Geotechnical Risk Management

A Statistical Framework to Assess the Effect of Sampling Density on Dredge Performance Forecasting

Sil Roelen
Geotechnical Risk Management

A Statistical Framework to Assess the Effect of Sampling Density on Dredge Performance Forecasting

by

Sil Roelen

to obtain the degree of Master of Science at the Delft University of Technology, to be defended publicly on Friday October 14, 2016 at 11:00 AM.

Student number: 4038339
Project duration: Dec 1, 2015 – October 7, 2016
Thesis committee: Prof. Dr. Michael A. Hicks, TU Delft (Chair)
Prof. Dr.-Ing. Jörg Benndorf, TU Bergakademie Freiberg
Dr. Mario Alvarez Grima, Royal IHC
Ir. Tom Wambeke, TU Delft & Royal IHC

An electronic version of this thesis is available at http://repository.tudelft.nl/.
Abstract

This thesis proposes a methodology towards quantifying the uncertainty resulting from the density and spatial arrangement of samples and assessing their effect on forecasting key performance indicators, related to dredging and (wet) mining operations. Aside from quantifying this uncertainty, another objective of this research is to develop a framework for optimising the sampling density, as the availability of samples for geotechnical site characterisation is often limited.

The development of the framework is based on real data, concerning a heavy mineral sand deposit: this data, however, proved to be insufficiently dense for uncertainty analysis. Therefore a decision was made, based on this observation, to create a virtual reality in high density which allows for performing the desired investigations.

The assessment of geotechnical risk consists of multiple steps: data interpretation and preparation, geostatistical simulation, data processing and, lastly, analysing the results. During preparation the geotechnical units were identified, suitable simulation methods based on literature were chosen and the decision was made to generate a virtual dataset. For continuous properties Sequential Gaussian Simulation was used; in case of a categorical or a discrete property, Sequential Indicator Simulation was adopted. The simulations result in 100 realisations, for each simulated property: unit top and bottom boundaries, SPT \( N \) values and the locations of indurated zones. These realisations then allow for analysing possible distributions and the uncertainty related to these parameters and their propagation through transfer functions.

Five different sampling strategies were implemented: three regularly spaced grids (100, 250 and 500 m) and two clustered grids (500 m regularly spaced with preferred zones). After combining the various simulation results, the effect of these grids is assessed by investigating the inferred variograms, RMSE, P90 and P10 values and distributions of production forecasts.

It is observed that the geostatistical methods are suitable for geotechnical risk analysis and, generally, an increase in sampling density leads a decrease in prediction uncertainty. However, the absolute number of samples is not directly related to this trend: preferred samples and their locations should also be taken into account here, as their effect proved to be less predictable.
About ten months ago I started my research on the subject of sampling uncertainty and its implications on predicting key performance indicators in dredging, resulting in this document and forming the conclusion of my MSc degree programme in Applied Earth Sciences. The project was carried out partly in cooperation with Royal IHC, whose support I would hereby like to gratefully acknowledge.

First of all I would like to express my gratitude towards ir. Tom Wambeke, my supervisor, for introducing me to both IHC and the thesis subject. His continuous help throughout the project, discussions and ideas were really helpful. He provided me with inspiration towards choosing the right direction, while still allowing me to work autonomously.

I would like to thank Dr. Mario Alvarez Grima, from Royal IHC, for his patience, the useful discussions and guidance he offered me over the course of this project. He was always available when I had questions; willing to listen and discuss any issues. I am also thankful to ir. Bart Hogeweg from Royal IHC, who has been very helpful in introducing me to the thesis topic and relating the theoretical methods to practical applications.

Furthermore, I would like to thank the students and employees at Royal IHC for the pleasant and motivating working environment.

Likewise, my thanks to Prof. Dr. Michael Hicks from TU Delft and Prof. Dr.-Ing. Jörg Benndorf from TU Bergakademie Freiberg (formerly from TU Delft). I am grateful for the discussions we had during the (committee) meetings and the advice, proving to be very helpful in the preparation of this document.

Lastly, I would like to express my thanks to my family and friends, for the motivation and encouragement over the duration of this research.

Sil Roelen
Delft, October 2016
Contents

List of Figures ix
List of Tables xi

1 Introduction 1
  1.1 Problem Statement .................................................. 2
  1.2 Research Objectives .................................................. 3
  1.3 Thesis Outline ......................................................... 3

2 Literature Review 5
  2.1 The Relevance of Geostatistics ....................................... 5
  2.2 Site Investigation ..................................................... 6
    2.2.1 In Situ Testing .................................................... 6
    2.2.2 Uncertainty in Sample Data ...................................... 9
  2.3 Estimation ............................................................. 10
    2.3.1 Random Variables .................................................. 10
    2.3.2 Inverse Distance Weighting ..................................... 10
    2.3.3 Kriging ............................................................ 12
    2.3.4 Estimation Uncertainty .......................................... 15
  2.4 Simulation ............................................................ 18
    2.4.1 Sequential Simulation ............................................ 19
    2.4.2 Simulation Uncertainty .......................................... 21
    2.4.3 Simulation Example Cases ....................................... 22
  2.5 Concluding Remarks ................................................ 25

3 Project Background 29
  3.1 Geology ............................................................... 29
  3.2 Geotechnical Investigation .......................................... 29
  3.3 Overview of the Dataset ............................................. 32
  3.4 Excavation Equipment ............................................... 36

4 Methodology 37
  4.1 Simulation Process Overview ........................................ 37
  4.2 A Virtual Exhaustive Dataset ........................................ 39
    4.2.1 Grid Definition ................................................... 39
    4.2.2 Layer Boundaries ................................................ 40
    4.2.3 Indurated Zones .................................................. 42
    4.2.4 Geotechnical Properties ....................................... 45
  4.3 Transfer Functions .................................................. 48
    4.3.1 Production Forecasting .......................................... 48
    4.3.2 Production Costs ................................................ 49
  4.4 Experimental Setup .................................................. 49
    4.4.1 Varying the Sampling Density ................................... 49
    4.4.2 Target Production ................................................. 53
    4.4.3 Incorporating Indurated Zones in Production .................. 54
  4.5 Software Implementation ............................................ 55

5 Results 57
  5.1 Spatial Correlation .................................................. 58
  5.2 Simulation Validation ............................................... 62
    5.2.1 P10, P90 and Sample Locations ................................ 62
    5.2.2 Root Mean Square Error ........................................ 64
Comparing the Sampling Strategies

6 Univariate Statistics

6.1 General Conclusions

6.2 Specific Conclusions

6.3 Practical Implications of the Observed Results

6.4 Recommendations for Further Research

A Exploratory Data Analysis

A.1 Error Check

A.2 Univariate Statistics

A.3 Bivariate Statistics

B SGeMS

B.1 Normal Score Transformation

B.2 2D Sequential Gaussian Simulation

C Variograms per Sampling Strategy

C.1 Experimental Variogram Parameters

C.2 Inferred Variogram Models

D 2D Slice Maps

D.1 Target at 600 m$^3$/h

D.2 Target at 900 m$^3$/h

D.3 Target at 1200 m$^3$/h

Nomenclature

Bibliography
List of Figures

1.1 Relation between available information and converting mineral resources to ore reserves, by JORC (2012) ........................................ 2
2.1 A schematic representation of the SPT (Kovacs et al., 1981) ..................... 7
2.2 An SPT-CPT correlation function (Robertson et al., 1983) .................... 8
2.3 Grain size estimation chart (Robertson et al., 1983) ............................ 8
2.4 Inverse Distance Weighting of surface level measurements: \( p = 2 \) (a) and \( p = 3 \) (b) .... 11
2.5 Simple Kriging estimation of surface level measurements (a) and kriging variance (b, in normal score domain) ..................................... 14
2.6 Indicator Kriging example of artificial binary data ................................ 15
2.7 Estimation vs. simulation; when investigating a non-linear transfer function, based on Martinez (2009), Wambeke (2013) ....................... 17
2.8 Sequential Gaussian Simulation of surface level measurements: four out of 100 equiprobable realisations ........................................ 20
3.1 Geotechnical units from simulation; vertical scale x20 ......................... 31
3.2 Borehole locations, according to their GPS coordinates ....................... 33
3.3 Uncorrected cone tip resistance (MPa) Zone 2 (a) & Zone 3 (b) ......... 33
3.4 SPT-CPT correlation: SPT \( N \) capped at 50 (a) and uncapped (b) ........ 35
3.5 CSD cutting process (van Rhee, 2012) ............................................. 36
4.1 General overview of the simulation and processing methodology .......... 38
4.2 Flowchart describing the different steps of the simulation process .......... 39
4.3 A representation of the search ellipsoid (or neighbourhood), as defined by the minimum, medium and maximum ranges ......................... 40
4.4 Variogram models inferred based on original samples ......................... 42
4.5 3D grid containing layers (geotechnical units 2 & 3 in yellow, unit 'C'); vertical scale x20. 43
4.6 Variogram model fitted to experimental points, indurated layers ............ 44
4.7 3D simulated grid, showing one realisation of indurations in orange (indicator '1'), to be constrained by unit boundaries; vertical scale x20 ............. 45
4.8 Variogram model fitted to experimental points, SPT \( N \) values ............... 46
4.9 3D simulated grid showing one realisation of SPT \( N \) values, to be constrained by unit boundaries; vertical scale x20 ............................. 47
4.10 3D simulated grid showing one realisation of SPT \( N \) values, layer boundaries and indurations; vertical scale x20 ............................. 47
4.11 Mining sequence used in production calculation; black arrows indicate the start of the next horizontal cut (increasing in depth over time) .......................... 49
4.12 Spatial representation of the selected sampling strategies .................... 51
4.13 Production forecasting based on geotechnical parameters ................. 53
4.14 Two dimensional slice of 3D grid containing production values: examples of indurated layers are the blue lines, indicating low production, surrounded by a red area (high production, i.e. soft material) .............. 54
5.1 Ergodic fluctuations in simulations (10 randomly chosen realisations); top boundary, based on coarse (a) and cross (b) sampling schemes ............... 58
5.2 Comparison of variograms for layer top, based on different sampling strategies; "original" referring here to the variogram used to generate the virtual exhaustive dataset ................. 59
5.3 Comparison of variograms for layer bottom, based on different sampling strategies; "original" referring here to the variogram used to generate the virtual exhaustive dataset ................. 59
5.4 Comparison of variograms for indurations, based on different sampling strategies; "original" referring here to the variogram used to generate the virtual exhaustive dataset.

5.5 Comparison of variograms for SPT \( N \) values, based on different sampling strategies; "original" referring here to the variogram used to generate the virtual exhaustive dataset.

5.6 Interval defined by the P10 and P90 values, along a one dimensional line in the x direction, y coordinate 1012.5; based on coarse sampling grids.

5.7 P10 and P90 plots along a one dimensional line; based on cross (a), rectangle (b), medium (c) and fine (d) sampling grids.

5.8 Average RMSE over all simulation grid locations; top boundary (a), bottom boundary (b) and SPT \( N \) values (c).

5.9 Volume contained within simulated top and bottom boundaries (a) and indurated zones as fraction of total layer volume (b).

5.10 Two dimensional slice of 3D grid containing discrete production probabilities; using a production target of 600 m\(^3\)/h.

5.11 Two dimensional slices; from left to right: discrete production target probability, the original deposit (where either the target is missed or reached) and continuous probability, all at 600 m\(^3\)/h target and a depth of -4.75 m. From top to bottom: varying the sampling strategy.

5.12 Effect of sampling density on production estimation; using a production target of 600 m\(^3\)/h (a), 900 m\(^3\)/h (b) and 1200 m\(^3\)/h (c).

5.13 Effect of number of samples on uncertainty reduction; using a production target of 600 m\(^3\)/h (a), 900 m\(^3\)/h (b) and 1200 m\(^3\)/h (c).

5.14 Volume excavated over complete life of mine. Note that values on the y-axis are multiplied by a factor of 10\(^7\).

5.15 Volume excavated during the first month of mine operation versus volumes based on production targets.

5.16 Costs of excavation during the first month of mine operation.

B.1 Histogram and cdf before (a) and after (b) transformation.

B.2 Running software with administrator privileges.

B.3 Importing data files.

B.4 Data visualisation.

B.5 Transforming data histogram.

B.6 Variogram modelling parameters.

B.7 Variogram modelling; graphical user interface.

B.8 Sequential Gaussian Simulation.

B.9 Cartesian grid definition.

B.10 2D simulation visualisation.

C.1 Variogram models; layer top.

C.2 Variogram models, layer bottom.

C.3 Variogram models, indurated layer locations.

C.4 Variogram models, SPT \( N \) values.

D.1 Two dimensional slices of discrete production target probability at 600 m\(^3\)/h target.

D.2 Two dimensional slices of discrete production target probability at 900 m\(^3\)/h target.

D.3 Two dimensional slices of discrete production target probability at 1200 m\(^3\)/h target.
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Overview of Methods</td>
<td>26</td>
</tr>
<tr>
<td>3.1</td>
<td>Geotechnical units description</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Correlation function parameters</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Simulation grid specifications</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>Experimental variogram parameters</td>
<td>41</td>
</tr>
<tr>
<td>4.3</td>
<td>2D SGS parameters</td>
<td>42</td>
</tr>
<tr>
<td>4.4</td>
<td>Indicator experimental variogram parameters</td>
<td>44</td>
</tr>
<tr>
<td>4.5</td>
<td>SIS parameters</td>
<td>44</td>
</tr>
<tr>
<td>4.6</td>
<td>SPT experimental variogram parameters</td>
<td>46</td>
</tr>
<tr>
<td>4.7</td>
<td>3D SGS parameters</td>
<td>46</td>
</tr>
<tr>
<td>4.8</td>
<td>Cutter head production transfer function</td>
<td>48</td>
</tr>
<tr>
<td>4.9</td>
<td>Production cost estimation parameters</td>
<td>49</td>
</tr>
<tr>
<td>4.10</td>
<td>Sampling strategies specifications</td>
<td>50</td>
</tr>
<tr>
<td>C.1</td>
<td>Variogram parameters for indurations; different sampling strategies</td>
<td>92</td>
</tr>
<tr>
<td>C.2</td>
<td>Variogram parameters for SPT values; different sampling strategies</td>
<td>93</td>
</tr>
<tr>
<td>C.3</td>
<td>Variogram parameters for elevations (top of layer); different sampling</td>
<td>94</td>
</tr>
<tr>
<td>C.4</td>
<td>Variogram parameters for elevations (bottom of layer); different</td>
<td>94</td>
</tr>
</tbody>
</table>

- Spacings in brackets indicate different sampling strategies.
- The tables are organized to provide a comprehensive overview of the methods and parameters used in the study.
Production estimation (based on geotechnical data), equipment selection and the resource model are important aspects forming the basis for a successful mining operation. The 3D models used contain properties of the desired material, like metal concentration, strength, locations of faults and other structures.

Information is obtained through site investigation techniques and laboratory tests. This information subsequently serves as input for the resource model. In case of soil sampling the borehole spacing may vary over the study area, depending on local structures, features and direction of variability.

This spacing of samples is an important factor: it affects not only the accuracy of the resulting predictions but also the required time and costs of performing them. Sampling is usually done according to a certain sampling grid/spacing and strategy. Eventually, geostatistical methods are applied to the data to estimate properties at unsampled locations.

Examples of these techniques are:

- Estimation: generate a single realisation of a possible model, giving a single value of each parameter per (unsampled) location.
- Simulation: generate multiple equiprobable realisations of the resource model which may be used to determine a probability density function of each parameter per location.

Taking this into consideration, certain properties may be sampled at different densities than others. For example, when metal grades are important in forecasting the resource potential and economic feasibility of its exploitation, geological properties may be sampled in a denser grid compared to geotechnical properties. However, similar to grade information which is valuable for profit estimations, geotechnical data is essential for the estimation of the production, forming another important part of determining the mining project’s feasibility.

Since the development of Kriging interpolation it has been used in estimation of, amongst others, resource potential, petroleum reserves and soil pollution, while also forming the basis of some (conditional) simulation methods. These applications usually result in a 3D model of the subsurface, e.g. a block model, as used in mining.

Applying these methods to geotechnical measurements is, however, not yet common practice (Bye, 2006), while creating such a 3D model containing the geotechnical properties of the in situ rock mass, may aid in the following operations: predict rock quality, optimise production, slope stability optimisation and planning and estimating costs.

For instance, in case of production optimisation, it may be useful to have a 3D model containing multiple realisations to assess the probability of occurrence for a certain property. However, input data for such a model requires a certain number of samples in order to attain the desired accuracy.

The relation between available information and associated confidence is described in the JORC Code, by JORC (2012), of which a schematic overview is given by figure 1.1: from top to bottom the amount
of and confidence in the available information increases; from left to right the so-called "Modifying Factors" are applied, converting resources to reserves. The effect of economic feasibility, processing and metallurgical factors, amongst others, are considered here. At this stage geotechnical properties are also important, as they affect the physical feasibility of the project, i.e. the excavatability, and thus subsequent equipment selection and production performance forecasting.

For the purpose of this thesis a methodology is developed to investigate the effect of sampling density on the uncertainty characterisation and accuracy of the model. Next, the propagation of this uncertainty through transfer functions for calculating key performance indicators is assessed.

![Figure 1 General relationship between Exploration Results, Mineral Resources and Ore Reserves.](image)

**Figure 1.1: Relation between available information and converting mineral resources to ore reserves, by JORC (2012).**

### 1.1. Problem Statement

A geotechnical dataset relating to a heavy mineral deposit forms the basis of this project. This deposit is assumed to be mined by using a cutter suction dredger in an artificial pond. This deposit poses several problems originating from its geotechnical properties: the failure exhibited by the material was not as desired, along with another problem of the deposit, being the presence of hard, indurated, lenses. Based on these experiences it was decided to perform a more comprehensive site investigation and develop a mining plan to handle this type of deposit.

Although the challenges of this particular deposit are not the main cause for this thesis, the geotechnical data resulting from the investigation mentioned above is used for the purpose of testing the methodology proposed in this document.

The relation of geotechnical properties to the failure behaviour itself and other types of soil behaviour are out of scope for this thesis. The emphasis lies on the uncertainty resulting from sampling and its effect on forecasting key performance indicators and associated project risk.

The aforementioned data is used as input for creating a subsurface model. However, since one of the goals is to investigate the effect of sampling density on uncertainty, a set of high density samples is required, which does not seem to be the case here. Therefore, the simulation is split into two parts: the first part consists of creating a dense (exhaustive) virtual data set by simulation based on the source data, whereas the second part consists of sampling a single realisation (selected randomly) of this virtual data and using the results as input for further modelling.
1.2. Research Objectives
As explained in section 1, deciding on the feasibility of exploiting a certain resource is based on limited data and should therefore be approached carefully. Due to these limitations of the available information, decisions on critical aspects like equipment design and selection have to be made while taking uncertainty into account. The goal of this research project is to evaluate the impact of geotechnical sampling density on the model accuracy and uncertainty of predicted equipment performance.

The objectives of this thesis are:

- Develop a methodology to quantify the spatial uncertainty on geotechnical parameters given the limited amount of data gathered during site investigation.
- Establish relationships to translate the modelled geotechnical parameters into key performance indicators such as energy usage, power requirement.
- Develop a process flow to compare the impact of different sample densities on the characterized geotechnical parameters and predicted equipment performance.
- Design a framework to select an optimal sampling density considering exploration costs on one hand and an acceptable level of risk on the other.

1.3. Thesis Outline
To illustrate the approach for reaching the objectives mentioned in section 1.2, the overall structure of the thesis may be described as follows:

1. Review current state of the art of geostatistics and find suitable simulation methods.
2. Define requirements for data based on exploratory data analysis.
3. Design a method to observe the impact of sampling density on geotechnical parameters.
4. Relate geotechnical parameters to transfer functions to determine key performance indicators.
5. Develop a framework, which describes the methodology and is flexible regarding to the type of input data. In this way the resulting framework should be applicable in a variety of situations and types of deposits.

Chapter 1 The reader is introduced to the thesis topic, the relevance of the subject and objectives of the investigation.

Chapter 2 The current state is described as found in literature: an overview of exploration and interpolation methods is given along with their applications; also, a modelling workflow is proposed.

Chapter 3 Information regarding the project background is provided: geology of the initially sampled area, site investigation and excavation methods.

Chapter 4 A methodology towards uncertainty assessment is proposed based on sampling density, explaining the choice for a virtual dataset and introducing the transfer functions and software implementation.

Chapter 5 The results of the proposed methodology are shown. This is not limited to processed results, but also contains intermediate and direct simulation output.

Chapter 6 In this chapter, first the main conclusion is given, followed by more specific conclusions and recommendations.
This chapter aims to provide the reader with an overview of the current state of the research fields related to the subject of this thesis. Several concepts related to site investigation, sampling and geostatistical estimation and simulation methods are introduced and evaluated.

2.1. The Relevance of Geostatistics

Observations in earth and natural sciences are usually distributed in either time or space, and possibly both. This spatial, or temporal, arrangement of sample points may be a result of practical considerations like access to a location or available time and may lead to a data set not entirely covering the study area: for instance, an area containing a mineral deposit might have been sampled in a regular grid, at 200 m intervals.

This degree of sample coverage could pose a problem if a property value is needed for all possible, locations in the study area. Subsequently, a problem like this requires the application of interpolation and extrapolation methods to estimate the property value at an unsampled location.

In general the degree of spatial dependency is related to the origin of the studied soil property (Chang et al., 1998). Intrinsic properties, resulting from diagenetic processes or differences in source rock present a strong spatial dependency, whereas extrinsic properties, e.g. resulting from human interference with soil, like fertilization, show weak spatial dependence.

This means that geostatistical methods could be a useful tool to investigate spatial dependency and its inherent uncertainty as a result from sampling, when dealing with data of a geological origin. Relatively simple methods exist to deal with such problems, an example being inverse distance weighting. This method takes into account the distance from the unsampled location to a known data point and applies a weighting factor based on that distance (Isaaks and Srivastava, 1989). This way values close to the location are given priority over values further away when determining the estimate.

Since the 1960s increasingly advanced methods have been in development, arising from the need to more accurately estimate recoverable reserves in an ore deposit (Goovaerts, 1997). The development was initiated by the South African mining engineer D. Krige, who developed a BLUE (best linear unbiased estimate) type of estimator (Cressie, 1990). This method takes not only the distance between unsampled and sampled locations into account, but also the correlation to the known data points.

It has found applications in a wide variety of fields: mining engineering, petroleum engineering, oceanography and environmental sciences, among others.

Aside from providing estimations, kriging also gives an indication of estimation uncertainty by the kriging variance. However, this is only based on the type of model and locations of samples and not on the actual sample data.

In order to assess the spatial uncertainty more advanced techniques are required (Goovaerts, 2001), such as stochastic simulation. Often based partly on kriging estimation, these simulation methods gen-
erate multiple possible and equally probable representations of the area of interest, called realisations. Whereas kriging only provides a single estimate per location, simulation provides multiple. This is useful for investigating the uncertainty of an estimate at a particular location: the variability of the data between realisations is described by a probability density function, instead of a single value.

2.2. Site Investigation
A site investigation, comprised by a variety of sampling methods, is an important part of the feasibility study process. The results of this investigation, after data processing, are helpful for decision making and, thus, have a major impact on the possible continuity of the project. Not only the economic potential is evaluated, but also the excavatability of the material containing the resource. In determining these factors a good balance should be found between reducing the uncertainty by sampling, and the consequence, or risk, of uncertain knowledge (Whitman, 2000).

The process of obtaining samples is often time-consuming and costly: in order to reduce expenditure, while considering the type of deposit, a certain sampling grid is used. The grid depends on the goal of the sampling: if larger structures are to be detected then a large spacing is necessary, whereas a decreased spacing applies to smaller structures. Another possibility is to combine several sampling densities. This could result in saving costs by only applying a high sampling density if needed: start with a coarse grid and, if necessary, adjust sampling density at important locations like high ore grade or areas posing excavation difficulties. Utilising varying sample spacings over the region could also benefit the estimation process (Marchant and Lark, 2007), as this may aid in variogram inference, especially at small lag distances.

Note that preferential sampling should be taken into account in subsequent geostatistical processing (explained in section 2.3.3).

2.2.1. In Situ Testing
The cone penetration test (CPT) and the standard penetration test (SPT) are two of the most widely used in situ tests. In North America the SPT is a popular method (Robertson et al., 1983), whereas in the Netherlands, and other European countries where it is more likely to encounter soft soils, the CPT is more popular. Different standards for the CPT have been developed, most of them based on the Dutch version (Lunne et al., 1997).

While the names of both tests imply similarity, there are significant differences between them. They do, however, both rely on the principle of penetrating the subsurface and measuring the resistance in the process.

**Standard Penetration Test** The SPT is performed by using a hollow tube to penetrate the soil (in case of a split-spoon barrel), which is able to obtain a sample. However, in case of gravelly soil or soft rock, a cone may be used. The penetration is induced by dropping a free falling hammer onto the drill rods, of standardised weight (63.5 kg or 140 lb). The resistance to penetration is defined as the number of blows, or $N$ value, necessary to attain a penetration depth of 30 cm (or 1 ft).

According to Robertson et al. (1983) results of the test may vary considerably due to differences in the test procedures and equipment used, counter to the observation that methods have been standardised. Several causes of this problem are identified:

- The size of the drill hole.
- The number of windings of the rope around the cathead (see figure 2.1).
- The way of borehole wall stabilisation: e.g. drilling mud, casing.
- The type of sampling tube.
- The penetration depth measured: 0 - 0.30 m or 0.15 - 0.45 m).

The factor having the largest influence on the reported $N$ value was found to be the amount of energy delivered to the drill rods. In order to solve this problem, further standardisation is required or correction factors may be applied to the data (refer for example to (NEN, 2005) for further details).
2.2. Site Investigation

Figure 2.1: A schematic representation of the SPT (Kovacs et al., 1981).

Cone Penetration Test  The CPT consists of pushing a cone into the soil while measuring resistance. The cone, or penetrometer, is attached to a series of push rods and is pushed into the soil at a constant rate of penetration. Usually measurements are taken continuously of the sleeve friction, \( f_s \), and cone resistance, \( q_c \). If a piezocene is used, it is also possible to measure the pore pressure around the cone, which is referred to as CPTu.

The dimensions of the cone are standardised NEN (2012), with a cone base area of 1000 mm\(^2\) and a nominal apex angle of 60°. The cone resistance is defined as \( q_c = Q_c / A_c \), with \( Q_c \) the force on the cone measured by a load cell and \( A_c \) the projected area of the cone.

Due to the continuous nature of the measurement it has certain advantages over the SPT: easier to detect stratification and other small features and the reproducibility of the test (the procedures are standardised and often automated).

SPT-CPT Correlation  Despite the disadvantages of the SPT, it is still one of the most widely used methods in North America and other areas. Geotechnical engineers have often developed correlation and interpretation methods based on their own experiences.

With an increasing popularity of the CPT, a reliable correlation between SPT and CPT could be a welcome development for using CPT data in existing SPT based design procedures. Consequently, a wide range of \( q_c / N \) ratios and correlation functions have been published, applying to different types of deposits.

By Robertson et al. (1983) an attempt is made at providing a historical overview of SPT-CPT correlations. Two main factors were identified, affecting the correlation:

- Grain size: when plotting \( q_c / N \) ratios, sourced from various publications, against the mean grain size \( d_{50} \) an increasing scatter is observed with increasing grain size.

- Energy transmitted: another major influence was found to be the hammer type, affecting the amount of energy transmitted to the rods.

- Stratification: since a CPT takes measurements at smaller depth intervals, transition between strata may result in rapid variations in data.
Considering the above, a correlation is proposed by Robertson et al. (1983) for determining the \( q_c/N \) ratio based on \( d_{50} \), under the assumption that the same hammer type is used (donut type hammer with rope and cathead):

\[
q_c/N = 4.70 - 0.05 \cdot c_f
\]  

(2.1)

In equation (2.1) \( N_{55} \) denotes the SPT \( N \) value for an energy efficiency of 55% and \( c_f \) the fines content. Aside from the scatter plot and the proposed function fitted, a statistical indication of correlation
2.2. Site Investigation

is, however, not presented. Using the fines content has the benefit of requiring only a single sieve size (#200 or 0.0074 mm, according to Chin et al. (1990)) to determine the percentage, whereas for $d_{50}$ a more elaborate sieve analysis is required. A disadvantage of using either of the previously mentioned correlations is that grain size data is required to be available either way. To overcome this issue estimation charts like figure 2.3 could be used in the field, to determine the $N$ value directly from $q_c$. The correlation is recommended to use for SM and SP soil types, equal to silty fine sand and fine sand respectively, as defined by the Unified Soil Classification System (Murthy, 2003).

A more recent correlation is proposed by Jarushi et al. (2015), which is also based on sandy soils. To be able to compare CPT and SPT, the difference in measurement intervals are taken into account: the SPT drive depth of 300 mm was taken as a reference for corresponding CPT values. The correlation is investigated between both $q_c$ and $N$ value and $f_s$ and $N$ value. Overall it was concluded that the correlation between SPT and CPT for sandy soils is weak (0.1 to 0.3). The best correlation was found for fine sand (class SP), with a linear function for $N$ against $q_c$ and a correlation coefficient of 0.60.

2.2.2. Uncertainty in Sample Data

When dealing with soil, and problems related to geology in general, uncertainty in data is often taken into account. This uncertainty is influenced by three factors (Phoon and Kulhawy, 1999):

- **Inherent variability**: geological processes continuously alter the subsurface.
- **Measurement error**: sampling at intervals cause uncertainty, as well as the equipment with a specific measurement error.
- **Transformation uncertainty**: (empirical) correlation functions used to convert measurements to parameters used in engineering practice.

In order to deal with uncertainty, research is currently ongoing in the application of geostatistics to geotechnical engineering problems. In several publications, e.g. by Fenton (1999) and Lloret-Cabot et al. (2014), the cone penetration test (CPT) is used to provide insight into the variability of soil properties. The reason for this is that the cone tip resistance, $q_c$, is considered closest to a point property: only the vertical variability in the soil is characterized, by an almost continuous measurement, usually at intervals of 1 - 2 cm. A parameter associated with soil variability is the scale of fluctuation, $\theta$. This parameter denotes the maximum distance over which a property is correlated. In an example by Phoon and Kulhawy (1999), the vertical scale of fluctuation for $q_c$ is less than 1 m and the horizontal approximately 40 - 60 m. Considering a sedimentary deposit with horizontal layering this would be a logical observation, since the thickness of the layers are small compared to the lateral extent. The scale of fluctuation used here is similar to the variogram range; where the variogram describes the spatial correlation based on separation distance.

Aside from the variability of the property under investigation, the variability itself may fluctuate according to a certain trend. For example, $q_c$ may depend on depth: as the cone penetrates deeper into the soil, displacement of the grains by the cone becomes more difficult due to increasing overburden stresses. For this reason, when the CPT data is used to produce a geotechnical model, the trend has to be removed to isolate the fluctuating component of the geotechnical property.

**Remarks** At this point several conclusions from the material presented so far may be drawn:

- **Correlation** Based on the available data and the goals of further geostatistical processing, a correlation function may be required. For the purpose of this thesis, both CPT and SPT are available. Since at this moment the transfer functions only accept SPT data, a correlation function to convert CPT to SPT has to be found.

- **Geostatistics** Resulting from the scarcity of available sample data and the requirement of the transfer functions for an evenly gridded 3D model, the need for spatial interpolation arises. This
method should be able to handle irregularly spaced samples, incorporate some form of spatial correlation and provide insight into modelling accuracy and data uncertainty.

2.3. Estimation
This section will elaborate further on the assumptions in geostatistical methods and describe the estimation and simulation methods.

2.3.1. Random Variables
The definition of a model is a simplified representation of reality. In geostatistics, this simplification is often a result of limited understanding of the (physical) processes governing the area under investigation. Two types of models may be distinguished (Goovaerts, 1997):

- Deterministic model: when queried for a value at an unsampled location, returns a single estimate
- Probabilistic model: this type of model returns probabilities for different values, for example by giving a probability density function

When dealing with spatial data (in this case geotechnical) it is possible to predict values at the desired locations. It is the accuracy of this prediction, however, that may pose a problem. The underlying processes, like sedimentation and other rock forming processes, would have to be known precisely. As this is not sufficiently the case at the moment, probabilistic methods may be employed to obtain possible values at unsampled locations.

As the underlying physical processes are unknown, a probability density function may be assumed instead. The random variable output then takes the form of a range of possible outcomes. These outcomes are governed by a certain distribution function related to the underlying process. Several types of random variables may be distinguished: continuous, discrete and indicator (Goovaerts, 1997, Remy et al., 2009). For example a geotechnical property, like cone tip resistance, could be represented by a continuous variable. At a certain location \( u \) the probability that the variable \( Z \) is no greater than a threshold \( z \) is then given by:

\[
F(u; z) = \Pr(Z(u) \leq z) \forall z
\]  
(2.2)

Equation (2.2) represents the cumulative distribution function (cdf) and may attain values between 0 and 1. Deriving this equation yields the probability density function (pdf), giving the probability for a certain value to occur.

In the case of a discrete variable, the number of outcomes is finite. This applies, for example, to geotechnical units. A probability is assigned to each discrete variable (or category): for example, the probability that one of those units occurs in the study area.

Another case of the discrete RV is the indicator random variable, which is a binary variant and thus results in only two possible outcomes. Considering (2.2), the same could be achieved by using an indicator variable (Goovaerts, 1997):

\[
F(u; z) = \Pr(Z(u) \leq z) = E[I(u; z)]
\]  
(2.3)

In which \( I(u; z) \) is defined as:

\[
I(u; z) = \begin{cases} 
1, & \text{if } Z(u) \leq z \\
0, & \text{otherwise} \\
\end{cases} \forall z
\]  
(2.4)

In case of an indicator RV, the probability that a category occurs at a certain location is defined as the expected value of the indicator RV at that particular location.

2.3.2. Inverse Distance Weighting
A basic interpolation method is the inverse distance weighting algorithm. This algorithm only takes into account the distance between sampled locations \( u_s \) and unsampled locations \( u \) (Shepard, 1968). Based on this, it determines weights which are allocated to the sample data. When iterating over the
grid points and sample data is encountered (i.e. the distance is 0) no weights are applied. In equation (2.5), the power parameter is denoted by $p$ and affects the influence of the distance to the weights $\lambda$. In order to ensure that the estimates are globally unbiased, the weights are normalized to sum to 1. Common values for $p$ are 2 (squared) and 3 (cubed). The squared variant is appropriate for situations dealing with smoothly varying properties, e.g. layer thicknesses, whereas the cubed variant applies a higher weight to sample data at short distances and is thus more appropriate for more erratic data (Deutsch and Pyrcz, 2014).

\[
Z^*(\mathbf{u}) = \begin{cases} 
\frac{\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) Z(\mathbf{u}_\alpha)}{\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u})}, & \text{if } d(\mathbf{u}, \mathbf{u}_\alpha) \neq 0 \\
Z(\mathbf{u}_\alpha), & \text{if } d(\mathbf{u}, \mathbf{u}_\alpha) = 0
\end{cases}
\]

\[
(2.5)
\]

The differences between the two power parameters are shown in figure 2.4: an increasing power parameter leading to more weight assigned to sample data at close proximity to the unknown location.

**IDW Example Case**  In the mining industry it is common to generate ore reserve models, aiding in mine planning and feasibility study activities. The data serving as input for the models are obtained from boreholes and other in situ and laboratory tests. Aside from geological properties like ore grade, this approach is extended to geotechnical parameters by Bye (2006), using inverse distance weighting. The resulting block model is obtained after interpolation, while constraining for rock types and important structures and features. The interpolation is mainly applied between rock face maps and exploration boreholes, where the rock faces were treated as horizontal boreholes. The geostatistical method used is IDW, since the data was observed to be evenly spaced (Bye, 2006), yielding a single model containing the geotechnical information. This way of interpolation may limit the possibilities for assessing the data since it is not possible to determine the uncertainty of the estimated data. However, the data does still provide insight into the variability over the modelled area in 3D (albeit smoothed), leading to the conclusion that the model is a viable method to assess the variability of the rock mass and determine locations which may require additional sampling or pose certain risks to the mining operation.

Another potential of a more comprehensive geotechnical model is proposed by Bye (2006): being one of the constituents of an ore reserve model, which is often required by financial organisations for investment decisions, an improved geotechnical model could be a helpful tool in determining project feasibility and performing mine evaluations.

![Figure 2.4: Inverse Distance Weighting of surface level measurements: $p = 2$ (a) and $p = 3$ (b).](image)
2.3.3. Kriging

The kriging estimation method is a type of least-squares linear regression method, which was developed in the 1960's and only later named after its developer D. Krige (Cressie, 1990). The initial method accounted only for data related to a single continuous attribute being estimated; later methods were developed which were able to account for secondary information as well. Similar to IDW, kriging is a solution to the problem of predicting the value of a continuous attribute at an unsampled location. The main difference is that the data-to-data and data-to-unknown correlations are taken into account. This means that when, for example, two sample points lie very close to each other but equally far away from an unsampled location they will not be given the same weight as is the case for IDW. Based on Goovaerts (1997) the “Simple Kriging” variant of the method is introduced in this section.

Assumptions & Considerations Geostatistical methods usually rely on a number of assumptions. These assumptions ensure that a method is valid under certain conditions. An example of such a condition is dealing with a limited number of samples. It is important to note that these assumptions are made prior to the modelling process and that they are not a property of the area under investigation itself.

- **Stationarity** This assumption presumes invariance of a property of a random function through space. An example is the assumption that a joint probability density function of any number of random functions is the same. Geostatistical methods often require stationarity of first and/or second order moments, i.e. mean and variance, and assume that correlation only depends on the distance between two points.

- **Ergodicity** Stochastic methods regard data as one of the many possible realisations. The assumption of ergodicity implies that the samples are representing the domain of interest. A definition given by Chilès and Delfiner (2012) is that a stationary RF is ergodic in the mean if the spatial average over study area \( \Lambda \) converges to the RF’s expected value, when \( \Lambda \) becomes infinitely large. In order to minimise the effects of multiple realisations, as a rule of thumb, the estimation field dimensions should therefore be larger (usually three times) than the variogram range.

- **Bias** This means that there is a tendency towards certain values. Possibly resulting in an estimator giving values which are systematically over or underestimated.

- **Preferential Sampling** Samples may be clustered at one location and be more wide spread at another location. Some methods, e.g. kriging, already take this into account. However, if parameters need to be inferred prior to estimation or simulation, clustering may lead to errors. As a solution, the data can be prepared by applying a declustering method like polygonal declustering or cell declustering.

Considering the attribute to be estimated is \( z \), at an unsampled location \( u \), accounting only for \( z \)-data available in the study area \( \Lambda \) with a number of samples \( n = \{z(u_{a}), a = 1, ..., n\} \). The estimator is defined as:

\[
Z^*(u) - m(u) = \sum_{a=1}^{n(u)} \lambda_{a}(u)[Z(u_{a}) - m(u_{a})] \tag{2.6}
\]

In equation (2.6) the variables denote:

- \( \lambda_{a}(u) \): weight assigned to \( z(u_{a}) \), which is the realisation of random variable \( Z(u_{a}) \).
- \( m(u_{a}), m(u) \): the mean of RVs \( Z(u) \) and \( Z(u_{a}) \), respectively.
- \( n(u) \): number of samples accounted for in the estimation, may vary for different locations depending on search neighbourhood \( W(u) \).
- \( Z^*(u) \): the basic linear regression estimator
The estimation error is defined as \( Z^*(\mathbf{u}) - Z(\mathbf{u}) \). This results from the assumption that the unknown value \( z(\mathbf{u}_a) \) and known values \( z(\mathbf{u}_b) \) may be interpreted as realisations from, respectively, \( Z(\mathbf{u}) \) and \( Z(\mathbf{u}_a) \).

One of the goals of kriging, and any of its variants, is to minimise the estimation of error variance \( \sigma^2_{est}(\mathbf{u}) \) since the estimator is assumed to be unbiased. This idea may be illustrated by the following equation:

\[
\sigma^2_{est}(\mathbf{u}) = \text{Var}(Z^*(\mathbf{u}) - Z(\mathbf{u}))
\] (2.7)

Consequently, equation (2.7) is minimised, leading to:

\[
E[Z^*(\mathbf{u}) - Z(\mathbf{u})] = 0
\] (2.8)

Several variants of the kriging estimator exist, based on different models for the random function \( Z(\mathbf{u}) \). As shown below, the random function may be decomposed into a residual component \( R(\mathbf{u}) \) and a trend component \( m(\mathbf{u}) \):

\[
Z(\mathbf{u}) = R(\mathbf{u}) + m(\mathbf{u})
\] (2.9)

The residual component is modelled as a stationary RF, with a mean equal to zero and a covariance based on distance \( h \):

\[
E[R(\mathbf{u})] = 0
\] (2.10a)

\[
\text{Cov}[R(\mathbf{u}), R(\mathbf{u} + \mathbf{h})] = E[R(\mathbf{u}) \cdot R(\mathbf{u} + \mathbf{h})]
\] (2.10b)

Considering (2.10), the expected value at the unsampled location \( \mathbf{u} \) is equal to the value of the trend component at the same location:

\[
E[Z(\mathbf{u})] = m(\mathbf{u})
\] (2.11)

As mentioned previously, different models for the RF \( Z(\mathbf{u}) \) constitute the variety of existing kriging estimators. In case of SK, it is assumed that the mean \( m(\mathbf{u}) \) is known and constant throughout the area \( A \) to be investigated:

\[
m(\mathbf{u}) = m, known \forall \mathbf{u} \in A
\] (2.12)

Recalling equation (2.6) and taking the above into consideration leads to the following equation:

\[
Z^*(\mathbf{u}) = \sum_{a=1}^{n(\mathbf{u})} \lambda_a(\mathbf{u})Z(\mathbf{u}_a) + \left[ 1 - \sum_{a=1}^{n(\mathbf{u})} \lambda_a(\mathbf{u}) \right] m(\mathbf{u})
\] (2.13)

The kriging weights \( \lambda_a \) sum to 1, in order to ensure the unbiasedness condition by minimising the error variance (equation (2.7)). This unbiasedness of the estimator itself is already ensured, since the mean error is assumed to be zero:

\[
E[Z^*_SK(\mathbf{u}) - Z(\mathbf{u})] = m - m = 0
\] (2.14)

The kriging weights are obtained by solving the following linear system of equations:

\[
\sum_{\beta=1}^{n(\mathbf{u})} \lambda^*_\beta \mathbf{u}\alpha \cdot C(\mathbf{u}_\alpha - \mathbf{u}_\beta) = C(\mathbf{u}_\alpha - \mathbf{u}), \text{ with } \alpha = 1, ..., n(\mathbf{u})
\] (2.15)

In equation (2.15) \( \lambda^*_\beta \) denotes the vector with SK weights, \( C(\mathbf{u}_\alpha - \mathbf{u}_\beta) \) is the data-to-data covariance and \( C(\mathbf{u}_\alpha - \mathbf{u}) \) data-to-unknown covariance.

The equation \( K_{SK} \lambda_{SK}(\mathbf{u}) = k_{SK} \) in matrix form:

\[
\begin{bmatrix}
C(\mathbf{u}_1 - \mathbf{u}_1) & \cdots & C(\mathbf{u}_1 - \mathbf{u}_n) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
C(\mathbf{u}_n - \mathbf{u}_1) & \cdots & C(\mathbf{u}_n - \mathbf{u}_n) & 1
\end{bmatrix}
\begin{bmatrix}
\lambda^*_1 \\
\vdots \\
\lambda^*_n
\end{bmatrix}
= \begin{bmatrix}
C(\mathbf{u}_1 - \mathbf{u}) \\
\vdots \\
C(\mathbf{u}_n - \mathbf{u})
\end{bmatrix}
\] (2.16)
For this system the covariance may also be represented by its semivariogram equivalent. The relation $C(h) = C(0) - \gamma(h)$ is then used (Goovaerts, 1997).

Equation (2.15) has a solution if the covariance matrix $K_{SK}$ is positive definite, which in practice is the case if (Goovaerts, 1997):

- No two data are located at the same coordinates (colocation): $u_\alpha \neq u_\beta$ for $\alpha \neq \beta$.
- The covariance model $C(h)$ is permissible.

In order to ensure the covariance model (or its semivariogram counterpart) is permissible, utilising a linear combination of semivariogram models which are already known to be permissible is possible; otherwise, this has to be verified. The most commonly used models are: nugget, spherical model, exponential model, gaussian model and the power model. With the exception of the power model, all of these have a covariance model counterpart.

A characteristic of kriging which should be considered when deciding on an estimation method is the smoothing effect. The global unbiasedness ensures that the mean remains similar to the original data, but the variability may be reduced resulting in "smoothing" of short range components (Goovaerts, 1997). With decreasing sampling density and/or an increased nugget effect, the smoothing increases. For example, this smoothing may result in minimum and maximum values, as observed in the sample data, are not present as frequently in the estimation results as expected.

A possibly advantageous property of kriging is its non-convexity. What this means is that, when kriging weights are negative, the estimated values do not necessarily lie in the interval defined by the minimum and maximum values of the original data (as expected from smoothing). However, problems may arise in situations where the estimated data is constrained physically. For example a property represented by a fraction (e.g. porosity, grade) might become negative or more than 1. Some ways for mitigating this issue are given by Goovaerts (1997), like forcing all the kriging weights to be positive or using indicator constraint intervals.

To illustrate some of these limitations, SK is applied to the same data as used in section 2.3.2, resulting in figure 2.5. What can be observed is that when a high value is surrounded by lower values, the high value is given a relatively low weight and when a location is far away from sample points (like in the bottom center) the algorithm behaves similar to IDW. Another noticeable effect of kriging is "smoothing". In this case visible by the absence of some extreme values as shown in IDW estimation (figure 2.4).

The kriging variance is also shown: here it is clear that it does not depend on actual sample data but only one the model type and spatial distribution.

Many other variants of kriging exist such as Ordinary Kriging and Indicator Kriging; with Ordinary Kriging being the variant which is probably the most widely used (de Souza et al., 2004). In case of
OK the stationarity of the mean is limited to the sample search neighbourhood or ellipsoid \( W(u) \), thus considering local variability of the mean. The OK estimator is defined as equation (2.17); see equation (2.13) for comparison with SK.

\[
Z^*(u) = \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) Z(u_\alpha) + \left[ 1 - \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) \right] m(u)
\]  (2.17)

The IK method, on the other hand, is useful for handling discrete data; besides binary data it is also possible to use more than one category. Similar to "normal" kriging, there is a distinction between Simple IK and Ordinary IK: in case of SIK the indicator mean is assumed to be known and stationary over the study area \( A \). The SIK estimator follows from a linear combination of \( n(u) \) indicator RVs \( I(u_\alpha; z_k) \) and the cdf value \( F(z_k) \) (Goovaerts, 1997):

\[
[F(u; z_k \mid n)]_{SIK} = \sum_{\alpha=1}^{n(u)} \lambda_\alpha^{SIK}(u; z_k) + \left[ 1 - \sum_{\alpha=1}^{n(u)} \lambda_\alpha^{SIK}(u; z_k) \right] F(z_k)
\]  (2.18)

To illustrate IK estimation a simplified example is shown next. In this example 28 samples with values of either 1 or 0 were drawn from a uniform distribution, resulting in a probability of occurrence for value 1 of 0.43. The sample spacing is large compared to the estimation domain, so variogram inference is difficult: a nugget of 0.05, a sill for the experimental model of 0.25 and a range of 800 is assumed. The IK search ellipsoid is kept equal to the variogram range, resulting in figure 2.6 (some samples located along the axes, 500m spacing).

![Figure 2.6: Indicator Kriging example of artificial binary data.](image)

What can be seen here is that the distribution of values over the grid is similar to the initial probabilities (i.e. an approximate coverage of 43% for value 1). Also worth noting is that although a location is close to a sample there is still a small probability of another value occurring, so the red and blue areas are not uniformly "filled" with a single value; nonetheless, the values of the actual samples are still retained at their location.

### 2.3.4. Estimation Uncertainty

The geostatistical methods introduced up to this point, along with their application, are not necessarily a goal in themselves: the results of these techniques should be processed in order to gain further understanding in aspects like local and spatial uncertainty of the data.

Local uncertainty refers to the uncertainty of a zone smaller than the total estimation domain, whereas spatial uncertainty applies to multiple locations distributed in space, e.g. representing a geological unit or other structural feature.
In case the uncertainty assessment of estimated data is solely based on kriging (without involving simulations) there are several methods available (Chang et al., 1998):

- Mean Square Error (MSE)
- Correlation coefficient
- Kriging variance

The MSE and correlation coefficient could be useful when real data is available. For example, reliability of an estimation method could then be investigated by comparing estimated values to their true data counterpart. The kriging variance, on the other hand, is one of the outputs of the estimation method and is sometimes used to measure uncertainty. There is, however, a limitation to this approach since this variance is only based on the spatial arrangement of samples (Goovaerts, 1997). Thus, it may not be the most suitable method to assess variability of data, since the actual data values are not taken into account.

A more reliable method lies in employing simulation algorithms to obtain a large number (usually in the order of 100) of realisations. This results in 100 estimated values at each grid cell which can be used to construct a cdf at each location.

By Chang et al. (1998) kriging is compared to sequential simulation techniques (SGS and SIS). In this case the tonnages of coal accumulation were estimated, followed by an analysis of measurement errors. These tonnages were found to be similar for both kriging and simulation, but the spread of the error was different: the error derived from the kriging variance was observed to be symmetric (resulting from an assumed normal distribution of error). This assumption does not apply to sequential simulation, resulting in an asymmetric error distribution.
The Flaw of Averages One of the main differences between the methods described in this chapter is the approach to modelling the uncertainty: this can be defined either locally or spatially. Kriging and its many variants yield a single realisation of estimated values, along with the kriging variance. However, this variance does not correlate very well with the actual data values estimated at unsampled locations. Therefore, kriging is useful for obtaining the best estimate at unknown locations but may not be appropriate for assessing the local uncertainty.

Figure 2.7: Estimation vs. simulation; when investigating a non-linear transfer function, based on Martinez (2009), Wambeke (2013).

The differences between estimation and simulation are illustrated in figure 2.7. The problem caused by only using a single estimate, e.g. resulting from one of the methods explained in section 2.3, is referred to as 'the flaw of averages', after Martinez (2009). What this means is that when applying a non-linear transfer function to an average value, the result is not the average of the process represented by this function; or \( F(E[X]) \neq E[F(X)] \).

Using simulation methods could be a better way of approaching this problem, as this provides multiple realisations which may be processed to obtain probability density functions of the values at unsampled locations.
2.4. Simulation

Some of the simulation methods have already been mentioned briefly in the previous section. This section will elaborate further on this subject; along with ways of processing the resulting data consisting of multiple realisations.

First a brief overview of current methods and new developments related to simulation is given:

- **Sequential Gaussian Simulation** The SGS method is a sequential method, visiting all the cells on a predefined grid. For each cell a kriging estimation is performed, over a certain search neighbourhood. Conditioning may be integrated in this process; treating the simulated values as samples in the next step of the sequence. As the name implies the algorithm assumes a Gaussian data distribution, resulting from the use of kriging. The method is easy to use and to implement; however, it is computationally intensive, due to the solving of a kriging system at each grid cell (Wambeke, 2013).

- **Sequential Indicator Simulation** SIS is similar to SGS, with the main difference being the use of Indicator Kriging. The method allows for simulation of either continuous or categorical variables. Although, in case of continuous variables, a transformation into categories based on predefined classes (e.g. ranges based on sample statistics) is required.

- **Turning Bands** The idea of this method is to obtain a multi-dimensional random field by simulating one-dimensional lines; it was developed by Matheron (1973). After transforming the input data to normal scores, the variogram or covariance being reproduced, an unconditional random field is generated. The result then has to be conditioned, e.g. by post-processing using kriging (Rossi and Deutsch, 2014). In early applications this method lead to artifacts caused by a limited number of lines, which in turn was a result of the available computational power. Currently this limitation does not apply anymore; although artifacts may still be present if the method is not implemented correctly (Rossi and Deutsch, 2014).

- **Local Average Subdivision** LAS generates realisations of random fields based on local averages, where the statistics are reproduced consistently with the grid dimensions (Wambeke, 2013); contrary to other methods where the statistics vary depending on sample support, e.g. block vs. point, referred to as dispersion variance. This is one of the advantage of the method, especially useful when simulating blocks. Other advantages are flexibility towards changing the resolution in specific regions and its ease of implementation (Wambeke, 2013). After simulation, the random fields are conditioned to actual data at sample locations.

- **Simulated Annealing** This optimisation procedure was proposed in the 1940s and recently caught new attention by the increased availability of computer power. It is based on thermodynamic processes related to the crystallisation of materials (e.g. cooling and annealing of metals). The minimisation refers to crystal structures slowly reverting to their minimum energy states, by lining up of atoms during cooling (Rossi and Deutsch, 2014). The method employs a trial-and-error process, based on the fact that this algorithm does not necessarily go immediately towards a nearest solution (similar to atoms being able to line up over distances a billion times their size). The method is conditional, honours sample data, distribution and variogram, and secondary data may be used as well (Srivastava, 2013). In the mining industry this method has not found many applications yet, due to the difficulty of setting up the objective function (thermodynamic energy analogue) and the control parameter (temperature analogue) and their sensitivity to artifacts (Rossi and Deutsch, 2014).

- **Multiple Point Statistics** This is a new development in geostatistics; comprising a method honouring spatial information beyond the capabilities of the two-point statistics variogram (Srivastava, 2013). Spatial structures are represented by a so-called training image, where MPS tries to match data values to one or more training images (Remy et al., 2009) and takes multiple data points into account during the process. Consequently, a training image sufficiently representative of the spatial structures is required, similar to a variogram for other methods. Modern MPS implementations are both pixel-based and object-based: a pixel-based algorithm builds a realisation one pixel at a time, allowing for conditioning to sample data (Remy et al., 2009), whereas an object-based algorithm inserts one or more complete objects or patterns into the simulation grid.
2.4.1. Sequential Simulation

Expanding on the definition of a random variable (section 2.3.1), a random function (RF) may be defined as a set of, usually, dependent random variables (Goovaerts, 1997). This means that each unsampled location is represented by an RV, for the area of interest $A$: $\{Z(u), \forall u \in A\}$. Recalling equation (2.2) and applying it in this situation, yields:

$$F(u_1, ..., u_N; z_1, ..., z_N) = \text{Prob}\{Z(u_1) \leq z_1, ..., Z(u_N) \leq z_N\}$$

(2.19)

In equation (2.19), which may be referred to as an N-point cdf, $N$ is the number of locations and $k = 1, ..., N$. It represents the joint uncertainty about the $N$ actual values. However, applying this equation in a large 3D grid could be impractical as such a grid may contain millions of locations depending on cell size (Remy et al., 2009). An exception to this is a distribution which may be specified by a small number of parameters, like the Gaussian RF (Goovaerts, 1997). This RF is only defined by two parameters: mean and variance. Thus, if only these two parameters are known, or reliably estimated, the Gaussian RF could form the basis of further estimation or simulation methods, along with a representation of spatial correlation.

When considering a joint simulation at only 2 locations, instead of $N$, the idea of sequential sampling is to trade the modelling of a 2-point ccdf (conditional cumulative distribution function) for the sequential sampling of two one-point ccdfs. Extrapolating this idea to $N$ locations, it is assumed that the $N$-point ccdf (equation (2.19)) can be approximated by a product of $N$ one-point ccdfs as shown in equation (2.20) (Goovaerts, 1997). The complete procedure of performing SGS is described by algorithm 1.

$$F(u'_1, ..., u'_N; z_1, ..., z_N|(n)) = \prod_{a=1}^{n(u)} F(u'_a; z_a|(n+(1-\alpha)))$$

(2.20)

Before applying the SGS algorithm on the area under investigation, the normality of the distribution should be assessed. The sample data could also be transformed to the normal score domain first, as a way to overcome this issue. When the normality can safely be assumed, the sequential simulation under the Gaussian RF model may be applied as follows (Goovaerts, 1997):

Algorithm 1 Sequential Gaussian Simulation

1: Transform data to normal scores.
2: for all realisations do
3: Determine random path. This should be done in such a way that each unsampled grid cell is only visited once.
4: for all grid cells $u'_1, ..., u'_N$ do
5: Model the ccdf at the first location $u'_1$ conditional to the $n$ original data $z(u_1)$:
$$F(u'_1; z|(n)) = \text{Prob}\{Z(u'_1) \leq z|(n)\}$$

(2.21)

Simple Kriging (see also section 2.3.3) is used in this step to determine the mean and variance: the parameters for the Gaussian ccdf $G(u'; y|(n))$, with the semivariogram $\gamma_y(h)$ based on the normal score data $y$. In this ccdf $(n)$ refers to the conditioning data, consisting of both sampled data and simulated values from previously visited locations.
6: From the ccdf (2.21) draw a value $z^{(1)}(u'_1)$, which also becomes part of the data used for conditioning subsequent drawings.
7: At the $i$th cell $u'_1$, the ccdf $Z(u'_1)$ is modelled which is conditional to the $n$ original data and all $(i−1)$ previously simulated values $z^{(i)}(u'_1)$:
$$F((u'_1)|(n+i−1) = \text{Prob}\{Z(u'_1) \leq z|(n+i−1)\}$$

(2.22)

8: end for
9: end for
10: Back-transform simulated values to original distribution.
After running the algorithm, a single realisation is the result. To obtain more realisations, a different random path for visiting unsampled locations should be chosen. These realisations should honour the data at sampled locations and represent the area statistics, meaning that the variogram and probability density function (histogram) should look similar. This algorithm is conditional, in the way that it includes simulated values in the conditioning data; although is also possible to perform SGS without conditioning.

In case the sample data was transformed to normal scores, a back-transformation has to be applied to obtain the data in its original domain.

Considering again the surface elevation data used in sections 2.3.2 and 2.3.3 and applying the SGS algorithm, gives the results shown in figure 2.8.

![Figure 2.8: Sequential Gaussian Simulation of surface level measurements: four out of 100 equiprobable realisations.](image)

In these representations a high degree of "randomness" is observed. This may indicate an insufficient sample spacing and/or number of samples, but to substantiate such a claim more detailed investigation is required. Also, when performing sequential simulations the inspection of a single realisation is often not the goal as the total number (e.g. 100) of realisations provides information by giving insight into probability of occurrence and spatial uncertainty.

Aside from simulating continuous variables, in some cases one may want to investigate categorical or discrete properties: e.g. geotechnical units, sedimentary facies. Sequential Indicator Simulation is such a method which is based on Indicator Kriging (see section 2.3.3) and able to handle both continuous and discrete variables. The cccdfs are modelled in a different way, based on the indicator approach. In case of continuous data, the range is divided into classes: each class represented by a probability and requiring separate variograms. The transformation applied to the data results in a vector of indicator data for each sample location (see equation (2.4)).

Algorithm 2 describes the application of SIS to a discretised grid, after Goovaerts (1997).

Note that to obtain multiple realisations, the random path should be changed before running the algorithm again. Also, in case of binary data (e.g. 1 = hard layer and 0 = soft layer), the SIS process is less complicated compared to simulating a continuous variable. When, for example, simulating a continuous variable the range of attainable values is subdivided into $K + 1$ classes with $K$ threshold values. Each class requires a specific variogram (with the exception of "Median IK", where a single representative variogram is chosen and scaled appropriately for each class).

However, in case of the previously mentioned binary data, a transformation is not necessary as it al-
Algorithm 2 Sequential Indicator Simulation

Ensure: Categorical or indicator data
1: Using IK, determine $K$ ccdfs. When $K + 1$ classes exist, $K$ is the number of threshold values. Each datum $z(u)$ is transformed into a vector of indicator values: $\mathbf{i}(\mathbf{u}_z) = [i(u_z, z_1), \ldots, i(u_z, z_K)]$ (Remy et al., 2009).
2: for all realisations do
3: Determine random path. This should be done in such a way that each unsampled grid cell is only visited once.
4: for all grid cells $u_1, \ldots, u_N$ do
5: Indicator kriging yields estimates of probabilities for each class. Per location, these probabilities should sum to 1.
   If the probabilities do not sum to 1 (order relation deviation), this may be mitigated by setting negative probabilities to zero and dividing the remaining probabilities by the sum (Deutsch, 2006).
6: A ccdf model is constructed by interpolation and extrapolation. Usually interpolation is done linearly between threshold values and for extrapolation a power model is used (Goovaerts, 1997).
7: A simulated value is drawn from the complete ccdf and added to conditioning data.
8: From this new ccdf, draw a value $z^{(0)}(u')$, which is also included in the conditioning data for the following step.
9: end for
10: According to the random path, go to next location and repeat steps 1 to 5.
11: end for

ready consists of 1s and 0s and a single variogram is sufficient, along with two probabilities (occurrence of either a soft or hard layer).

2.4.2. Simulation Uncertainty

A number of methods for dealing with uncertainty have been explained in section 2.3. When dealing with multiple realisations resulting from a particular simulation, different ways of processing this data and assessing their uncertainty are available, compared to estimation.

A distinction may be made between single-location uncertainty and multi-location uncertainty. The single-location uncertainty applies to a specific location, or grid cell, and refers to the probability of the estimate exceeding a certain threshold $z_c$ (Juang et al., 2004):

$$\text{Prob}_{\text{sis}}[z(x') > z_c] = \frac{n(x')}{N}$$ (2.23)

In this case the number of realisations is $N$, $z(x')$ the data estimate and $n(x')$ the number of realisations in which the estimate exceeds the threshold.

The single-location uncertainty can be used to create a probability map. On the other hand, the multi-location uncertainty can be used to assess the probability of multiple values exceeding a limit, which may be used to create another type of probability map. This probability is defined as:

$$\text{Prob}_{\text{sis}}[z(x'_1) > z_c, z(x'_2) > z_c, \ldots, z(x'_m) > z_c] = \frac{n(x'_1, x'_2, \ldots, x'_m)}{N}$$ (2.24)

The number of realisations taken into account is again denoted by $N$ and $m$ refers to the number of locations.

Aside from probability maps, Goovaerts (1997) presents two other methods of processing and visualizing the simulation results:

- Quantile map: instead of showing the probability of exceeding a certain threshold, it is possible to map the $p$-quantile. For example the 0.1-quantile would show the value for which 90% of the simulated values are higher. In petroleum reservoir estimation this also referred to as the $P_{10}$ case (usually accompanied by its $P_{50}$ and $P_{90}$ counterparts).
- Map of spread: a visualization representing the local variability of realisations. For this purpose, the interquartile range may be employed to indicate the spread.
2.4.3. Simulation Example Cases
As mentioned previously, sampling is an important aspect of decision making in several fields dealing with excavation of soil and/or rock: e.g. soil pollution remediation, extraction of mineral resources. Decisions have to be made based on production estimates of ore recovered or pollution remediated; even when only limited data is available. In support of this decision process, it is important to utilise the available information to the largest possible extent.

Sequential Simulation  Although soil remediation generally deals with a different material (excavating top soil vs. the mining of material located deeper in the subsurface), the general approach presented by Demougeot-Renard et al. (2004) may be applicable in both situations. The method is iterative and was developed to determine the sampling density based on costs and the risks of pollution involved. The framework consists of five steps, aiming to model the impact of extra samples on the uncertainty:

1. The initial data is processed to estimate the remediation volume and uncertainty.
2. Remediation costs and uncertainty are estimated, solely based on actual samples.
3. When the financial and volumetric accuracy is insufficient, a conditional simulation is used to generate input for virtual sampling.
4. Using both forecasted and real data, the remediation volume and uncertainty are forecasted.
5. Finally, the remediation costs and uncertainty are forecasted, to indicate the effect of extra samples.

When an increasing number of additional data is required, steps 3 to 5 are repeated and when the financial and volumetric estimates are deemed acceptable, the iteration is stopped.

A conditional method is used for simulation (SIS, in this case), honouring the initial data and experimental data from successive samples. For each simulated value, the probability is estimated that it exceeds a certain cut-off.

Three effects resulting from sampling are described:

• When dealing with the clean-up of a polluted site, the division of the site is made up of blocks based on the degree of pollution. Similar to a mining situation the data originates from boreholes, thus resulting in a different support: the data from the boreholes may be considered point data, whereas the production units are blocks with volumes in the range of cubic meters.

• The actual measurement of pollutant concentrations is performed on a composite of samples. These samples are distributed over a remediation block, and may give different block concentrations compared to reality. This is a result from the information effect; a term used in mining to describe that more information is present at the time of ore/waste selection, compared to the information available during resource modelling (Rossi and Deutsch, 2014).

• Sampling and measurement errors, depending on the types of laboratory and field equipment used, as well as the experience of the operator.

The method described by Demougeot-Renard et al. (2004) to account for the change of support is based on sample averaging. First, the blocks are discretised in a grid, then each block value is calculated by taking the mean of the samples contained within. The measurement error is accounted for by adding an error \( \epsilon \), drawn from a uniform distribution between -1 and 1.

In case the results from the simulation indicate that the uncertainty of production volumes and remediation costs are still too high, the impact of extra samples may be forecasted. The suggested approach is to randomly select a realisation of pollutant concentration and use this as a reference state. The extra samples are taken in the uncertain area and are used for conditioning, together with the original data, in a new simulation. Based on the new estimates, new volumes and costs may be calculated.

The advantages of this method are its flexibility towards handling sampling effects and the results being close to real values encountered in soil remediation processes.
However, some limitations are also present; the property under investigation should show a spatial structure, meaning that if there is only a nugget effect present extra sampling will not improve the estimates. Another disadvantage is that the decision for additional sampling is based on a single realisation, chosen randomly. This means that the simulation results cannot be used as estimates of concentration, but merely provide an indication of the possible benefits of additional sampling.

**Other methods**

This paragraph describes some applications of simulation methods different from those introduced earlier, with the main difference being the way of conditioning: first random fields are generated, then they are conditioned using kriging. The relevance to this thesis lies in the fact that the applications are related to the effect of sampling density and uncertainty reduction.

*The method proposed by Goria et al. (2001) is aimed more towards the extraction of mineral resources* and also assesses the effect if additional sampling. Here the goal is to decide whether or not to extend a small mining pit, possibly requiring extra boreholes to reduce the uncertainty in the extension area. The methodology employed is comparable to the one shown by Demougeot-Renard et al. (2004), in the way that it consists of conditional simulation to assess the effect of additional virtual samples. The specific simulation method is not mentioned, though, aside from OK being used for conditioning after simulation.

1. The number of additional boreholes is defined.
2. Run 100 point simulations of values on drillholes.
3. Choose five specific cases for further simulation, based on average grades.
4. Run 50 block simulations for each of the five cases and the original data.
5. Analyse the results.

In this example, prior to simulation, the locations and the number of boreholes is defined. Then 100 simulations are performed on the locations of the boreholes only. The ore body in this case study shows high variability in ore grades. Therefore, the ore body is divided into two areas of low and higher average grades. Based on these grades, the next step is to consider five cases defined by different combinations of high and low grade areas. For these five cases another block simulation is performed, giving 50 realisations, while doing the same for the original data. To assess the effect of the extra samples, the simulation based on only the original data is compared to the simulation based on both the original data and virtual samples for conditioning.

After evaluating the results, based on simulations and a literature review, it was found that the uncertainty decreases with an increase of available information. However, this relation only holds up to a certain point: in the occasion that unexpected information is found, the uncertainty may not decrease. The cause of this is based on assumptions for the simulations performed: a base case is defined, including expected values like the average grade and confidence limits. In some cases the additional drilling may reveal new information, considerably different from the original data, changing the base case and thereby the average value and uncertainty.

*Another case is the investigation into the effect of varying sample spacing on the development of a marine gold deposit* by Saravanakumar et al. (2014). The approach used may be outlined as follows:

1. Borehole data is selected as input for simulation. Prior to simulation this data is transformed to Gaussian, declustered and used for variogram inference.
2. The Turning Bands algorithm (Mantoglou and Wilson, 1982) is used for generating 100 unconditional 2D realisations.
3. Sampling is performed on these realisations, according to three different regularly spaced grids (25, 50 and 100 m).
4. Each sample set is estimated using OK into equal blocks of 50 x 50 m².
5. The actual and estimated blocks are compared (i.e. confidence of sampling strategy).

Costs were determined per sample, so with increasing sampling density, the sampling costs were observed to rise accordingly: 50 m spacing leading to costs 4 times as high and 25 m spacing to a 16 times increase.
The ore grade estimate, using a 100 m x 100 m sampling interval, was found sufficient for generating an Indicated Resource estimate. With decreased sample spacing, however, the accuracy increased, yielding more accurate estimates: the amount of actual ore was lower, but the waste treated as ore was lower as well. This is a non-linear effect which could be important and should, thus, be taken into consideration when processing simulation results.

An example of applying geostatistical methods in a geotechnical situation is the investigation of the effect of CPT spacing on global uncertainty reduction by Lloret-Cabot et al. (2014). Similar to the previously mentioned approach, data conditioning is not integrated with random field modelling, but performed after simulation:

1. Using Local Average Subdivision (LAS), the inherent soil variability is modelled by a random field.
2. Kriging is used to condition this to a certain number of CPTs.
3. The reduction in uncertainty is determined by comparing the variance of the original data to the kriging variance, for different numbers of CPTs (and thus spacings).

At sample location this method should result in variance reductions to 0, in an ideal situation with continuous measurements. However, since CPTs provide the input data, the vertical sample separation is 2 cm.

When comparing the results of increasing the number of CPTs within the domain it was found that the uncertainty decreased; while the uncertainty reduction increases as the number of samples increases, as is expected. Besides the sampling density, the spacing was investigated. In this case the effect of varying the distance between only two CPTs on the global uncertainty is assessed. After a certain distance the influence of the second CPT is at a minimum. This distance was observed to be 35 m, with the CPTs located 17.5 m away from the domain boundary. The horizontal scale of fluctuation (similar to variogram range) was found to be 7 m, and was estimated by inferring a variogram.

Remarks Several simulation methods have been presented in this section. Based on this information, the following observations are made:

- **Realisations** All the algorithms mentioned provide a way of generating multiple realisations. This is a requirement of risk and uncertainty analysis, as this provides a range of possible values at each unsampled location.

- **Correlation** Each of the methods requires some form of representing spatial correlation. The variogram is easy to implement but its inference is not entirely objective: the resulting model depends partly on experience and knowledge of the person performing the task.
  In case of MPS another type of representation is required: a training image, representing both possible structures and spatial correlation.

- **Implementation** The ease of implementation varies with the choice of simulation method. Some of these, e.g. TB and LAS, require an additional step for obtaining conditioned realisations; in the sequential algorithms this is integrated in the simulation procedure.

- **Application** The adoption of the more involved methods seems to start in the field of petroleum engineering (Srivastava, 2013), followed by mining engineering. The application to geotechnical engineering problems seems less common, except for some 2D cases, as explained by Li et al. (2016), who introduce a novel 3D method as well. Judging by the examples introduced previously though, the application of geostatistical methods to, for example, CPT data seems to be a viable option for assessing the uncertainty based on sampling density. The propagation of this uncertainty into subsequent transfer functions is not covered in these examples and is therefore part of the research objectives of this thesis.
2.5. Concluding Remarks

So far, in this chapter several theories and methods have been described which are related, and could be useful, to designing an approach towards the research objectives. In general, the topics that were covered are:

- Estimation and simulation methods.
- Assessing uncertainty of simulated values.
- Sampling density.
- Propagation of uncertainty into transfer functions.
- Correlating CPT to SPT.

The subject of this thesis lies in the fields of both geotechnical and resource engineering. Within these fields, the estimation methods are used in a similar way, sometimes with small differences (e.g. use of covariance model instead of semivariogram). The methods are used to model the (spatial) uncertainties related to soil and rock properties.

When assessing the feasibility of exploiting an ore deposit, it is common practice to apply geostatistical methods and obtain models of ore grades and other properties. In case of geotechnical parameters, this is not the case at the present time. Some steps have been taken towards this goal, resulting in, for example, a 3D geotechnical model (Bye, 2006) or an assessment of spatial uncertainty of geotechnical parameters (Lloret-Cabot et al., 2014).

In order to reach the goals introduced in section 1.2 a methodology, or framework, has to be developed to integrate the assessment of spatial uncertainty of geotechnical data, sampling density and decision making related to additional sampling and project risks. Some of these methods, like the geostatistical simulation techniques, can be based on existing ones described in literature.

Based on the findings so far, the following gaps are identified:

- **Geotechnical Application** The application of the simulation and estimation methods presented in this chapter to geotechnical engineering problems has been investigated, but only in limited cases: research has been done into the effect of sampling density on uncertainty reduction, in 2D cases. Examples of 3D cases exist, but are related to other fields of research and apply to different properties.

- **Uncertainty Propagation** Uncertainty propagation through transfer functions is a common way of assessing (financial) project risk in mining and petroleum engineering. However, these methods are usually applied to properties directly related to economic feasibility, like metal concentrations. A different approach would be to apply the same idea to geotechnical parameters and investigate the effect on key performance indicators. Transfer functions may rely on a wide range of input parameters, and thus result in a non-linear function. Consequently, it is important to know how the uncertainty propagates through such a transfer function; in contrast to a linear function, where propagation is more obvious.

- **Integrated Framework** Considering these gaps, it would be useful to have a framework, or workflow, according to which the effect of sampling density on uncertainty reduction and project risk may be assessed. An attempt has been made by Wambke (2013), using LAS applied to a 2D exhaustive dataset.

Regarding the scientific contribution, it is the goal of this thesis to fill these gaps and develop an approach based on actual sample data and resulting in quantified estimations of production uncertainty and associated (financial) project risks.
Modelling Workflow  As shown in this chapter, a wide variety of estimation and simulation methods is available and some variant may suit a particular application better than another. By Goovaerts (1997) four criteria are given, on which a decision for a certain method can be based:

1. The algorithm that best fits the goals of the study.
2. The manual work and CPU time required for generating a given set of realisations.
3. The amount of relevant information accounted for (conditioning data).
4. The precision and accuracy of probabilistic prediction.

The available methods (see table 2.1) rely on different input data, e.g. categorical or continuous, and distribution assumptions. The time required to process this data by carrying out these methods may be lowered by creating scripts to automate certain tasks. Another factor influencing this time is the amount of information taken into account, also depending partly on the chosen algorithm: for example, a method may be used which is also capable of including secondary information. Lastly, the precision (proximity of estimate to real value) and accuracy (spread of estimate) may be assessed by analysing the resulting realisations.

Table 2.1: Overview of Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>$\mathbb{R}^n$</th>
<th>RF Model</th>
<th>Systematic Grid Required</th>
<th>Conditioning</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OK</td>
<td>All $n$</td>
<td>Gaussian</td>
<td>No</td>
<td>N/A</td>
<td>Variogram</td>
</tr>
<tr>
<td>IK</td>
<td>All $n$</td>
<td>Discrete</td>
<td>No</td>
<td>N/A</td>
<td>Variogram$^1$</td>
</tr>
<tr>
<td>SGS</td>
<td>All $n$</td>
<td>Gaussian</td>
<td>No</td>
<td>Direct</td>
<td>Variogram</td>
</tr>
<tr>
<td>SIS</td>
<td>All $n$</td>
<td>Discrete</td>
<td>No</td>
<td>Direct</td>
<td>Variogram$^1$</td>
</tr>
<tr>
<td>TB</td>
<td>$n &gt; 1$</td>
<td>Gaussian</td>
<td>No</td>
<td>Kriging</td>
<td>Variogram</td>
</tr>
<tr>
<td>LAS</td>
<td>All $n$</td>
<td>Gaussian$^2$</td>
<td>Yes</td>
<td>Kriging</td>
<td>Variogram</td>
</tr>
<tr>
<td>SA</td>
<td>All $n$</td>
<td>Any</td>
<td>No</td>
<td>Direct</td>
<td>Variogram</td>
</tr>
<tr>
<td>MPS</td>
<td>All $n$</td>
<td>Any</td>
<td>No</td>
<td>Direct</td>
<td>Training Image</td>
</tr>
</tbody>
</table>

When the spatial uncertainty of estimated data has been quantified, the next step would be to determine its effect on transfer functions and decision making. An example of this is described in section 2.4.3: the method uses conditional simulations to assess spatial uncertainty and defines a methodology to propagate this uncertainty into a decision making process, while considering additional sampling.

$^1$Depending on number of categorical variables or classes: in case of more than one, a variogram for each is required or, alternatively, a single variogram representative of all (Median IK).

$^2$May be based on any RF model. However, Gaussian is the easiest to implement (Wambeke, 2013).
Regarding workflows, again numerous ways exist in stochastic simulation, depending on the scientific field and the requirements set by the user. An example of a field in which (stochastic) simulation techniques are widely used is petroleum engineering: when trying to assess reservoir potential, a combination of sample data from various sources is used (Deutsch and Pyrcz, 2014). As this data is obtained by different measurement techniques (e.g. seismic, core data, well log data) it represents properties and structures at varying scales.

A general approach for handling such data is presented by Deutsch and Pyrcz (2014):

1. Establish stratigraphic layering and subset data.
2. Large scale modelling (discrete).
3. Facies modelling (discrete).
4. Porosity/permeability modelling (continuous).
5. Loop over layers.
6. Obtain model of geometry and rock/fluid properties.
7. Loop for multiple realisations.

The idea of this approach could also be applied in other fields, such as geotechnical, mining or environmental engineering. In case of geotechnical engineering: data may be obtained from CPT, SPT (including disturbed soil samples), core samples and others. From this data it is possible to derive soil information at different scales, like layering, porosity and particle size distribution.

In geotechnical engineering it is common to divide the area of interest into geotechnical units. These units are not necessarily the same as stratigraphic or lithological units, but are based on geotechnical properties: each geotechnical unit is expected to exhibit similar behaviour under loading or excavation.

Adapting the previously mentioned workflow:

Algorithm 3 Simulating Geotechnical Properties

1: Define geotechnical units and subunits.
2: for all realisations do
3: for all geotechnical units do
4: Large scale modelling (e.g. outer domain and unit boundaries, discrete).
5: Subunit modelling (discrete).
6: SPT $N$ value simulation (continuous).
7: end for
8: Obtain model of geometry and soil/rock behaviour properties.
9: end for

In case of this research project, both discrete and continuous data will serve as input for the simulation workflow. For the discrete properties the SIS algorithm is chosen. For continuous properties numerous methods exist; however, SGS is chosen here for its ease of implementation and (partial) similarity to SIS, thereby reducing the time needed for implementing the simulation software.

Note that this is only the general approach and that it will be further developed and elaborated on in the following chapters.
In this chapter the background from which the project originates is described. The data serving as a basis for this thesis is related to a heavy mineral sand deposit which is mined using one of the conventional dredging methods: a cutter suction dredger.

3.1. Geology
The main process related to the origin of the deposit is a cycle of transgression and regression of the middle to late Pleistocene sea. Marine clay was deposited in the initial stage of sea transgression. In the northern part of the deposit, close to a mountain range, a lag deposit remained after transgression containing heavy minerals. The area was protected from extreme weather and subsequently, during constant sea level rise, sediments were supplied which lead to a prograding strandline sand formation. These strandline sands now form the deposit, containing heavy minerals, like ilmenite.

During a geological and resource estimation study the following stratigraphic units have been identified:

- Transgressional Lag Zone: irregularly deposited in northern part, high HM content.
- Mineralised Sands: divided in three subunits, with varying degrees of sorting and grain sizes. Bottom subunit contains induration caused by organic processes in water saturated material.
- Clay Floor: paleochannels in clay result in undulating floor of mineral deposit. Contact between mineralised sands and clay defined at 10% slimes content.
- Bed Rock: gneiss underlies the clay floor and the mineralised sands in shallow peripheral areas of deposit.

Although variation in geotechnical properties are partly a result of differences in geology, the geotechnical zones defined during the investigation are not necessarily correlated to the stratigraphic units.

3.2. Geotechnical Investigation
An elaborate geotechnical site investigation was performed after the dredger used at the site was not performing as expected. In situ tests were carried out at 55 locations, consisting of SPTs at multiple depths. At 25 of these locations a CPTu was performed approximately 4 m away from the borehole. At multiple depths the hard indurated layers were found during testing. If this was the case, both the SPT and CPTu were not able to progress further and were assisted by hydrovibracore drilling. The SPT was performed at a vertical spacing of 1.5 m (30 cm sample size) whereas the CPT results in a more or less continuous profile with a sample size of 1 or 2 cm (NEN, 2012).

Several different geotechnical units have been identified across the deposit (table 3.1). The identification is based on borehole core logs, SPT, lab data and CPTu. In total five units have been identified: aeolian sands, strandline sands, strandline sands and indurations, deposit base and residual sandy clay.
The aeolian sands consist of wind blown sands, placed over the strandline sand deposit. These sands make up the top layer, although they are not found everywhere in the deposit. The last two units mentioned form the base of the deposit and are not intended to be mined. They were found to contain more silt and clay than the overlying sand layers. As explained previously the boundary is set at 10% fines content. The second and third units are regarded as the most important, containing the mineralised sands and only smaller quantities of silt and clay. The second layer, strandline sands, is divided into seven sub-zones and the third is divided into three subzones, as shown in table 3.1.

Table 3.1: Geotechnical units description

<table>
<thead>
<tr>
<th>Zone ID</th>
<th>Name</th>
<th>Subzone</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aeolian Sands</td>
<td>A1</td>
<td>Clean, usually freeflowing sands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2</td>
<td>Same as A1, alternating medium/coarse/medium coarse layers</td>
</tr>
<tr>
<td>2</td>
<td>Strandline Sands</td>
<td>B1</td>
<td>Generally non-freeflowing, slightly organic sands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2</td>
<td>Generally non-freeflowing, clean with organic beds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B3</td>
<td>Generally non-freeflowing, organic rich with clean beds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B4</td>
<td>Generally non-freeflowing, alternating medium and silty/clayey sand beds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>Non-freeflowing, ferruginised sands</td>
</tr>
<tr>
<td>3</td>
<td>Strandline Sands and Indurations</td>
<td>A</td>
<td>Like zone 2, with indurated layers (spacing ≤1.0 m)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>Like Zone 2, with indurated layers (spacing &gt;1.0 m)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>Ferruginised sandstone</td>
</tr>
<tr>
<td>4</td>
<td>Deposit Base</td>
<td></td>
<td>Fine grained sand, silt and clay</td>
</tr>
<tr>
<td>5</td>
<td>Residual Sandy Clays</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Zone 2 is subdivided based on organic content, layering, sorting and grain sizes. The consistency varies from loose to very dense, but is geotechnically classified as soil. In case this soil is very dense, this could lead to SPT refusal but would not necessarily indicate an induration.

Zone 3 represents the geotechnical unit containing the hard, sandstone-like, indurations. It is observed to be a mixture of Zone 2 sands with interbedding of indurated layers, caused by organic/fungal growth. Because of this, the large-scale behaviour may be different than the small-scale: meaning that the material, although consisting of soil, behaves different than expected during dredging.

In figure 3.1 the geotechnical units are shown, based on a simulation. In this case 'D' (red) represents the deposit base and sandy clays (geotechnical units 4 and 5), 'C' (yellow) represents the to be mined material (geotechnical units 2 and 3), 'B' (light blue) represents aeolian sands on the top (geotechnical unit 1) and 'A' is air (the boundary between 'A' and 'B' representing the surface elevation).
3.3. Overview of the Dataset
The dataset which is used as the basis for the simulation is provided in Microsoft Access database format and includes:

- CPTu measurements
- SPT measurements
- Borehole info (GPS coordinates, borehole IDs etc.)
- Borehole logs
- Borehole interpreted lithologies
- Geotechnical zone reference
- Geotechnical core lab results
- Dissipation test data
- Permeability test results
- Soil lab results

The SPT measurements are performed at 55 locations (see figure 3.2). At 25 of these locations CPT measurements were taken as well, at a distance of approximately 4 m (however, recorded x, y and z coordinates from GPS are the same for both CPT and SPT). The SPT is accompanied by rotary core drilling resulting in discontinuous measuring, as opposed to CPT which records values at 1 or 2 cm depth intervals.

A report accompanies the measurement data, containing some interpretations of the data. For example, geotechnical units have been identified based on both the in situ and laboratory test results. These have been further divided into sub-units which are correlated with both the CPT and SPT data.

As described in table 3.1, five main zones were identified: two of these (zones 2 and 3) comprise the most important and economically feasible part of the ore body. These zones mainly consist of sandy material but also include indurated lenses, caused by biological activity.

The main differences between geotechnical units 2 and 3 is the presence of indurated material: in unit 3 this was found to be more likely to occur.

By browsing through the data “manually” errors, mistakes and other observations may already be found. In this case, when analysing the CPT and SPT data:

- CPT sample size varies from borehole to borehole (either 1 or 2 cm).
- For one borehole (“CPT13”) horizontal coordinates were present, but the elevation was unknown.
- Float and string values are sometimes mixed, e.g.: ‘Refusal’ is sometimes found in the SPT data.
- In some cases, depth measurements are still present but other data is omitted due to inaccuracy.
- When indurated layers are encountered vibracore drilling is used leading to entries with depth intervals larger than usual (more than 1 or 2 cm, in case of CPT).
- Both commas and dots are sometimes used as decimal separator: Excel may handle these differently depending on system settings, whereas Pandas (a Python package) does not seem to have a problem with this.
- Depth is measured relative to local elevation: elevation of borehole location is required to obtain true depth.
3.3. Overview of the Dataset

Figure 3.2: Borehole locations, according to their GPS coordinates

Figure 3.3: Uncorrected cone tip resistance (MPa) Zone 2 (a) & Zone 3 (b)

Plotting the borehole locations in a 2D grid illustrates the distribution and the space separating them (see figure 3.2): the spacing looks uniform, but the shape of the area coverage seems not to be ideal for simulation (e.g. as opposed to a rectangular or similar shape).

A histogram of the sampled data is shown in figure 3.3. Here the difference can be seen between units 2 and 3: the absolute number of measurements is lower in unit 3, but high $q_c$ values are more frequently present.

In order to save on memory usage and focus only on the mined part of the deposit, the data is transformed and filtered in several ways.

As shown in figure 3.2 most of the data is located within an elongated shape along an azimuthal direction of approximately 60°. In case of a regular (rectangular) simulation grid, there is a large area not covered by samples. To overcome this issue, the coordinates are rotated and transformed according to the following algorithm:

1. Find bounding rectangle, by calculating the convex hull.
2. Determine angle of bounding rectangle with respect to either x or y axis.
3. Transform coordinates by rotating over found angle.
4. Optionally, translate coordinates to obtain (0,0) origin.

Note: if required, the coordinates of the simulation results may be back-transformed for visualisation on their original location.
### Table 3.2: Correlation function parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Capped</th>
<th>Uncapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman</td>
<td>0.521</td>
<td>0.513</td>
</tr>
<tr>
<td>Pearson</td>
<td>0.502</td>
<td>0.496</td>
</tr>
<tr>
<td>Function slope</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Function y-intercept</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

One of the observations mentioned earlier was that the depth was measured with respect to the surface level: starting at 0 and positively increasing with depth. Since the surface level varies significantly over the sampled area, this should be taken into account when defining the simulation input data. This process is quite straightforward: determine surface elevation at sampled location, then multiply each recorded depth value by -1 and subtract the surface elevation.

Next, the data is filtered by zones of interest; in this case these are geotechnical units 2 and 3. Also some of the data was found to be unusable, and is not included in input data: the surface elevation of borehole 'CPT13' is unknown, so the actual depth cannot be determined.

In section 2.2.1 the relevant in situ tests were introduced, namely the SPT and CPT. Several correlation methods (e.g. $q_c/N$ ratio, linear trend) between these tests were described: many different ways exist, depending on local deposit type and experience of the user. The references mentioned show applications in different circumstances but a single "best" method does not seem to be available.

In case of this project both CPT and SPT data is available, and to benefit the most of this fact a suitable correlation should be found. The first attempt was based on depth-corrected data: the depth trend was removed (as described by Lloret-Cabot et al. (2014)) from the $q_c$ values and, in case of SPT, the overburden corrected values were used (see also (NEN, 2005) for more details). This, however, lead to negative $q_c$ values in some cases, while the SPT correction factors were not observed to increase correlation (as stated in the geotechnical investigation report as well).

In the geotechnical report a first attempt at correlation is described, without corrections applied to either of the test results: it was found that in 58% of the cases the correlation coefficient was "acceptable" (0.50 - 0.94), with the remainder showing correlation coefficients from -0.53 to 0.50. Also worth noting is that in this calculation the SPT $N$ values were cut off at 50, as higher values were believed to be inaccurate (although (NEN, 2005) indicates that values from 50 - 100 may still be used, in case of soft rock). Correlation was also attempted with corrected SPT values (e.g. energy, overburden etc.) but this was not observed to improve correlation significantly. Other reasons given for insufficient correlation were inaccurate SPT depth measurement, soil disturbance by vibracore drilling, distance between SPT and CPT (4 m) and the number of samples.

In the available, already processed, data SPT values higher than 50, including 'Refusal', were set to 50 prior to applying correction factors, possibly removing any trend in the higher values. For this reason the original field data was used to preserve this trend: the value for 'Refusal', was chosen to be 75, considering the high values above and below the depth at which refusal occurred. Correlation was attempted for uncorrected $q_c$ and $N$ values, while also comparing the values "capped" at 50 and the original values. The only modification of data prior to correlation was correcting for support differences: over the penetration length for each SPT the $q_c$ values were averaged. Another comparison is made between the type of correlation: Spearman (rank) and Pearson correlation coefficients.
From table 3.2 it can be determined that the capped values yield a slightly higher correlation coefficient. However, the uncapped values which preserved the trend towards higher values still show a Pearson correlation coefficient higher than 0.50.

When figure 3.4 is also taken into consideration it can be seen that the function fitted to the capped values leads to lower maximum $N$ values, compared to the other one. Since this is not expected to be realistic, the function derived from the uncapped $N$ values is chosen. An overview of existing SPT-CPT correlations is presented by Tarawneh (2014). The proposed correlations vary significantly; mainly due to different soil types and deposit histories. For sand and silty sand the $q_c/N$ ratio is on average 0.5. This is consistent with the value shown in table 3.2: the slope of 1.8 results in a $q_c/N$ ratio of 0.56.
3.4. Excavation Equipment

Dredging could be defined as excavating rock or soil, while underwater. This method has been used for centuries and over the course of its existence various methods have been developed. Depending on the type of project and in situ conditions, one method may be more suitable than another. At the moment, the following methods are commonly used:

- **TSHD**: Trailing Suction Hopper Dredger.
- **CSD**: Cutter Suction Dredger.
- **Backhoe**: excavating bucket, similar to excavators used onshore.
- **Clamshell**: clamshell bucket, mounted to an arm or dragline.

The largest dredging vessels are usually TSHDs and CSDs. The main differences between the two are the method of excavation and manoeuvrability. The TSHD is sailing while dredging and “drags” the drag head over the seabed, loosening the material and transporting it vertically by pumps. The CSD on the other hand, is stationary while excavating and its cutter head is moved by winches.

The CSD is able to cut a wide variety of soils and rock, with a maximum compressive strength of 30 to 40 MPa (VBKO, 1998). As the vessel remains stationary while cutting, it is sensitive to wave conditions and could pose a problem towards shipping lanes. The cutting depth is limited by the ladder to which the cutter head is mounted, with a maximum of approximately 30 m.

When considering a wet mining operation consisting of an artificial pond to dredge from, the CSD could be an appropriate choice: no waves and shipping activities, limited depth (depending on mineral deposit). The CSD excavation process relies on both the breaching behaviour of soil in which the slope fails continuously in a controlled manner and the actual cutting of material. The CSD cutting process consists of the following steps:

1. The cutter disintegrates material, rotating the cutter head by the side winches around the main spud.
2. The material mixture is transported by pumps from the suction mouth.
3. Preparing for the next cut, the dredger is pushed forward from the main spud.
4. When in center-line (see figure 3.5), main spud is hoisted up and the auxiliary spud is lowered.
5. Main spud (red) is moved forward by carrier and lowered, then the auxiliary spud is hoisted up.
6. The cutting process continues.

![Figure 3.5: CSD cutting process (van Rhee, 2012).](image)

The properties of the soil affect the feasibility of any dredging operation, for example: grain size distribution, SPT $N$ value, permeability, porosity and internal angle of friction. The SPT is an important test as it can give an indication, by empirical correlations, of several of those soil properties.
As described in section 2.5, the desired simulation framework should be able to handle both discrete and continuous data, i.e. 3D subunits and rock/soil properties; therefore a combination of methods will be used. Furthermore, to ensure the modelling results are realistic and appropriate for further processing, like propagation through transfer functions, they should be validated as well. This may be done by comparing variograms, histograms or deviations of the simulation output from the source input data. In case there is not sufficient original data present to facilitate these validation procedures, a simulation or estimation based on these samples may be performed to obtain a "virtual reality", on which to build further processing steps and demonstrate their possibilities. These processes are described in this chapter.

4.1. Simulation Process Overview

In section 2.1, as part of the literature review, it was described how the definition of a simulation workflow depends on the input data, as samples may be obtained by different methods. The flowchart shown in figure 4.1 proposes the general process, consisting of data input, simulation, processing and applying transfer functions.

The decision was made to generate a virtual "exhaustive" dataset for facilitating the desired validation steps, based on the question whether sufficient sample data is available for reliably performing the investigation; this was also outlined in problem statement (section ). In that case, however, only the validity of the process is ensured, based on the model, which is then not necessarily related to the actual measurements but still provides insight into the effect of sampling on uncertainty reduction and its propagation through transfer functions (Demougeot-Renard et al., 2004).

In figure 4.2 the general steps are proposed for performing a geostatistical simulation (box "Simulation" in figure 4.1), along with a brief description (further details follow in section 4.2). Usually the data preparation consists of an Exploratory Data Analysis (EDA), to obtain a general overview of the available data. Also, part of the process is iterative: for example, the variogram inference is a process partly based on knowledge and experience of the person performing the analysis, and might be repeated if the simulation results do not sufficiently honour the original sample data or correlation characteristics.

Error Check  The input data may be obtained by various in situ and/or laboratory procedures. Before starting to actually use the data it might be useful to check for errors, like missing data columns or rows and mixed use of string and integer/float values.

Exploratory Data Analysis  An important step in performing a geostatistical investigation is the exploratory data analysis, or EDA, which is used to get acquainted with the dataset and obtain statistics. Some common operations are generating a histogram, Q-Q plot, scatter plot. The results, in this case, will be statistics like mean, variance and frequency distribution. Depending on the simulation method
and spatial distribution of sample locations, declustering of preferential samples may be necessary, ensuring that data points used as input for simulation do not result in biased output.

**Histogram Transformation** Some simulation methods assume a certain distribution of input data. For example SK, and therefore SGS, assume a normal distribution. In this case the data has to be transformed prior to inferring variogram models and executing simulation algorithms. The cdf used for transformation is interpolated between sample extremes and the tails are extrapolated (e.g. by using a power function).

In case this forward transformation is applied, the simulation output has to be back-transformed to obtain values in a similar range as the input data.

**Variography** The experimental variogram is based on the sample data; it is a two-point statistic describing the correlation between sample values at different lag distances. After experimental variogram calculation, a model is fitted, to adhere to the requirement of permissibility (section 2.3.3).

**Simulation** This part of the flowchart represents the application of a geostatistical simulation method; in this case either SGS or SIS, depending on the type of input data (i.e. continuous or discrete).

**Output Data** The set of realisations, forming the output from the simulations, are subjected to several analyses: first, the data should be compared to the original data to make sure that they honour the sample data and original distribution and statistics. If this is not the case, there might be a problem with the semivariogram models, which describe the spatial correlation, or the simulation and transformation methods used. Next, the transfer functions are applied and the results are analysed.
4.2. A Virtual Exhaustive Dataset

As described in sections 2.5 and 4.1, the approach is based on the type of deposit and goal of the simulation. The deposit is of a layered sedimentary type, resulting in a large lateral extent compared to the maximum vertical extent. The approach consists of multiple steps: simulate boundaries, simulate indurated zones (contained within boundaries) and finally simulating the geotechnical properties within the layer and indurated lens boundaries.

As shown in figure 4.1, the available sample data may not be sufficient for performing the desired investigations. Therefore, a virtual, high-density (exhaustive, i.e. well known) dataset is generated which will serve as input for subsequent virtual sampling and simulation procedures.

4.2.1. Grid Definition

The grid is a regularly spaced array of cells onto which the simulation algorithm is applied. Depending on the situation the grid is either two or three dimensional. The decision for spatial extent and cell dimensions (and thus number of cells) is based on the requirements set by the application of the simulated data. For example, if the deposit to be modelled has a certain shape, it might be appropriate to de-
fine the cell dimensions accordingly to be able to honour spatial correlation of structures and properties.

As shown in table 2.1, the SGS algorithm does not necessarily require a regular grid, but, for the purpose of this thesis, a regular grid is required. The cell dimensions are selected to comply with the shape of the mineral deposit: long in horizontal directions and short in the vertical direction (cell dimensions: 25 x 25 x 0.50 m\(^3\)). This cell size may seem small, considering the original sample spacing. However, a sufficient number of cells should be available to accommodate the following investigation of varying sampling densities. Furthermore, the cell size is also related to the minimum block size mined by the excavator, also referred to as SMU (Smallest Minable Unit). The resulting grid based on these requirements is shown in table 4.1.

<table>
<thead>
<tr>
<th>Grid property</th>
<th>Value (x, y, z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cells</td>
<td>132, 66, 82</td>
</tr>
<tr>
<td>Cell size</td>
<td>25.0, 25.0, 0.5</td>
</tr>
<tr>
<td>Center coordinate of lower left cell</td>
<td>12.5, 12.5, -26.75</td>
</tr>
</tbody>
</table>

In figure 4.3 the search ellipsoid (or neighbourhood) is shown in three dimensions; if anisotropy in a certain direction is to be taken into account, this ellipsoid can be transformed and/or translated: e.g. if the azimuth is set to 90°, the x-direction switches from the medium range to the maximum range. In case of 3D variograms, this ellipsoid also illustrates the ranges, which are then the maximum distances of spatial correlation. The specific characteristics of each search ellipsoid related to one of the simulated properties is shown in the next sections.

4.2.2. Layer Boundaries

The bounding surfaces are an important part of stochastic simulation (Rossi and Deutsch, 2014): the volume is a stationary domain, having a first-order impact on reserves, production planning and economic forecasting.

An assumption made in the first step is that the boundaries do not overlap or cross each other. This results in a more simplified simulation: each boundary can be simulated individually in two dimensions, with x and y coordinates and a single property (depth) to be simulated. If necessary, these can later be combined in a 3D grid for visualisation purposes or for data processing applied to a specific zone. The touching of boundaries is, however, included, as this facilitates pinching-out behaviour of layers.
The simulation of a boundary is performed according to the following process:

1. Obtain layer boundaries at each sample location (based on SPT, CPT, borehole and lab data).
2. Transform the data into the normal score domain.
3. Determine variogram in normal score domain.
4. Define simulation grid.
5. Choose sample search ellipsoid appropriate for available data (see figure 4.3).
6. Perform SGS.
7. Export data.
8. Combine the separate 2D boundaries to form a single 3D grid.

During the data preparation (section 3.3) the depth of the various boundaries have been extracted from the data. CPT measurements were preferred, as their vertical sample spacing is smaller compared to SPT. However, at locations where only SPT data is available, these have been used (SPT was performed at intervals of approximately 1.5 m).

As SGS assumes a normal distribution, normal score transformation is required if the data is not distributed accordingly. Part of this is defining the minimum and maximum values for cdf lower and upper tail extrapolation, if the range of the cdf needs to be increased (Goovaerts, 1997). Since a previous assumption states that boundaries are not allowed to intersect, these values should not vary significantly from the sample data. For example, in case of the top boundary of combined geotechnical units 2 and 3, the sample minimum and maximum were -0.665 m and 10.604 m, respectively. The values used for extrapolation: -1 m and 11 m (see table 4.3).

The variogram is modelled in order to obtain a representation of spatial correlation. The horizontal spacing of the original measurements is large, as will be illustrated by the resulting variogram: the spacing ranging from approximately 300 to 500 m with a horizontal domain extent of about 3200 by 1600 m. The experimental variogram calculation parameters are chosen in such a way that sufficient sample data is captured in a single lag distance (Isaaks and Srivastava, 1989), shown in table 4.2.

<table>
<thead>
<tr>
<th>Variogram</th>
<th>Omnidirectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>12</td>
</tr>
<tr>
<td>Lag separation</td>
<td>300 m</td>
</tr>
<tr>
<td>Lag tolerance</td>
<td>100 m</td>
</tr>
<tr>
<td>Number of directions</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>45°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
</tr>
<tr>
<td>Tolerance</td>
<td>180°±1</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100 m</td>
</tr>
</tbody>
</table>

In this case an omnidirectional variogram is used to model the correlation. If more data would be available it is also possible to combine several directions into one variogram to take anisotropy into account (Isaaks and Srivastava, 1989), both horizontally and vertically. However, in this 2D case of a layer boundary, insufficient data is available for determining directions of anisotropy (note the lag separation and tolerance); therefore and omnidirectional model is chosen.

Considering the experimental variogram output it may seem logical to define a large nugget effect (see figure 4.4). However, the nugget effect accounts for measurement inaccuracies and geological and spatial variability. The division of geotechnical units is (partly) defined by sedimentary layers: the

\[A\] tolerance exceeding 90° indicates an omnidirectional variogram.
depth of the boundary between such layers is not expected to vary much between two sample points: e.g. at a grid spacing of 25 m, a large difference between two neighbouring points is not logical, since sedimentary layers are deposited horizontally. Therefore, the nugget is kept small but a linear combination of two exponential structures is used. This results in a steep incline in the variogram at small lags and a decreasing incline with increasing lag distance. Thus, instead of large nugget, a short-scale structure is used, resulting in a continuous transition between locations instead of a more sudden increase.

The search ellipsoid radii are based on both the sample spacing and variogram ranges. Since correlation between data points is limited to a certain range, as defined by the variogram, the ranges for the ellipsoid should not exceed these.

The result of performing a simulation for each boundary is a set of 2D realisations. These can be combined into a single 3D grid to help visualisation, as shown in figure 4.5.

### 4.2.3. Indurated Zones

From the site investigation it is known that the layers containing hard material vary in thickness from 10 to 20 cm and that their lateral extent is limited due their lens-like shape and pinch-out behaviour. Their locations are derived from both the CPT and SPT data (see section 3.3).

In the case of CPT data, the locations are obtained by finding the measurement intervals larger than the standard interval of 1 or 2 cm. This is based on the observation from the data that hard indurations are preceded by high CPT values, finally resulting in the cone not being able to penetrate further and the assistance of drilling, during which no CPT is performed and thus no measurements are present. For SPT, the layers are indicated by a ‘Refusal’ entry: searching the database for these entries will then give the required locations.
4.2. A Virtual Exhaustive Dataset

However, this method is based on some assumptions:

- 'Refusal' is only caused by soil induration.
- Although the thickness of indurations is known (0.1 - 0.2 m), vertical extent derived from sample data seems much higher (e.g. 1 m).
- The hard layers occur in the form of lenses, pinching out horizontally.

The sample derived thickness is higher than expected; this might be a result of multiple thin layers closely spaced vertically or a safety margin applied while drilling (CPT tests were accompanied by vibracore drilling, for penetrating the harder layers). Therefore, the data is prepared prior to simulation:

1. Determine vertical extent of hard layers at each sample location.
2. Assign indicator value:
   - 1 for a hard layer
   - 0 for other material
3. Find probability of occurrence.
4. Export data.

The modelling of the harder, rock-like, layers is conducted using SIS, which is able to handle categorical or indicator data. From the geotechnical investigation it is clear that CPT and SPT data is only available to a limited extent. From this data, however, it is still possible to determine the location of the layers. Also, samples of this particular material were collected and subjected to laboratory experiments; in this case UCS and BTS tests.

Similarly to determining the depth at which boundaries are located, the CPT depth measurements are preferred in this case as well. However, the depth from CPT is not as accurate in this case: during the execution of the CPTs, when encountering a hard layer, a vibracore drill was used to penetrate the layer, allowing the CPT to continue further. This means that the tested soil may include material from partially collapsed borehole walls or that the drill penetrated further than necessary.
Table 4.4: Indicator experimental variogram parameters

<table>
<thead>
<tr>
<th>Variogram</th>
<th>Horizontal</th>
<th>Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>300</td>
<td>1</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>100</td>
<td>0.5</td>
</tr>
<tr>
<td>Number of directions</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>45°</td>
<td>0°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>90°</td>
</tr>
<tr>
<td>Angular Tolerance</td>
<td>180°</td>
<td>5°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100</td>
<td>0.5</td>
</tr>
</tbody>
</table>

![Variogram model fitted to experimental points, indurated layers.](image)

The thickness of these layers was quantified and found to range from 10 to 20 cm. From the number of known locations of hard layers occurring along the length of CPT and SPT tests, the probability for this occurrence can be determined.

The probability of occurrence for each categorical variable, or "indicator", is a required input value for the SIS method. In this case there are only two possibilities: a 1 for a hard layer, and a 0 for the rest. The probability is derived from the total observed layer thickness and borehole lengths, resulting in the following probability: 0.064 or 6.4% for encountering an induration.

Contrary to the simulation of boundaries, this part is performed in three dimensions. Therefore, if horizontal versus vertical anisotropy is considered, this has to be accounted for in a single variogram. In table 4.4 the parameters for determining this experimental variogram are listed.

Table 4.5: SIS parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of realisations</td>
<td>100</td>
</tr>
<tr>
<td>Random seed</td>
<td>14071789</td>
</tr>
<tr>
<td>Indicator values</td>
<td>2</td>
</tr>
<tr>
<td>Corresponding marginal probabilities</td>
<td>0.936 0.064</td>
</tr>
<tr>
<td>Search ellipsoid range (see figure 4.3)</td>
<td>(350, 350, 0.20)</td>
</tr>
<tr>
<td>Variogram range</td>
<td>(468, 468, 8)</td>
</tr>
<tr>
<td>Type</td>
<td>Full IK^2</td>
</tr>
</tbody>
</table>
As explained earlier, these layers were observed to have a thickness in the range of 10 - 20 cm. This is taken into account during simulation by choosing an anisotropic search ellipsoid with dimensions 350 x 350 x 0.20 m³. The result is a distribution of cells (per realisation) representing possible occurrences of hard layers; an example of which is shown in figure 4.7. The z-direction of the grid is smaller than the cell size here: the reason for this is that smaller cell sizes lead to a rapid increase in cell number and therefore calculation time and another reason is that, based on the geotechnical investigation, the layer thickness is limited; this way, preference is given to horizontal correlation and extent, as opposed to vertical correlation.

The simulation output contains the locations of these layers: (geotechnical) properties may be assigned to them, as will be described in the following sections.

![Indurated Zones - 3D SIS Simulation](image)

**Figure 4.7: 3D simulated grid, showing one realisation of indurations in orange (indicator ‘1’), to be constrained by unit boundaries; vertical scale x20.**

### 4.2.4. Geotechnical Properties

The division of geotechnical units was determined during a previous site investigation, as described in section 3.2. The simulation of these properties is restricted to units 2 and 3, since these contain the material to be mined, as well as the indurated zones. Both CPT and SPT data is used as input for the simulation; CPT is in this case converted to SPT using a correlation function.

Since the spatial arrangement of samples is still the same as described in section 4.2.2, the experimental variogram parameters are also similar, shown in tables 4.2 and 4.6.

The resulting experimental variogram is shown in figure 4.8, together with the inferred model. Again, a possible lack of samples may be observed here: at small lag distances (i.e. 0 - 250 m) there are no samples present, thus making it difficult to substantiate the inferred model.

Considering algorithm 1 the parameters needed for setting up the simulation are as shown in table 4.7, with sample data and variogram serving as input. The sample search ellipsoid and variogram ranges reflect the shape of the deposit and expected structures here: a relatively small ratio between thickness and horizontal extent.

---

2 Full IK requires a variogram for each category; contrary to Median IK where a single one is used. However, because of the single category in this case (valued either 0 or 1), a single variogram is used here as well.
Table 4.6: SPT experimental variogram parameters

<table>
<thead>
<tr>
<th>Variogram</th>
<th>Horizontal</th>
<th>Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>12</td>
<td>82</td>
</tr>
<tr>
<td>Lag separation</td>
<td>300</td>
<td>0.5</td>
</tr>
<tr>
<td>Lag tolerance</td>
<td>100</td>
<td>0.2</td>
</tr>
<tr>
<td>Number of directions</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>45°</td>
<td>0°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>90°</td>
</tr>
<tr>
<td>Angular tolerance</td>
<td>180°</td>
<td>10°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Figure 4.8: Variogram model fitted to experimental points, SPT N values.

Table 4.7: 3D SGS parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of realisations</td>
<td>100</td>
</tr>
<tr>
<td>Random seed</td>
<td>14071789</td>
</tr>
<tr>
<td>Search ellipsoid range (see figure 4.3)</td>
<td>(500, 500, 10)</td>
</tr>
<tr>
<td>Variogram range</td>
<td>(360, 360, 11)</td>
</tr>
<tr>
<td>Lower tail extrapolation</td>
<td>-</td>
</tr>
<tr>
<td>Upper tail extrapolation</td>
<td>-</td>
</tr>
</tbody>
</table>
For the purpose of using the results of the different simulations in further processing, such as production forecasting and uncertainty analysis, they are combined into a single model leading to multiple realisations each containing the layer boundary (also referred to as “constraining”), induration and SPT information. A realisation of the 3D model obtained here is shown in figure 4.10, based on the exhaustive data presented up to this point. In this figure the indurated zones can be seen, indicated by the light blue horizontal layers surrounded by darker blue, continuous, SPT $N$ values.

Remarks

The simulations performed in this chapter provide several results and insights:

- **Samples** At first glance, the limited availability of sample data may not be obvious (e.g. observing figure 3.2). However, when attempting to fit a variogram model it became clear that sufficient data for modelling the correlation at small lag spacings was not available. This observation further substantiates the choice for generating a virtual reality, on which to base subsequent simulations and processing.
• **Variogram range** The ranges from the variograms inferred are, on average, similar: approximately 500 m. Most likely this is a result of previous observations, i.e. insufficient samples to safely assume ergodicity.

• **Simulation** Each simulation gives a set of realisations for each property (top boundary, bottom boundary, induration location and SPT $N$ value). Since the available number of samples is insufficient for further uncertainty analysis based on varying sampling density, a single realisation of the virtual reality per property will serve as input for performing virtual sampling, simulations and transfer functions.

### 4.3. Transfer Functions

For the purpose of this thesis, transfer functions are applied to the simulation output allowing for the assessment and forecasting of dredge performance. Two functions are used here: one for obtaining cutter head production in $\text{m}^3/\text{h}$ and another for evaluating production costs per ton dry material excavated.

#### 4.3.1. Production Forecasting

The effect of sampling uncertainty on key performance indicators (KPI) forms one of the objectives of this thesis. An example of such a KPI is the cutter head production from a CSD. Aside from the soil and rock behaviour of the local deposit conditions, this production value is affected by many parameters related to the CSD vessel type and other equipment used. Within the scope of this thesis the production is calculated by propagating the simulated soil properties, as well as vessel parameters selected based on the deposit, through a transfer function.

The input values for this transfer function are shown in table 4.8.

<table>
<thead>
<tr>
<th>Input Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total power installed</td>
<td>2819</td>
<td>$[kW]$</td>
</tr>
<tr>
<td>Cutter head diameter</td>
<td>2.2</td>
<td>$[m]$</td>
</tr>
<tr>
<td>Cutter power at shaft</td>
<td>550</td>
<td>$[kW]$</td>
</tr>
<tr>
<td>Cutter speed</td>
<td>30</td>
<td>$[rpm]$</td>
</tr>
<tr>
<td>Swing winch power</td>
<td>70</td>
<td>$[kN]$</td>
</tr>
<tr>
<td>Dredging depth</td>
<td>15</td>
<td>$[m]$</td>
</tr>
<tr>
<td>Pump limit</td>
<td>3000</td>
<td>$[m^3/h]$</td>
</tr>
<tr>
<td>Spill factor</td>
<td>25</td>
<td>$[%]$</td>
</tr>
<tr>
<td>Dry density</td>
<td>1.71</td>
<td>$[t/m^3]$</td>
</tr>
<tr>
<td>SPT $N$ value</td>
<td>location dependent</td>
<td>$[\cdot]$</td>
</tr>
</tbody>
</table>

| Output                           | location dependent | $[m^3/h]$ |

The production is assumed based on a certain excavation sequence, as illustrated by figure 4.11. The direction of excavation is changed in the x-direction, with one row having the width of one block (note that for illustration, the block dimensions are exaggerated in the figure).
4.3.2. Production Costs

Based on the produced volumes forecasted during a certain timeframe and the variability in soil conditions across the study area, a cost estimate is calculated in order to get an overview of the effect of the different sampling strategies on the costs involved in dredge production. The cost estimation is based on the values shown in table 4.9. Furthermore, the power requirement (i.e. from the cutter head) is expected to depend on soil/rock strength: for the effect of soil strength on power requirement a linear relation between SPT and required power is assumed. The operating costs are in this case assumed to only depend on the required time for excavation, whereas the maintenance cost estimation is based on the amount of material excavated. These factors are also affected by material strength: for example, stronger material could require more time for excavation and thus increased costs per m$^3$.

Table 4.9: Production cost estimation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>0.10</td>
<td>[euro/kWh]</td>
</tr>
<tr>
<td>Dredge Operator</td>
<td>50000</td>
<td>[euro/yr]</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.20</td>
<td>[euro/t]</td>
</tr>
</tbody>
</table>

4.4. Experimental Setup

In this section the investigation methods for several cases are described. The experiments are conducted by performing geostatistical simulations, with varying input data.

4.4.1. Varying the Sampling Density

Samples are obtained from a realisation of the virtual exhaustive dataset described in section 4.2. Since it is based on simulations it is possible to compare various sample spacings and preferential sampling methods. For example, samples do not necessarily have to be obtained from a regular grid (Goovaerts, 1997): local topography may limit the accessibility of certain locations or specific areas might require more densely spaced samples based on prior knowledge: in which case a preferential sampling grid...
consisting of a combination of a regularly spaced grid and more closely spaced zones may be used. In order to compare these possibilities a variety of grids is chosen, as defined in table 4.10.

Table 4.10: Sampling strategies specifications

<table>
<thead>
<tr>
<th>ID</th>
<th>Main grid Spacing</th>
<th>Main grid Type</th>
<th>Preferred zone Spacing</th>
<th>Preferred zone Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse</td>
<td>500 m</td>
<td>Regular</td>
<td>25 m</td>
<td>Cross</td>
</tr>
<tr>
<td>Cross</td>
<td>250 m</td>
<td>Regular</td>
<td>50 m</td>
<td>Rectangle</td>
</tr>
<tr>
<td>Medium</td>
<td>100 m</td>
<td>Regular</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The grids resulting from these specifications are shown in figure 4.12, in which the square and cross shaped preferred zones share the same centre coordinates: (500,500), (2000,1200) and (2500,300). The choice for these coordinates is not based on prior knowledge, like certain zones of interest, but are randomly distributed over the study area in order to assess their effect, compared to a regular grid, on simulation performance.
4.4. Experimental Setup

Figure 4.12: Spatial representation of the selected sampling strategies.
The execution of the simulation algorithms is similar to the description in section 4.2; the relevant parameters are shown in appendix C. The number of realisations is 100, based on the findings by Goovaerts (1999) illustrating that from a number of 20 realisations and onwards the decrease in modelled uncertainty becomes slower; important to note here is that this depends on the type of simulation (e.g. short or long-term) and the algorithm used. Another cause is the density of samples: more closely spaced samples lead to less variability in possible simulation outcomes. Also post-processing may require a high number of realisations: when optimistic and pessimistic scenarios are defined (i.e. the P10 and P90 cases) sufficient realisations should be available to accommodate this process (Goovaerts, 1999).

Regarding the software used here: although SGeMS provides a user interface to define input and simulation parameters, this was not attempted at this stage of the investigation. Instead a Python script was written for both saving time and automating the process, of which some important features are shown below:

- **Random seed** Define unique random seed for each simulation; resulting in different random paths for visiting grid locations. As recommended by Remy et al. (2009) this value is a large odd integer; refer to table 4.3 for an example.

- **CDF extrapolation** Calculate values until which to extrapolate the cdf, based on the range of input values from each sampling grid.

- **Variogram** Import variogram model parameter files, which were manually inferred prior to simulation.

- **Simulation grid** For each simulation, automatically switch either to a 2D or 3D grid, based on input sample data requirements.

In this case the output of the script consists of 20 GSLIB files, from the simulation of 5 sampling grids and 4 parameters, which are combined in the same manner as described in section 4.2.
4.4.2. Target Production

The calculation method for obtaining production values based on vessel and soil parameters is introduced in section 4.3. In order to make use of this data it will need to be processed, based on e.g. requirements from the ore processing plant and subsequent production targets. In such a case, the target production ensures an approximately continuous flow of material entering the processing plant. Within the scope of this thesis, however, the production target is only based on the CSD pump limit and the estimated production value statistics:

- The mean production value observed, based on the various sampling strategies, is approximately 1200 m$^3$/h.
- The maximum pumping limit of the CSD is assumed to be 3000 m$^3$/h (see figure 4.13).
- Consequently, three production targets are chosen: 600, 900 and 1200 m$^3$/h.

The effect of uncertainty from varying sampling densities on production forecasts is quantified based on these production targets. In this investigation, the following cases are considered:

- A probability of 80% or higher that the production estimated at a location (a single grid cell) meets or exceeds the target value.
- A probability of 80% or higher that the production estimated at a location misses the target value.
- Uncertain: the probability of attaining target production lies between 20% and 80%.

Important to note is that when considering these cases, discrete values are produced. This means that when interpreting this data, it may not be entirely clear to which extent an "uncertain" location is actually uncertain: for example using this method, a probability of 75% is marked as "uncertain" (and vice versa) since it is close to, but still below, the chosen probability of 80%.

In this case, when plotting this data in two dimensions, a combination of quantile maps is obtained (see section 2.4.2). By only showing three categories it is clear in which zones target production is reached and in which it is not, thereby possibly indicating zones requiring additional sampling.

An alternative to this approach could be the calculation of continuous probabilities (resulting in a probability map as introduced in section 2.4.2), or considering multiple probabilities (e.g. 50% or 70 %, instead of 80%) using the discrete method. This way the "behaviour" of the uncertain zones may be investigated.

Figure 4.13: Production forecasting based on geotechnical parameters.
4.4.3. Incorporating Indurated Zones in Production

The lenses consisting of hard material distributed over the deposit may cause problems during excavation: if the used equipment is designed based on only considering the soft material, production could be negatively affected by trying to excavate this hard material.

A production estimation method based on UCS is used here, as shown in algorithm 4.

**Algorithm 4** Incorporating Indurations in Production Estimation

**Require**: Cutter head diameter and maximum layer thickness.

1: for all realisations do
2: for all y coordinates do
3: for all x coordinates do
4: Obtain induration thickness.
5: if thickness < cutter head diameter then
6: Calculate production from UCS.
7: else if thickness > cutter head diameter then
8: Assume production of 0 m$^3$/h.
9: end if
10: end for
11: end for
12: end for

The algorithm used here is based on recurring characters in a GSLIB file, indicating the number of cells sharing the same indicator value in the vertical direction, thereby calculating the thickness needed for comparison with the cutter head diameter.

![Figure 4.14: Two dimensional slice of 3D grid containing production values: examples of indurated layers are the blue lines, indicating low production, surrounded by a red area (high production, i.e. soft material).](image)

As described by algorithm 4 the thickness of indurated zones is taken into account when calculating the production. An example of the location and thickness of these zones is shown in figure 4.14. This slice is taken from the virtual dataset, used as input for sampling. Here it can be seen that although indurations are present in the area, and are surrounded by soft material (indicated by the high production forecasts, in red), their thickness is limited in comparison to the cutter head diameter and they would therefore, most likely, not have a large influence on production continuity.
4.5. Software Implementation

During this research project a combination of readily available software and programming are used. Many implementations exist of the current variety of geostatistical estimation and simulation algorithms. An example of such a software package is SGeMS, or Stanford Geostatistical Modeling Software. This software provides a graphical user interface and is able to handle different types of input files, like CSV and GSLIB. Both data in two and three dimensions can be processed and visualized. In case tasks have to be repeated or automated the software may be accessed from a command line interface or Python scripts. A user’s guide has been written by the creators Remy et al. (2009), containing tips for using the software and also explaining some of the underlying geostistical theories.

SGeMS Several steps of the simulation process will make use of the SGeMS software:

- Fitting semivariogram models to the data.
- Data transformation: e.g. normal score transform in case of continuous data.
- Execution of the sequential simulation algorithms.
- Back-transformation of the data to its original domain.

Python Another tool that will be used is the Python programming language. It is convenient for automating certain tasks and, also, additional libraries exist to help with some of them. An example of such a library is Pandas: this is useful for importing large amounts of data, inspect this data, obtain information and statistics. Using Python in combination with packages like Pandas and Matplotlib, it is then possible to import data, export it, apply SGeMS operations and finally visualise and analyse the results.
In this chapter the results are presented, which are obtained during several stages of the simulation process: variograms inferred prior to simulation, direct simulation output (e.g. boundary depths, SPT) and processed results.

The spatial correlation is an important factor in geostatistical simulation: when comparing the simulation output to the sampled area statistics the resulting variograms should look similar and thus represent the same structures. In case of this investigation, however, different sampling strategies are compared, possibly resulting in different correlation ranges.

The performance of the simulation is assessed by comparing the original data to the P10 and P90 quantiles from the simulation output. The separation between P90 and P10 quantiles (or spread) should, ideally, approach zero at sample locations, since the algorithms honour sample data; on the contrary, when moving away from this location the spread is expected to increase.

Another indicator of the simulation performance is the similarity between the input and output distribution. This comparison is performed by calculating the root mean square error (RMSE) between the sampled, original, dataset and the output realisations.

The processing steps following the simulations consist of applying transfer functions to the simulation output. This way the effect of varying sampling densities on dredge production forecasting is investigated, considering both produced volumes and production costs.
5.1. Spatial Correlation

In the most ideal case, the simulations performed based on samples reproduce the spatial structures present in the study area. However, as explained by Goovaerts (1997), these deviations by the realisations from the model statistics, called “ergodic fluctuations” (see section 2.3.3), are controlled by multiple factors:

- The simulation algorithm used.
- The number of available samples for conditioning.
- The semivariogram parameters and the simulation grid dimensions.

![Experimental Variograms](image_url)

Figure 5.1: Ergodic fluctuations in simulations (10 randomly chosen realisations); top boundary, based on coarse (a) and cross (b) sampling schemes.

A comparison between two sampling strategies is presented in figure 5.1. Here it is observed that the coarse sampling grid leads to a relatively high fluctuation in reproduced statistics, similar to the fluctuations at high lag distances for the preferential sampling grid (with cross-shaped zones). However, when considering small lag distances (i.e. lower than 300 m) the realisations exhibit smaller fluctuations, which can be attributed to the preferential zones. These zones consist of cross-shaped areas containing samples densely spaced at 25 m, combined with a coarse grid spaced at 500 m; this way the variograms are able to represent both small and large scale structures.
5.1. Spatial Correlation

Figure 5.2: Comparison of variograms for layer top, based on different sampling strategies; “original” referring here to the variogram used to generate the virtual exhaustive dataset.

Figure 5.3: Comparison of variograms for layer bottom, based on different sampling strategies; “original” referring here to the variogram used to generate the virtual exhaustive dataset.
5. Results

Figure 5.4: Comparison of variograms for indurations, based on different sampling strategies; “original” referring here to the variogram used to generate the virtual exhaustive dataset.

Figure 5.5: Comparison of variograms for SPT N values, based on different sampling strategies; “original” referring here to the variogram used to generate the virtual exhaustive dataset.
5.1. Spatial Correlation

Discussion In this section the variograms have been presented for the different simulated properties and a comparison is made for the different sampling strategies, also including the original simulation input data (referring to the virtual exhaustive data discussed in section 4.2).

Several observations:

- **Nugget** With increasing sample spacing the nugget also increases. This illustrates the effect of a small number of samples: fewer samples making it more difficult to identify short-scale structures.

- **Sill** The variograms are calculated in the normal score domain. Therefore, the sill is expected to be 1, equal to the variance of the Gaussian distribution. When considering the "original" variogram it may be observed that this value is not reached, possibly a result of the area size: if the area used for variogram calculation is too small compared to the study area, the observed variance could be lower than 1.

- **Range** The range, or the lag separation distance at which the sill is reached, varies with sampling strategy: the spatial structures found depend on sample spacing; meaning that with increased density and differences in preferential samples a different correlation distance is found. This is also illustrated by the experimental variograms (see section C.2): when including preferred zones, the inference at small lag distances (25 and 50 m) is improved and the modelling of smaller structures is supported by the more closely spaced sample data, as indicated by the experimental points. However, when the modelling is based on larger sample spacings, upwards of 100 m on a regular grid, the variogram modelling process becomes more arbitrary and is only able to detect larger spatial structures.

- **Inference** It is worth noting here that variogram model inference may be prone to subjectivity and thus result in varying results, depending on the person carrying out the process. This may be a cause for the SPT variograms to look similar, as shown in figure 5.5. When taking into account the experimental variograms from section C.2, it can be seen that no clear structure is present in, for example, the variogram based on the coarse sampling strategy; therefore a different model could also have been likely.

In literature a similar relation between correlation range and sampling density is found: an investigation by Basarir et al. (2010), concerned with a stockpile site, consists of sampling by SPT at spacings ranging from approximately 20 to 100 m. Although OK is used here for estimation, the process of variogram inference is identical when using simulations. The ranges found, based on smaller sample spacings compared to ones described in this thesis, are fluctuating around 7 m. The calculated semivariance values by Basarir et al. (2010) cannot be compared to the variograms presented in this section as normal scores are used here and it is likely that the local geology is different. The reason for using normal scores is to allow for easier variogram inference since the effect of outliers is reduced and variograms of different properties (i.e. elevation (m) and SPT N values can be compared.

Another study by Altun et al. (2013) is also based on SPT testing in an area of the same approximate dimensions as the one introduced in section 3.2. The lateral separation of boreholes varies from 100 to 500 m, comparable to the samples used for the purpose of this thesis. The correlation ranges found by Altun et al. (2013) are on average 350 m, which is consistent with the values found here, shown in figure 5.5.
5.2. Simulation Validation

An important aspect when conducting research by performing simulations is the validation of the output results: as a simulation is an approximation of reality it is vital to know how well the simulation performs and whether the results are realistic.

5.2.1. P10, P90 and Sample Locations

Discussion  In figure 5.6 the interval defined by the P10 and P90 cases is shown, along with the mean and original values at the specified locations. The observations made here are as follows:

- **Samples** In the example of the coarse sampling grid, with a 500 m regular spacing, it can be seen that at sample locations the sample data is honoured and, when moving away from these locations, the spread of possible values increases.

- **Spreading** The amount of spread defined by the P10 and P90 varies with sampling strategies: in case of the coarse grid, the spreading increases rapidly when moving away from a sampled location; here the rate at which the spreading increases also differs per sampling strategy. Considering the cross-shaped clustered zones, it can be seen that there is no spread (see figure 5.7a); this fact can be attributed to the sample spacing within the cross being equal to the sampled virtual exhaustive dataset. Consequently, in the rectangle-shaped zones with a slightly higher sample spacing (25 vs. 50 m) there is an increased spreading visible, albeit still relatively small.

- **Uncertainty** Although the sample spacing is high for the coarse grid, the general trend is still preserved and approximately 75% of the original data lie within the interval defined by the P10 and P90 values. This means that the simulated values are consistent with the original values, although the values should ideally lie within an 80% interval (as expected, taking into account that the interval is limited by the P90 and P90 values). This is likely to be a result from simulation limitations: for example, CPT is converted to SPT using a correlation function, possibly introducing additional variability.
Figure 5.7: P10 and P90 plots along a one dimensional line; based on cross (a), rectangle (b), medium (c) and fine (d) sampling grids.
5.2.2. Root Mean Square Error

The mean square error is used here for evaluating the simulation performance, i.e. quantifying the proximity of the simulation output to the original data. In this case the MSE is the mean of the squared differences of values (100 realisations) at each grid location, between original data and simulation output. The root of the MSE is taken after averaging over the complete grid, resulting in the RMSE, in order to obtain values of the same unit (a single value per sampling strategy, for comparison).

Discussion  The average RMSE is calculated for the top boundary, the bottom boundary and the SPT \( N \) values. In general, the trend observed when moving from the coarse towards the fine grid is a decrease in RMSE. This seems logical, since more samples lead to more information upon which to base forecasting. However, when looking more closely, (small) deviations from this trend are observed: the fine grid at 100 m spacing results in the lowest error for all sample strategies, but when comparing the preferential grids to the medium (250 m) a clear trend is less obvious.

Although the medium grid contains less samples in absolute number, it results in a lower error for all three cases shown in figure 5.8, compared to the preferential strategies. This is possibly a result of the preferential grid set-up: these are combinations of a coarse and a very fine grid. Comparing the variograms (see section 5.1) it is clear that this combination leads to a better reproduction of spatial correlation at small lag distances, but judging from the RMSE results, an early conclusion that may be drawn at this point is that evenly spaced conditioning data results in a lower error compared to preferential sampling. This is similar to the observations made by Goria et al. (2001), in which additional sampling leads to more information and increased knowledge of the range of possible values, thereby changing the base case and in some cases leading to even more uncertainty.
5.3. Comparing the Sampling Strategies

In this section the effect of various sampling strategies is assessed by comparing the simulation output based on these samples. Boxplots are used here to illustrate the possibilities of simulations: 100 realisations provide a distribution of equiprobable values per location. A brief explanation:

- **blue** line = median value of distribution (identical to \( q_{0.5} \) quantile and \( Q_2 \) quartile).
- **red** box = IQR, the range from the first to the third quartile (\( Q_1, Q_3 \)); containing 50% of the data centred around the median.
- whiskers ‘- - ’ = \( Q_1 - 1.5 \times \text{IQR} \) and \( Q_3 + 1.5 \times \text{IQR} \).
- markers ‘+’ = outliers.

5.3.1. Excavated Volumes

![Boxplot of Excavated Volumes](image)

(a) Volume contained within simulated top and bottom boundaries (a) and indurated zones as fraction of total layer volume (b).

**Figure 5.9**

**Discussion** Considering the layer volume (figure 5.9a), some observations can be made for the sampling strategies:

- The coarse grid results in a large spread of possible values and the highest number of outliers. In comparison, the fine grid exhibits less spread and a close proximity to the original value.
- The medium grid, with a smaller number of samples than the preferential grids, shows the second smallest spread in results, with a similar deviation from the original values as the coarse grid.
- The preferential grids are closer to the original value but also show a relatively large spreading.

The indurations are expressed as a fraction of the total layer volume, displayed by figure 5.9b. Here, a trend is observed as well. The trend is similar to the total layer volume, but only up to the fourth sampling strategy (medium, at 250 m spacing). In case of the 100 m spacing, as indicated by the IQRs in figure 5.9b, the spreading is the least pronounced, but the median value lies closest to the one obtained by the coarse sampling strategy.

This observation, together with the relatively large deviations from the original, is probably a result of the variogram inference combined with the random nature of SIS applied in this particular case: the range of the variogram model for the coarse sampling is the most identical to the original (see figure 5.4), whereas for the other strategies the inferred ranges are smaller leading to differently sized lenses of hard material across the deposit. Regarding the probability of occurrence: as described in section 4.2.3, the probability for encountering indurations is 6.4% for the original data and is similar for the samples taken from this dataset.
5.3.2. Production Targets
Below the results are shown, based on the investigation presented in section 4.4.2: a comparison between both sampling strategies and production targets, at a fixed probability of 80%. The 2D plot shows a horizontal slice through the 3D grid, indicating zones where the production target is either missed or reached and zones where reaching the target is uncertain.

Figure 5.10: Two dimensional slice of 3D grid containing discrete production probabilities; using a production target of 600 m$^3$/h.

The 2D slice shown in figure 5.10 is taken from a simulated 3D grid (100 realisations), based on the 500m + 50m (rectangle) sampling grid; in this particular case there are not many regions where missing the target is certain. More examples with varying sampling strategy are shown in figure 5.11, the slices for the remaining production targets are shown in appendix D. Aside from the discrete plots, “continuous” versions are shown: the production target is the same, with the colour now indicating the probability of reaching it (varying between 0 and 1).

The plots shown in the centre of figure 5.11 are based on the virtual dataset, used as input for sampling (all five plots are identical, shown for comparison). Here it is observed that the certainty of zones missing or reaching the target, shown on the right, are consistent with the original dataset.

The bar graphs shown in figure 5.12a are based on the complete 3D grid, of which the plot in figure 5.10 is a 2D slice with the colours indicating the same cases. The percentage indicates the proportion of blocks allocated to one of those three cases.
5.3. Comparing the Sampling Strategies

| Production Target: 600 m³/h - Depth: -4.75 m |
| Sampling Strategy: 100m |
| 80% Below | Uncertain | 80% Above |
| Case |

| Production Target: 600 m³/h - Depth: -4.75 m |
| Sampling Strategy: 250m |
| 80% Below | Uncertain | 80% Above |
| Case |

| Production Target: 600 m³/h - Depth: -4.75 m |
| Sampling Strategy: 500m + 50m (rectangle) |
| 80% Below | Uncertain | 80% Above |
| Case |

| Production Target: 600 m³/h - Depth: -4.75 m |
| Sampling Strategy: 500m |
| 80% Below | Uncertain | 80% Above |
| Case |

Figure 5.11: Two dimensional slices; from left to right: discrete production target probability, the original deposit (where either the target is missed or reached) and continuous probability, all at 600 m³/h target and a depth of -4.75 m. From top to bottom: varying the sampling strategy.
Discussion  In all cases of the production targets it can be seen that the number of uncertain blocks (yellow) decreases from left to right, varying with sampling strategy. Another observation, considering the production targets, is that the number of “certain blocks” (in red, indicating 80% probability or higher of reaching the target), increases with a decreasing production target. Conversely, the same applies to the blocks in blue and an increasing production target, respectively.

Note that the sample locations are visible in the slice maps, either coloured blue or red, because they are known and thus by definition either the target is missed or reached. Judging by the continuous plots, it is clear how the definition of uncertainty (below 80% in this case) may affect the resulting discrete plots. Already, in the discrete plots, “red” is more predominantly present than “blue”. When lowering the probability for defining the uncertainty of reaching the target, this might skew the ratio further. This effect may be mitigated by increasing the sampling density (e.g. more “blue” is visible in the fine grid) or changing the locations for preferred sampling (now they are all located in predominantly “red” zones).

Considering the discussion following the RMSE results (section 5.2.2), a similar observation can be made here: although the preferential sampling strategies contain more samples compared to the medium sized grid, the resulting uncertainty still lies between the coarse and medium grids.
5.3.3. Uncertainty Reduction

In this section the uncertainty reduction compared to the number of samples per sampling strategy is shown, based on the uncertainty indicated by the figures in section 5.3.2: the uncertainty is defined as the proportion of blocks marked yellow in the bar graphs.

![Uncertainty Reduction and Sample Spacing](image)

Figure 5.13: Effect of number of samples on uncertainty reduction; using a production target of 600 m$^3$/h (a), 900 m$^3$/h (b) and 1200 m$^3$/h (c).

**Discussion** The graphs depicted by figure 5.13a confirm the observations made in the previous sections: in general a higher sampling density leads to a decrease in simulation uncertainty. However, a high density, compared to a lower one, does not necessarily lead to a lower uncertainty, as this depends also on the spatial arrangement of samples. This means that the decision for a preferential strategy should be approached carefully; possibly by starting with a relatively coarse grid and, based on the results, choose the locations and associated sample spacing for the preferred zones accordingly.
5.3.4. Monthly Production and Associated Costs

Figure 5.14: Volume excavated over complete life of mine. Note that values on the y-axis are multiplied by a factor of $10^7$.

Figure 5.15: Volume excavated during the first month of mine operation versus volumes based on production targets.
Discussion  As explained in section 5.3.2 the preferential sample zones do not seem to have a substantial effect on production target uncertainty reduction. However, when considering the volume excavated, different observations are made:

• **Total volume** Over the full life of mine, the estimated volumes produced are closer to the true value (original) in case of preferential sampling.

• **First month** When only considering the first month of production, a similar trend is observed; with the difference that coarse and medium sampling grids result in underestimated volumes instead of overestimations. This is probably related to the location, as in the first month only a relatively small part is mined compared to the total volume. In this case the dredged volume resulting from a 600 m³/h target is almost identical to the original value. This means that, depending on the chosen target, preferential sampling can in some cases lead to a higher forecasting precision.

• **Costs** The costs are calculated for the first month as well; based on power requirement, maintenance and operation. The preferential samples lead to a decreased spreading compared to the coarse grid and are closer to the original as the medium grid: here the effect of preferential sampling is visible, contrary to previous observations. This is most likely caused by the way of calculating the production costs, which is based on costs per kWh, cutter power requirement based on soil and rock strength, effect of indurations and the variability in excavated layer volume; the observed trend is similar to the life of mine production, although the preferential sampling has a more pronounced effect on cost forecasting accuracy here.
Conclusions and Recommendations

In this chapter the conclusions of this research project are presented, following from the results and discussions described in chapter 5. The main conclusion is that the combination of geotechnical site investigation and geostatistical simulation is a useful tool for quantifying the uncertainty based on sampling density and assessing the project risks involved in wet mining and dredging operations. As mentioned in chapter 2 a 2D approach towards this goal already existed. The methodology presented in this thesis, however, can be used to generate 3D models, containing both geotechnical properties and 3D structural information like geotechnical units and sub units.

6.1. General Conclusions

A methodology is proposed in this document for the assessment of sampling uncertainty. The spatial uncertainty is quantified in different ways, based on the source data, relating to the sampled geotechnical parameters and further propagation through transfer functions. Based on the number and distribution of initial samples it was decided to generate a virtual exhaustive dataset upon which to base further sampling and processing steps. A set of sampling strategies is determined, serving as input for different geostatistical simulations which are performed to generate 100 realisations for each location in the study area and each of the required parameters.

The framework can be used for determining whether a sampling scheme is appropriate, based on the number of samples, the reduction of uncertainty and excavation costs: the choice for a certain sampling strategy is a balance between the costs of exploration and the project risk resulting from this sampling. For example, a low sampling density may initially not come with high costs but in the long term pose higher risks as production forecasts at unsampled locations are less accurate and thus result in increased project risk.

Within the scope of this thesis the KPIs are defined as cutter head production and the associated costs. These are calculated from the simulation output, based on the sampling strategies, and thus result in various distributions, instead of single values, allowing for project risk analysis.

If an optimal sampling density is to be determined, this would depend on how the optimum is defined, which is explained as follows: considering the reduction in simulation error (RMSE), decrease in uncertainty based on production targets, the number of samples and the production costs, the following observations are made: the 250 m spacing leads to a larger RMSE decrease than the preferred zones, the uncertainty reduction is higher with a lower number of samples compared to preferential sampling and the number of samples increases rapidly when moving from 250 m to a spacing of 100 m. However, considering the production forecasts, the accuracy is comparable to preferred sampling, whereas the precision is lower. Also, the small scale variogram reproduction is better in case of preferred sampling. Therefore, only if the RMSE, the uncertainty related to production targets and the number of samples
are considered the most important aspects, then the 250 m (or "medium") sampling strategy can be identified as the optimum. In that case the optimal spacing is half the variogram range, depending on which property is investigated (variograms shown in section 5.1), which is consistent with the findings by Li et al. (2016).

6.2. Specific Conclusions
The conclusions from the observations (as shown in chapter 5), may be subdivided into several categories:

Conclusions regarding the process implementation and used methods:

- The method used for modelling layers is based on simulating the depths of the boundaries between layers. In some of the realisations this led to very thin layers, close to overlapping.
- Based on the literature review and observations during data processing it can be concluded that the CPT provides data in a better quality compared to SPT.

Intermediate conclusions, based on results obtained during the simulation process:

- An increased sampling density leads to a decrease in ergodic fluctuation and better variogram reproduction at small lag distances.
- A change (either increase or decrease) in sampling density leads to different correlation ranges observed, and thus spatial structures found.

Direct conclusions, based on results generated by the simulations:

- Sample data is honoured at sample locations, which can be seen in figure 5.7, where the spreading at sample locations is the lowest.
- The overall RMSE over the complete study area decreases with increasing sampling density.
- In the results it was observed that, when not taking into account the spatial structure of the sampling scheme, an increase in absolute number of samples (i.e. additional information) does not necessarily cause a decrease in prediction error.

Indirect conclusions, based on results from processing:

- The accuracy and precision of predicted excavated volumes increase with sampling density, also taking into account clustered samples.
- The uncertainty based on production targets decreases with increasing sampling density. However, with increasing available information the number of blocks not meeting target production increases as well.
- Similar to the production volume calculation, the production in the first month shows an increased accuracy and precision in forecasting with higher sampling density and the use of preferred sampling. When also considering excavation costs the trend is similar, with an increased accuracy which can be attributed to clustered samples.

6.3. Practical Implications of the Observed Results
Aside from the scientific contribution and the academic relevance of the information presented in this document, it is useful to know in which way the results affect the current practice in dredging and (wet) mining operations. As explained in chapter 1, the geotechnical investigation is not always given similar priority compared to geological and grade information, although its effect on assessing excavatability and therefore production performance could be equally important.

The sampling strategy definition, and thereby the number and spatial arrangement of samples, is part of the feasibility stage executed prior to the actual exploitation or initiation of the project. Using the methodology proposed here it is possible to optimise the sampling density, which can be performed in an iterative sense: first the area could be sampled in a coarse grid, then potential zones of interest are defined which can subsequently be sampled in a higher density to decrease prediction uncertainty and/or improve the spatial correlation at certain distances (depending on the desired scale of structures to be found).
Regarding the equipment selection, the framework could be used for investigating the effect of, for example, the combination of indurated zones and soft material on predicting production performance as well as the influence of varying vessel parameters, e.g. cutter power requirement, cutter head diameter and dredging depth. Furthermore, if more elaborate financial forecasts are included, this approach would lead to a probabilistic method describing the uncertainty resulting from sampling density, while also propagating this through (financial) transfer functions for assessing both project and financial risk.

Within the scope of this thesis, for example, it can be seen that the propagation of this uncertainty into forecasted KPIs is not always as intuitively expected. What this means is that depending on the property investigated, the trend shown by the spreading of forecasts varies with the choice of transfer function: in general more samples lead to a decrease in forecasting uncertainty, but in some cases the adoption of additional sampling in preferred zones shows a different result. This was observed clearly in the cost estimation, where the preferred sampling lead to more accurate results. However, when only considering the direct simulation output of SPT $N$ values, this was not as obvious (the simulation error, in this case RMSE, was even higher in case of preferred sampling); whereas combining the SPT $N$ values with other simulated parameters for obtaining cost estimates resulted in a different trend (as shown in figure 5.16), showing the possible benefits of an iterative sampling approach with additional drilling in preferred zones.

### 6.4. Recommendations for Further Research

**Modelling Subsurface Structures**  In order to avoid the issue of very thin or overlapping layers, it is recommended to change this part of the simulation procedure: model the first boundary occurring between layers and then, instead, model the thickness for each of the following layers. Another improvement, regarding the SIS simulation of lenses and thus applying to the indurations here, is proposed by Deutsch (2006). As the SIS algorithm may result in small, geologically unrealistic, structures arising from short-scale variability it could be useful to perform "clean-up" operations, mitigating the presence of these small artifacts. Different levels of cleaning are proposed, each increasing in intensity. Important to note is that these cleaning operations are not part of SGeMS; although they are compatible with the GSLIB file format, which is used during this project.

**Induration Thickness**  As described in section 4, it was chosen to model the indurated zones using a cell size larger than the expected layer thickness, based on required simulation time and expected vertical versus horizontal extent of a layer. In case a more realistic model is needed, it is recommended to choose a different (smaller) cell size. Depending on available CPU power this may increase calculation time but could, additionally, also yield more detailed and realistic results.

**Accounting for Secondary Information**  When performing site investigations, often multiple methods are required to assess different properties of the subsurface. For example, when performing an SPT this results in the $N$ value (blow count) representative of the test location. However, a disturbed soil sample is also obtained by this test, depending on the barrel, which could provide additional information like a grain size distribution. In geostatistics a variety of methods exists for dealing with this kind of information. An application of such methods could be useful when secondary information is sampled in a higher density, compared to the primary data (Goovaerts, 1997). A requirement of the secondary data is that it must be correlated to the primary data, where the degree of correlation is one of the factors affecting the weight assigned to the secondary data. Different variants of kriging exist to account for secondary information, such as cokriging. A more advanced approach which could be useful when dealing with spatial uncertainty is the use of conditional simulation. For example, SGS may be adapted to make use of both primary and secondary data. Kriging is then used, as part of SGS, to estimate mean and variance of the local cdf, whereas with secondary data one could employ cokriging. When using this approach, semivariograms are required for both primary and secondary data, as well as a cross semivariogram describing the correlation between them (Goovaerts, 1997).
Conclusions and Recommendations

CPT Based Transfer Functions  Comparing CPT to SPT: CPT is closer to a point value, it provides a higher vertical density of samples, the procedures are more standardised and the execution is less prone to human error and it is more easily automated. Because of these benefits and assuming that the SPT and CPT are used in similar soil types, it is advised to investigate or further develop transfer functions and correlations based on CPT data.

Sampling Grids  Within the scope of this thesis, five different sampling grids were chosen to define which locations are sampled to obtain input values for subsequent simulation and data processing. To study the effect of preferred samples in more detail, it is recommended to compare more sampling strategies: ideally in such a way that the preferential grids contain an equal number of samples as one of the regular grids used, thereby isolating the effect of zones containing clustered samples.

Optimal Sampling Density  In the conclusion an indication of the optimal sampling density is given. However, due to time limitations, a more detailed investigation was not performed. It is therefore recommended to implement such a procedure based on, for example, multiple-criteria decision analysis.

Bayesian Updating  Depending on the application of a geotechnical model, the adoption of a Bayesian updating technique may prove beneficial. For example, when considering a mining operation, the geotechnical model is part of reserve definition (see figure 1.1): new measurements may be taken throughout the life of the mine as the excavation progresses. These measurements could provide valuable insights and may be used to update the initial subsurface model. An example of such an updating technique is Kalman Filtering, which can be applied in a sequential manner allowing for the continuous updating as new information is obtained, while weighing the new measurement against the initial model, taking measurement error and uncertainty into account as well.

Simulation Software  Over the course of this project, some limitations were found: error messages, handling of point vs. block variance and process automation. The error messages and crashes from SGeMS appeared to be random at first, since no description was given. After more use it became clear that certain variogram inference parameters (e.g. a lag spacing too small to contain sample locations) led to such errors and that certain operations require administrator rights for the software to write multiple realisations to the hard drive. Not being able to handle measurements with different support (e.g. blocks and points) is a disadvantage here, since this is often an issue in mining related projects. Therefore, it is recommended to either adapt the software (it is open source), develop new software or choose alternative (commercial) software, if this issue arises.

The difficulty of automation is another drawback. This has been partly overcome by creating Python scripts which are able to run multiple simulations sequentially; however these could still benefit from more optimisation: one example is converting GSLIB files to SGeMS binary files, which was tested and allows for quicker and automated importing of input files, but is not fully implemented in the scripts yet.
Exploratory Data Analysis

Before getting involved with processing the data using estimation and simulation software, it is consid-
ered good practice to perform an EDA (Exploratory Data Analysis). As the name implies, this process
is meant to obtain an overview of the data and get familiar with its properties and their distribution.
An important first step is to look for errors and missing data. Using Python in combination with Pandas,
or Excel, the following common errors can be found and corrected:

- **Missing data**: compare number of entries in column. For example, in CPT data, ensure that every
depth value is accompanied by a cone tip resistance value.
- **Data type**: float, integer and string values. When dealing with SPT data, string and float values
may be present in the same column, e.g.: number of blow count (float) and 'Refusal' (string).
- **Outliers**: compare minimum, maximum and average values.

Visualising the data is also a useful tool to get insight in the data. When plotting the data in two
dimensions (or even three), using colour-coding, a large amount of data is easily inspected: extreme
values, spatial distribution and structures may be assessed.

Another thing to consider is the source or measurement type of the data. For example:

- **Point**: a small localized specimen, or a measurement with a small support
- **Linear**: small sample, similar to point, but with one very long dimension. Examples: drill cores,
cuttings.
- **Planar**: Regular array of small samples, e.g. from a rock face.

If the data has sources of different origin, the volume representing the sample (data support) is
likely to vary as well. In this case, the data cannot be treated as a single population and will have to
be corrected. This process, called compositing, will result in a uniform support size for all samples and
may be performed by (weighted) averaging. When determining compositing size the boundaries of the
geological domain, the type of deposit and mine specific parameters (bench height, smallest selective
mining unit) should be taken into account.

An example is CPT versus SPT: aside from the difficulty of correlating them, their sample data support
is also significantly different: a CPT is measured more or less continuously at intervals of 1 or 2 cm,
whereas an SPT is performed over approximately 30 cm (1 ft).

**A.1. Error Check**

Before starting to work with the data it is useful to know whether or not errors and gaps are present.
Common errors could include: missing data, extreme values, points where commas are expected as
decimal separators, unexpected characters in numerical values. To obtain an overview of the data,
first the data needs to be imported. This can be done by opening the Excel file in Pandas and adding
it to a new dataframe. Then, several functions are available to inspect the database:
• Show the data types per column, e.g. float, boolean, integer, character string (object).
• The number of values in each column.
• List the unique values per column.
• A description per column showing mean, average, standard deviation and extreme values.

For example, these functions could be used to find all the unique borehole ID’s, then filter the data for a certain borehole and obtain a description of this borehole. The same can be done for another column or data value, e.g. a certain geotechnical unit. When errors are encountered they should be mitigated, resulting in a new dataset, the corrected data, serving as input for the EDA.

A.2. Univariate Statistics

Frequency distribution A common way to depict the frequency distribution of a certain measurement is a histogram. In case of a continuous variable it shows the relative frequency of values within a certain data range (bin or class).

When the proportion of data values below a certain threshold is required, a cumulative frequency distribution can be used.

Central Value Often the arithmetic mean \(m\) or \(\mu\) is taken as the central value of a distribution (Goovaerts, 1997):

\[
m = \frac{1}{n} \sum_{a=1}^{n} z(\alpha)
\]  

(A.1)

However, depending on the presence of outliers, alternatives like the median and mode may be used. Considering the frequency distribution: the median \(M\) is the value with a corresponding cumulative frequency of 0.5 and the mode \(D\) is the value with the highest frequency of the distribution.

Measure of Spread The variance is a measure of spread around the mean of a distribution and is defined as:

\[
\sigma^2 = \frac{1}{n} \sum_{a=1}^{n} (z(\alpha) - m)^2
\]  

(A.2)

The square root of the variance is the standard deviation, \(\sigma\). The ratio between the standard deviation and the mean is called the coefficient of variation:

\[
CV = \frac{\sigma^2}{m}
\]  

(A.3)

Skew and Kurtosis Aside from the mean and variance, higher order moments are sometimes used in statistics. The third order moment describes the skewness of a distribution: a distribution may be skewed left or right, in comparison to a normal distribution. Skewness is defined as follows, with \(z(\alpha)\) being a random variable:

\[
\gamma_1 = E\left\{\left(\frac{z(\alpha) - m}{\sigma}\right)^3\right\}
\]  

(A.4)

The kurtosis is a fourth order moment, also describing the shape of distribution. It is related to the "thickness" of the tails; again relative to a normal distribution which has a kurtosis of 3.

\[
\kappa = \frac{E((z(\alpha) - m)^4)}{(E((z(\alpha) - m)^2))^2}
\]  

(A.5)
A.3. Bivariate Statistics

Covariance  The covariance is a measure of the joint variation of two variables around their respective means:

\[
\sigma_{ij} = \frac{1}{n} \sum_{a=1}^{n} (z_i(a) - m_i) \cdot (z_j(a) - m_j)
\]  (A.6)

Interpretation of the covariance depends on the measurement scales of the data values. An independent measure of covariance is the correlation coefficient \( \rho \):

\[
\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} \quad \in [-1, 1]
\]  (A.7)
B.1. Normal Score Transformation

The process of data transformation in SGeMS is shown in appendix B.2. A transformation which is often performed is the normal score transformation, which results in a Gaussian data distribution. The dataset is sorted and ranked, subsequently an equivalent rank from the Gaussian distribution is found for each rank in the data. The transformed dataset consists of the normally distributed values associated to these ranks.

![Histogram and cdf before (a) and after (b) transformation.](image)

Figure B.1: Histogram and cdf before (a) and after (b) transformation.
B.2. 2D Sequential Gaussian Simulation

In this appendix the process of performing a 2D Sequential Gaussian Simulation, using SGeMS, will be described. The data is assumed to be available and prepared for use as simulation input, meaning that the data is filtered and/or cleaned of errors and the data location coordinates are in an appropriate format.

Running SGeMS from Microsoft Windows  When using Windows as operating system and in case the software needs write access to the hard drive, it is necessary to run SGeMS "as Administrator" (see figure B.2).

An example is when a simulation outputs more than one realisation: in this case, if SGeMS has no administrator rights, it will show an error message and stop running.

Figure B.2: Running software with administrator privileges.
Importing Data  The next step is importing the data. As shown in figure B.3, this is done by navigating to "Load Object" from the "Objects" drop-down menu. Then a file, e.g. in GSLIB or CSV format, can be loaded and a type may be selected (either point set, cartesian grid or masked grid).

Figure B.3: Importing data files.
**Viewing Data** When the data is imported, a list of the properties is shown in the centre "Objects" window (see figure B.4). The black area on the right shows the distribution of sample locations, in 2D or 3D, coloured according to the range of property values. When right-clicking a property in the centre window, or by navigating to the "Data Analysis" drop-down menu, a histogram can be generated along with other useful statistics.

![Figure B.4: Data visualisation.](image)

**Data Transformation** Any of the available simulation or estimation algorithms may require a certain distribution for the input data (e.g., Gaussian). If the data is not conforming to this distribution it is necessary to transform it. This can be done by using the built-in function "trans", for histogram transformation (accessed from the top left "Algorithms" window.

The three successive tabs in the function’s window are shown in figure B.5. When the input distribution is chosen to be "Non-parametric", the back-transformed data (after simulation), will be distributed exactly like the input data. Another important aspect is lower and upper tail extrapolation: the inter- or extrapolation of the cdf outside the input data limits. Setting these values too high or too low may result in unrealistic results.

An example when the "trans" function is used: the sGs algorithm depends on a variogram model in the normal score domain. This means that the data may have to be transformed before performing the variography.

**Variogram Modelling** Aside from the input data, another requirement for estimation and simulation methods is a representation of the spatial correlation, e.g., in the form of a variogram, correlogram or covariance.

The menu is accessed by navigating to the "Data Analysis" drop-down menu (see figure B.6). First, the head and tail properties are selected. Next, the parameters like lag distance and orientation should be entered. Important to note is that these should be appropriate for the sample spacing: a single lag should be represented by sufficient data. If for example the lag spacing is too small, resulting in no data found, SGeMS might crash without showing a specific error message.

The next window shows the experimental variogram along with an interactive menu for variogram modelling (see figure B.7). Several structures may be combined, with possibly different angles to account
for anisotropy. When right-clicking in the plotting window, the number of samples per lag are shown which may help in the modelling process. When the model is finished, it can be saved to a file for later use.

**Executing a Simulation Algorithm** A variety of simulation algorithms is available from the “Simulation” sub-menu in the “Algorithm” window, as shown in figure B.8. In this example, the Sequential Gaussian Simulation is selected. The algorithm consists of a menu with three tabs and is started by clicking “Run Algorithm”. Various parameters may be entered manually or loaded from parameter files. For example, the variogram model from the previous step can be loaded from the saved file.

Two important aspects: the search ellipsoid and the "Use Target Histogram" option. The search ellipsoid defines the area from which (conditioning) data are taken into account. If the angles are not entered, the ellipsoid definition remains at the default setting: "Maximum" and "Minimum" indicate the horizontal ranges, while "Minimum" represents the vertical range. Since the simulation in this case is performed in 2D, the minimum range is left blank. Selecting "Use Target Histogram", means that the input data is transformed to Gaussian both in a forwards and backwards manner, resulting in a simulation output in its original (non-parametric or otherwise) distribution. The process is similar to step 4.

Another requirement for running a simulation is the grid in which the simulated values are placed.
For this purpose a new cartesian grid may be created, by navigating to the "Objects" drop-down menu and selecting "New Cartesian Grid". Next, the properties like cell dimensions and spacing should be defined (see figure B.9). In this case the data is transformed in such a way that the origin is located at (0,0,0) for ease of use.
B.2. 2D Sequential Gaussian Simulation

Figure B.7: Variogram modelling; graphical user interface.

Figure B.8: Sequential Gaussian Simulation.
Figure B.9: Cartesian grid definition.
Simulation Results  When the simulation is finished the various realisations are added to simulation grid, shown in the centre window. The results may be shown in the visualisation window on the right, along with a colourbar (see figure B.10).

Figure B.10: 2D simulation visualisation.
C

Variograms per Sampling Strategy

C.1. Experimental Variogram Parameters
Table C.1: Variogram parameters for indurations; different sampling strategies (spacings in brackets).

<table>
<thead>
<tr>
<th></th>
<th>Indurations (500m + cross)</th>
<th>Indurations (500m + rectangle)</th>
<th>Indurations (100m)</th>
<th>Indurations (250m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>144</td>
<td>144</td>
<td>144</td>
<td>144</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>250.5</td>
<td>250.5</td>
<td>100</td>
<td>50.5</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>100.1</td>
<td>100.1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Number of directions</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Azimuth</td>
<td>90°</td>
<td>90°</td>
<td>90°</td>
<td>90°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>Angular Tolerance</td>
<td>5°</td>
<td>5°</td>
<td>5°</td>
<td>5°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Note: Table C.1: Variogram parameters for indurations; different sampling strategies (spacings in brackets).
Table C.2: Variogram parameters for SPT values; different sampling strategies ( spacings in brackets).

<table>
<thead>
<tr>
<th>Variogram</th>
<th>SPT (500m + cross)</th>
<th>SPT (500m + rectangle)</th>
<th>SPT (500m)</th>
<th>SPT (250m)</th>
<th>SPT (100m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Horizontal</td>
<td>Vertical</td>
<td>Horizontal</td>
<td>Vertical</td>
<td>Horizontal</td>
</tr>
<tr>
<td>Number of lags</td>
<td>144</td>
<td>82</td>
<td>72</td>
<td>82</td>
<td>9</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>25</td>
<td>0.5</td>
<td>50</td>
<td>0.5</td>
<td>400</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>10</td>
<td>0.1</td>
<td>20</td>
<td>0.1</td>
<td>160</td>
</tr>
<tr>
<td>Number of directions</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>90°, 180°</td>
<td>180°</td>
<td>90°, 180°</td>
<td>180°</td>
<td>180°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>90°</td>
<td>0°</td>
<td>90°</td>
<td>0°</td>
</tr>
<tr>
<td>Angular Tolerance</td>
<td>5°</td>
<td>10</td>
<td>5°</td>
<td>10</td>
<td>180°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>50</td>
<td>1</td>
<td>100</td>
<td>1</td>
<td>160</td>
</tr>
</tbody>
</table>
Table C.3: Variogram parameters for elevations (top of layer); different sampling strategies (spacings in brackets).

<table>
<thead>
<tr>
<th>Variogram</th>
<th>500m + cross</th>
<th>500m + rectangle</th>
<th>500m</th>
<th>250m</th>
<th>100m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>144</td>
<td>72</td>
<td>9</td>
<td>14</td>
<td>36</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>25</td>
<td>50</td>
<td>400</td>
<td>250</td>
<td>100</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>10</td>
<td>20</td>
<td>160</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>Number of directions</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>Angular Tolerance</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>50</td>
<td>100</td>
<td>160</td>
<td>100</td>
<td>40</td>
</tr>
</tbody>
</table>

Table C.4: Variogram parameters for elevations (bottom of layer); different sampling strategies (spacings indicated).

<table>
<thead>
<tr>
<th>Variogram</th>
<th>500m + cross</th>
<th>500m + rectangle</th>
<th>500m</th>
<th>250m</th>
<th>100m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lags</td>
<td>144</td>
<td>72</td>
<td>9</td>
<td>14</td>
<td>36</td>
</tr>
<tr>
<td>Lag Separation</td>
<td>25</td>
<td>50</td>
<td>400</td>
<td>250</td>
<td>100</td>
</tr>
<tr>
<td>Lag Tolerance</td>
<td>10</td>
<td>20</td>
<td>160</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>Number of directions</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Azimuth</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>Dip</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
<tr>
<td>Angular Tolerance</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
<td>180°</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>50</td>
<td>100</td>
<td>160</td>
<td>100</td>
<td>40</td>
</tr>
</tbody>
</table>
C.2. Inferred Variogram Models

Figure C.1: Variogram models; layer top.
Figure C.2: Variogram models, layer bottom.
C.2. Inferred Variogram Models

Figure C.3: Variogram models, indurated layer locations.
Figure C.4: Variogram models, SPT N values.
In this appendix the results are shown from the procedure described in section 4.4.2; a probability of 80% is used here (see figure D.1) and the production target is varied (600 m$^3$/h, 900 m$^3$/h and 1200 m$^3$/h).

From left to right, depth: -4.75 m, 0.25 m and 5.25 m.
From top to bottom, sampling strategy: coarse, cross, rectangle, medium and fine.
Colour indicators:

- **blue** = 80% probability of missing the target.
- **yellow** = uncertain.
- **red** = 80% probability meeting or exceeding target.
D.1. Target at 600 m³/h

Figure D.1: Two dimensional slices of discrete production target probability at 600 m³/h target.
D.2. Target at 900 m$^3$/h

Figure D.2: Two dimensional slices of discrete production target probability at 900 m$^3$/h target.
D.3. Target at 1200 m³/h

Figure D.3: Two dimensional slices of discrete production target probability at 1200 m³/h target.
### Nomenclature

#### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTS</td>
<td>Brazilian Tensile Strength.</td>
<td>[MPa]</td>
</tr>
<tr>
<td>CPT</td>
<td>Cone Penetration Test.</td>
<td></td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit.</td>
<td></td>
</tr>
<tr>
<td>CSV</td>
<td>Comma Separated Values.</td>
<td></td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory Data Analysis.</td>
<td></td>
</tr>
<tr>
<td>GSLIB</td>
<td>Geostatistical Software Library.</td>
<td></td>
</tr>
<tr>
<td>IDW</td>
<td>Inverse Distance Weighting.</td>
<td></td>
</tr>
<tr>
<td>IK</td>
<td>Indicator Kriging.</td>
<td></td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile Range.</td>
<td></td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator.</td>
<td></td>
</tr>
<tr>
<td>LAS</td>
<td>Local Average Subdivision.</td>
<td></td>
</tr>
<tr>
<td>MPS</td>
<td>Multiple Point Statistics.</td>
<td></td>
</tr>
<tr>
<td>OK</td>
<td>Ordinary Kriging.</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>Random Function.</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error.</td>
<td></td>
</tr>
<tr>
<td>RV</td>
<td>Random Variable.</td>
<td></td>
</tr>
<tr>
<td>SGeMS</td>
<td>Stanford Geostatistical Modeling Software.</td>
<td></td>
</tr>
<tr>
<td>SGS</td>
<td>Sequential Gaussian Simulation.</td>
<td></td>
</tr>
<tr>
<td>SIS</td>
<td>Sequential Indicator Simulation.</td>
<td></td>
</tr>
<tr>
<td>SK</td>
<td>Simple Kriging.</td>
<td></td>
</tr>
<tr>
<td>SMU</td>
<td>Smallest Minable Unit.</td>
<td></td>
</tr>
<tr>
<td>SPT</td>
<td>Standard Penetration Test.</td>
<td></td>
</tr>
<tr>
<td>TB</td>
<td>Turning Bands.</td>
<td></td>
</tr>
<tr>
<td>UCS</td>
<td>Unconfined Compressive Strength.</td>
<td>[MPa]</td>
</tr>
<tr>
<td>UTM</td>
<td>Universal Transverse Mercator.</td>
<td></td>
</tr>
<tr>
<td>VED</td>
<td>Virtual Exhaustive Dataset.</td>
<td></td>
</tr>
<tr>
<td>WGS</td>
<td>World Geodetic System.</td>
<td></td>
</tr>
</tbody>
</table>

#### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>exhaustive</td>
<td>Well known; as in an exhaustively sampled, and therefore completely known, area.</td>
</tr>
<tr>
<td>experimental variogram</td>
<td>Variogram calculation based only on sample data.</td>
</tr>
<tr>
<td>kriging variance</td>
<td>Variance resulting from spatial sample distribution and chosen model; not related to actual sample data.</td>
</tr>
<tr>
<td>P10</td>
<td>The 10% quantile ($q_{0.1}$) of a distribution.</td>
</tr>
<tr>
<td>P50</td>
<td>The 50% quantile ($q_{0.5}$) or median of a distribution.</td>
</tr>
<tr>
<td>P90</td>
<td>The 90% quantile ($q_{0.9}$) of a distribution.</td>
</tr>
<tr>
<td>Q2</td>
<td>The second quartile of a distribution, identical to the median and $q_{0.5}$</td>
</tr>
<tr>
<td>realisation</td>
<td>One of the equiprobable, possible, outcomes of an RV.</td>
</tr>
<tr>
<td>sampling strategy</td>
<td>Spatial representation of sampling: either an evenly spaced or preferential sampling grid.</td>
</tr>
</tbody>
</table>
## Greek Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ</td>
<td>Semivariance; depending on ( u ).</td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>Skewness of a distribution.</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>Kurtosis of a distribution.</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>Weight applied to sample values in geostatistical estimation.</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>Standard deviation.</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>Variance.</td>
</tr>
</tbody>
</table>

## Latin Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( h )</td>
<td>Lag distance vector.</td>
<td>[m]</td>
</tr>
<tr>
<td>( u )</td>
<td>Location vector.</td>
<td>[m]</td>
</tr>
<tr>
<td>Cov</td>
<td>Covariance operator.</td>
<td></td>
</tr>
<tr>
<td>( d_{50} )</td>
<td>Median grain size of grain size distribution.</td>
<td></td>
</tr>
<tr>
<td>( E )</td>
<td>Expectation operator.</td>
<td></td>
</tr>
<tr>
<td>( m )</td>
<td>Mean.</td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>Number of hammer blows from an SPT.</td>
<td></td>
</tr>
<tr>
<td>( Q )</td>
<td>Production.</td>
<td>([m^3/h])</td>
</tr>
<tr>
<td>( q_c )</td>
<td>Cone tip resistance from a CPT.</td>
<td>([MPa])</td>
</tr>
<tr>
<td>( q_p )</td>
<td>The p-quantile of a distribution</td>
<td></td>
</tr>
<tr>
<td>( R )</td>
<td>Residual component of a random function.</td>
<td></td>
</tr>
<tr>
<td>( Var )</td>
<td>Variance operator.</td>
<td></td>
</tr>
<tr>
<td>( Z )</td>
<td>Random variable.</td>
<td></td>
</tr>
<tr>
<td>( z )</td>
<td>Realisation of a random variable.</td>
<td></td>
</tr>
<tr>
<td>( Z^* )</td>
<td>Estimator.</td>
<td></td>
</tr>
</tbody>
</table>

## Subscripts

<table>
<thead>
<tr>
<th>Subscript</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>Sample location.</td>
</tr>
<tr>
<td>( \beta )</td>
<td>Secondary sample location.</td>
</tr>
</tbody>
</table>


