \cdot \mathbb{E}[\cos(\omega_k x + \Phi_k)] = 0 \quad (\text{A-11})$$

$$\text{Var}[Z_k(x)] = \mathbb{E}[Z_k(x)^2] = \mathbb{E}[C_k^2] \mathbb{E}[\cos^2(\omega_k x + \Phi_k)] = \frac{1}{2} \mathbb{E}[C_k^2]. \quad (\text{A-12})$$

Notice that these results could only be obtained due to the independence of the amplitude and phase angle, and the fact that the phase angle is uniformly distributed between 0 and

Figure A-1: Two-sided spectral density function $S(\omega)$ (Griffiths and Fenton, 2007).

π . Thanks to Priestley (1981), one also knows that the sinusoid components $Z_k(x)$ are independent of each other. Using this property, all k components can be put together to obtain the mean and variance of the complete random process:

$$E[Z(x)] = \mu + \sum_{k=-N}^N E[Z_k(x)] = \mu, \quad (\text{A-13})$$

$$\text{Var}[Z(x)] = \sum_{k=-N}^N \text{Var}[Z_k(x)] = \sum_{k=-N}^N \frac{1}{2} E[C_k^2]. \quad (\text{A-14})$$

In other words, the prescribed mean of $Z(x)$ is preserved by the spectral representation and the variance of the sum is the sum of the variances of each component frequency (Fenton and Griffiths, 2008, p. 97). The power in the sinusoid amplitude, $\frac{1}{2} E[C_k^2]$, indicates how much each frequency component contributes to the overall variance of the process. Due to the application of some trigonometry relations, it is also possible to express the covariance function in the spectral domain. The mathematical derivation which is reproduced in appendix A-4 finally results in:

$$C(h) = \sum_{k=-N}^N \frac{1}{2} E[C_k^2] \cos(\omega_k h) \quad (\text{A-15})$$

Notice the similarities between the variance (Eq. (A-14)) and the covariance formulation in the spectral domain (Eq. (A-15)). For a zero lag distance, the covariance formula reduces to the a-priori variance of the random process (Eq. (A-14)).

Now time has come to introduce an additional technique, which is needed to calculate the power of each sinusoid amplitude in both the variance and covariance formulation. The power of each amplitude component can be related to a specific area under the spectral density function. Let us define a two-sided spectral density function $S(\omega)$, such that the power $\frac{1}{2} E[C_k^2]$ can be calculated as the area $S(\omega_k) \Delta\omega$ of a rectangle centered at ω_k (Figure A-1). The variance of a Gaussian process can thus be written as:

$$\text{Var}[Z(x)] = \sigma^2 = \sum_{k=-N}^N S(\omega_k) \Delta\omega = \int_{-\infty}^{\infty} S(\omega) d\omega, \quad (\text{A-16})$$

where the last equality holds for $\Delta\omega \rightarrow 0$ and $N \rightarrow \infty$. The variance of the process $Z(x)$ thus equals the area under the two-sided spectral density function. In a similar way, the covariance

Figure A-2: One-sided spectral density function $G(\omega)$ (Griffiths and Fenton, 2007).

of the random process can be expressed as:

$$C(h) = \sum_{k=-N}^N S(\omega_k) \cos(\omega_k h) \Delta\omega = \int_{-\infty}^{\infty} S(\omega) \cos(\omega h) d\omega. \quad (\text{A-17})$$

The covariance function $C(h)$ can thus be written as the Fourier transform of the spectral density function, $S(\omega)$. The inverse operation can be applied to find $S(\omega)$ from $C(h)$:

$$S(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} C(h) \cos(\omega h) dh. \quad (\text{A-18})$$

Since the covariance function is even, i.e. $C(h) = C(-h)$, and $\cos(x) = \cos(-x)$, the spectral density function $S(\omega)$ must also be an even function (Figure A-1). The fact that $S(\omega)$ is symmetric around the origin, means that only the positive half side of spectrum needs to be known. This motivates the introduction of the one-sided spectral density function (Figure A-2): $G(\omega) = 2 \cdot S(\omega)$ $\omega \geq 0$. The factor two is included to preserve the correct power for the sinusoid amplitude components, when only positive frequencies are considered. The Fourier transformation pair becomes:

$$\begin{aligned} C(h) &= \int_0^{\infty} G(\omega) \cos(\omega h) d\omega, \\ G(\omega) &= \frac{2}{\pi} \int_0^{\infty} C(h) \cos(\omega h) dh. \end{aligned} \quad (\text{A-19})$$

The spectral density representation is particularly useful when the frequency domain is an integral part of the problem under consideration. For example, the results of earthquake ground motion or geophysical measurements are quite often represented in the spectral domain. The ground motion is largely sinusoidal and the power corresponding to the discretized frequency components is dictated by the resonance of the soil and rock through which the waves are traveling.

In addition, the spectral representation is also suited to simulate a stationary Gaussian process. Such a simulation can be performed by generating independent realizations of the amplitudes C_k and phase angles Φ_k , and then recombining them together according

to Eq. (A-10). The insights gained in the spectral representation of a one-dimensional Gaussian process $Z(x)$, can now easily be extended to a two dimensional process $Z(\mathbf{x})$, where $\mathbf{x} = \{\mathbf{x}(1), \mathbf{x}(2)\}$ denotes a location vector.

Spectral Representation of Two Dimensional Process

The above derived formulas can now easily be extended to a two dimensional setting. The formulation of a two-dimensional stationary Gaussian process can be obtained by including in Eq. (A-10) a double sum:

$$Z(\mathbf{x}) = \mu + \sum_{k_1=-N_1}^{N_1} \sum_{k_2=-N_2}^{N_2} C_{k_1 k_2} \cos(\omega_{k_1} x_1 + \omega_{k_2} x_2 + \Phi_{k_1 k_2}), \quad (\text{A-20})$$

where, as in the one-dimensional case, $C_{k_1 k_2}$ is a random amplitude and $\Phi_{k_1 k_2}$ a random phase angle. The one-dimensional variance formulation Eq. (A-14) then needs to be adjusted to fit the two-dimensional setting:

$$\text{E}[Z(\mathbf{x})] = \sigma^2 = \sum_{k_1=-N_1}^{N_1} \sum_{k_2=-N_2}^{N_2} \frac{1}{2} \text{E}[C_{k_1 k_2}^2] \quad (\text{A-21})$$

Subsequently, a two-dimensional spectral density function $S(\omega_1, \omega_2)$ can be defined such that the power of each amplitude component $\frac{1}{2} \text{E}[C_{k_1 k_2}^2]$ equals the volume $S(\omega_{k_1}, \omega_{k_2}) \Delta\omega_{k_1} \Delta\omega_{k_2}$ of a beam centered at $\{\omega_{k_1}, \omega_{k_2}\}$.

$$S(\omega_{k_1}, \omega_{k_2}) \Delta\omega_1 \Delta\omega_2 = \frac{1}{2} \text{E}[C_{k_1 k_2}^2] \quad (\text{A-22})$$

The overall variance of the process can now be expressed as the total volume under the two-dimensional spectral density function.

$$\sigma^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(\omega_1, \omega_2) d\omega_1 d\omega_2 \quad (\text{A-23})$$

Finally, the two dimensional Fourier transformation pair becomes:

$$\begin{aligned} C(h_1, h_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S(\omega_1, \omega_2) \cos(\omega_1 h_1 + \omega_2 h_2) d\omega_1 d\omega_2 \\ S(\omega_1, \omega_2) &= \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C(h_1, h_2) \cos(\omega_1 h_1 + \omega_2 h_2) dh_1 dh_2 \end{aligned} \quad (\text{A-24})$$

where $C(h_1, h_2)$ represent the two dimensional covariance function.

A-4 Spectral Representation of the Covariance Function of a one-dimensional Process

This section shall proof that the covariance function can be translated to the spectral domain, and therefore that the spectral density function is equivalent description of the spatial

continuity of a random process. Assuming a process mean of zero to simplify the algebra (the end result does not depend on the value of μ), this is done as follows:

$$\begin{aligned} C(h) &= \text{Cov}[Z(0), Z(h)] \\ &= \sum_{k=-N}^N \sum_{j=-N}^N \text{E}[Z_k(0)Z_j(h)], \end{aligned} \quad (\text{A-25})$$

This result was obtained by inserting Eq. (A-14) into the covariance expression and by compacting the k^{th} frequency component $C_k \cos(\omega_k x + \Phi_k)$ to $Z_k(x)$. Taking account of the independence between the frequency components, this relation can be further simplified to:

$$C(h) = \sum_{k=-N}^N \text{E}[Z_k(0)Z_k(0)] \quad (\text{A-26})$$

Then, since $Z_k(0) = C_k \cos(\Phi_k)$ and $Z_k(h) = C_k \cos(\omega_k h + \Phi_k)$, and due to following trigonometry relation $\cos(\alpha) \cos(\beta) = \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$ we get:

$$\begin{aligned} C(h) &= \sum_{k=-N}^N \text{E}[Z_k(0)Z_k(0)] \\ &= \sum_{k=-N}^N \text{E}\left[C_k^2 \cos(\Phi_k) \cos(\omega_k h + \Phi_k)\right] \\ &= \sum_{k=-N}^N \text{E}\left[C_k^2\right] \text{E}[\cos(\Phi_k) \cos(\omega_k h + \Phi_k)] \\ &= \sum_{k=-N}^N \text{E}\left[C_k^2\right] \text{E}\left[\frac{1}{2} \{\cos(\omega_k h + 2\Phi_k) + \cos(\omega_k h)\}\right] \end{aligned} \quad (\text{A-27})$$

Finally the covariance function can then be written as:

$$C(h) = \sum_{k=-N}^N \frac{1}{2} \text{E}\left[C_k^2\right] \cos(\omega_k h) \quad (\text{A-28})$$

Appendix B

Derivation of Ordinary Kriging Equations

B-1 Weighted Linear Combination and the Unbiasedness Condition

In theory, the properties of an ore body could be measured at an infinite number of places. It is rather obvious, that such a sampling strategy is not feasible, since it would result in the excavation of the whole deposit. Mainly due to reasons of economy, the number of samples gathered during a (pre-)feasibility study is limited. Nevertheless, engineers, analysts and managers would already like to have a complete characterization of the deposit, during this early stage in the project. In order to comply with the stringent requirements of these stakeholders, the properties of the deposit at unsampled locations need to be estimated from the available data.

Returning to the probabilistic modelling approach, each measurement at a sampling location can be considered as an outcome of a random variable. It is now sensible, to estimate an unknown true value at a particular location $\mathbf{x}_\eta$, as the most probable outcome of the random variable $Z(\mathbf{x}_\eta)$. This random variable can be modeled as a linear combination of other random variables $Z(\mathbf{x}_\alpha)$, which are positioned at the sampling locations $\mathbf{x}_\alpha$. The modeled random variable is denoted by $\hat{Z}(\mathbf{x}_\eta)$ and is written as:

$$\hat{Z}(\mathbf{x}_\eta) = \sum_{\alpha=1}^{n_k} \lambda_\alpha Z(\mathbf{x}_\alpha). \quad (\text{B-1})$$

To obtain a good estimate of the geological property at a specific location, it would suffice to calculate the expectation of the random variable described in Eq. (B-1). The expected value of this weighted linear combination of random variables is equal to the weighted linear combination of the individual expected values (Isaaks and Srivastava, 1989):

$$E[Z(\mathbf{x}_\eta)] = E\left[\sum_{\alpha=1}^{n_k} \lambda_\alpha Z(\mathbf{x}_\alpha)\right] = \sum_{\alpha=1}^{n_k} \lambda_\alpha E[Z(\mathbf{x}_\alpha)]. \quad (\text{B-2})$$

Figure B-1: Data configuration in an ordinary kriging example. The available sample data is numbered from 1 to 7 and their measured value is written immediately to the right of the plus sign. The plus sign indicates the sample location $\mathbf{x}_\alpha$. The unknown true value $z(\mathbf{x}_\eta)$ at location $\mathbf{x}_\eta$ will be modeled as a linear combination of the surrounding sample data (after Isaaks and Srivastava (1989)).

This expression is even valid if there is some dependence between the variables involved. It is obvious that in practice, in order to calculate an estimate at an unsampled location, each expected value $E[Z(\mathbf{x}_\alpha)]$ in Eq. (B-2), is replaced by the corresponding measured value $z(\mathbf{x}_\alpha)$. The estimate at location $\mathbf{x}_\eta$ is thus a linear combination of measurement data in the neighborhood (Figure B-1). Although the notations with expected values offer no significant advantage for the calculation of the estimate, it is rather convenient for the formulation of the estimation variance.

The minimization of this estimation variance, which is achieved by setting various partial derivatives to zero, will eventually result in the ordinary kriging equations. These kriging equations finally will lead to the calculation of the weights in Eq. (B-1). Before formulating an expression of the modeled error variance and calculating its derivatives, it is necessary to introduce the unbiasedness condition. The unbiasedness condition states that the expected error calculated over all prediction points should be zero. Under the assumption that the mean of the random process is unknown, the unbiasedness condition further implies that the kriging weights in Eq. (B-1) have to sum up to one. This additional constraint is distinctive for ordinary kriging and holds during the remainder of this chapter. Ordinary kriging is by far the most common type of kriging in practice and therefore following discussion focuses on its theory only (Webster and Oliver, 2007).

In order to protect this extra requirement, it will be necessary to use the technique of the Lagrange parameter which can convert the constrained minimization problem into an unconstrained one. The constraint results from the practical intention of having the actual average error of a particular set of estimates to be zero (Goovaerts, 1997).

If we define the error R at location $\mathbf{x}_\eta$ as a random variable which is defined as the difference between the true unknown random variable and the one, modelling the true value at that

particular location:

$$R(\mathbf{x}_\eta) = \sum_{\alpha=1}^{n_k} \lambda_\alpha Z(\mathbf{x}_\alpha) - Z(\mathbf{x}_\eta) \quad (\text{B-3})$$

then, the average error of a set of n_p estimates is:

$$\text{average error} = R_{Av} = \frac{1}{n_p} \sum_{\eta=1}^{n_p} R(\mathbf{x}_\eta). \quad (\text{B-4})$$

In this equation, the random variables $R_1, R_2, \dots, R_{n_p}$ model the errors at the n_p prediction locations. Obviously, it is desirable that the average estimation error over all these prediction points would be equal to zero. Similar to Eq. (B-2), the expected average error can be expressed as a linear combination of the individual expected errors:

$$E[R_{Av}] = E\left[\frac{1}{n_p} \sum_{\eta=1}^{n_p} R(\mathbf{x}_\eta)\right] = \frac{1}{n_p} \sum_{\eta=1}^{n_p} E[R(\mathbf{x}_\eta)] = 0. \quad (\text{B-5})$$

To ensure that the expected average error is zero, one must insist that each of the individual expected errors has a value of zero. Mathematically, this statement can be written as:

$$E[R(\mathbf{x}_\eta)] = \sum_{\alpha=1}^{n_k} \lambda_\alpha E[Z(\mathbf{x}_\alpha)] - E[Z(\mathbf{x}_\eta)] = 0 \quad \forall \eta = 1, \dots, n_p. \quad (\text{B-6})$$

Assuming that the random function is stationary, Eq. (B-6) can be further simplified to:

$$E[R(\mathbf{x}_\eta)] = \sum_{\alpha=1}^{n_k} \lambda_\alpha E[Z] - E[Z] = 0. \quad (\text{B-7})$$

The expected value of the individual error $E[R(\mathbf{x}_\eta)]$, at any particular location, is known as the *bias*. Setting this expected value to zero ensures unbiasedness and results in the conclusion that n_k weights should sum to one:

$$\sum_{\alpha=1}^{n_k} \lambda_\alpha = 1. \quad (\text{B-8})$$

Please notice that the previously discussed theoretical principles were derived under the specific assumption that the mean of the random process is assumed to be unknown. The unbiasedness condition, which safeguards the desire of having an expected error of zero, thus results in a constraint on the so called ordinary kriging weights (Eq. (B-8)) (Isaaks and Srivastava, 1989; Goovaerts, 1997).

An expected error of zero alone, however, does not guarantee that the linear combination is a good estimator. Alternating errors, with large absolute values but opposite signs, still can average to zero, while it is clear that such a situation should definitely be avoided. Likely, the problem can be solved by calculating the ordinary kriging weights in such a way, so that they minimize the error variance.

B-2 The Random Function Model and Error Variance

Previously, it was already hinted, that in order to obtain the set of kriging equations, it is necessary to formulate the error variance and calculate its derivatives. Subsequently, the text

shall explain how this variance should be formulated and why it should be minimized.

The error variance is actually the expected squared difference between the error and its mean. For one particular location, the error variance σ_R^2 of a set of k estimates can be written as:

$$\sigma_R^2 = \frac{1}{k} \sum_{i=1}^k (r_{\eta,i} - m_R)^2 = \frac{1}{k} \sum_{i=1}^k (r_{\eta,i})^2, \quad (\text{B-9})$$

where $r_{\eta,i}$ corresponds with a realization of the error variable R_η . The last equality holds, since the unbiasedness condition enforces the expected or mean error at an individual location to be zero.

Taking account of Eq. (B-3), the error variance at a certain location can also be modeled as the variance of the difference between the estimate and the random variable modelling the true value (Isaaks and Srivastava, 1989; Webster and Oliver, 2007)

$$\sigma_R^2 = \text{var}[R(\mathbf{x}_\eta)] = \text{Var}[\hat{Z}(\mathbf{x}_\eta) - Z(\mathbf{x}_\eta)]. \quad (\text{B-10})$$

The variance of this difference is nothing more then a simplified version of the variance of a weighted linear combination of several random variables. Let's say that interest lies in a random variable, which is created by the linear combination $\sum_{i=1}^k w_i Z_i$, then the variance of this variable can be calculated as:

$$\text{Var}[\sum_{i=1}^k w_i Z_i] = \sum_{i=1}^k \sum_{j=1}^k w_i w_j \text{Cov}[Z_i Z_j]. \quad (\text{B-11})$$

This relation can now be applied to further develop the variance of the random variable $R(\mathbf{x}_\eta)$ displayed in Eq. (B-10):

$$\sigma_R^2 = \text{Cov}[\hat{Z}(\mathbf{x}_\eta) \cdot \hat{Z}(\mathbf{x}_\eta)] + \text{Cov}[Z(\mathbf{x}_\eta) \cdot Z(\mathbf{x}_\eta)] - 2\text{Cov}[\hat{Z}(\mathbf{x}_\eta) \cdot Z(\mathbf{x}_\eta)]. \quad (\text{B-12})$$

The first term of Eq. (B-12) is the covariance of $\hat{Z}(\mathbf{x}_\eta)$ with itself. The covariance of a certain variable with itself can be replaced by the variance of the same variable. Remember also that the estimate $\hat{Z}(\mathbf{x}_\eta)$ is a linear combination of other random variables. Eventually, the first term can thus be written as the variance of a linear combination. Taking account of Eq. (B-11), the first becomes:

$$\text{Cov}[\hat{Z}(\mathbf{x}_\eta) \cdot \hat{Z}(\mathbf{x}_\eta)] = \text{Var}[\hat{Z}(\mathbf{x}_\eta)] = \sum_{\alpha=1}^{nk} \sum_{\beta=1}^{nk} \lambda_\alpha \lambda_\beta C_{\alpha\beta}. \quad (\text{B-13})$$

The symbol $C_{\alpha\beta}$ is the covariance of Z between the data points $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$. The second term of Eq. (B-12) is the covariance of $Z(\mathbf{x}_\eta)$ with itself, which is equal to the variance of $Z(\mathbf{x}_\eta)$. Assuming that all the random variables have the same variance σ^2 , which corresponds with the point variance of the random function, then the second term can be written as:

$$\text{Cov}[Z(\mathbf{x}_\eta) \cdot Z(\mathbf{x}_\eta)] = \sigma^2. \quad (\text{B-14})$$

The third term of Eq. (B-12) is the covariance between the estimate $\hat{Z}(\mathbf{x}_\eta)$ and the random variable $Z(\mathbf{x}_\eta)$ modelling the true value. This term can be further elaborated:

$$\begin{aligned}
 2\text{Cov}[\hat{Z}(\mathbf{x}_\eta) \cdot Z(\mathbf{x}_\eta)] &= 2\text{Cov}\left[\sum_{\alpha=1}^{nk} \lambda_\alpha Z(\mathbf{x}_\alpha) \cdot Z(\mathbf{x}_\eta)\right] \\
 &= 2\text{E}\left[\sum_{\alpha=1}^{nk} \lambda_\alpha Z(\mathbf{x}_\alpha) \cdot Z(\mathbf{x}_\eta)\right] - 2\text{E}\left[\sum_{\alpha=1}^{nk} \lambda_\alpha Z(\mathbf{x}_\alpha)\right] \cdot \text{E}[Z(\mathbf{x}_\eta)] \\
 &= 2 \sum_{\alpha=1}^{nk} \lambda_\alpha \text{E}[Z(\mathbf{x}_\alpha) \cdot Z(\mathbf{x}_\eta)] - 2 \sum_{\alpha=1}^{nk} \lambda_\alpha \cdot \text{E}[Z(\mathbf{x}_\alpha)] \cdot \text{E}[Z(\mathbf{x}_\eta)] \\
 &= \sum_{\alpha=1}^{nk} \lambda_\alpha C_{\alpha\eta}.
 \end{aligned} \tag{B-15}$$

The symbol $C_{\alpha\eta}$ is the covariance of Z between the α^{th} data point and the target point $\mathbf{x}_\eta$. Recombining these three terms together, results in a final formulation of the error variance:

$$\sigma_R^2 = \sigma^2 + \sum_{\alpha=1}^{nk} \sum_{\beta=1}^{nk} \lambda_\alpha \lambda_\beta C_{\alpha\beta} - 2 \sum_{\alpha=1}^{nk} \lambda_\alpha C_{\alpha\eta}. \tag{B-16}$$

Once the random function model is defined, i.e. a value has been assigned to the point variance σ^2 and a covariance relationship has been modelled, the previous equation provides us with a tool to calculate the error variance in function of n_k unknown variables. At this point, it comes down to finding the weights, $\lambda_1, \lambda_2, \dots, \lambda_{n_k}$, which minimize this error variance.

B-3 Ordinary Kriging Equations

The minimization of a function of n_k variables is usually accomplished by setting the n_k partial derivatives to zero. This produces a system of n_k equations with n_k unknowns which can be solved easily by some standard techniques in the linear algebra. Unfortunately, such a solution of the n_k equations might violate the unbiasedness condition. To ensure that Eq. (B-1) is an unbiased estimator, an additional equation has to be added to the existing system of equations. This additional equation, previously derived from the unbiasedness condition (Eq. (B-8)), however does not introduce any new variable. The situation of having $n_k + 1$ equations with n_k variables is far from favorable, even frightening. To overcome this awkward problem, the Lagrange parameter ψ is introduced into the equation of the error variance. The formulation of the error variance has to be adjusted (Webster and Oliver, 2007):

$$\sigma_R^2(\lambda_\alpha, \psi) = \sigma^2 + \sum_{\alpha=1}^{nk} \sum_{\beta=1}^{nk} \lambda_\alpha \lambda_\beta C_{\alpha\beta} - 2 \sum_{\alpha=1}^{nk} \lambda_\alpha C_{\alpha\eta} + 2\psi \left(\sum_{\alpha=1}^{nk} \lambda_\alpha - 1 \right) \tag{B-17}$$

Notice that the last term, which introduces the Lagrange parameter into the equation, does not upset the equality. This is because the unbiasedness condition was embedded into the

adjusted formulation in such a way that it would cancel out once the weights sum up to one. This small trick is demonstrated by the calculation of following partial derivative:

$$\frac{\delta \sigma_R^2}{\delta \psi} = 2 \sum_{\alpha=1}^{nk} \lambda_{\alpha} - 2. \quad (\text{B-18})$$

To minimize the error variance, the right hand side of this result has to be zero. It is clear that this eventually result in an equation that will protect the unbiasedness condition. The other n_k equations belonging to the kriging system are obtained by taking the partial derivatives of the error variance with respect to the other remaining variables and setting them to zero:

$$\frac{\delta \sigma_R^2}{\delta \lambda_{\alpha}} = 2 \sum_{\beta=1}^{nk} \lambda_{\beta} C_{\alpha \beta} - 2 C_{\alpha \eta} + 2 \psi = 0 \quad \forall \alpha = 1, \dots, n_k. \quad (\text{B-19})$$

Finally, the weights of the proposed linear combination Eq. (B-1) can be calculated by solving the following $n_k + 1$ equations:

$$\begin{aligned} \sum_{\beta=1}^{nk} \lambda_{\beta} C_{\alpha \beta} + \psi &= C_{\alpha \eta} \quad \forall \alpha = 1, \dots, n_k \\ \sum_{\alpha=1}^{nk} \lambda_{\alpha} &= 1 \end{aligned} \quad (\text{B-20})$$

The obtained ordinary kriging equations can also be presented in matrix form:

$$\mathbf{C} \lambda = \mathbf{b} \quad (\text{B-21})$$

or

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n_k} & 1 \\ C_{21} & C_{22} & \cdots & C_{2n_k} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C_{n_k 1} & C_{n_k 2} & \cdots & C_{n_k n_k} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n_k} \\ \psi \end{bmatrix} = \begin{bmatrix} C_{1\eta} \\ C_{2\eta} \\ \vdots \\ C_{n_k \eta} \\ 1 \end{bmatrix}. \quad (\text{B-22})$$

The ordinary kriging weights are obtained by inverting Eq. (B-21):

$$\lambda = \mathbf{C}^{-1} \mathbf{b}. \quad (\text{B-23})$$

Mathematical Principles of the Local Average Subdivision Method

C-1 Local Average Theory and Variance Reduction

This section mathematically captures the change of support effect for a one dimensional random process. The information, summarized in this section, is based on work of Griffiths and Fenton (2007) and (Fenton and Griffiths, 2008).

Let's consider a point random process $Z(\mathbf{x})$, which is averaged in a moving window with length T . This moving local average can be defined as:

$$Z_T(x) = \frac{1}{T} \int_{x-T/2}^{x+T/2} Z(\xi) d\xi, \quad (\text{C-1})$$

where $Z_T(x)$ is the local average of $Z(x)$ over length T . As the window is moved along in time or space x , the local average $Z_T(x)$ changes more slowly.

An example of such a local averaging process is displayed in Figure C-1. The upper part of this figure corresponds with the motion of a saw dust particle on the surface of the ocean. The motion of this particle is influenced by each wave and is highly variable. The lower plot, on the other hand, displays the motion of an ocean liner in the same waves. The liner does not bounce with every wave. Its motion is rather averaged out over its length, reducing the vertical variability in the record.

According to this figure, it becomes clear that local averaging can be considered as a low-pass filter. The ocean liner damps out the high-frequency components leaving wavelengths of the order of the length of the ship and longer. If the ocean waves are composed of just long-wavelength swells, then the motion of the sawdust particle and the ocean liner are the same. Conversely, if the waves are composed of short-wavelength components then the ocean liner does not move at all, while the sawdust particle moves up and down like a bouncing ball.

The amount of variance reduction is thus proportional to the high-frequency content present

Figure C-1: Effect of local averaging on variance. The upper plot displays the variability of a 1D point random process. The lower part is obtained by averaging this random process over a moving window with length T . Notice the reduction in the variability which typically can be described by the local average theory (Griffiths and Fenton, 2007, p. 60).

in the initial point signal. In other words, the variance reduction increases with an increased independence in the random field. To summarize, the two main effects of local averaging are the reduction of the variance in the signal and the damping of the high frequency components.

The mean of a certain signal however remains constant when changing the support. Figure C-1 clearly indicates that the mean elevation of the ocean linear and the saw dust particle are the same. Intuitively this is logical, since the average of local averages is the same as the average of the initial field. Mathematically, for stationary random fields, this can be proven as follows:

$$\begin{aligned}
 E[Z_T(x)] &= E\left[\frac{1}{T} \int_{x-T/2}^{x+T/2} Z(\xi) d\xi\right] \\
 &= \frac{1}{T} \int_{x-T/2}^{x+T/2} E[Z(\xi)] d\xi \\
 &= E[Z(x)].
 \end{aligned} \tag{C-2}$$

Putting it another way, local averaging preserves the mean of a random process. Previously a clear example already indicated that local averaging both smooths the process and reduces its variance. This observation will now be translated into a mathematical formula, which needs to express the variance of a local average as a function of its size. For a one dimensional local

average, this formula can be obtained as follows:

$$\begin{aligned}
 \text{Var}[Z_T(x)] &= \text{E}[(Z_T(x) - \mu_T)^2] \\
 &= \text{E}\left[\frac{1}{T} \int_0^T [Z(\xi) - \mu] d\xi \cdot \frac{1}{T} \int_0^T [Z(\psi) - \mu] d\psi\right] \\
 &= \frac{1}{T^2} \int_0^T \int_0^T C(\xi - \psi) d\xi d\psi \\
 &= \frac{\sigma^2}{T^2} \int_0^T \int_0^T \rho(\xi - \psi) d\xi d\psi \\
 &= \sigma^2 f(T),
 \end{aligned} \tag{C-3}$$

where $C(h)$ is the covariance function of $Z(x)$, and $\rho(h)$ is the correlation function of $Z(x)$ such that $C(h) = \sigma^2 \rho(h)$. In the last expression, $f(T)$ is the so-called variance reduction formula. For a specified length T , the value of $f(T)$ can be considered as a variance adjustment factor, which relates the amount of reduction to the length over which the local average was calculated.

When the length T is equal to zero, the adjustment factor is equal to one. As T increases the adjustment factor decreases towards zero. The factor is actually calculated as the average correlation coefficient over all possible pairs of points in the $[0, T]$ interval.

If the correlation function $\rho(h)$ falls off rapidly, indicating that there exists a greater independence in the random field, then the variance adjustment factor will be small. If on the other hand, there is a strong correlation between pairs of points separated by a relatively large distance, then $f(T)$ will be larger. The maximum value of the adjustment factor will be one, since in the ultimate case all possible pair of points are perfectly correlated, implying that $\rho(h)$ is equal to one for all distances. In this case, the field displays no variance reduction under local averaging.

The double integral to calculate the variance adjustment factor of Eq. (C-3) can further be simplified by considering the geometrical conditions of the integration over the square region $[0, T] \times [0, T]$ in the (ξ, ψ) space. Since the correlation coefficient $\rho(\xi - \psi)$ remains constant over each diagonal, it becomes possible to collapse the double integral to a single one. This single integration has to be performed in a direction perpendicular to these diagonal lines. For each integration interval, the correlation coefficient $\rho(\xi - \psi) = \rho(\tau)$ has to be multiplied by the differential area of a diagonal, which has a length $\sqrt{2}(T - |\tau|)$ and width $d\tau/\sqrt{2}$. Eventually the integration to calculate the variance adjustment factor can be written as:

$$f(T) = \frac{2}{T^2} \int_0^T (T - \tau) \rho(\tau) d\tau \tag{C-4}$$

In practice, it is neither convenient, nor feasible to calculate this integral analytically. An efficient and accurate approximation can be found in a Gaussian Quadrature evaluation:

$$f(T) \approx \frac{1}{T} \sum_{i=1}^{n_g} w_i (A - x_i) \rho(x_i), \tag{C-5}$$

where $x_i = \frac{1}{2}T(1 + z_i)$ and the Gauss points x_i and weights w_i are given in appendix C-5. The described local average theory is applied to formulate a correct variance for each different cell size encountered during the different calculation stages in the LAS simulation.

C-2 Variance and Covariance of Local averages

This section briefly illustrates how the covariance between two different local averages can be calculated, using the efficient and accurate Gaussian Quadrature approximation. The calculation method can be used to determine:

- the covariance between two cells of different size,
- the covariance between two cells of the same size,
- the variance of a single cell (two dimensional equivalent of Eq. (C-5)).

Consider the two general local averages A and B. The covariance between domain A and B can be approximated numerically as a Gaussian quadrature:

$$\begin{aligned}
 \text{Cov}[X_A, X_B] &= \frac{\sigma^2}{AB} \int_{a_{1y}}^{a_{2y}} \int_{a_{1x}}^{a_{2x}} \int_{b_{1y}}^{b_{2y}} \int_{b_{1x}}^{b_{2x}} \rho(x - \xi, y - \psi) d\xi d\psi \\
 &\approx \frac{\sigma^2}{16} \sum_{i=1}^{n_g} w_i \sum_{j=1}^{n_g} w_j \sum_{k=1}^{n_g} w_k \sum_{L=1}^{n_g} w_L \rho(x_j - \xi_L, y_j - \psi_k) \\
 \text{where } x_j &= a_{1x} + \frac{1}{2} A_x (1 + z_j), \quad \xi_L = b_{1x} + \frac{1}{2} B_x (1 + z_L) \\
 y_i &= a_{1y} + \frac{1}{2} A_y (1 + z_i), \quad \psi_k = b_{1y} + \frac{1}{2} B_y (1 + z_k) \quad (\text{C-6})
 \end{aligned}$$

where the weights w_i and the Gauss Points z_i can be found in Appendix C-5, A and B are the areas of both domains and $\rho(h_x, h_y)$ is the two-dimensional correlation function ($C(h_x, h_y) = \sigma^2 \rho(h_x, h_y)$).

C-3 One Dimensional Local Average Subdivision

This section shall introduce the concepts and the mathematical formulas necessary to generate a one-dimensional random process. Although final interest lies in the generation of a two-dimensional random field, the elaboration of the one-dimensional LAS technique will result in an improved understanding of the simulation methodology.

Notice that this section assumes a zero mean random process, which is not necessarily a restriction. The LAS simulation technique merely focuses on the reproduction of the spatial variability in a certain deposit. The mean value of the process is added later during the conditioning step.

The construction procedure of a one dimensional local average process, via the LAS technique, can best be compared with the structure of a tree diagram. Each level in the tree corresponds with a calculation stage during which each existing parent cell is further being subdivided into two child domains. The values in the child domains are generated in such a way as to ensure that the global average remains constant throughout the subdivision. This property is ensured by requesting that each generated pair is equivalent to the parent value. The procedure start at stage 0, where a global average is generated for the region of interest. Next, in stage 1, the global region is being subdivided into two domains and a value for each

Figure C-2: Top-down approach to generate one dimensional local average process. Stage 0, generate a global average; stage 1, subdivide domain into two regions whose values must in turn average to the global value; stage 2, further subdivide the stage 1 parent cells and generate local averages for the child cells. Subsequent stages are obtained by subdividing each parent cell and generating values for the resulting regions, while preserving upwards averaging (after Fenton and Vanmarcke (1990)).

child domain is generated. This iterative process is repeated until the size of the final child domains correspond with the specified resolution. The simulation methodology for a one-dimensional process is illustrated in Figure C-2. A more detailed description of the algorithm is printed below:

1. Generate a realization of the random variable which models the global average of the random process (Z_1^0). This unknown global average is modeled as a normal distribution with mean zero and a variance obtained from the local average theory (Eq. (C-3)).
2. Subdivide the global field Z_1^0 into two equal parts Z_1^1 and Z_2^1 .
3. The unknown local average of each part is modeled as a normally distributed random variable with mean and variance selected to satisfy the following three criteria:
 - (a) Each local average random variable has a variance which respects the local average theory.
 - (b) Both local average random variables are properly correlated with one another.
 - (c) Both local average random variables average to the parent value: $\frac{1}{2}(Z_1^1 + Z_2^1) = Z_1^0$.

Satisfying these three criteria will ensure that both distributions of the child variables are conditioned on the parent one. Generate, for each cell in this stage, a realization from the corresponding modeled distribution.

4. Subdivide each cell in stage 1 into two equal domains.
5. The unknown local averages of domain Z_1^2 and Z_2^2 are modeled as a normally distributed random variables with mean and variance selected to satisfy the following four criteria:
 - (a) Each local average random variable has a variance which respects the local average theory.
 - (b) Both local average random variables are properly correlated with one another.

- (c) Both local average random variables average to the parent value: $\frac{1}{2}(Z_2^1 + Z_2^2) = Z_1^1$.
- (d) Both local average random variables are properly correlated with Z_3^2 and Z_4^2 .

These four criteria will result in a distribution for both child variables which is conditioned on the direct and neighboring parent variables. Generate, for each cell in this stage, a realization from the corresponding modeled distribution.

The top-down approach is repeated until the size of the cells corresponds with the desired resolution. If the discretization size becomes small enough, the generated local average process will even become indistinguishable from the limiting continuous process. Often such a high resolution will not be necessary. Previous algorithmic description now has to be translated into mathematical formulas. Before starting with the mathematics, it is already possible and necessary to introduce two approximations.

- The first approximation states that a child value will be only explicitly conditioned on the direct and neighboring parent variables.
- The second approximation limits the range of parent cells on which to condition the distribution to a restricted neighborhood.

Both assumptions are remarkable, because it might appear that they violate criteria four of the last step. However, since child values are not always known during the generation of a new stage, it is necessary to resort to conditioning on parent cells. Much of the remainder of this section is devoted to the construction of these conditional distributions.

At this point, the previous story needs to be translated into a numerical algorithm. Since the numerical scheme proceeds in a recursive fashion that exhibits a lot of similarities, the numerical formulations will only be derived for a single representative cell. By looping over the proper indices, it is very straightforward to simulate the whole sequence of cells for all the different stages. Each cell in stage i has a width which is defined as:

$$T^i = \frac{T}{2^i}, \quad i = 0, 1, 2, \dots, L, \quad (\text{C-7})$$

such that the final realization contains $N = 2^L$ local averages. The k^{th} cell in this i^{th} stage is centered at location $x_k = (k - \frac{1}{2})T^i$ and extends over the interval $(k - 1)T^i < x \leq kT^i$. The average of this cell is calculated as:

$$Z_k^i = \int_{(k-1)T^i}^{kT^i} Z(\xi) d\xi. \quad (\text{C-8})$$

Previous equations should be sufficient to understand the indexing system applied throughout the rest of the section. Let us now consider the general case where the conditional distribution of cell $2j$ in stage $i + 1$ needs to be constructed based upon the available information of the previous stage i (Figure C-3). The probabilistic mechanism of the random variable, which models the uncertainty in the local average of that cell, can be written as the sum of a mean component m_{2j}^{i+1} and a zero mean discrete white noise $c^{i+1}U_j^{i+1}$:

$$Z_{2j}^{i+1} = m_{2j}^{i+1} + c^{i+1}U_j^{i+1}, \quad (\text{C-9})$$

	$j - 1$		j		$j + 1$		
			$2j - 1$	$2j$			

Figure C-3: (after Fenton and Vanmarcke (1990)).

c^{i+1} equals the standard deviation of the conditional distribution and U_j^{i+1} corresponds with the standard normal distribution $N(0,1)$. The best linear estimate for the mean m_{2j}^{i+1} is calculated as a linear combination of the direct and adjacent generated parent values. To reduce computation time and memory usage, while preserving an acceptable level of accuracy, the stage i values included in the summation are extracted from a limited neighborhood: $j - n, \dots, j + n$. The mean is calculated as:

$$m_{2j}^{i+1} = \sum_{k=j-n}^{j+n} a_{k-j}^i z_k^i. \quad (\text{C-10})$$

Notice the switch to the small letter notation z_k^i , which indicates that this is a linear combination of single generated values. The next step in this analysis is the construction of the simple kriging equations which are used to determine the coefficients $a_l^i, l = -n, \dots, n$. The derivation of the necessary simple kriging equations could be performed in a similar manner as explained in Chapter 2 and Appendix B. However, in this case there is a more convenient and fairly straightforward way of obtaining these equations.

First start with the formulation of the conditional distribution of cell $2j$ as a linear combination of parent random variables:

$$Z_{2j}^{i+1} = \sum_{k=j-n}^{j+n} a_{k-j}^i Z_k^i. \quad (\text{C-11})$$

The second step consists of the multiplication of this equation with each Z_m^i parent variable located in the selected neighborhood. The third and final step consists of taking the expectation of each obtained equation. This procedure results in following system of equations from which the coefficients $a_l^i, l = -n, \dots, n$ can be solved:

$$E[Z_{2j}^{i+1} Z_m^i] = \sum_{k=j-n}^{j+n} a_{k-j}^i E[Z_k^i Z_m^i] \quad m = j - n, \dots, j + n \quad (\text{C-12})$$

The obtained set of simple kriging equations differs in three ways from the previously formulated ordinary kriging equations (Eq. (2-11)). The first obvious difference is the lack of the Lagrange parameter and the associated additional equation. Both are no longer necessary since the weights are not anymore constrained to sum up one. This difference mainly results from the fact that the mean of the process is assumed to be known.

The assumption of a constant zero mean brings us to the second remarkable difference. Instead of a formulation containing covariances, Eq. (C-12) is build up from expectations of products. Both formulations are merely the same since:

$$\begin{aligned} \text{Cov}(A, B) &= E[(A - E[A])(B - E[B])] \\ &= E[AB] - E[A]E[B] \\ &= E[AB] \end{aligned} \quad (\text{C-13})$$

and both $E[A]$ and $E[B]$ are zero. Thus each expectation in Eq. (C-12) represents the covariance between either the cell of interest and a certain parent cells and or between a pair of parent cells. The third difference with the previous derived ordinary kriging equations is that covariances in this case needs to be considered over two specified regions, which do not necessarily need to be of equal size.

To finalize the one-dimensional LAS simulation methodology, it is necessary to formulate some mathematical equations which can be used in a convenient way to calculate the covariances between the different regions. The covariance between two cells of equal size in stage i separated by a distance mT^i between centers is:

$$\begin{aligned} E[Z_k^i Z_{k+m}^i] &= E \left[\frac{1}{(T^i)^2} \int_{(k-1)T^i}^{kT^i} \int_{(k+m-1)T^i}^{(k+m)T^i} Z(\xi) Z(\psi) d\xi d\psi \right] \\ &= \frac{1}{(T^i)^2} \int_{(k-1)T^i}^{kT^i} \int_{(k+m-1)T^i}^{(k+m)T^i} C(\xi - \psi) d\xi d\psi \end{aligned} \quad (C-14)$$

This double integral can be collapsed to a single integral, in a similar way as was used to move from Eq. (C-3) to Eq. (C-4). The resulting single integral can subsequently be approximated by a Gaussian quadrature:

$$\begin{aligned} E[Z_k^i Z_{k+m}^i] &\approx \frac{1}{4} \sum_{\nu=1}^{n_g} w_\nu [(1 + z_\nu) C(r_\nu) + (1 - z_\nu) C(s_\nu)] \\ \text{where } r_\nu &= T^i(m - \frac{1}{2}(1 - z_\nu)) \\ s_\nu &= T^i(m + \frac{1}{2}(1 + z_\nu)), \end{aligned} \quad (C-15)$$

and the weights w_ν , and positions z_ν can be found in Appendix C-5 for n_g Gauss Points. Previous equation provides us with a efficient and accurate way to calculate the covariance between the parent cells in the right hand side of Eq. (C-12). To be able to calculate the covariance between the cell of interest and a parent cell, a simple trick can be introduced which enables us to keep on using the fast Gaussian Quadrature approximation. The trick lies in the fact that adjacent cell Z_{2j-1}^{i+1} is determined by the requirement that the global average is preserved, i.e. that each child value pair in stage $i+1$ averages to its corresponding parent i parent value. The value of the adjacent cells is thus:

$$Z_{2j-1}^{i+1} = 2Z_j^i - Z_{2j}^{i+1}. \quad (C-16)$$

This requirement gives a mean of evaluating the cross-stage covariances in Eq. (C-12):

$$E[Z_{2j}^{i+1} Z_m^i] = \frac{1}{2} E[Z_{2j}^{i+1} Z_{2m-1}^{i+1}] + \frac{1}{2} E[Z_{2j}^{i+1} Z_{2m}^{i+1}]. \quad (C-17)$$

Now that all the desired covariances can be calculated, it becomes possible to formulate the m kriging equations with m unknown weighting coefficients (Eq. (C-12)). These kriging equations can be solved quite easily by some basic techniques in linear algebra. The resulting set of weights can then be used to calculate the mean of the conditional distribution of the cell $2j$. By using following equation and plugging in the required covariances, one can calculate the conditional variance of the cell:

$$c^{i+1,2} = E[(Z_{2j}^{i+1})^2] - \sum_{k=j-n}^{j+n} a_{k-j}^i E[Z_{2j}^{i+1} Z_k^i]. \quad (C-18)$$

Generating a random value from the standard normal distribution $N(0, 1)$ and plugging it into Eq. (C-9) together with the calculated mean and variance will result in a simulated value of cell $2j$ of stage $i + 1$. This process has to be repeated in the different calculation stage for all child pairs. There are however some overhead calculations which have to be performed only once. For stationary processes, the set of coefficients a_l^i and variance c^i are independent of absolute position and thus need to be evaluated only once. The complete simulation framework can be summarized as follows:

1. Iterate over the necessary stage while performing the overhead calculations: for $i = 0, 1, 2, \dots, L$ compute the set of coefficients $\{a_l^i\}, l = -n, \dots, n$ using Eq. (C-12) and the conditional variances c^{i+1} using Eq. (C-18).
2. For stage 0, generate a global mean using Eq. (C-1) and Eq. (C-3).
3. Subdivide the domain.
4. For each $j = 1, 2, 3, \dots, 2^i$ calculate m_{2j}^{i+1} using Eq. (C-10) and subsequently generate a realization for Z_{2j}^{i+1} using Eq. (C-9). Then calculate Z_{2j-1}^{i+1} using Eq. (C-16).
5. increment i and if not greater than L , return to step 3.

C-4 Conditioning of the Random Field

The previously introduced LAS simulation methodology is ideally suited to simulate the spatial variability of the modeled random field. Although it is important to model the correct absolute difference between certain local averages in space, it is also essential to model the absolute value of the individual local averages correctly. What actually needs to happen is an elevation, some sort of local translation, of the modeled zero mean random process to the measured values in the field. This procedure is called the conditioning of the random field. The conditioning step also ensures that the simulated random field takes on certain known values at specific locations in the field.

This section seeks to explain how the simulated spatial variability field $Z_u(\mathbf{x})$ has to be conditioned on the measurement data $\mathbf{x}_\alpha$, $\alpha = 1, \dots, n_k$, where n_k is the number of measurement locations. This operation will result in a conditional field $Z_c(\mathbf{x})$ and can conceptually be written as:

$$Z_c(\mathbf{x}) = [Z_u(\mathbf{x}) | z(\mathbf{x}_\alpha), \alpha = 1, 2, \dots, n_k]. \quad (\text{C-19})$$

The vertical bar in previous formula is similar to the well known conditioning operator of Bayesian statistics. To accomplish the conditioning operation, the final simulated collection of local average cells needs to be partitioned into two categories: (1) $\mathbf{x}_\alpha$, $\alpha = 1, \dots, n_k$ being those locations at which measurement data is available and at which the random field takes on the deterministic values $z(\mathbf{x})_\alpha$, and (2) $\mathbf{x}_\eta$, $\eta = 1, 2, \dots, N - n_k$ being those locations at which random realizations need to be generated from the modeled conditional distribution. That is the subscript α denotes the locations with known values, while the subscript η denotes the locations with unknown values which need to be simulated. N is the total number of cells

in the field to be simulated. The necessary conditioning operation is simply formed by three components:

$$Z_c(\mathbf{x}) = Z_u(\mathbf{x}) + [Z_k(\mathbf{x}) - Z_s(\mathbf{x})] \quad (\text{C-20})$$

where,

- $Z_c(\mathbf{x})$ = desired conditional simulation,
- $Z_u(\mathbf{x})$ = unconditional simulation,
- $Z_k(\mathbf{x})$ = best linear unbiased estimate of the field based on the known (measured) values at $\mathbf{x}_\alpha$,
- $Z_s(\mathbf{x})$ = best linear unbiased estimate of the field based on the unconditional simulation values at $\mathbf{x}_\alpha$.

Is is fairly straightforward to prove that this conditioning procedure would return for each known location the correct measured value. Due to the properties of the BLUE calculation $Z_k(\mathbf{x}_\alpha)$ equals $z(\mathbf{x}_\alpha)$ at the measurement location, while $Z_s(\mathbf{x}_\alpha)$ simply is given as $Z_u(\mathbf{x}_\alpha)$. Inserting these obtained results in Eq. (C-20) will prove the formulated statement:

$$\begin{aligned} Z_c(\mathbf{x}) &= Z_u(\mathbf{x}_\alpha) + [Z_k(\mathbf{x}_\alpha) - Z_s(\mathbf{x}_\alpha)] \\ &= Z_u(\mathbf{x}_\alpha) + [z(\mathbf{x}_\alpha) - Z_u(\mathbf{x}_\alpha)] = z(\mathbf{x}_\alpha) \end{aligned} \quad (\text{C-21})$$

In other words, at each sampling location, the conditioning algorithm will always generate the corresponding measured value.

In order to obtain an conditional simulated value at one of the remaining $N - n_k$ simulation locations $\mathbf{x}_\eta$, Eq. (C-20) needs to be evaluated. The first component on the right hand side of this conditioning formula is generated by an unconditional simulation technique. In this research, the unconditional simulated component $Z_u(\mathbf{x})$ is obtained through the application of LAS method, and represents the spatial variability of the modeled random field.

The second component is estimated as the sum of the unconditional mean μ_η at location $\mathbf{x}_\eta$ and a linear combination of the differences between each measured value and the corresponding unconditional mean μ_α :

$$Z_k(\mathbf{x}_\eta) = \mu_\eta + \sum_{\alpha=1}^{n_k} \beta_\alpha (z(\mathbf{x}_\alpha) - \mu_\alpha) \quad (\text{C-22})$$

for $\eta = 1, 2, \dots, N - n_k$. In particular, this BLUE field is based on the measured value at each sampling locations.

The third and final component is calculated as the sum of the unconditional mean μ_η at location $\mathbf{x}_\eta$ and a linear combination of differences between the unconditional simulated value and the unconditional mean at each $\mathbf{x}_\alpha$:

$$Z_s(\mathbf{x}_\eta) = \mu_\eta + \sum_{\alpha=1}^{n_k} \beta_\alpha (Z_u(\mathbf{x}_\alpha) - \mu_\alpha), \quad (\text{C-23})$$

for $\eta = 1, \dots, N - n_k$. In particular, this BLUE field is based on the unconditional simulated value at each sampling location. The only substantial difference between $Z_k(\mathbf{x})$ and $Z_s(\mathbf{x})$,

is that the first is based on actual measurement data, while the latter is calculated using simulated cell values for the same locations. The difference between both components can be evaluated in a more efficient way, since the unconditional mean terms would cancel out. This leads to:

$$Z_k(\mathbf{x}_\eta) - Z_s(\mathbf{x}_\eta) = \sum_{\alpha=1}^{n_k} \beta_\alpha (z(\mathbf{x}_\alpha) - Z_u(\mathbf{x}_\alpha)) \quad (\text{C-24})$$

for $\eta = 1, \dots, N - n_k$. The weights β_α , necessary to model this linear combination of differences, are determined from:

$$\beta = \mathbf{C}^{-1} \mathbf{b}, \quad (\text{C-25})$$

where β is the vector containing the n_k weighting coefficients β_α , $\mathbf{C}$ is the $n_k \times n_k$ matrix containing the covariances between local average cells located over sampling points. This matrix $\mathbf{C}$ is build up from following components:

$$C_{\alpha\beta} = \text{Cov}[Z(\mathbf{x}_\alpha), Z(\mathbf{x}_\beta)] \quad (\text{C-26})$$

for $\alpha, \beta = 1, \dots, n_k$. The $\mathbf{b}$ vector in previous equation, is a vector of length n_k containing the covariance between the local average cells located at the current prediction point of interest $\mathbf{x}_\eta$ and the other sampling points:

$$b_\alpha = \text{Cov}[Z_u(\mathbf{x}_\eta), Z(\mathbf{x}_\alpha)] \quad (\text{C-27})$$

for $\alpha = 1, \dots, n_k$. The covariances of Eq. (C-26) and Eq. (C-27) can be calculated efficiently using the Gaussian quadrature formulation of Eq. (C-6). Since the matrix $\mathbf{C}$ is dependent on the covariances between known points, its inverse only needs to be calculated once and can repeatedly being used in Eq. (C-25). Moreover since both matrices $\mathbf{C}$ and $\mathbf{b}$ are not related with any measured or simulated values, they can repeatedly being used over several simulation loops. Clearly stated, this means that the matrix $\mathbf{C}$ needs to be constructed once and that the vector $\mathbf{b}$ needs to be constructed for each specific cell. This means that each cell has a specific set of weighting coefficients which remains constant over several simulation loops. The whole conditioning operation can be summarized by following task list:

1. Partition field into the known (α) and unknown η cells
2. Form the matrix $\mathbf{C}$ which contains the covariances between the cells located over the known locations. Invert the matrix to determine $\mathbf{C}^{-1}$.
3. Simulate unconditional random field $Z_u(\mathbf{x})$ at all cells.
4. For each unknown cell value, $\eta = 1, \dots, N - n_k$:
 - (a) Form the vector $\mathbf{b}$ which contains the covariances between the prediction cell and the other cell located over sampling locations.
 - (b) Calculate the weighting coefficients $\beta_\alpha, \alpha = 1, 2, \dots, n_k$ by solving Eq. (C-25).
 - (c) Compute the difference $Z_k(\mathbf{x}_\eta) - Z_s(\mathbf{x}_\eta)$ using Eq. (C-24).
 - (d) Compute the conditional random field $Z_c(\mathbf{x}_\eta)$ at location $\mathbf{x}_\eta$ using Eq. (C-20).
 - (e) Increment η and if not greater than $N - n_k$, jump to sub-step (a).

C-5 Weights and Gauss Points

The weights w_i and standardized locations z_i , necessary to evaluate the Gaussian quadrature approximation, are displayed in Table C-1.

Table C-1: Weights and Gauss points to evaluate the Gaussian quadrature approximation.

	w_i	z_i
0.027152459411754	-0.989400934991649	
0.062253523938648	-0.944575023073232	
0.095158511682493	-0.865631202387831	
0.124628971255533	-0.755404408355003	
0.149595988816576	-0.617876244402643	
0.169156519395002	-0.458016777657227	
0.182603415044923	-0.281603550779258	
0.189450610455068	-0.095012509837637	
0.189450610455068	0.095012509837637	
0.182603415044923	0.281603550779258	
0.169156519395002	0.458016777657227	
0.149595988816576	0.617876244402643	
0.124628971255533	0.755404408355003	
0.095158511682493	0.865631202387831	
0.062253523938648	0.944575023073232	
0.027152459411754	0.989400934991649	

Appendix D

Detailed Discussion on the Data and Area Statistics

The inference process, embedded in each “area and data statistics” box aims at estimating some relevant summary statistics of the studied area or population. Remember that, during the analysis, the population is represented by the same collection of data points with a changing support and projection space.

The use of sample statistics, as estimates for population parameters, requires that the collected data is representative for the area under investigation. Such a representativeness can be achieved by carefully designing the sampling scheme (Goovaerts, 1997; Isaaks and Srivastava, 1989). However, such a sampling scheme is rarely implemented, since the interest of the explorers deviates substantially from the theoretical requirements of an optimal sampling design.

Possible bias, due to preferential sampling, can be removed using a declustering algorithm. This algorithm returns a set of weights, which are subsequently used in each “area and data statistics” box to convert certain sample statistics into representative population parameters. The outcome of the first box (box A, Figure 4-1) gives a primary insight into the characteristics of the investigated mineral deposit. This primary investigation is also known as the Exploratory Data Analysis (EDA).

This section contains a brief explanation on the calculation of several relevant global and spatial statistics. The text discusses how the conventional formulas have to be adjusted in order to be able to compute representative population parameters for both gridded and declustered data. The global and spatial statistics are generally calculated as some sort of weighted linear combination of collected data values. The distinction with the conventional sample statistics lies in the fact that each data point now receives a location dependent weight, which does not necessarily have to be constant anymore. Later on, it will become clear why and how these declustering weight have to be calculated. Now, the focus just lies on the set of tools stored in each “area and data statistics” box. The following discussion is subdivided into an characterization of the general data properties and an analysis of the spatial relationships.

D-1 Data or Global Statistics

Probably, the most well known statistical value is the mean. The mean or arithmetic average is calculated as an equal-weighted linear combination of all available data values. Each data value receives a constant weight, which amounts the inverse of the total number of data points. The resulting linear combination is thus nothing more than the sum of all available data values divided by the total number of measurements made. To correct for possible bias, each weight $1/n_k$ in the conventional formula, needs to be replaced by λ_α , a weight that has the potential of becoming location dependent. This correction yields following formula for the (declustered) mean:

$$m = \sum_{\alpha=1}^{n_k} \lambda_\alpha \cdot z(\mathbf{x}_\alpha). \quad (\text{D-1})$$

Remark, that the conventional formula of the mean reappears, if each data point receives an equal weight $1/n_k$.

By analogy, the variance formula can also be adjusted. Conventionally, the variance is calculated as the average squared difference of the observations from their mean. Generally this average is calculated with each component receiving an equal weight. To obtain the formula for the (declustered) variance, the equal weights $1/n_k$ need to be replaced by the variable declustering weights λ_α :

$$\sigma^2 = \sum_{\alpha=1}^{n_k} \lambda_\alpha \cdot [z(\mathbf{x}_\alpha) - m]^2, \quad (\text{D-2})$$

where m corresponds with the (declustered) mean, defined as in Eq. (D-1).

Since the calculation involves squared differences, the variance is very sensitive to erratic or disproportionate high values. Immediately, it becomes clear why a correction needs to be applied, if e.g. high values were preferentially sampled. Before discussing the histogram and the cumulative distribution function (cdf), it is important to mention that the n_k weights λ_α need to have values between zero and one, and that they need to sum up to one as well.

The histogram or frequency distribution The distribution of continuous data values is generally depicted by a histogram, with the range of data values discretized into a specific number of classes, each of equal width. The proportion of data falling within each class is expressed by the height of the bars (Figure ??). A frequency distribution or histogram is created as follows:

- Arrange the collected data values in ascending order.
- Group data values together in their corresponding classes.
- Calculate the proportion or frequency of each class by summing the weights λ_α of the data points belonging to that particular class. Multiply with hundred to express the frequency as a percentage.
- Construct the bars as rectangles of equal width. The heights of the bars must equal the calculated frequency. Position each bar within its specific class range.

Figure D-1: Frequency distribution of Walker Lake Exploration Data. The proportion of data values falling within each class is expressed by the heights of the bars. The width of each bar is 50 ppm and represents the range of the classes. Each data point received an equal weights during the construction of this frequency distribution.

Normally during the construction of the conventional histogram, each data point receives an equal weights, $1/n_k$. In contrast, the previous described methodology can also be used to construct a representative histogram for biased data. The applied correction, implemented to remove the negative effects of preferential sampling, just involves location dependent declustering weights λ_α .

Figure ?? shows a histogram of the Walker Lake Exploration Data. To construct the displayed frequency distribution, each data value received an equal weight. That is, each λ_α equals $1/n_k$. Later on, it will become clear that this conventional frequency distribution is not representative for the underlying population. Instead, a short analysis will indicate that a declustering algorithm is needed to correct the weights and make them location dependent.

cumulative distribution function Another convenient illustration, that is able to capture and summarize certain interesting characteristics of the data distribution, is the plot of the cumulative frequencies. A cumulative frequency can be regarded as a value-proportion pair that indicates the amount of data points that do not exceed the specified value. Those frequencies are very useful to study the proportion of the population, in this case the area of investigation, that has an attribute value below or above a critical threshold value. Such a threshold value may indicate the grade below which mining is not profitable, or a concentration of a pollutant above which remedial actions have to be taken.

The resulting cumulative distribution function (cdf) does not only provide an answer to the threshold question, but also gives an insight into the spread of the underlying data or population. Moreover, the well-known p -quantile values, which are cumulative frequencies corresponding to certain specified proportions, are ideally suited to report confidence intervals. A p -quantile value of the distribution, denoted q_p , is thus the value that is not exceeded by a proportion p of the data, i.e. q_p is such that $F(q_p) = p$, where $F()$ denotes the cdf. It is obvious that the median is the same as the 0.5 quantile. Other quantiles of widespread use are the nine deciles that correspond to cumulative frequencies 0.1, 0.2, ..., 0.9. To calculate

Figure D-2: Cumulative Distribution Function of Walker Lake Exploration Data. The cumulative frequency indicates the proportion (y-axes) of the exploration data that does not exceeds the corresponding value (x-axis). The dashed line connects the cumulative frequencies of all available sorted data points. The crosses further indicate the specific p-quantiles, ranging from 0.05, with steps of 0.05, to 0.95. During the calculation of this cdf, each value received a weight of $1/470$, where 470 is the total number of data points.

the cdf, previously denoted as $F()$, followings steps have to be taken:

- Arrange the collected data values in ascending order.
- Calculate for each data value, included in the ordered set, the (declustered) cumulative frequency (Goovaerts, 1997):

$$F(z_\beta) = \sum_{\alpha=1}^{\beta} \lambda_\alpha - 0.5\lambda_\beta \quad \text{for } \beta = 1, \dots, n_k. \quad (\text{D-3})$$

- Multiply with hundred to express the frequency as a percentage.
- Plot the calculated cumulative frequencies versus the sorted data values.
- Calculate, using linear interpolation, the set of requested p-quantiles.

Normally, the construction of a conventional cdf requires that each data point receives an equal weight $1/n_k$. This implies that, in order to calculate for example the cumulative frequency of the fifth value of the ordered set, the sum of five equal weights $1/n_k$ needs to be taken. However, in the case of preferential sampling, it is vital that each data point receives a location dependent declustering weights λ_α . Eq. (D-3) is written in such a way that it can be used to calculate both the declustered as the conventional cdf.

Figure ?? illustrates the calculated cdf of the Walker Lake exploration data. This cdf is calculated without the use of declustering weights. As mentioned before, the Walker Lake exploration data set is biased and thus any calculated statistic that involves equal weights will not be representative of the underlying population.

The discussed group of global statistics probably contains sufficient techniques to fully characterize the general properties of a data set. Now time has come to introduce some procedures to investigate the spatial correlation structures inherently present in the area under investigation.

Figure D-3: Basic spatial description of the Walker Lake exploration data. Color coded data posting; Each specific measurement is indicated by a color coded dot. The color of the dot corresponds with the measured value (left). Moving window statistics; Posting of statistics obtained from 25 non-overlapping $60 \times 60 \text{ m}^2$ moving windows. The mean of each moving window is plotted above the '+', and the standard deviation is plotted below. The number of data points within each moving window varies from 7 to 55. During this analysis, the data points within each window all received equal weights (right).

D-2 Area or Spatial Statistics

An earth science data set distinguishes itself from others, in that each measured value belongs to some location in space. Spatial features, such as the location(s) of extreme values, the overall trend, or the degree of continuity are often of considerable interest. None of the previously introduced descriptive global estimates are able to capture these spatial features. Hence, this subsection is devoted to some techniques which are frequently used to investigate the spatial continuity and variability of exploration data.

Data posting and moving window statistics The most effective tools to discover the spatial features hidden in the exploration data are based on visual interpretation. The simplest visualization of the spatial data is a data posting, a map on which each data location is plotted as a color coded dot. The color of each dot corresponds with the measured value at that specific location.

The left hand side of Figure D-3 shows an example of such a posting map. The construction of such a map really fits well in a quality assurance philosophy. A posting maps reveals obvious errors in the data locations, and often draws attention to data values that might be erroneous. A single isolated high value surrounded by multiple low values, and vice versa, are worth checking. Often, a posting map also gives a clue on how the data was collected.

During the analysis of exploration data, one is often interested in spatial anomalies, e.g. high grade zones or strong rock formations. It is definitely eligible to focus on such anomalous zones, since they have a direct positive or negative impact on the mining process. A posting map or contour plot will definitely help to locate these anomalous zones.

However, interest not only lies in the characterization of areas with anomalous average values,

but also in the identification of regions with a high variability. A strong variability might have serious practical implications, and therefore it needs to be characterized correctly. The moving window procedure is frequently used to investigate anomalies, both in average value as in variability (Isaaks and Srivastava, 1989). The technique is based on the calculation of a few global summary statistics within a moving window or local neighborhood. The procedure is fairly straightforward and proceeds as follows:

1. Dimension a window or local neighborhood, specify the height and width.
2. Design a discrete sequence according which the window has to move over the area under investigation.
3. Collect the β data points which are located within the borders of the defined window. Store the β data values $z(\mathbf{x}_{\alpha'})$ and their corresponding weights $\lambda_{\alpha'}$ in a separate array. The apostrophe indicates that the current collection of data points is just a fraction of the complete data set.
4. Normalize the weights stored in the array: $\lambda_{\alpha'} = \lambda_{\alpha} / \sum_{\alpha'=1}^{\beta} \lambda_{\alpha'}$ such that $\sum \lambda_{\alpha'} = 1$.
5. Calculate the (declustered) mean and the standard deviation of the current collection of data values. Insert the weights $\lambda_{\alpha'}$ and the values $z(\mathbf{x}_{\alpha'})$ into Eq. (D-1) and Eq. (D-2).
6. Translate the window to the next position in the sequence.
7. Repeat step 3 to 6 until all positions of the designed sequence are visited.
8. Construct a posting map of the calculated moving window statistics.

Rectangular windows are often used, mainly for reasons of computational efficiency and ease of implementation. The dimensioning step is the most critical one and requires a good compromise between the amount of data needed to calculate reliable summary statistics and the number of windows needed to ensure the detailed localization of anomalous areas. The size of the window eventually depends on the average spacing of the sampling locations and the overall dimension of the investigated area.

Sometimes, it is useful to design the sequence in such a way that subsequent windows partially overlap. Such a strategy results in a set of reliable statistics for a sufficient number of windows, ensuring the required level of detail (Isaaks and Srivastava, 1989). In such a situation, multiple adjacent local neighborhoods might have data in common. The right hand side of Figure ?? gives an example of the results obtained from a non-overlapping moving window analysis.

Once, some moving window statistics are calculated, it is recommended to check for the presence of a feature known as heteroscedasticity. Such a feature implies a changing local variability across the investigated area. The proportional effect, which is a specific form of heteroscedasticity, states that the local variance of the data is related with the local mean. For a positively skewed distribution, the variance often increases in function of the local mean (direct proportional effect) (Goovaerts, 1997). A proportional effect, in general, can be easily detected through the construction of a scatterplot displaying local means versus local variances. Figure ?? shows such a scatterplot, displaying data obtained from the 25 local neighborhoods shown in the right hand side of Figure ??.

It is important to consider the proportional effect, since it will have an immediate effect on

Figure D-4: Proportional effect of Walker Lake exploration data; Plot of local standard deviation versus local mean computed from 25 non-overlapping $60 \times 60 \text{ m}^2$ local neighborhoods laid across the study area. The number of data points used to calculate each statistic varies from 7 to 55. The data points within each window all received equal weights.

the accuracy of local estimations. Intuitively, it is apparent that local estimations in areas with very uniform values have a higher change to be successful than estimations in areas with very fluctuating values.

The h-scattergram Another convenient tool, to explore the spatial relations hidden in the measurement data, is also based on the principles of the scatterplot. The so called h-scattergram displays all possible data pairs which are separated by a specified displacement vector $\mathbf{h}$. This displacement vector is characterized by its unique combination of distance h and direction θ . In order to build the h-scattergram on a sufficient number of pairs, the displacement vectors are grouped together into classes of similar distances (lags) $[h \pm \Delta h]$ and directions $[\theta \pm \Delta \theta]$. By convention, the value at the start of the displacement vector $\mathbf{h}$ is called the tail value and is displayed on the abscissa, whereas the value at the end is called the head value and is displayed on the ordinate. Because the two axes are related to the same variable, though at a different location, a perfect correlation would entail that all the points lie on the 45° line.

When the points, displayed in the h-scattergram, become more dissimilar the cloud gets fatter and more diffuse. The shape of the cloud thus tells us how continuous the data values are over a certain distance in a particular direction. Figure D-5 shows an example of a scatterplot, displaying data pairs which are separated by about 40m.

By constructing several h-scattergrams for a specific selection of distances and directions, it is possible to investigate the spatial relation in function of the displacement vector. Normally, the inflation of the cloud would increase with increasing separation distance. Such an inflation would reflect the increasing dissimilarity between measurements farther apart.

Now time has come to introduce some techniques, which can summarize the visual conclusions into quantitative information. The common idea, behind the proposed techniques, is to capture the fatness of each cloud by a single correlation related number. This calculation can be repeated for each class of displacement vectors and finally results in a graph that visualizes the spatial relation in function of distance and direction.

Figure D-5: h-scattergram of gridded normal scores located 40m apart (angle tolerance = 180°, distance tolerance = 5 m). The shape of the cloud of points on the h-scattergram tells us how continuous the data values are over a certain distance in a particular direction. The vertical solid line represent the mean tail value, while the horizontal solid line indicates the value of the mean head value. The dashed lines indicate the boundaries of the respective 68% confidence intervals.

Semi-variance, covariance and correlogram As pointed out previously, one of the essential features of an h-scattergram is the fatness of the cloud. This fatness can be captured by the moment of inertia around the first bisector. The moment of inertia is actually nothing more than the average of all squared orthogonal distances between the points and the bisector. The resulting geometrical tedious formulation can be further simplified to a formula stating that the dissimilarity of the data can be expressed as half the average squared difference between the components of every data pair. This so called experimental semi-variance expresses how much the data points drift away from the 45° line. The semi-variance for a particular displacement vector $\mathbf{h}$ can thus be calculated as:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{x}_{\alpha}) - z(\mathbf{x}_{\alpha+\mathbf{h}})]^2, \quad (\text{D-4})$$

where $N(\mathbf{h})$ represents the number of data points plotted in the corresponding h-scattergram. An ordered set of calculated semi-variances, displayed in function of their corresponding lag distance, is called an experimental semi-variogram. In general, the semi-variance increases with decreasing dissimilarity. The left hand side of Figure D-6 shows two examples of such semi-variograms.

An alternative index to express the fatness of the cloud or the spatial continuity is the covariance. The conventional covariance coefficient was already introduced in Appendix A-1 and can now be adapted to measure the similarity between data points. The covariance between $N(\mathbf{h})$ data pairs separated by a vector $\mathbf{h}$ is computed as follows:

$$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} z(\mathbf{x}_{\alpha}) \cdot z(\mathbf{x}_{\alpha+\mathbf{h}}) - m_{-\mathbf{h}} \cdot m_{+\mathbf{h}}, \quad (\text{D-5})$$

Figure D-6: Spatial description of Walker Lake data set. Experimental semi-variance function or semi-variogram showing the calculated semi-variances in function of the lag distance (left). Experimental (auto)covariance function showing the calculated covariances in function of the lag distance (right). A green colored symbol (+) indicates a calculated statistic which was obtained from regularly spaced gridded data points. The red symbols (x), on the other hand, correspond with calculated statistics of the complete exploration data set, also including clustered data points.

where m_{-h} and m_{+h} are the means of the tail and head values belonging to the specified class of distance and direction. An ordered set of computed covariances for different lags is called an experimental (auto)covariance function. In general, the covariance coefficient decreases with decreasing similarity. Eventually the covariance coefficient reaches a value of zero, indicating that after a certain distance, the data becomes uncorrelated. The right hand side of Figure ?? shows an example of two series of calculated experimental covariance functions.

Finally, also a unit free measure of similarity is available. This measure is derived from the well known correlation coefficient and is actually a standardized form of the covariance function. The resulting function expresses the relation between the correlation coefficient of an h-scattergram and the displacement vector h . A correlation coefficient of one corresponds with perfect correlation, while a value of zero denotes that the set of values is spatially uncorrelated. An ordered set of computed correlation coefficients for different lags is called the experimental (auto)correlation function or correlogram and is calculated using following formula:

$$\rho(h) = \frac{C(h)}{\sqrt{\sigma_{-h}^2 \cdot \sigma_{+h}^2}}, \quad (D-6)$$

where σ_{-h}^2 and σ_{+h}^2 are the variance of the tail and head values. Figure ?? shows an example of two series of calculated experimental correlation functions.

Lag mean and lag variance Not only the fatness of the cloud displayed in a h-scattergram is important, also the location and the extent of the cloud need to be investigated. The position of the cloud can be characterized by a coordinate pair consisting of the mean tail value and the mean head value. This coordinate pair indicates the location of the cloud's midpoint. The

Figure D-7: Experimental (auto)correlation function or correlogram of Walker Lake data set; A green colored symbol (+) indicates a calculated statistic which was obtained from regularly spaced gridded data points. The red symbols (x), on the other hand, correspond with calculated statistics of the complete exploration data set, also including clustered data points.

declustered tail mean ($m_{-\mathbf{h}}$) and head mean ($m_{+\mathbf{h}}$) are calculated using following formulas:

$$m_{-\mathbf{h}} = \sum_{\alpha=1}^{N(\mathbf{h})} w_{\alpha} \cdot z(\mathbf{x}_{\alpha}) \quad (\text{D-7})$$

$$m_{+\mathbf{h}} = \sum_{\alpha=1}^{N(\mathbf{h})} w_{\alpha+\mathbf{h}} \cdot z(\mathbf{x}_{\alpha} + \mathbf{h}). \quad (\text{D-8})$$

The weights w_{α} and $w_{\alpha+\mathbf{h}}$ correspond with the standardized declustering weights of respectively the tail and head values. The previously introduced declustering weights λ_{α} are normalized by the sum of all the weights belonging to the corresponding side of the displacement vector.

$$w_{\alpha} = \frac{\lambda_{\alpha}}{\sum_{\alpha=1}^{N(\mathbf{h})} \lambda_{\alpha}} \quad w_{\alpha+\mathbf{h}} = \frac{\lambda_{\alpha+\mathbf{h}}}{\sum_{\alpha=1}^{N(\mathbf{h})} \lambda_{\alpha+\mathbf{h}}} \quad (\text{D-9})$$

Remark that these weights would remain constant if the analysis of the h-scattergram is based on a selection of non-declustered data. According to the stationarity assumptions, it is preferential that these mean values remain constant with changing lag distance.

Subsequently, one can average both calculated mean values and plot them against their corresponding lag distance. An example of the resulting graph is displayed in left hand side of Figure D-8. The graph actually illustrates the results of three lag mean computations, calculated according different declustering strategies, but all based on the same initial data set. The usefulness of such a plot will be explained later.

The vertical and horizontal extent of the cloud, on the other hand, can be defined using the standard deviation of the tail and head values. These standard deviations can subsequently being used to construct a confidence interval around the corresponding mean value (Figure ??). The declustered standard deviations of the tail and head values are

Figure D-8: Lag mean and lag variance of Walker Lake exploration data. Lag mean as a function of lag distance (left). Lag variance as a function of lag distance (right). The three displayed curves were calculated according to three different declustering strategies. The green (+) and red curve (x) show the results calculated from respectively the gridded and complete equal-weighted data set. The orange curve (o) on the other hand shows the results based on the complete declustered data set.

computed as follows:

$$\sigma_{-\mathbf{h}}^2 = \sum_{\alpha=1}^{N(\mathbf{h})} w_{\alpha} [z(\mathbf{x}_{\alpha}) - m_{-\mathbf{h}}]^2 \quad (\text{D-10})$$

$$\sigma_{+\mathbf{h}}^2 = \sum_{\alpha=1}^{N(\mathbf{h})} w_{\alpha+\mathbf{h}} [z(\mathbf{x}_{\alpha} + \mathbf{h}) - m_{+\mathbf{h}}]^2. \quad (\text{D-11})$$

Both resulting standard deviations are then averages and plotted against their corresponding lag distance. The resulting graph of the lag variance expresses some sort of uncorrelated variability present in the different h-scattergrams. An example of such a plot can be found in the right hand side of Figure ??.

The plots of the lag mean and lag variance are very useful to determine whether preferential sampling and the proportional effect have profound negative effects on the calculation of correlation indicators. The preferential sampling of high valued areas entails that most data pairs contributing to small lags originated from clustered high valued areas. The corresponding lag mean is high and due to the presence of a proportional effect, the lag variance is also large. Fortunately, as the lag distance increase, the data pairs contributing the h-scattergram become more representative of the entire area and the lag mean and lag variance decreases. Such a trend may render especially the estimation of the spatial correlation indicators at short lags unreliable. This is due to an overestimation of the semi-variances and underestimation of covariances and correlation coefficients corresponding with those short lag distances. The nugget effect might also be severely overestimated.

The best approach to remove the unwanted impact of the proportional effect and clustering, is to use only regularly spaced data. However, this would result in ignoring most of the critical short-range information. Sometimes, due to data scarcity, it is almost impossible to use only the regularly spaced data. In such cases, an alternative that also takes short-range information into account would be very useful. In principle, the existing alternatives are based on the correction of the lag mean variation by computing relative correlation indicators.

Figure D-9: Pairwise relative experimental semi-variance function of Walker Lake exploration data; The red curve (+) represents the calculated relative semi-variances of the complete Walker Lake data set. The green curve (+) represents the calculated relative semi-variances of the gridded Walker Lake data set.

Pairwise relative semi-variance An example of such an alternative is the pairwise relative semi-variance, which is part of the relative semi-variance family. In general, a relative semi-variance scales each calculated semi-variance value by a function of the lag mean. This function might be determined from the scatter plot of the local mean versus the local standard deviation (Figure D-4). Often, for positively skewed data, one usually takes this function to be the squared lag mean. The relative semi-variance then scales each h-increment by the squared average of the head and tail values. In this case, the computed correlation indicator is called the pairwise relative semi-variance:

$$\gamma_{PR}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \frac{[z(\mathbf{x}_{\alpha}) - z(\mathbf{x}_{\alpha} + \mathbf{h})]^2}{\left[\frac{z(\mathbf{x}_{\alpha}) + z(\mathbf{x}_{\alpha} + \mathbf{h})}{2}\right]^2}. \quad (\text{D-12})$$

This calculated statistic reduces the influence of each large value in the computed semi-variogram and is thus ideally suited to deal with data from preferentially sampled high valued areas. Figure D-9 shows an example of two series of computed pairwise relative semi-variances.

The introduced group of spatial statistics are subsequently being used to investigate the spatial continuity and variability features inherently present in the area under investigation. Before proceeding with a detailed explanation of the main operations, a very short recapitulation of the complete set of exploratory and summary tools is formulated.

D-3 Summary

A good way to close the previous discussion is to present a conclusive list, that gives an clear overview of the available tools contained in the “area and data statistics” boxes (Figure ??). The discussion clearly indicated, as does the name, that the set of tools contained in these boxes can be subdivided into two groups, dependent on their final intention. The first group aims at estimating some general data properties. While the second group rather focuses on the characterization of the spatial continuity. Remember that, in order to be capable of handling calculations involving preferentially sampled data, some conventional formulas

had to be adapted. The implemented adjustments were necessary to obtain representative statistics which are able to capture the most essential features of the underlying population. Following list gives a final overview of the content presented in the section.

- data or global statistics
 - mean
 - variance and standard deviation
 - probability density function (histogram)
 - cumulative distribution function and quantiles
- area or spatial statistics
 - Data posting and moving window statistics
 - h-scattergram
 - semi-variance, covariance and correlation
 - lag mean and lag variance
 - pairwise relative semi-variance

The listed tools turn out to be very useful during the calibration and validation of the simulation model. At the beginning of this chapter, it was already mentioned that the developed practical process regularly takes advantage of the discussed tools to investigate the effects of certain transformations and support adjustments. The results of such an investigation are subsequently used to decide whether certain model parameters have to be adjusted. Moreover, the introduced tools are also extensively used during both validation procedures.

The great value and purpose of the introduced set of tools will become more evident during the discussion of the main operations and the validation procedures in this and the next chapter.

Appendix E

Declustering of Clustered Preferential Sampling

The first operation aims to remove some unwanted effects caused by preferential sampling. Preferential sampling occurs when the data locations are neither regularly, nor randomly distributed over the study area. This often happens in practice, since the exploration scheme is rarely designed with the aim of reducing its negative influence on the statistical analysis. Preferential sampling is not necessarily connected with a thoughtless exploration strategy. In certain circumstances, the surveyors deliberately choose for clustered data locations (Goovaerts, 1997):

- *Conditions of accessibility*; During land exploration programs, it is quite common to drill the exploration holes at locations that are easier to reach. Areas bordering roads or farmland are easier accessible than dense forests or rugged terrain. A similar sort of situation might also occur during deep sea exploration programs. Consider for example an SMS deposit; it is probably easier to drill an exploration hole on a plain than on a steep slope.
- *Expected attribute value*; Sampling is often denser in areas that are deemed critical for the success of the mining operation. Figure E-1 shows a realistic situation of a deposit that is explored over a sequence of exploration phases. During the first phase, relatively few exploration holes were drilled on a coarse grid. The subsequent phases of the campaign focused on the areas which proved to be highly mineralized. Such an exploration campaign is based on an economically sound strategy. Often, the strategy is to discover and delineate the anomalous zones in the deposit at the lowest possible cost. Although this is a sound exploration strategy, it results in a clustered and biased data set.
- *Sampling strategy*; A proper feasibility study requires a detailed study of the local geological and grade continuity (Dominy et al., 2002). To obtain the necessary confidence, clustered locations are deliberately sampled in order to improve the characterization of the short-range variability of the deposit.

Figure E-1: Preferential sampling; Sampling is often denser in areas that are deemed critical for the success of the mining operation. The gray shaded areas indicate zones with anomalous attribute values. During the first exploration phase, relatively few exploration holes were drilled on a coarse grid (black points). The second phase was designed by placing four additional drill holes in the blocks which proved to contain high mineralization (points with wave pattern). Finally, the last phase was developed to delineate the boundary of the anomalous zone (white points).

Great attention should be paid to preferential sampling, since it may skew the results of any data analysis. Previous factors clearly illustrate, that a proper documentation of the sampling strategy already can give a good indication whether declustering is needed. However, even if high or low valued areas were not purposely targeted, preferential sampling may be present and is likely to impact the sample statistics.

The need for declustering can be evaluated by comparing the quantiles of the gridded data, q_p^g , with those of the complete data, q_p^{g+c} . The collection of gridded data is formed by the data points that are located on a regular sampling grid, whereas the complete data is defined as the combination of gridded and clustered data. Recall that q_p , the p -quantile value of a distribution, is the value below which a proportion p of the data falls. Expressed mathematically: $F(q_p) = p$. Notice, that in this case, both cdfs are calculated using constant weights (Eq. (D-3)). For the gridded data, all λ_α are equal to the inverse of the number of gridded data points. While for the complete data, the weights λ_α are all equal to the inverse of the total number of data points.

The comparison of both sets of quantiles, q_p^{g+c} and q_p^g , will give a clear indication whether certain value classes were deliberately or purposely targeted. Similar distributions should have similar quantiles and the graph of q_p^{g+c} versus q_p^g should appear as a straight line. An example of this so-called Q-Q plot is displayed in Figure E-2. The figure clearly illustrates that the calculated quantiles of the complete data all exceed the corresponding quantiles of the gridded data. This phenomena most likely indicates that the exploration campaign focused on highly mineralized zones (Figure E-1). Such an exploration strategy would indeed cause the overestimation of all the quantiles. Representative statistics can be calculated using the declustered complete data set or the equal-weighted gridded data. Important to realize is that major estimation errors might occur if no proper declustering is performed.

One procedure to correct for preferential sampling is to only retain the regularly spaced or gridded data. This approach is straightforward and might be appropriate for data sets

Figure E-2: Q-Q plot of the distribution of the complete data versus that of the gridded data. Each blue cross corresponds with the position of a specific p -quantile pair (q_p^g, q_p^{g+c}) . In total 17 pairs are displayed, with a proportion p ranging from 0.05 with steps of 0.05 to 0.90. The 0.95 quantile pair is located at (802.4, 918.0) and thus lies outside the boundaries of the plotted space. Similar distributions should plot on the 45° line. All the quantiles in this plot were calculated with constant weights (see text). The figure clearly indicates that the conventional quantiles of the complete data give a wrong impression of the underlying population. This figure provides a strong indication that declustering is needed or that only the gridded data has to be used to calculate representative statistics.

that include enough gridded sampling points to ensure reliable inference. However, when data sparsity does not allow one to ignore the clustered values, other procedures have to be implemented. Intuitively, data points in densely sampled areas should receive less weights than those in sparsely sampled areas (Figure E-1).

Now, the objective is to discuss certain algorithms which can perform such an adjustment of the weights. Two commonly used declustering techniques are the polygon and the cell declustering method. Please notice that neither of these methods can hope to replace actual samples from the population, all they do is make intelligent use of the available samples.

E-1 Polygon declustering

The polygon declustering technique is fairly intuitive and assigns declustering weights to datum locations based on their polygons of influence. Each sample in the data set has a polygon of influence, that is the area constituted by all locations $\mathbf{x} \in \mathcal{A}$ closer to $\mathbf{x}_\alpha$, than any other datum location. Figure E-3 shows a sketch of the polygons of influence of an arbitrary data set. The polygons of influence are constructed as follows:

- Connect neighboring data points with line segments.
- Construct the perpendicular bisector for each line segment. A perpendicular bisector is a line on which points are equidistant from either end of the line segment.
- Perpendicular bisectors between a sample and its neighbors form the boundaries of the polygon of influence.

Figure E-3: Polygon Declustering; The declustering weights are proportional to the area of the polygon of influence of each datum location. Clustered data with small polygons of influence receive less weight than isolated locations with large polygons of influence. A polygon of influence of a particular datum location is defined as the area that is constituted of all locations that are closer to that specific datum location than to the remaining ones.

- Truncate the bisectors so that only the polygons of influence remain.

The declustering weights for each datum location $\mathbf{x}_\alpha$ are subsequently calculated as the ratio of the area of the corresponding polygon of influence, ω_α , divided by the total area of interest, $|\mathcal{A}|$. The total area can be defined as $\int d\mathbf{x}_\alpha = \sum_\alpha \omega_\alpha$. Finally, the formula of the resulting declustering weights can be written as $\lambda_\alpha = \omega_\alpha / \sum_\alpha \omega_\alpha$.

Clustered samples will tend to get small weights corresponding to their small polygon of influence. On the other hand, samples with large polygons of influence can be thought of as being representative of a larger area and are therefore entitled to a larger weight.

This declustering technique generally performs well, makes sense and is easily understandable. The polygons of influence can be conveniently constructed by hand on a piece of paper. The areas of the resulting polygons are however not that easy to calculate. Also the implementation of this declustering algorithm in computer software is not that straightforward. On the other hand, the technique is easy to use since no calibration parameters are required.

The application of the polygon declustering approach needs to be treated with care, since erratic extreme isolated datum values might receive weights that are relatively large. These large weights would increase the relative importance of the erratic values and thus would lead to biased global estimates (Goovaerts, 1997; Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978).

E-2 Cell declustering

In the cell declustering approach, the entire study area $\mathcal{A}$ is divided into rectangular cells. Each sample receives a weight that is inversely proportional to the number of samples that are located within the corresponding cell. The approach calls for counting the number B of cells containing at least one sample and the number n_b of datum locations falling within each cell

Figure E-4: Cell Declustering; The declustering weights are inversely proportional to the product of the number of cells containing at least one sample and the number of samples within the cell. n indicates the number of samples within each cell, while w indicates the declustering weight of each sample located in that cell. Clustered samples generally receive lower weights because the cells in which they are located will also contain several other samples. To obtain optimal declustering weights following parameters have to be calibrated correctly: cell width, cell height, origin and orientation of the grid.

b. Finally, each datum location, $\mathbf{x}_\alpha$ receives a declustering weight λ_α that equals $1/(B \cdot n_b)$. Such a declustering weight generally gives less importance to clustered data points, since the cell in which they are located also contain several other samples.

Figure E-4 shows a grid of such cells superimposed on a number of clustered datum locations. The lines show the boundaries of the four declustering cells ($B = 4$). Each of the eight clustered samples located in the southeastern cell receive a weight $\lambda_\alpha = 1/(4 \cdot 8) = 1/32$, while the single sample in the northwestern cell receives a weight of $\lambda_\alpha = 1/(4 \cdot 1) = 1/4$. The total weight of the fifteen datum locations sums up to one as it should.

The cell declustering approach can actually be seen as a two-step procedure involving non-overlapping moving windows. Since all the samples within a particular cell receive an equal weight and each cell also receives an equal weight, the representative population mean is calculated as the average value of all the window means. The first step thus involves the calculation of some relevant window statistics (mostly the mean), while the second step subsequently uses this window statistics to obtain global estimates.

In contrast to polygon declustering, cell declustering requires some calibration parameters. First of all, it is important to select a proper origin and orientation for the declustering grid. Secondly, the cell width and height needs to be optimally calibrated. Often the initial phase of the exploration campaign is based on an underlying pseudo-regular grid. The spacing of this sampling grid usually provided a good size.

If the sampling pattern does not suggests a natural size, the effect of several cell sizes and orientations need to be investigated. It is common practice to select the set of parameters that results in the smallest or largest declustered mean according to whether high- or low-valued areas where preferentially sampled. If the cell size is as large as the complete study area, each datum location will fall into the same cell and thus receives an equal weight. On the

other hand, if the cell sizes are so small that each cell just contains one sample, then again all the samples will be equally important. The appropriate medium is located somewhere in between these two extremes. To correct for erratic extreme values located in specific cells, it might be useful to even average results for several grid configurations.

The cell declustering algorithm is fairly easy to implement into computer software. The parameter calibration of this approach might provide some difficulties, but is generally considered to be straightforward if the initial sampling phase is based on a pseudo-regular grid. In this case, the cell configuration needs to be adjusted in such a way that the grid nodes are more or less located around the centers of the cells. The author believes that such a calibration procedure is advantageous, since it forces the user to investigate the possible consequence of several declustering strategies. The lack of a plug and play algorithm prevents that errors are made without noticing. On the other hand, if the data was not regularly spaced during any phase of the complete exploration campaign, it might better to consider the polygon declustering technique (Goovaerts, 1997; Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978).

E-3 Quality control of the declustering operation

One of the main goals of this chapter is to introduce a transparent framework that reduces the likelihood of errors. In order to avoid errors, that go undetected, several quality control methodologies are implemented at some strategic locations in the constructed practical process. Since the cdf plays a key role in the remainder of the geostatistical analysis, it is important to estimate it right. Therefore, two quality control points, which ensure the correctness of the cdf analysis, are introduced into the declustering phase.

Figure E-5 gives a clear overview of the sequence of actions that need to be completed during the first phase of the geostatistical analysis. The gray shaded diamonds indicate the positions and the methodologies of both inserted quality checks.

The first quality check supports the decision of running the declustering algorithm. Global estimates derived from both the complete and the gridded data set, are compared. To ascertain that declustering is not needed, the calculation of all kind of global statistics for both sets should yield similar results. As mentioned before, a convenient quality check is based on the QQ-plot of the distribution of complete data versus that of the gridded data set (Figure E-2). To be representative for the underlying population and under the condition that enough gridded data points are available, the distribution of the complete data set should be similar to that of the gridded data set. As long as the selection of p -quantile pairs are not located on the straight 45° line, bias will be introduced into any further calculated estimates.

The second quality check, on the other hand, is designed to validate the declustering operation. During this check, global estimates of the complete declustered data are compared with those of the gridded data. If the results are not acceptable, it is necessary to recalibrate the model (Figure E-5).

Figure E-6 shows an example of the proposed comparisons that need to be made during both quality checks. The graph of the left hand side clearly indicates the presence of bias due to preferential sampling. The graphs shows that especially anomalous high valued zones were targeted during the exploration campaign. The graph support the decision of proceeding along the left side of the previously introduced flowchart (Figure E-5). The graph on the

Figure E-5: Quality control in declustering operation; The gray shaded diamonds indicate the positions of two quality checks inserted into the declustering operation. The first quality check supports the decision of running the declustering algorithm. The second quality check evaluates the performance of the declustering operation. If the results of the second check are not satisfactory, the declustering model needs to be recalibrated.

Figure E-6: Quality Control of Declustering Operation; Q-Q plot of the distribution of the complete data versus that of the gridded data. Each blue cross corresponds with the position of a specific p -quantile pair (q_p^g, q_p^{g+c}) (left). Q-Q plot of the distribution of the complete declustered data versus that of the gridded data. Each blue cross corresponds with a specific p -quantile pair (q_p^g, q_p^d) (right). In total 18 pairs are calculated ranging from a proportion p of 0.05 with steps of 0.05 to 0.95. Due to the chosen limits of the axes, the 0.95 quantile pair is not displayed in the left graph (802.4,918.0). Similar distributions should plot on the 45° line. The figure clearly indicates that declustering improves the global estimates of the underlying population.

right hand side illustrates that the declustering algorithm was able to neutralize the effect of preferential sampling. This graph results from the second quality check which is located at the end of the iterative process of calibrating and declustering.

If after some attempts, the results are still not improving, a detailed study should be performed to reveal the reasons behind the difference in global estimates of the gridded and the declustered data. If the quality of the declustering operation can not be confirmed, it is better to use only the gridded data to obtain global estimates.

Once the declustering operation is performed, the computed weights are stored into the central memory of developed software. In this way, they are accessible from all the “area and data statistics” boxes where representative global and spatial statistics are computed. One of these calculated global statistics, the cdf, is subsequently being used during the normal score transformation.

Appendix F

Input Data and Design of Simulation Grid

In order to safeguard the reproducibility of the performed research, it is wise to give a brief explanation on the origin of the input data. The investigated collection of spatial measurements is actually a modified version of a generally available and well known data set. In order to fit the complete data set nicely within the physical simulation boundaries, it was necessary to redefine the origin of the coordinate system. The performed modifications were only confined to some translations of the sampling locations, and have thus no impact at all on the calculation result and possible conclusions.

Moreover, the applied uniform translation made it also possible to design a simulation mesh with a convenient cell size of one square meter. Due to the cleverly chosen mesh configuration, it became possible to reblock these small cells, within their second generation parent boundaries, to $16 \times 16 \text{ m}^2$ blocks. It is clear that the performed translation was intended to lead to a more conveniently dimensioned simulation grid.

The text, presented in this chapter, explains how the modifications were performed and how this aided the construction of a conveniently dimensioned simulation grid. First, the original data set and its corresponding exhaustive counterpart are introduced. Then, a short discussion on the design of the simulation grid follows, together with a brief explanation of the performed adjustments. Finally, a possible underlying sampling strategy is presented and a clear distinction is made between the gridded and complete data set.

F-1 Original Data Set

The discussed geostatistical analysis is based on a modified version of the well known Walker Lake data set. This data set was introduced in 1989 by Isaaks and Srivastava and has ever since being used during multiple academic case studies. The Walker Lake database distinguishes itself from many others in that it also contains the completely known reality.

This known reality makes the database particularly useful to investigate the performance of newly developed estimation and simulation methods.

The entire database is composed of two tables, one representing an exploration campaign, the other approximating the known reality.

The first table contains 470 data points, which supposedly were collected during a sequence of exploration campaigns. All the measurements were made within a 243 *m* wide and 283 *m* high rectangle. The *x* coordinate of the sampling points ranges from 8 *m* to 251 *m*, while the *y* coordinate varies between 8 *m* and 291 *m*. The actual measured concentrations fluctuate between 0 *ppm* and 1528,1 *ppm*. The extent of this original exploration data is also summarized in the upper part of Table F-1. Furthermore, Figure F-1 displays the original locations of all the data points as green dots.

The completely known reality, on the other hand, is captured by a collection of 78.000 data points. This huge amount of measurements is located within a 259 *m* x 299 *m* rectangle, where each data point is positioned on one of the nodes of a 1 *m* by 1 *m* grid. The *x* coordinate of these sampling points ranges from 1 *m* to 260 *m*, while the *y* coordinate varies between 1 *m* and 300 *m*. The original spatial extent of the known field is illustrated in Figure ?? by the green rectangle and is also presented in the upper part of Table ?. Also interesting to mention is that the concentrations in the know field fluctuate between 0 *ppm* and 1631.16 *ppm*.

The performed investigation of the spatial extent of the data really aids the optimal design of the simulation grid, which in turn influences the selection of proper translation parameters.

F-2 Design of Simulation Grid and Performed Data Adjustment

Once the spatial extent of the original data is investigated and captured, time has come to design an optimal simulation mesh. This mesh, however, can not be designed without any specific considerations. Due to the applied LAS methodology and the corresponding sequence of cell quartering, the dimensions of the simulation grid have to comply with certain restrictions. Once a suitable mesh is created, it might be necessary to translate and crop the original data so that all the selected data points nicely fit in the physical simulation boundaries.

The restrictions, dictating the design of the mesh, are inherent to the selected LAS simulation methodology. The previously discussed restrictions instruct that the number of cells in the *x* and *y* directions (N_1, N_2) have to be a multiple of the same power of two (Eq. (3-12)). These restrictions do not necessarily have to be very limiting, since the user can always adopt flexible and unconventional cell dimensions.

However, mainly for reasons of convenience, the actual geostatistical simulation is performed on a cell size of 1 x 1 *m*². After all, this facilitates the whole dimensioning procedure, since now the number of cells in a particular direction equals the actual length of the grid in that direction.

This specific choice implies that the dimensions of the simulation grid are connected with the number of subdivisions (*m*) and the lay-out of the primary cell lattice (k_1, k_2), see Section 3-2. The remaining challenge lies in a proper selection of these parameters such that the resulting grid nicely encloses the exploration data. This means that the grid at least needs to have a width of 243 *m* and a height of 283 *m*. Suitable grid dimensions are obtained by choosing

Table F-1: Walker Lake data adjustment; The upper part of the table summarizes the spatial extent of the original Walker Laker exploration data and the known field (top). The middle part illustrates the effect of the translation on the spatial extent of both data sets. Each individual location was shifted 1.5 *m* westwards and 5.5 *m* southwards (middle). The bottom part shows that 4282 data points, belonging to the known field, were discarded because they were located outside the physical simulation boundaries (bottom)

Original			
	Xloc [m]	Yloc[m]	Conc [ppm]
Exploration: 470 Samples			
<i>min</i>	8	8	0
<i>max</i>	251	291	1528,1
Walker Lake: 78.000 Samples			
<i>min</i>	1	1	0
<i>max</i>	260	300	1631,16
Adjustment 1: translation			
	Xloc [m]	Yloc [m]	Conc.[ppm]
Exploration: 470 Samples			
<i>min</i>	6,5	2,5	0
<i>max</i>	249,5	285,5	1528,1
Walker Lake: 78.000 Samples			
<i>min</i>	-0,5	-4,5	0
<i>max</i>	258,5	294,5	1631,16
Adjustment 2: cropping			
	Xloc [m]	Yloc [m]	Conc.[ppm]
Exploration: 470 Samples			
<i>min</i>	6,5	2,5	0
<i>max</i>	249,5	285,5	1528,1
Walker Lake: 73.728 Samples			
<i>min</i>	0,5	0,5	0
<i>max</i>	257,5	287,5	1631,16

Figure F-1: Walker Lake data adjustment; The translation procedure of the exploration data is visualized by both sets of dots. The old locations are indicated as green dots, while the new locations are shown as blue dots. The plot clearly indicates that each individual point underwent the same translation. The physical simulation boundaries are delineated by the red rectangle. The two other rectangles illustrate the translation of the completely known reality. The green rectangle delineates the extent of the initial known field, while the blue rectangle delineates the extent of the translated known field. Finally the known field is cropped, so that only points that are located within the simulation boundaries are retained.

for 5 subdivisions and selecting an initial 8×9 lattice of local averages, each having an area of $32 \times 32 \text{ m}^2$. This particular set of parameters finally results in a mesh that is 256 m wide and has a height of 288 m .

As mentioned in the introduction of this section, the resulting 73.728 square meter cells can eventually be reblocked, within their respective second generation parent boundaries, to $288 \times 16 \text{ m}^2$ blocks.

The boundaries of the newly designed simulation mesh are able to enclose the spatial extent of the introduced Walker Lake exploration data. However, this does not yet mean that all the Walker Lake exploration data points are located within these boundaries. Deviations are to be expected since both the simulation grid as the exploration data are expressed in euclidean coordinate systems with different origins. The measurement locations of the exploration data need to be translated such that all the data points are located within the physical simulation boundaries.

It is known that the horizontal extent of the data amounts 243 m and that the minimum x coordinate is 8 m . Now, in order to position all the data points in the foreseen horizontal 256 m , each x coordinate needs to be shifted $1,5 \text{ m}$ west. This value is deliberately chosen, so that a vertical empty border of $6,5 \text{ m}$ wide appears at both sides of the exploration data. The vertical extent, on the other hand, amounts 283 m and starts at y coordinate of 8 m . In order to position all the data points in the foreseen vertical 288 m , each y coordinate needs to be shifted $5,5 \text{ m}$ south. This vertical translation results in two horizontal empty borders, with a height of $2,5 \text{ m}$, above and below the exploration data.

The complete translation is visualized in Figure F-1. The old locations are indicated as green dots, while the improved new locations are shown as blue dots. The physical simulation boundaries are indicated by the red rectangle. The plot also shows that each individual data point underwent the same translation. Similar information is summarized in Table F-1.

It is obvious that also the 78.000 data points, which represent the known reality, have to be translated accordingly. Figure F-1 and Table F-1 illustrate the effect of the translation on the delineation of the known field. The blue rectangle, which represents the extent of the translated known field, shows that not all of the 78.000 data points are located within the physical simulation boundaries. This is because the total area of the known field is slightly larger than that of the simulation grid. The bottom part of Table F-1 shows that this problem can be easily overcome by discarding the 4282 data points that are located outside the simulation boundaries.

The remaining 73.728 data points will be used later on to validate the simulation procedure. Notice that each data point belonging to the know field is exactly located in the middle of a corresponding square meter cell.

F-3 Underlying Sampling Strategy

The previously discussed modifications finally result in the data set, hereafter referred to as the Walker Lake exploration data, that will be used during the further course of following geostatistical analysis. The entire Walker Lake exploration data is listed in Table G-1

Before proceeding with any statistical analysis, it might be wise to have a quick look at the spatial configuration of the 470 data points. The spatial configuration is useful to find out which specific sampling strategy has been used to collect all the samples.

Figure F-2: Color coded plot of Walker Lake exploration data. The color of each dot corresponds with the measured value at that specific location. The primary exploration points were positioned on a 20 m by 20 m grid. This initial sampling campaign resulted in 195 regularly spaced data points (left). The subsequent exploration cruises focused on the characterization and the delineation of the highly mineralized zones. This strategy added 275 clustered data points to the initial exploration data (right). The gray lines indicate the initial 20 m by 20 m sampling grid. The origin of this sampling grid is located at $(-2; -4, 5)$ and was chosen such that most of regularly spaced sampling points are located around the midpoint of the corresponding grid cell

The right hand side of Figure F-2 shows a color coded posting of the complete Walker Lake exploration data. This plot immediately reveals that the complete collection of data points was not measured along the same single sampling structure. One of the first things to notice is that a certain part of the data is collected along a regularly spaced pattern. Obviously, this feature is illuminated by superimposing the sampling grid over the plotted data points. Another striking feature are the high concentrations which are measured in most of the clustered data points. Both observations support the idea that the data set was collected through a sequence of exploration stages.

In most cases, the sampling strategy should be documented or can be requested from the involved exploration geologist. Since following analysis is just based on an academic case study, some hypothetical exploration strategy is formulated. The following description of the entire exploration campaign seems to be reasonable and definitely would explain the observed spatial configuration.

- *Stage 1:* Superimpose the designed 20 m by 20 m sampling grid on the area of investigation. Then take one measurement at the center of each cell. This first exploration step results in the 195 gridded data points.
- Deploy the exploration tools in the most economically responsible way. This means that the further exploration steps are only focused on profitable sub areas (preferential sampling).
 - *Stage 2:* Intensive sampling of the grid cells where previously high concentrations were measured. Such dense sampling is necessary to correctly estimate the value and the potential of the deposit.

- *Stage 3:* The last exploration step focuses on the delineation of the boundaries or extent of the rich zones in the deposit. This step further improves the estimation of the deposit value.

The last two exploration steps result in an additional 275 clustered data points.

The data set is thus regarded to be a combination of regularly spaced and clustered data points. The 195 regularly spaced data points are shown in the left hand side Figure F-2 and are listed in Table G-2.

Since this case study started with only the availability of the complete data set, some tricks needed to be applied in order to obtain the gridded data points. These gridded data points were collected according to the displayed 20 *m* by 20 *m* sampling grid. The origin of this sampling grid is located at $(-2; -4, 5)$ and was chosen such that most of regularly spaced sampling points are located around the midpoint of the corresponding grid cell. The 195 gridded data points were extracted from the complete data set through the application of a fairly straightforward algorithm, which is implemented in the declustering software. Essentially, the algorithm just runs through the sequence of grid cells and, for each cell, only the data point that is located closest to the midpoint is retained.

Later on, it will become clear that this gridded data set has a significant impact on the success of the geostatistical analysis.

Appendix G

Walker Lake Exploration Data

This chapter contains the Walker Lake Exploration Data, obtained after some minor translations of the original data set. A detailed discussion on the performed data adjustments can be found in Appendix F.

Two tables are presented. One contains the entire collected of sampled concentrations, while the other just present the data points sampled on a regular grid.

G-1 Walker Lake Exploration Data: Gridded + Clustered

Table G-1: Walker Lake Exploration Data

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
1	9.5	2.5	0
2	6.5	24.5	0
3	7.5	42.5	224.4
4	6.5	63.5	434.4
5	7.5	84.5	412.1
6	8.5	104.5	587.2
7	7.5	123.5	192.3
8	9.5	144.5	31.3
9	8.5	164.5	388.5
10	6.5	182.5	174.6
11	7.5	203.5	187.8
12	8.5	225.5	82.1
13	9.5	244.5	81.1
14	8.5	263.5	124.3
15	6.5	282.5	188
16	29.5	5.5	28.7
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
17	27.5	23.5	78.1
18	26.5	45.5	292.1
19	29.5	62.5	895.2
20	26.5	82.5	702.6
21	28.5	104.5	490.3
22	26.5	124.5	136.1
23	26.5	144.5	335
24	28.5	165.5	277
25	26.5	184.5	206.1
26	29.5	203.5	24.5
27	26.5	223.5	198.1
28	28.5	244.5	60.3
29	29.5	263.5	312.6
30	29.5	283.5	240.9
31	47.5	5.5	653.3
32	47.5	23.5	96.4
33	49.5	42.5	105
34	47.5	62.5	37.8
35	48.5	82.5	820.8
36	49.5	103.5	450.7
37	46.5	123.5	190.4
38	47.5	145.5	773.3
39	49.5	162.5	971.9
40	46.5	184.5	762.4
41	48.5	205.5	968.3
42	47.5	225.5	394.7
43	49.5	244.5	343
44	48.5	262.5	863.8
45	49.5	284.5	159.6
46	69.5	3.5	445.8
47	69.5	23.5	673.3
48	68.5	45.5	252.6
49	66.5	64.5	537.5
50	67.5	84.5	0
51	66.5	104.5	329.1
52	66.5	122.5	646.3
53	67.5	142.5	616.2
54	67.5	163.5	761.3
55	68.5	185.5	918
56	67.5	202.5	97.4
57	67.5	223.5	0
58	66.5	244.5	0
59	69.5	262.5	0
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
60	69.5	282.5	2.4
61	89.5	5.5	368.3
62	89.5	23.5	91.6
63	88.5	43.5	654.7
64	89.5	62.5	645.5
65	89.5	85.5	907.2
66	89.5	105.5	826.3
67	87.5	124.5	975.3
68	86.5	143.5	551.1
69	87.5	164.5	155.5
70	87.5	182.5	10.7
71	88.5	205.5	0
72	88.5	224.5	0
73	86.5	243.5	0
74	86.5	263.5	12.1
75	86.5	282.5	62.2
76	107.5	5.5	399.6
77	109.5	25.5	176.6
78	106.5	43.5	402
79	107.5	62.5	260.6
80	106.5	82.5	192
81	108.5	103.5	237.6
82	107.5	123.5	702
83	108.5	142.5	38.5
84	109.5	163.5	22.1
85	109.5	185.5	2.7
86	108.5	202.5	17.9
87	107.5	224.5	174.2
88	107.5	243.5	12.9
89	107.5	262.5	187.8
90	109.5	285.5	268.8
91	128.5	3.5	572.5
92	129.5	25.5	29.1
93	128.5	42.5	75.2
94	126.5	64.5	399.9
95	127.5	84.5	243.1
96	129.5	103.5	0
97	127.5	122.5	244.7
98	129.5	142.5	185.2
99	129.5	163.5	26
100	127.5	185.5	0
101	126.5	203.5	100.3
102	128.5	225.5	530.3
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
103	129.5	242.5	107.4
104	126.5	263.5	159.3
105	129.5	282.5	70.7
106	146.5	2.5	260.2
107	147.5	23.5	326
108	148.5	43.5	332.7
109	149.5	63.5	531.3
110	148.5	83.5	547.2
111	148.5	103.5	482.7
112	148.5	123.5	84.1
113	148.5	145.5	4.7
114	147.5	163.5	180.6
115	149.5	184.5	0
116	146.5	202.5	342.4
117	148.5	222.5	602.3
118	147.5	245.5	209.1
119	147.5	265.5	79.4
120	146.5	285.5	104.1
121	166.5	2.5	446
122	169.5	23.5	189.9
123	167.5	43.5	280.4
124	166.5	63.5	0
125	166.5	85.5	499.3
126	169.5	103.5	457.3
127	166.5	125.5	341.2
128	169.5	144.5	0
129	169.5	165.5	208.3
130	167.5	185.5	99.7
131	168.5	204.5	636.6
132	168.5	224.5	173.1
133	167.5	243.5	17
134	166.5	265.5	283.1
135	166.5	284.5	30.9
136	188.5	5.5	348.5
137	189.5	22.5	222.4
138	189.5	42.5	59.1
139	188.5	63.5	0
140	188.5	83.5	326
141	186.5	105.5	325.1
142	189.5	123.5	114.7
143	187.5	143.5	481.6
144	188.5	163.5	324.1
145	188.5	183.5	10.9
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
146	186.5	204.5	332.9
147	189.5	225.5	184.4
148	188.5	242.5	146.6
149	187.5	264.5	92
150	187.5	284.5	2.5
151	209.5	5.5	358.1
152	207.5	24.5	473.3
153	209.5	43.5	308.8
154	208.5	64.5	406.8
155	207.5	84.5	812.1
156	208.5	105.5	339.7
157	209.5	124.5	223.9
158	206.5	145.5	673.5
159	207.5	162.5	141
160	206.5	185.5	61.8
161	208.5	205.5	258.3
162	209.5	222.5	590.3
163	209.5	244.5	166.9
164	206.5	262.5	125.2
165	206.5	283.5	29.3
166	229.5	4.5	617.6
167	229.5	22.5	425.9
168	228.5	44.5	295.7
169	228.5	65.5	224.9
170	227.5	85.5	31.7
171	227.5	104.5	377.4
172	228.5	125.5	333.3
173	226.5	142.5	351
174	227.5	163.5	0
175	229.5	185.5	137.6
176	229.5	202.5	451.2
177	227.5	222.5	639.5
178	229.5	243.5	119.9
179	229.5	262.5	27.2
180	228.5	285.5	2.1
181	247.5	3.5	167.7
182	248.5	24.5	147.8
183	247.5	42.5	442.7
184	249.5	63.5	487.7
185	249.5	85.5	0
186	246.5	103.5	28.2
187	247.5	124.5	0
188	246.5	144.5	18.3
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
189	248.5	163.5	266.3
190	248.5	184.5	502.3
191	249.5	202.5	0
192	249.5	223.5	240.9
193	247.5	245.5	234.4
194	246.5	264.5	22.4
195	248.5	285.5	45.6
196	38.5	65.5	76.2
197	19.5	63.5	284.3
198	26.5	74.5	606.8
199	27.5	53.5	772.7
200	39.5	75.5	269.5
201	16.5	74.5	1036.7
202	37.5	54.5	238.6
203	16.5	54.5	519.4
204	39.5	84.5	414.9
205	19.5	84.5	601.4
206	29.5	95.5	579.2
207	39.5	94.5	601.4
208	19.5	94.5	594.6
209	58.5	2.5	550.1
210	38.5	5.5	99.4
211	49.5	12.5	233.6
212	57.5	14.5	14.4
213	39.5	15.5	115.9
214	57.5	84.5	506.2
215	49.5	95.5	502.4
216	48.5	75.5	608
217	57.5	95.5	363.9
218	58.5	75.5	385.6
219	58.5	145.5	1521.1
220	36.5	142.5	340.9
221	48.5	154.5	879.1
222	48.5	132.5	413.4
223	59.5	152.5	868.9
224	37.5	154.5	657.4
225	59.5	133.5	477
226	36.5	134.5	268.5
227	59.5	164.5	806.4
228	37.5	164.5	914.4
229	47.5	173.5	811.5
230	56.5	173.5	1113.6
231	37.5	175.5	1008
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
232	58.5	185.5	1528.1
233	38.5	184.5	970.9
234	49.5	192.5	1109
235	58.5	192.5	1203.9
236	38.5	194.5	641.3
237	56.5	202.5	720.6
238	36.5	203.5	665.3
239	48.5	215.5	543.3
240	59.5	214.5	101.1
241	37.5	215.5	615.9
242	57.5	262.5	543.1
243	39.5	265.5	868.8
244	47.5	272.5	583
245	49.5	254.5	670.7
246	57.5	275.5	148.8
247	37.5	273.5	798
248	57.5	252.5	194.9
249	36.5	254.5	635.2
250	76.5	22.5	781.6
251	58.5	23.5	238.6
252	68.5	35.5	472
253	68.5	15.5	58.1
254	76.5	35.5	600.3
255	59.5	35.5	64.9
256	76.5	14.5	505.9
257	78.5	125.5	801.6
258	56.5	122.5	158.8
259	69.5	134.5	606.3
260	68.5	115.5	30.7
261	77.5	132.5	730.1
262	78.5	113.5	421.2
263	59.5	115.5	104.8
264	77.5	143.5	44.1
265	69.5	154.5	801.1
266	76.5	153.5	742
267	78.5	162.5	689.1
268	67.5	175.5	424.6
269	77.5	175.5	184.3
270	78.5	182.5	245.2
271	68.5	192.5	630
272	79.5	194.5	0
273	98.5	42.5	48.7
274	78.5	43.5	757.4
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
275	88.5	52.5	739.8
276	86.5	33.5	520.7
277	98.5	54.5	0
278	78.5	53.5	0
279	99.5	32.5	730.5
280	99.5	62.5	383.1
281	77.5	64.5	508.8
282	88.5	73.5	573.3
283	98.5	72.5	372.4
284	79.5	75.5	585.8
285	98.5	85.5	397.2
286	78.5	83.5	614.5
287	89.5	93.5	734.9
288	99.5	94.5	599.3
289	79.5	92.5	181.2
290	96.5	105.5	744.8
291	79.5	102.5	1022.3
292	88.5	114.5	899.3
293	98.5	112.5	363.7
294	96.5	124.5	513.2
295	88.5	134.5	648.8
296	88.5	132.5	645.4
297	119.5	125.5	13
298	109.5	134.5	190.3
299	106.5	115.5	893
300	118.5	135.5	104.7
301	117.5	112.5	150.4
302	156.5	222.5	558.4
303	138.5	223.5	558
304	148.5	235.5	318.5
305	149.5	212.5	394.3
306	159.5	235.5	141.9
307	139.5	234.5	112.5
308	158.5	212.5	580.4
309	137.5	214.5	535.9
310	176.5	205.5	398.2
311	157.5	203.5	517.3
312	167.5	215.5	427.2
313	168.5	192.5	367.6
314	178.5	212.5	374.7
315	176.5	195.5	144.8
316	156.5	192.5	169.8
317	217.5	82.5	235.1
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
318	196.5	84.5	611.7
319	209.5	94.5	746.4
320	206.5	74.5	436.6
321	219.5	93.5	540.9
322	197.5	92.5	801
323	218.5	75.5	272.1
324	196.5	72.5	204.1
325	218.5	144.5	543.9
326	198.5	144.5	606.2
327	206.5	153.5	356
328	208.5	134.5	440.9
329	219.5	154.5	301.8
330	196.5	155.5	369.4
331	217.5	133.5	166.8
332	198.5	133.5	230.9
333	237.5	2.5	240.3
334	216.5	2.5	737.1
335	227.5	13.5	518.6
336	237.5	12.5	390.7
337	216.5	12.5	797.4
338	236.5	223.5	602.6
339	216.5	222.5	430.8
340	229.5	233.5	354.1
341	228.5	215.5	602.4
342	238.5	233.5	172.6
343	219.5	235.5	324.8
344	237.5	214.5	420.1
345	216.5	212.5	763.5
346	33.5	65.5	687.8
347	22.5	65.5	735.8
348	32.5	82.5	86.9
349	21.5	85.5	817
350	52.5	4.5	637.9
351	44.5	5.5	512.3
352	53.5	83.5	423.4
353	43.5	83.5	569.6
354	51.5	144.5	858
355	44.5	142.5	234
356	53.5	162.5	876
357	41.5	164.5	1082.8
358	53.5	185.5	1392.6
359	42.5	185.5	646.6
360	53.5	205.5	889.5
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
361	44.5	205.5	509.2
362	52.5	263.5	613.1
363	41.5	265.5	767.8
364	71.5	23.5	649.4
365	62.5	25.5	235.4
366	73.5	123.5	782.8
367	62.5	123.5	227.3
368	71.5	143.5	722.9
369	62.5	145.5	974.5
370	73.5	165.5	512.2
371	61.5	162.5	1215.8
372	71.5	182.5	687.1
373	62.5	185.5	1259.9
374	91.5	42.5	687.5
375	84.5	42.5	471.9
376	91.5	64.5	512.1
377	82.5	63.5	963.9
378	91.5	84.5	874
379	84.5	83.5	582.4
380	94.5	105.5	553.2
381	83.5	102.5	937.3
382	91.5	125.5	883.6
383	84.5	125.5	879.9
384	112.5	125.5	268.4
385	104.5	124.5	651.5
386	153.5	223.5	386.4
387	143.5	224.5	333.2
388	172.5	202.5	339.2
389	164.5	205.5	600.3
390	213.5	83.5	595.2
391	203.5	83.5	809.6
392	213.5	142.5	293.3
393	202.5	145.5	697.3
394	234.5	3.5	515.9
395	221.5	3.5	613.2
396	234.5	223.5	665.3
397	224.5	224.5	813.6
398	33.5	74.5	174.8
399	22.5	73.5	891.8
400	34.5	55.5	699.6
401	24.5	52.5	39.5
402	14.5	74.5	915.6
403	41.5	54.5	584
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
404	13.5	82.5	610
405	44.5	93.5	566.8
406	34.5	93.5	38.1
407	52.5	74.5	483
408	44.5	75.5	542.6
409	52.5	155.5	959.3
410	41.5	155.5	631.9
411	63.5	154.5	928.3
412	31.5	154.5	431
413	34.5	164.5	672.3
414	51.5	173.5	1003.4
415	42.5	174.5	876.4
416	63.5	175.5	734.1
417	32.5	174.5	366
418	31.5	185.5	296.5
419	53.5	193.5	1069.2
420	44.5	192.5	804.3
421	61.5	195.5	731.1
422	32.5	195.5	318.1
423	63.5	204.5	238.6
424	33.5	202.5	428.9
425	44.5	214.5	734.4
426	34.5	213.5	429.1
427	33.5	211.5	597.4
428	51.5	252.5	442.6
429	44.5	254.5	765.2
430	43.5	275.5	605.5
431	33.5	272.5	795.9
432	33.5	253.5	235
433	82.5	24.5	562
434	82.5	35.5	411.4
435	73.5	34.5	696.7
436	71.5	135.5	790.9
437	61.5	134.5	696.5
438	82.5	132.5	687.3
439	74.5	153.5	597.5
440	82.5	155.5	437.4
441	84.5	163.5	317.4
442	71.5	193.5	470.7
443	74.5	45.5	498.7
444	92.5	55.5	778.7
445	83.5	54.5	523.3
446	102.5	32.5	617.1
Continued on next page			

Table G-1 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
447	91.5	35.5	395.5
448	73.5	84.5	518.9
449	92.5	95.5	383.7
450	83.5	94.5	704.1
451	102.5	103.5	562.3
452	73.5	104.5	655.3
453	93.5	115.5	823.6
454	81.5	113.5	847.7
455	92.5	134.5	607.5
456	101.5	133.5	491.2
457	112.5	114.5	319.5
458	102.5	112.5	594
459	194.5	85.5	433.5
460	213.5	95.5	209.6
461	202.5	95.5	533.8
462	194.5	95.5	592.4
463	193.5	143.5	478.7
464	214.5	5.5	660.2
465	223.5	13.5	832.2
466	212.5	13.5	242.5
467	243.5	225.5	161.2
468	231.5	214.5	626
469	224.5	215.5	800.1
470	211.5	212.5	482.6

G-2 Walker Lake Exploration Data: Gridded

Table G-2: Walker Lake Exploration Data gridded

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
1	9.5	2.5	0
2	29.5	5.5	28.7
3	47.5	5.5	653.3
4	69.5	3.5	445.8
5	89.5	5.5	368.3
6	107.5	5.5	399.6
7	128.5	3.5	572.5
8	146.5	2.5	260.2
9	166.5	2.5	446
10	188.5	5.5	348.5
Continued on next page			

Table G-2 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
11	209.5	5.5	358.1
12	229.5	4.5	617.6
13	247.5	3.5	167.7
14	6.5	24.5	0
15	27.5	23.5	78.1
16	47.5	23.5	96.4
17	69.5	23.5	673.3
18	89.5	23.5	91.6
19	109.5	25.5	176.6
20	129.5	25.5	29.1
21	147.5	23.5	326
22	169.5	23.5	189.9
23	189.5	22.5	222.4
24	207.5	24.5	473.3
25	229.5	22.5	425.9
26	248.5	24.5	147.8
27	7.5	42.5	224.4
28	26.5	45.5	292.1
29	49.5	42.5	105
30	68.5	45.5	252.6
31	88.5	43.5	654.7
32	106.5	43.5	402
33	128.5	42.5	75.2
34	148.5	43.5	332.7
35	167.5	43.5	280.4
36	189.5	42.5	59.1
37	209.5	43.5	308.8
38	228.5	44.5	295.7
39	247.5	42.5	442.7
40	6.5	63.5	434.4
41	29.5	62.5	895.2
42	47.5	62.5	37.8
43	66.5	64.5	537.5
44	89.5	62.5	645.5
45	107.5	62.5	260.6
46	126.5	64.5	399.9
47	149.5	63.5	531.3
48	166.5	63.5	0
49	188.5	63.5	0
50	208.5	64.5	406.8
51	228.5	65.5	224.9
52	249.5	63.5	487.7
53	7.5	84.5	412.1
Continued on next page			

Table G-2 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
54	26.5	82.5	702.6
55	48.5	82.5	820.8
56	67.5	84.5	0
57	89.5	85.5	907.2
58	106.5	82.5	192
59	127.5	84.5	243.1
60	148.5	83.5	547.2
61	166.5	85.5	499.3
62	188.5	83.5	326
63	207.5	84.5	812.1
64	227.5	85.5	31.7
65	249.5	85.5	0
66	8.5	104.5	587.2
67	28.5	104.5	490.3
68	49.5	103.5	450.7
69	66.5	104.5	329.1
70	89.5	105.5	826.3
71	108.5	103.5	237.6
72	129.5	103.5	0
73	148.5	103.5	482.7
74	169.5	103.5	457.3
75	186.5	105.5	325.1
76	208.5	105.5	339.7
77	227.5	104.5	377.4
78	246.5	103.5	28.2
79	7.5	123.5	192.3
80	26.5	124.5	136.1
81	46.5	123.5	190.4
82	66.5	122.5	646.3
83	87.5	124.5	975.3
84	107.5	123.5	702
85	127.5	122.5	244.7
86	148.5	123.5	84.1
87	166.5	125.5	341.2
88	189.5	123.5	114.7
89	209.5	124.5	223.9
90	228.5	125.5	333.3
91	247.5	124.5	0
92	9.5	144.5	31.3
93	26.5	144.5	335
94	47.5	145.5	773.3
95	67.5	142.5	616.2
96	86.5	143.5	551.1
Continued on next page			

Table G-2 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
97	108.5	142.5	38.5
98	129.5	142.5	185.2
99	148.5	145.5	4.7
100	169.5	144.5	0
101	187.5	143.5	481.6
102	206.5	145.5	673.5
103	226.5	142.5	351
104	246.5	144.5	18.3
105	8.5	164.5	388.5
106	28.5	165.5	227
107	49.5	162.5	971.9
108	67.5	163.5	761.3
109	87.5	164.5	155.5
110	109.5	163.5	22.1
111	129.5	163.5	26
112	147.5	163.5	180.6
113	169.5	165.5	208.3
114	188.5	163.5	324.1
115	207.5	162.5	141
116	227.5	163.5	0
117	248.5	163.5	266.3
118	6.5	182.5	174.6
119	26.5	184.5	206.1
120	46.5	184.5	762.4
121	68.5	185.5	918
122	87.5	182.5	10.7
123	109.5	185.5	2.7
124	127.5	185.5	0
125	149.5	184.5	0
126	167.5	185.5	99.7
127	188.5	183.5	10.9
128	206.5	185.5	61.8
129	229.5	185.5	137.6
130	248.5	184.5	502.3
131	7.5	203.5	187.8
132	29.5	203.5	24.5
133	48.5	205.5	968.3
134	67.5	202.5	97.4
135	88.5	205.5	0
136	108.5	202.5	17.9
137	126.5	203.5	100.3
138	146.5	202.5	342.4
139	168.5	204.5	636.6
Continued on next page			

Table G-2 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
140	186.5	204.5	332.9
141	208.5	205.5	258.3
142	229.5	202.5	451.2
143	249.5	202.5	0
144	8.5	225.5	82.1
145	26.5	223.5	198.1
146	47.5	225.5	394.7
147	67.5	223.5	0
148	88.5	224.5	0
149	107.5	224.5	174.2
150	128.5	225.5	530.3
151	148.5	222.5	602.3
152	168.5	224.5	173.1
153	189.5	225.5	184.4
154	209.5	222.5	590.3
155	227.5	222.5	639.5
156	249.5	223.5	240.9
157	9.5	244.5	81.1
158	28.5	244.5	60.3
159	49.5	244.5	343
160	66.5	244.5	0
161	86.5	243.5	0
162	107.5	243.5	12.9
163	129.5	242.5	107.4
164	147.5	245.5	209.1
165	167.5	243.5	17
166	188.5	242.5	146.6
167	209.5	244.5	166.9
168	229.5	243.5	119.9
169	247.5	245.5	234.4
170	8.5	263.5	124.3
171	29.5	263.5	312.6
172	48.5	262.5	863.8
173	69.5	262.5	0
174	86.5	263.5	12.1
175	107.5	262.5	187.8
176	126.5	263.5	159.3
177	147.5	265.5	79.4
178	166.5	265.5	283.1
179	187.5	264.5	92
180	206.5	262.5	125.2
181	229.5	262.5	27.2
182	246.5	264.5	22.4
Continued on next page			

Table G-2 – continued from previous page

No.	Xloc [m]	Yloc [m]	Conc. [ppm]
183	6.5	282.5	188
184	29.5	283.5	240.9
185	49.5	284.5	159.6
186	69.5	282.5	2.4
187	86.5	282.5	62.2
188	109.5	285.5	268.8
189	129.5	282.5	70.7
190	146.5	285.5	104.1
191	166.5	284.5	30.9
192	187.5	284.5	2.5
193	206.5	283.5	29.3
194	228.5	285.5	2.1
195	248.5	285.5	45.6

Bibliography

- Alvarez Grima, M., van Gelder, R., Heeren, J., Verichev, S. and van Wijk, J. (2011), Into the deep: a risk based approach for research to deep sea mining, *in* ‘CEDA Dredging Days 2011: Dredging and Beyond’, MTI Holland (member of IHC Merwede), Rotterdam.
- Anderson, T. (1958), *An introduction to multivariate statistical analysis*, John Wiley & Sons, New York.
- Banerjee, S., Gelfand, A. E. and Carlin, B. P. (2004), *Hierarchical Modelling and Analysis for Spatial Data*, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, London. ISBN: 978-1-58488-410-1.
- Benndorf, J. (2005), Efficient sequential simulation methods with implications to long-term production scheduling, Technical report, WH Bryan Mining Geology Research Centre.
- Benndorf, J. (2012a), ‘Adding project value by managing geological risk in mining’, *Natural Resource* **15**(1).
- Benndorf, J. (2012b), ‘Resource modeling and mine planning - part 1’, University Lecture.
- Borgman, L., Taheri, M. and Hagan, R. (1984), ‘Three-dimensional, frequency domain simulations of geological variables’, *Verley et al., eds., Geostatistics for Natural Resource Characterization, Part 1*.
- Caers, J. (2011), *Modeling Uncertainty in the Earth Sciences*, John Wiley & Sons Ltd., Chichester. ISBN: 978-1-119-99263-9.
- Chiles, J.-P. and Delfiner, P. (1999), *Geostatistics: modeling spatial uncertainty*, John Wiley & Sons Inc., Hoboken N.J. ISBN: 9780470183151.
- Clark, I. (1979), *Practical Geostatistics*, Elsevier Applied Science, London. ISBN: 085334843.
- Cooley, J. and Turkey, J. (1965), ‘An algorithm for the machine calculation of complex fourier series’, *Math. Comput.* **19**(90).
- David, M. (1977), *Geostatistical ore reserve estimation*, Elsevier, Amsterdam.

- Diggle, P. J. and Ribeiro Jr., P. J. (2007), *Model-based Geostatistics*, Springer Series in Statistics, Springer Science, New York. ISBN-13: 978-0-387-32907-9.
- Dimitrakopoulos, R. and Luo, X. (2004), 'Generalized sequential gaussian simulation on group size v and screen-affect approximations for large field simulations', *Mathematical Geology* **36**(5).
- Dimitrakopoulos, R. (1998), 'Conditional simulation algorithms for modelling orebody uncertainty in open pit optimisation', *International Journal of Surface Mining, Reclamation and Environment* **12**(4).
- Dimitrakopoulos, R., Farelly, C. T. and Godoy, M. (2002), 'Moving forward from traditional optimization: Grade uncertainty and risk effect in open-pit design', *Mining Technology* **111**(1).
- Dominy, S. C., Noppe, M. A. and Annels, A. E. (2002), 'Errors and uncertainty in mineral resource and ore reserve estimation: The importance of getting it right', *Journal of Exploration and Mining Geology* **11**(1-4).
- Evans, I. and Pomeroy, C. D. (1966), *The Strength, Fracture and Workability of Coal*, Pergamon Press, New York.
- Fenton, G. A. (1990), Simulation and Analysis of Random Fields, PhD thesis, Princeton University, Princeton.
- Fenton, G. A. and Griffiths, D. (2008), *Risk Assessment in Geotechnical Engineering*, John Wiley & Sons, Inc., Hoboken. ISBN: 978-0-470-17820-1.
- Fenton, G. A. and Vanmarcke, E. H. (1990), 'Simulation of random fields via local average subdivision', *Journal of Engineering Mechanics* **116**(8).
- Glacken, I. M. (1996), Change of support and use of economic parameters for block selection, in 'Geostatistics Wollongong'.
- Godoy, M. (2003), A multi-stage approach to profitable risk management for strategic planning in open pit mines, PhD thesis, WH Bryan Mining Geology Research Centre, The University of Queensland, Brisbane.
- Godoy, M., Dimitrakopoulos, R. and Costa, J. (2001), 'Economic functions and geostatistical simulation applied to grade control', *Mineral Resource and Ore Reserve Estimation - The AusIMM Guide to Good Practice (ed; A. C. Edwards)*.
- Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Applied Geostatistics Series, Oxford University Press, New York. ISBN: 0-19-511538-4.
- Griffiths, D. V. and Fenton, G. A., eds (2007), *Probabilistic Methods in Geotechnical Engineering*, Vol. International Centre for Mechanical Sciences (CISM), Courses and Lectures, Springer Wien New York, Udine. ISBN: 978-3-211-73365-3.
- Isaaks, E. H. and Srivastava, R. M. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press, New York. ISBN: 0-19-505012-6.

- Journel, A. and Huijbregts, C. (1978), *Mining Geostatistics*, Blackburn Press, London. ISBN-13: 978-1930665910.
- Kolmogorov, A. N. (1950), *Foundations of the theory of probability*, Chelsea Publishing Co., New York.
- Lipton, I. (2012), Mineral resource estimate: Solwara project, bismarck sea, png, Technical report, Golder Associates.
- Mantoglou, A. and Wilson, J. (1981), Simulation of random fields with the turning bands method, report #264, Technical report, Massachusetts Institute of Technology, Dept. Civil Eng.
- Martinez, L. A. (2009), Why accounting for uncertainty and risk can improve final decision-making in strategic open pit mine evaluation, in 'Project Evaluation Conference', Melbourne.
- Matheron, G. (1971), *The theory of Regionalized Variables and Its applications*, Vol. Les Cahiers du Centre de Morphologie Mathematique, Ecole Nationale Superieure des Mines de Paris, Fontainebleau.
- Matheron, G. (1973), 'The intrinsic random functions and their applications', *Advances in Applied Probability*.
- Murton, B. J., Parsons, L. M., Hunter, P. and Miles, P. (2001), Global non-living resources on the extended continental shelf: Prospects at the year 2001, Technical report, International Seabed Authority.
- Nautilus Minerals Inc. (2012), 'Resource extraction: How do we plan to extract resources at solwara 1', www.nautilusminerals.com/s/resourceextraction.asp.
- Nowak, M. and Verly, G. (2004), A practical process for simulation, with emphasis on gaussian simulation, in 'Orebody Modelling and Strategic Mine Planning', November.
- OceanfLORE (2012), 'Deep sea mining markets', www.oceanflore.com/6.
- Priestley, M. B. (1981), *Spectral Analysis and Time Series. Volume 1: Univariate Statistics*, Probability and Mathematical Statistics, Academic Press, New York. ISBN: 0125649010.
- Rademaker, M. and Koororshy, J. (2010), The global challenge of mineral scarcity, in 'Enriching the Planet, Empowering Europe - Optimising the use of natural resources for a more sustainable economy', The Hague Centre for Strategic Studies (HCSS), Houston.
- Savage, S. (2003), *Decision Making with Insight*, South-Western College Pub. ISBN-13: 9780534386399.
- Searle, S. and Smit, H. (2011), The challenges of developing and operating an underwater mine, in 'Offshore Technology Conference', IHC Marine and Mineral Projects, Houston.
- Smith, G. (2011), Deepwater seafloor resource production - the bismarck sea development project, in 'Offshore Technology Conference', Nautilus Minerals Inc., Houston.

- Soil Machine Dynamics Ltd. (2012), 'Mining products', <http://smd.co.uk/products/mining/list.htm>.
- Vallee, M. (2000), 'Mineral resource + engineering, economic and legal feasibility = ore reserve', *CIM Bulletin* **93**(1038).
- van de Ketterij, R. (2010), Final conclusions femp workshop, Technical report, IHC Merwede & Federation of European Mineral Programs (FEMP).
- Vann, J., Jackson, S., Bye, A., Coward, S., Moayer, S., Nicholas, G. and Wolff, R. (2012), Scenario thinking - a powerfull tool for strategic planning and evaluation of mining projects and operations, in 'Project Evaluation Conference', Melbourne.
- Webster, R. and Oliver, M. A. (2007), *Geostatistics for Environmental Scientists*, Statistics in Practive, John Wiley & Sons, Ltd., Chichester. ISBN-13: 978-0-470-02858-2.
- Yaglom, A. (1962), *An introduction to the theory of stationary random functions*, Prentice-Hall, Englewood Cliffs N.J.

Glossary

List of Acronyms

SMS	Seafloor Massive Sulphides
NPV	Net Present Value
DSM	Deep Sea Mining
SPT	Seafloor Production Tools
RALS	Riser and Lifting System
SSLP	Subsea Slurry and Lifting Pump
PSV	Production Support Vessel
OPS	Offshore Production System
UCS	Uniaxial Compressive Strength
BTS	Brazilian Tensile Strength
cdf	cumulative distribution function
pdf	probability density function
SMU	Smallest Mineable Unit
ARMA	Autoregression Moving Average
BLUE	Best Linear Unbiased Estimate
EDA	Exploratory Data Analysis
MA	Moving Average
CMD	Covariance Matrix Decomposition
DFT	Discrete Fourier Transform

FFT	Fast Fourier Transform
TB	Turning Bands
LAS	Local Average Subdivision
SGS	Sequential Gaussian Simulation
GSGS	Generalized Sequential Gaussian Simulation
DBSim	Direct Block Simulation
DAS	Data and Area Statistics
NPV	Net Present Value }

List of Symbols

\$	Dollars
%	Symbol used to indicate fraction (percentage)
$\gamma(\mathbf{h})$	Semi-variance as a function of lag vector
	Modeled, expected random variable
B	Covariance matrix, covariance between child cells
h	Lag vector, distance and direction between two points
L	Lower triangular matrix (Choleksy decomposition)
R	Covariance matrix, covariance between parent cells
S	Covariance matrix, covariance between parent and child cells
U	Vector with independently drawn random components
$\mathbf{x}_\alpha$	Sampling location
$\mathbf{x}_\eta$	Location of interest
$\mathbf{Z}^{i+1}$	Vector with child values
$\mathbf{Z}^i$	Vector with parent values
A	Study area
E	Collection of prediction points
ρ	Rock density
$\rho(\mathbf{h})$	Correlation as a function of lag vector
σ_0	Nugget
$\sigma_0 + \sigma_1$	Sill, point variance
σ_R^2	Error variance
a_0	Short range of nugget structure
a_1	Range of spherical model, correlation length
b_α	Covariance between prediction location $\mathbf{x}_\eta$ and sampling location $\mathbf{x}_\alpha$
$C(\mathbf{h})$	Covariance as a function of lag vector

C_m	Mining costs
C_p	Processing costs
$C_{\alpha\beta}$	Covariance between prediction locations $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$
CF_t	Cash flow at period t
$E[\]$	Expectation
f	Variance adjustment factor
$F(\mathbf{x}_\eta; b)$	Cumulative density function
$f(\mathbf{x}_\eta; b)$	Probability distribution function
g/ton	Concentration, grams per ton
$I(\mathbf{x}_\eta; b)$	Indicator variable
kg	Kilogram
kJ	Kilojoule
kN	Kilonewton
kt	A unit of weight, equal to 1000 metric tons (kiloton)
kWatt	Kilowatt
m	The standard unit of length (meter)
m^+	Mean grade if block is ore
m^-	Mean grade if block is waste
MPa	Megapascal
N_1	Number of grid cells in x direction
N_2	Number of grid cells in y direction
p	Metal price
P_o	Probability that block is ore
P_w	Probability that block is waste
ppm	Parts per million
Pr_{ore}	Profit for block of ore
Pr_{waste}	Costs for block of waste
q	Quantile
R	Discount rate
r	Recovery
$z(\mathbf{x}_\alpha)$	Realization of random variable at sampling location
$Z(\mathbf{x}_\eta)$	Random variable at location of interest
Z_c	Conditional field
Z_k	BLUE based on measurement data
Z_s	BLUE based on simulated values at sampling locations
Z_u	Unconditional field
$z_{cut-off}$	Economic cut-off grade