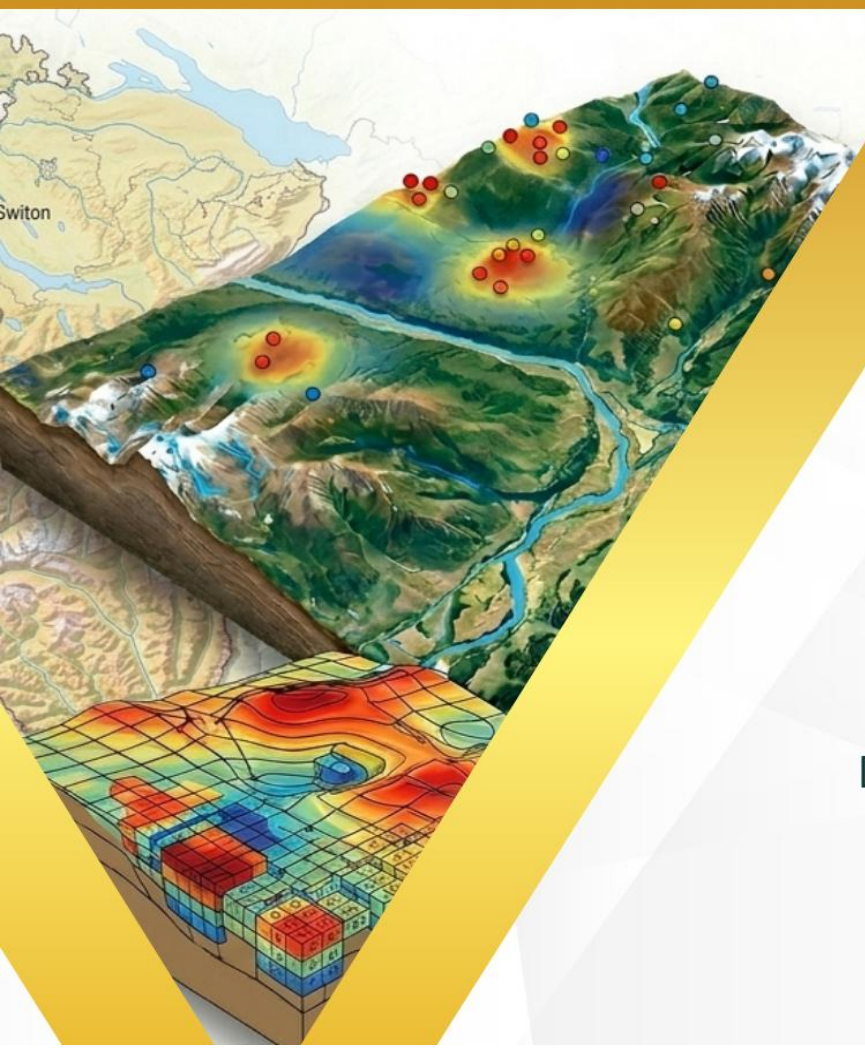
Ministry of Higher Education and Scientific Research
Djillali Liabes University of Sidi Bel Abbes
Faculty of Technology
Department of Civil Engineering and Public Works
Option: Geotechnical Engineering
Level: First Year Master's

GEOSTATISTICS

COURSE & EXERCISES



2026

Prepared by
Dr.E. **MOSTEFA KARA**

GEOSTATISTICS

The Geostatistics course is delivered in the second semester of the Master 1 program in Geotechnical Engineering and forms part of the Fundamental Teaching Unit (UEF 1.2.2) within the national harmonized curriculum framework. The teaching unit carries 10 credits with a coefficient of 5, while the Geostatistics course itself is assigned 4 credits and a coefficient of 2. The course consists of 1 hour and 30 minutes of lectures and 1 hour and 30 minutes of tutorials per week, totaling 45 contact hours and an overall student workload of 55 hours. Assessment is based on 40% continuous evaluation and 60% final examination.

This course aims to provide students with the theoretical foundations and methodological tools necessary for analyzing and modeling the spatial variability of geotechnical parameters.

It develops a solid understanding of regionalized variables, spatial dependence structures, and variographic analysis, and ensures proficiency in kriging-based estimation methods.

Upon completion, students will be able to quantify soil variability, model and interpret variograms, perform spatial estimation, assess associated uncertainty, and critically use specialized geostatistical software in geotechnical engineering applications.

Dr. Esma MOSTEFA KARA

Associate Professor, Lecturer and Researcher

Department of Civil Engineering and Public Works
Faculty of Technology
Djillali Liabes University, Sidi Bel Abbès 22000, Algeria
Civil Engineering and Environment Laboratory (L.G.C.E.)
email : mkesma@yahoo.fr/esma.sekkal@univ-sba.dz

2026

GEOSTATISTICS

Esma MOSTEFA KARA

2026

Preface

Training in geotechnical engineering confronts engineers with a fundamental reality: soil is a natural material, inherently variable, whose properties can only be known through limited sampling. This variability, often perceived as a constraint, is in fact an essential characteristic of the geological environment. The rational consideration of this heterogeneity now represents a major challenge for the reliability and safety of engineering structures.

Long approached through deterministic methods, geotechnical engineering is progressively evolving toward approaches that explicitly incorporate uncertainty and spatial variability. In this context, geostatistics emerges as a structuring scientific tool, providing a coherent probabilistic framework that integrates observation, modelling, and estimation.

This course manual has been developed within this perspective. It aims to support Master 1 students in Geotechnical Engineering in understanding both the theoretical foundations and practical applications of geostatistics. The objective is not merely to present computational techniques, but to cultivate an analytical approach that enables the interpretation of spatially distributed phenomena and the management of their engineering implications.

The adopted approach follows a logical progression: from the statistical foundations necessary for understanding regionalized variables, to variographic analysis, then to the theory of kriging, and finally to the use of modern computational tools. Each stage is designed to combine mathematical rigor with physical interpretation, thereby preventing any disconnect between theory and practice.

This document is aligned with the new national harmonized curriculum framework and contributes to the evolution of pedagogical practices toward an engineering approach grounded in uncertainty quantification and risk management. It represents a step in the training of engineers capable of addressing contemporary geotechnical challenges with discernment, methodological rigor, and critical thinking.

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

A. Statistical and probabilistic symbols

- X, Y: random variables
- x, y: realizations (observations) of random variables
- n: number of observations (sample size)
- $E[\cdot]$: expectation operator
- μ_X: mean of X(population mean)
- $\hat{\mu}_X$: sample mean of X
- σ_X^2: variance of X
- $\hat{\sigma}_X^2$: sample variance
- σ_X: standard deviation of X
- $\hat{\sigma}_X$: sample standard deviation
- $\delta_X$or COV: coefficient of variation of X
- ψ_X: skewness coefficient
- $Cov(X, Y)$: covariance between $X$and Y
- ρ_{XY}: correlation coefficient between $X$and Y
- $p_X(x)$: probability mass function (PMF)
- $f_X(x)$: probability density function (PDF)
- $F_X(x)$: cumulative distribution function (CDF)
- $\Phi(\cdot)$: standard normal CDF
- Z: standardized normal variable, $Z = \frac{X-\mu_X}{\sigma_X}$

B. Spatial statistics and geostatistics

- s, x: spatial location (position vector)
- h: separation (lag) vector
- $\| h \|$: lag distance (magnitude of h)

- $Z(s)$: random function (random field)
- $z(s)$: realization of $Z(s)$(regionalized variable)
- m: mean of the random function (assumed constant under stationarity)
- $C(h)$: covariance function
- $\rho(h) = \frac{C(h)}{C(0)}$: autocorrelation function
- $\gamma(h)$: variogram (semivariogram)
- C_0: nugget effect
- a: range (correlation range)
- C: structured (spatial) variance contribution
- $C_0 + C$: sill (total variance for a variogram with sill)
- $N(h)$: number of pairs in lag class h

C. Kriging (estimation)

- $Z^*(x_0)$: kriging estimate at location x_0
- x_0: estimation (target) location
- λ_i: kriging weights
- μ: Lagrange multiplier (ordinary kriging constraint)
- $\sigma_k^2(x_0)$: kriging variance (estimation variance)
- KS: simple kriging
- OK: ordinary kriging
- UK: universal kriging
- CoK: cokriging

D. Reliability and uncertainty (when used)

- p_F: probability of failure
- $g(X)$: limit state function
- $FORM$: First-Order Reliability Method
- MC: Monte Carlo simulation
- σ_c: calibration error standard deviation (empirical correlation)

E. Geotechnical symbols (common parameters used in the manual)

- c': effective cohesion
- φ': effective friction angle
- C_u: undrained shear strength
- γ: unit weight (or bulk unit weight; check your convention)
- ρ: density (if used explicitly)
- k: permeability
- m_v: coefficient of volume compressibility
- C_c: compression index
- e: void ratio
- $w$or w_L: water content / liquid limit (as applicable)
- D_r: relative density
- SPT: Standard Penetration Test
- Δp: stress increment
- p_0: initial effective stress

F. General abbreviations

- EDA: Exploratory Data Analysis
- GIS: Geographic Information System
- PDF: Probability Density Function
- PMF: Probability Mass Function
- CDF: Cumulative Distribution Function

General Introduction

Geotechnical engineering is distinguished by the specific nature of its material of interest: soil. Unlike manufactured materials, soil is a natural geomaterial formed through complex geological processes and characterized by intrinsic heterogeneity. Its physical and mechanical properties exhibit significant spatial variability, making exhaustive characterization at the scale of a site impracticable (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

The design of geotechnical structures therefore relies on point-based information obtained from limited site investigation programs. This inevitably introduces uncertainty into the parameters selected for design. Traditionally, such uncertainty has been addressed through deterministic approaches incorporating global safety factors. However, the evolution of engineering practice and the need to optimize investigation efforts have led to the adoption of probabilistic approaches that explicitly quantify variability and its impact on structural reliability (Baecher & Christian, 2003).

Within this context, geostatistics provides a theoretical framework particularly well suited to the analysis of spatially distributed phenomena. Originating from Matheron's foundational work (Matheron G. , 1971), it is based on the formalism of random functions and enables the modelling of spatial dependence in regionalized variables. Variographic analysis represents the central step in this approach, as it characterizes the spatial correlation structure of the studied phenomenon (Chile's & Delfiner, 2012).

One of the major contributions of geostatistics is kriging theory, an optimal estimation method that explicitly incorporates the variogram model and provides both a point prediction and a quantitative measure of the associated uncertainty (Cressie N. A., 1993). This capability to combine estimation with error quantification is essential in a discipline where risk control is critical.

Recent developments in spatial statistics and spatio-temporal modelling have further expanded the scope of geostatistics, notably through Bayesian approaches and hierarchical models (Cressie & Wikle, 2011). In geotechnical engineering, these advances allow soil variability to be incorporated more effectively into reliability analyses and risk management.

Prepared in accordance with the new national harmonized curriculum framework and intended for Master 1 students in Geotechnical Engineering (Semester 2), this course manual aims to provide the scientific foundations required to understand and apply geostatistical methods. It follows a structured progression from statistical fundamentals to practical applications, through variographic analysis and kriging theory.

Beyond learning computational tools, this document seeks to develop a rigorous analytical approach, enabling students to treat the natural variability of soils as an intrinsic component of the engineering problem rather than a constraint to be merely compensated. In this way, geostatistics is fully aligned with a modern geotechnical engineering perspective grounded in explicit uncertainty quantification and rational technical decision-making.

Chapter 1

Theoretical Basis of Geostatistics

Theoretical Basis of Geostatistics

Geostatistical analysis is built upon a conceptual framework grounded in probability theory and mathematical statistics. Before addressing spatial dependence modelling and estimation techniques, it is essential to master the fundamental notions related to random variables and random functions.

This first chapter introduces the theoretical foundations required to understand the spatial variability of geotechnical parameters. It establishes the mathematical framework within which soil properties can be interpreted as realizations of a regionalized variable characterized by a spatial dependence structure.

The objective is to provide students with the key conceptual tools needed to rigorously approach subsequent developments devoted to variographic analysis and kriging theory. This chapter therefore constitutes the scientific basis of the overall geostatistical methodology presented in this course manual.

1.1 Introduction

Geostatistics originates from the engineering sciences. Initially developed to address mineral resource estimation problems, it has, over recent decades, found applications in many fields (Chile's & Delfiner, 2012) (Wackernagel, 2003).

The design of a geotechnical structure first requires a sound understanding of the foundation soil. The initial step of any project is therefore the geotechnical site investigation program, which aims to determine representative values of the soil properties needed for design calculations.

However, selecting these representative values is a delicate task. Geotechnical parameters measured in situ exhibit spatial dispersion, which leads to uncertainty in the representative quantities used in design. Indeed, it is impossible to determine soil properties at every point of a site; consequently, representative parameters are generally derived from a limited number of laboratory tests on specimens sampled in a quasi-random manner and from in situ tests performed on a more or less dense investigation grid.

The physical and mechanical information obtained from such investigations is used to define representative design parameters; these values remain globally uncertain, which in turn generates uncertainty in predicting the subsequent behaviour of structures.

A new approach in soil mechanics therefore became necessary in order to address, at least more effectively, these issues of soil heterogeneity and behavioural modelling. Statistical and probabilistic methods have thus emerged to complement classical deterministic approaches.

The application of statistical and probabilistic analysis techniques aims to quantify the influence of the natural variability of soils on structural performance. Geostatistics provides a set of tools specifically oriented toward the estimation of these parameters.

Matheron (1971) defined geostatistics as follows: “geostatistics is the application of the formalism of random functions to the investigation and estimation of natural phenomena.”

The purpose of this chapter is to introduce the theoretical foundations required to understand and implement geostatistical methods. It successively addresses the following topics:

- discrete and continuous random variables;
- descriptive and quantitative analysis of variability;
- concepts of correlation and dependence;
- definitions of random functions and regionalized variables;
- stationarity assumptions;
- covariance and autocorrelation functions;
- The fitting of theoretical structure models.

These foundations provide the essential conceptual basis for variogram analysis and optimal estimation techniques, particularly kriging, which will be developed in the subsequent chapters.

1.2 Random Variables

Random variables $(X)$are variables whose values are not known with certainty. A probability is associated with the event that a random variable takes a given value. Random variables may be either discrete or continuous.

Consider first the case of a discrete random variable, which takes values from a finite or countable set of possible outcomes. For example, the number of potential slip circles in a slope stability problem may be modelled as a discrete random variable. The event $N = 3$corresponds to the occurrence of three slip circles. The associated set of probabilities is called the probability mass function (pmf), defined as $p_N(n) = P(N = n)$, for $n = 0,1,2, \ldots$.

Continuous random variables, on the other hand, correspond to measured parameters obtained either in the laboratory or in situ and affected by measurement uncertainty and other sources of variability, such as water content, shear strength, and permeability. When a random variable is continuous, probabilities are derived from a continuous function called the probability density function (pdf), denoted $f_X(x)$. Unlike the pmf, the pdf is not a probability; it must be multiplied by a length element to yield a probability, which explains the use of the term density.

Thus, for continuous random variables, probabilities are associated with intervals. The probability that a continuous random variable lies between $x$and $x + dx$is $f_X(x)\,dx$. For example, the shear strength of a soil may lie between 90 and 100 kPa. The probability that a specimen has a shear strength of exactly 98.00 kPa is exceedingly small. Since $f_X(x)$varies with x, the probability that Xlies between $a$and $b$is obtained by integrating the density over the interval (Ang & Tang, 2007).

$$\boldsymbol{P[a < X \leq b] = \int_a^b f_X(x)dx} \qquad \textit{(1.1)}$$

1.3 Graphical Analysis of Variability

Five methods can be used to analyse variability: histograms, frequency plots, frequency density plots, cumulative frequency plots, and scatter plots.

Variability often leads to uncertainty. For instance, the unit weight of a soil at a given location is generally unknown unless it has been measured at that specific point. This uncertainty arises because unit weight varies from one location to another within the soil mass.

As an example, unit weight measurements from a borehole are reported in Table 1.1. These boreholes were drilled offshore in the Gulf of Mexico at the site of an oil platform project. The soil consists of normally consolidated clay throughout the investigated depth. The unit weight varies with depth from 1521.75 to 2002.31 kg/m³. (Ang & Tang, 2007) (Baecher & Christian, 2003).

Table 1.1 Soil density values obtained from boreholes drilled in the Gulf of Mexico (Baecher & Christian, 2003)

Raw data					Sorted Data	
Borehole ID	**Depth (m)**	**Soil Density, x (kg/m³)**	**$(x-\mu)^2$**	**$(x-\mu)^3$**	**Depth (m)**	**Soil Density kg/m³**
1	0,15	1681,9	1857,6	-80063,0	51,60	1521,8
2	0,3	1906,2	32833,4	5949419,3	2,25	1537,8
3	0,45	1874,2	22260,6	3321287,5	1,50	1585,8
4	1,5	1585,8	19376,6	-2697228,3	6,60	1585,8
5	1,95	1617,9	11470,4	-1228480,9	13,50	1585,8
6	2,25	1537,8	35043,8	-6560206,8	30,60	1585,8
7	4,95	1826,1	10221,2	1033364,3	5,70	1601,8
8	5,7	1601,8	15178,2	-1869959,2	8,25	1601,8
9	6,6	1585,8	19376,6	-2697228,3	11,25	1601,8
10	7,5	1633,9	8299,2	-756058,0	15,00	1601,8
11	8,25	1601,8	15178,2	-1869959,2	24,45	1601,8
12	9,3	1617,9	11470,4	-1228480,9	36,45	1601,8
13	10,35	1617,9	11470,4	-1228480,9	1,95	1617,9
14	11,25	1601,8	15178,2	-1869959,2	9,30	1617,9

15	12	1617,9	11470,4	-1228480,9	10,35	1617,9
16	13,5	1585,8	19376,6	-2697228,3	12,00	1617,9
17	15	1601,8	15178,2	-1869959,2	18,60	1617,9
18	18,15	1649,9	5640,0	-423564,8	36,60	1617,9
19	18,6	1617,9	11470,4	-1228480,9	39,60	1617,9
20	21,45	1698,0	729,0	-19683,0	7,50	1633,9
21	21,6	1746,0	441,0	9261,0	27,45	1633,9
22	24,45	1601,8	15178,2	-1869959,2	33,60	1633,9
23	24,6	1665,9	3492,8	-206425,1	45,75	1633,9
24	27,45	1633,9	8299,2	-756058,0	18,15	1649,9
25	30,45	1698,0	729,0	-19683,0	24,60	1665,9
26	30,6	1585,8	19376,6	-2697228,3	42,75	1665,9
27	33,6	1633,9	8299,2	-756058,0	96,60	1665,9
28	36,45	1601,8	15178,2	-1869959,2	0,15	1681,9
29	36,6	1617,9	11470,4	-1228480,9	48,60	1681,9
30	39,6	1617,9	11470,4	-1228480,9	21,45	1698,0
31	42,75	1665,9	3492,8	-206425,1	30,45	1698,0
32	45,75	1633,9	8299,2	-756058,0	81,60	1698,0
33	48,6	1681,9	1857,6	-80063,0	60,45	1714,0
34	51,6	1521,8	41290,2	-8390176,8	84,45	1730,0
35	57,45	1858,1	17715,6	2357947,7	21,60	1746,0
36	60,45	1714,0	121,0	-1331,0	75,45	1746,0
37	63,45	1794,1	4774,8	329939,4	81,45	1746,0
38	72,45	1826,1	10221,2	1033364,3	78,45	1762,0
39	75,45	1746,0	441,0	9261,0	87,60	1778,0
40	78,45	1762,0	1369,0	50653,0	63,45	1794,1
41	81,45	1746,0	441,0	9261,0	93,45	1794,1
42	81,6	1698,0	729,0	-19683,0	102,45	1794,1
43	84,45	1730,0	25,0	125,0	123,45	1794,1
44	87,6	1778,0	2809,0	148877,0	129,60	1794,1
45	90,45	2002,3	76895,3	21323063,9	99,45	1810,1
46	93,45	1794,1	4774,8	329939,4	102,60	1810,1
47	96,6	1665,9	3492,8	-206425,1	4,95	1826,1
48	99,45	1810,1	7242,0	616295,1	72,45	1826,1
49	102,45	1794,1	4774,8	329939,4	111,45	1826,1
50	102,6	1810,1	7242,0	616295,1	117,45	1826,1
51	105,6	1858,1	17715,6	2357947,7	120,60	1826,1
52	108,45	1986,3	68277,7	17840960,4	114,45	1842,1
53	108,6	1874,2	22260,6	3321287,5	117,60	1842,1
54	111,45	1826,1	10221,2	1033364,3	123,60	1842,1
55	114,45	1842,1	13712,4	1605723,2	126,45	1842,1

56	117,45	1826,1	10221,2	1033364,3	132,60	1842,1
57	117,6	1842,1	13712,4	1605723,2	57,45	1858,1
58	120,6	1826,1	10221,2	1033364,3	105,60	1858,1
59	123,45	1794,1	4774,8	329939,4	0,45	1874,2
60	123,6	1842,1	13712,4	1605723,2	108,60	1874,2
61	126,45	1842,1	13712,4	1605723,2	0,30	1906,2
62	129,6	1794,1	4774,8	329939,4	135,45	1906,2
63	132,6	1842,1	13712,4	1605723,2	108,45	1986,3
64	135,45	1906,2	32833,4	5949419,3	90,45	2002,3
	∑	110399,2	834886,0	28880499,8		
	μ	1725,0				

1.3.1 Histogram

A histogram is constructed by dividing the data range into classes (bins). The minimum number of classes can be estimated using two empirical methods, including:

1. Sturges' rule:

$$N = 1 + (3.3 log n) \tag{1.2}$$

where ndenotes the sample size.

2. Yule's rule:

$$N = 2.5\sqrt[4]{n} \tag{1.3}$$

The class width for each interval

$$I = \frac{(X_{max} - X_{min})}{N} \tag{1.4}$$

The count for category $X_i$corresponds to the number of observations in that class, denoted n_i.

The relative frequency f_i(in %) is defined as:

$$f_i = \frac{Number\ of\ occurrences\ n_i}{\sum n_i} \tag{1.5}$$

The frequency density $f_{i\ cum}$ is defined as:

$$f_{i\ cum} = \frac{relative\ frequency}{(X_{max} - X_{min})} \tag{1.6}$$

cumulative frequencies $f_{i+1\ c}$ are defined as:

$$f_{i+1\ c} = \left(\frac{f_{i+1\ cum}}{\sum f_{icum}}\right) + f_{i\ c} \tag{1.7}$$

The data used in this example are adapted from (Ang & Tang, 2007) (Table 1.2).

The number of classes is $N = 7$according to both empirical rules; however, we adopt $N = 10$. The class width is therefore $I = (2100 - 1450)/10 = 65$, i.e., $I = 65$. Table 1.2 summarizes the class analysis using 10 classes, a class width of 65, a minimum value of 1450, and a maximum value of 2100. For instance, there are no values in the interval 1450-1515 kg/m³ (Table 1.1), two values in the interval 1515-1580 kg/m³, and so on.

A bar chart showing the number of occurrences in each interval is called a histogram. The histogram of soil density for the present example is shown in Figure 1.2.

A histogram provides important information on the variability of a dataset. It highlights the data range, the most frequent values, and the degree of dispersion around the mean. Several considerations should be taken into account when selecting the number of bins. First, the number of bins should generally depend on the sample size: as the number of observations increases, the number of bins should also increase. Second, the choice of binning affects the interpretation of the data: using too few or too many bins may obscure the distribution and its dispersion. Unfortunately, there is no universally accepted rule for selecting the optimal number of bins; exploring different bin widths is therefore a practical approach.

Table 1.2 Frequency analysis

Class	Min	Max	Count	Relative frequency (%)	Frequency Density (% kg/m³)	Cumulative Frequency (%)
1	1450	1515	0	0	0,00	0
2	1515	1580	2	3	0,05	3
3	1580	1645	21	33	0,50	36
4	1645	1710	9	14	0,22	50
5	1710	1775	6	9	0,14	59
6	1775	1840	13	20	0,31	80
7	1840	1905	9	14	0,22	94
8	1905	1970	2	3	0,05	97
9	1970	2035	2	3	0,05	100
10	2035	2100	0	0	0,00	100
			64	100	1,54	

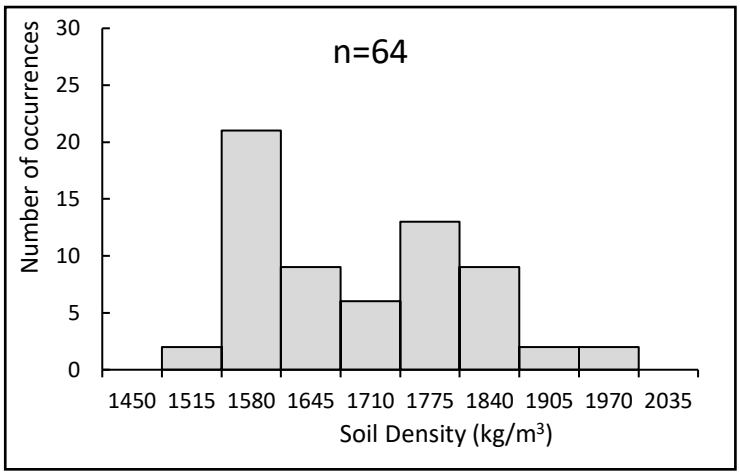


Figure 1.1. *Number of occurrence plot of Soil density*

1.3.2 Frequency Plot

The relative frequency in each histogram bin is obtained by dividing the number of occurrences in that bin by the total number of data points. A bar chart displaying the relative frequency for each bin is called a frequency plot. The bin frequencies for the soil density dataset are computed in Table 1.2, and the resulting frequency plot is shown in Figure 1.3.

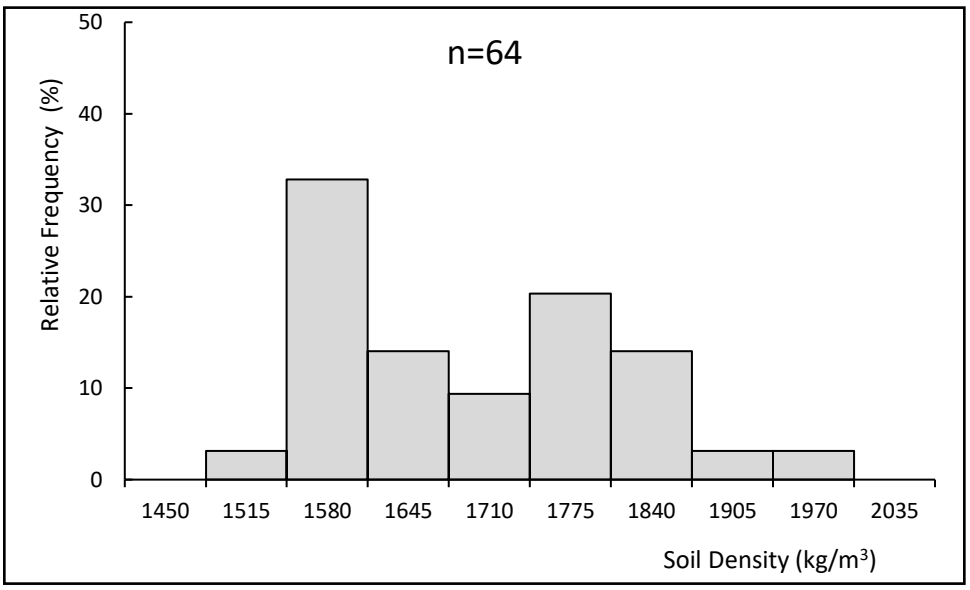


Figure 1.2. *Frequency plot of soil density*

Note that the histogram and the frequency plot have the same shape and convey the same information. The frequency plot is simply a normalized version of the histogram. Because it is normalised, the frequency plot is useful for comparing different datasets. Frequency plots are illustrated in Figures 1.3 to 1.6. Figure 1.3 presents the soil density data, which vary spatially.

Figure 1.4 provides an example of data that vary over time. The dataset corresponds to the average monthly pumping rate measured as a function of time for a leak detection system at a hazardous waste disposal facility. The values vary from month to month due to changes in leachate generation rates and waste placement. The plot shows the frequency distribution of the flow rate in m³/h.

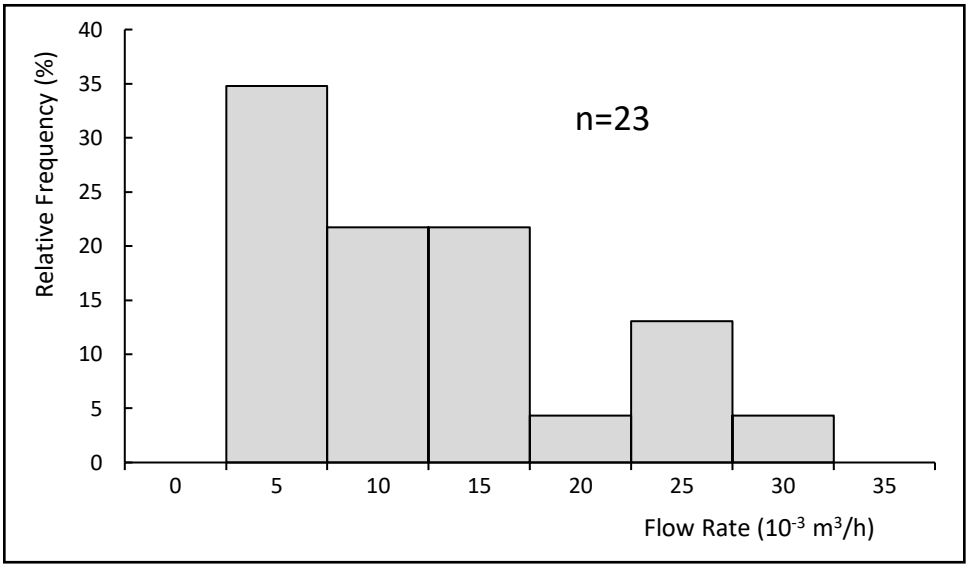


Figure 1.3. *Flow rate plot (m^3/h)*

Figure 1.5 provides an example of data that vary across construction projects. The data represent the ratio of actual cost to estimated cost for the remediation of Superfund sites (contaminated environments). These ratios vary from site to site due to differences in site conditions, weather, contractors, technology, and regulatory constraints. Note that the majority of projects exhibit cost ratios greater than 1.0 (Baecher & Christian, 2003).

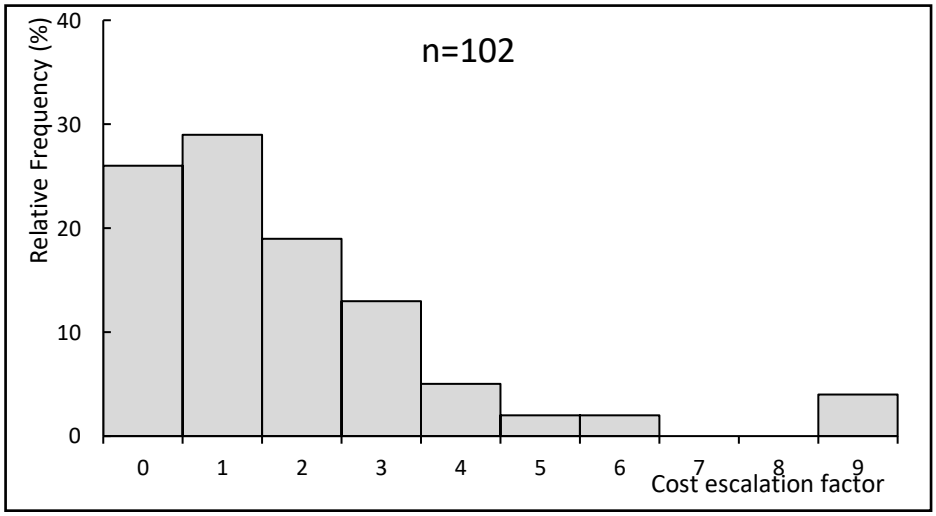


Figure 1.4. *Relative frequency plot of the cost escalation factor.*

Figure 1.6 presents an example of data that vary across geotechnical testing laboratories. The data correspond to friction angles measured on bulk Ottawa sand specimens. Although Ottawa sand is considered a relatively uniform material and only minor variations in specimen density were observed, the test results exhibit noticeable variability. Most of this variability is attributed to differences in testing equipment and procedures among the laboratories. (Baecher & Christian, 2003).

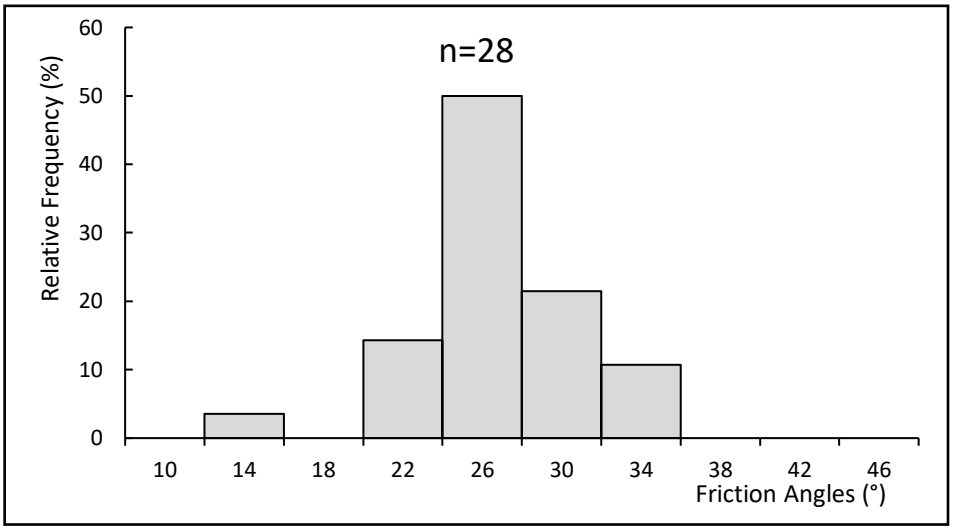


Figure 1.5. *Relative frequency of the friction angles*

1.3.3 Frequency Density Plot

Another plot related to the histogram is the frequency density plot. Frequency density is obtained by dividing the bin frequencies by the corresponding bin widths. A bar chart of frequency density is referred to as a frequency density curve (or frequency density plot). The purpose of dividing the frequency by the bin width is to further normalize the histogram: the area under the frequency density curve (obtained by multiplying each bar height by its width) is equal to 100%. This normalization is useful when fitting theoretical random-variable models to the data. The frequency densities for the soil density dataset are computed in Table 1.2. Note that the unit of frequency density is percent per unit of the measured variable, i.e., % per kg/m³ in the case of soil density. The resulting frequency density curve is shown in Figure 1.7.

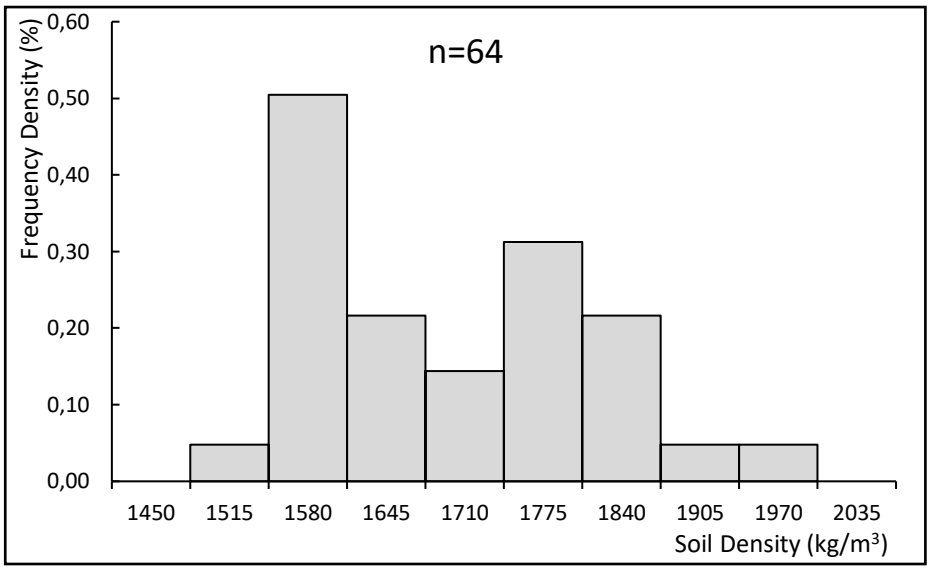


Figure 1.6. *Frequency density of the soil density*

1.3.4 Cumulative Frequency Plot

The cumulative frequency plot is a graphical tool used to analyse variability. The cumulative frequency is defined as the frequency of data points whose values are less than or equal to the upper bound of a given class interval in the frequency plot. It is obtained by summing (accumulating) the bin frequencies for all intervals up to that upper bound. A plot of cumulative frequency versus the upper-class limit is called a cumulative frequency curve.

The cumulative frequencies for the soil density data are computed in Table 1.2. For example, the cumulative frequency at an upper bound of 1645 kg/m³ is $0\% + 3\% + 33\% = 36\%$. The resulting cumulative frequency curve is shown in Figure 1.8. The percentile of a dataset corresponds to the value associated with a given cumulative frequency. For instance, the 50th percentile of the soil density dataset is 1710 kg/m³ (i.e., 50% of the values are $\leq$ 1710kg/m³), whereas the 90th percentile is 1874 kg/m³ (Figure 1.8).

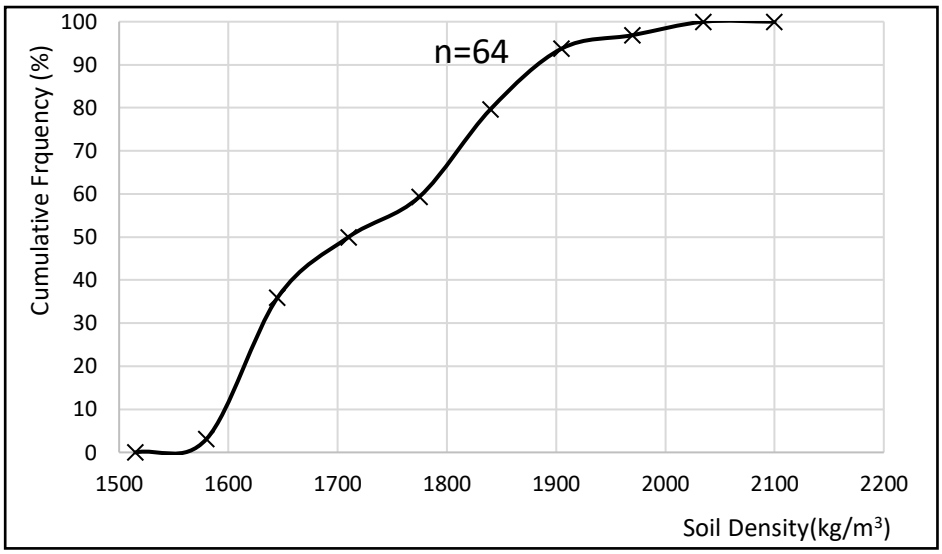


Figure 1.7. *Cumulative frequency plot*

1.4 Data Transformation

In some cases, it is useful to transform the data before plotting them. One example is a dataset of permeability values measured in a compacted clay layer (Baecher & Christian, 2003). The frequency plot for these data is shown in Figure 1.9. It provides limited insight because the permeability values span several orders of magnitude. A more informative representation is obtained by plotting the frequency distribution of the logarithm of permeability, as shown in Figure 1.10. From this plot, it can be observed that the most likely range lies between $10^{-8.4}$and $10^{-8.2}$cm/s, and that most of the data are less than or equal to 10^{-7}cm/s.

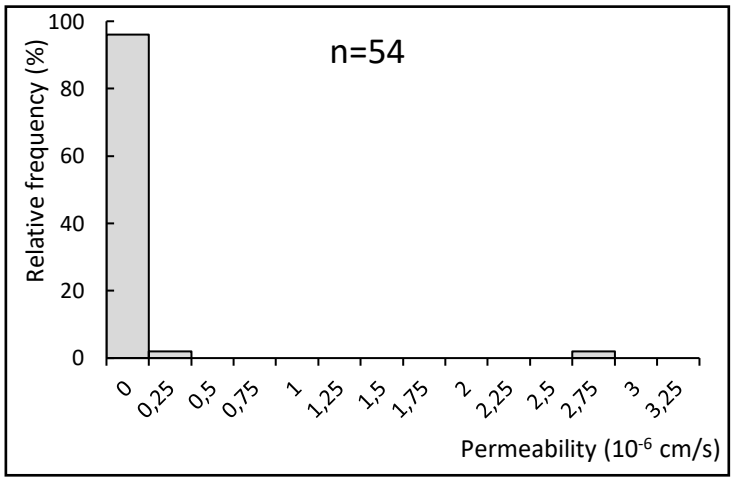


Figure 1.8. *Relative frequency of the permeability*

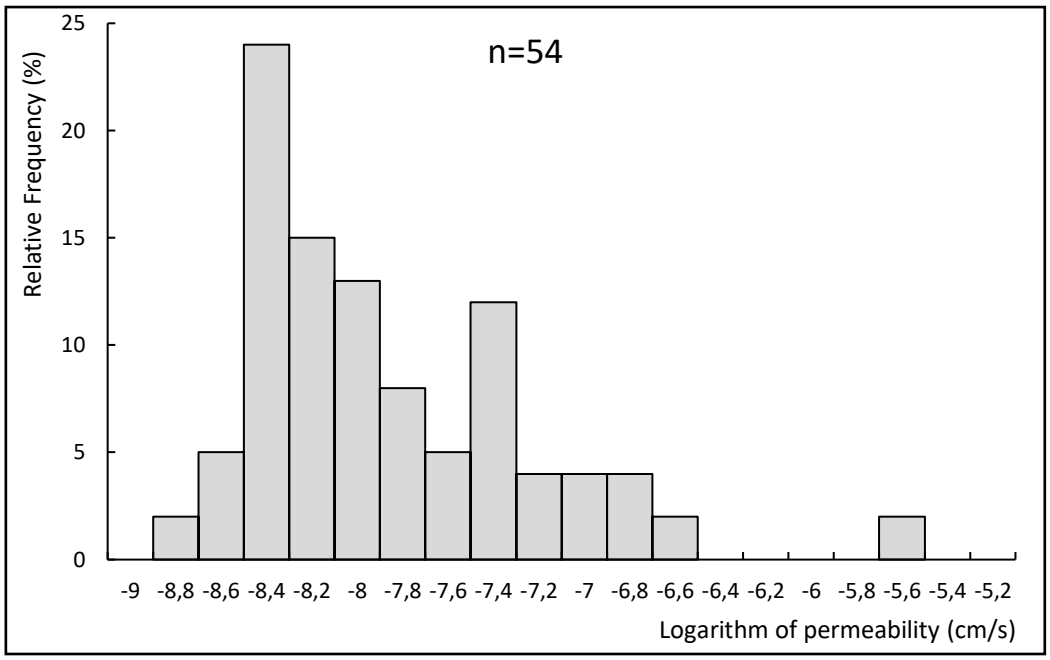


Figure 1.9. *Relative frequency of the Logarithm of permeability*

A second example where a transformation is useful concerns undrained shear strength data for a normally consolidated clay (Baecher & Christian, 2003). A frequency plot of these borehole measurements from an offshore site in the Gulf of Mexico is shown in Figure 1.11.

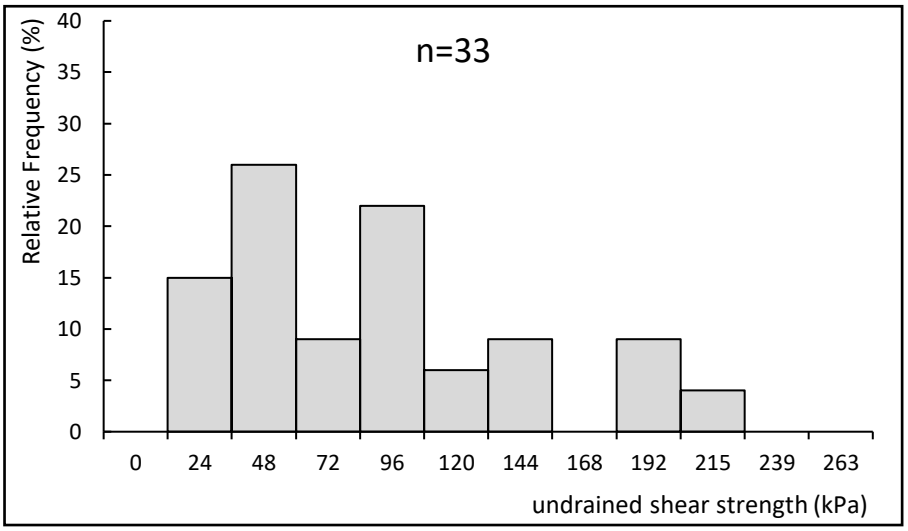


Figure 1.10. *Relative frequency plot of the undrained shear strength*

The data exhibit substantial variability with depth, ranging from 24 to 240 kPa. However, this frequency plot is misleading because a large part of the variability can be attributed to the increase in shear strength with depth.

To illustrate this trend, a scatter plot of undrained shear strength versus depth is presented in Figure 1.12. A more informative measure is obtained by normalizing the undrained shear strength by depth, as shown in Figure 1.13. This scatter plot indicates that the depth-related trend has been removed and that the variability of the normalized undrained shear strength is much lower than that of the raw undrained shear strength. A frequency plot of the undrained shear strength-to-depth ratio is shown in Figure 1.14.

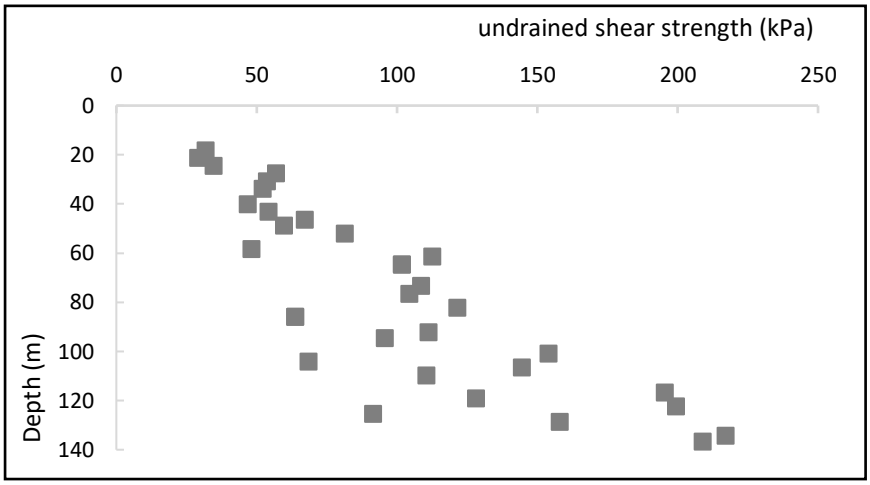


Figure 1.11. *Variability of undrained shear strength as a function of depth*

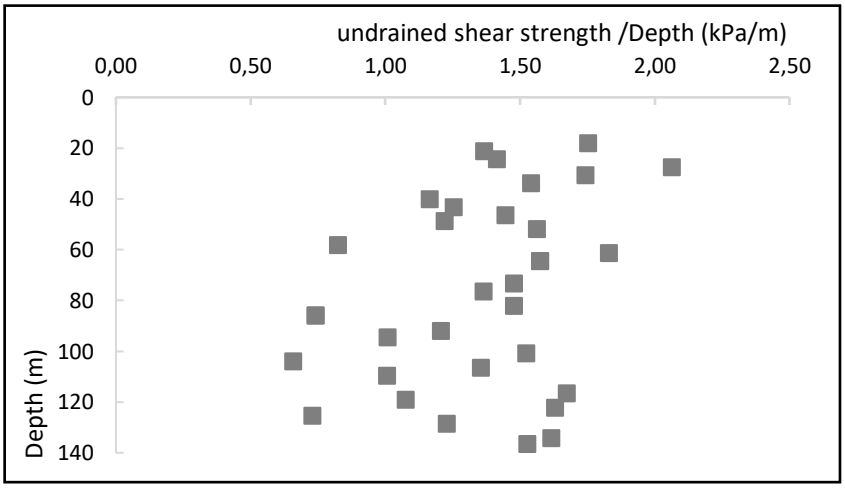


Figure 1.12. *Variation of the undrained shear strength-to-depth ratio as a function of depth*

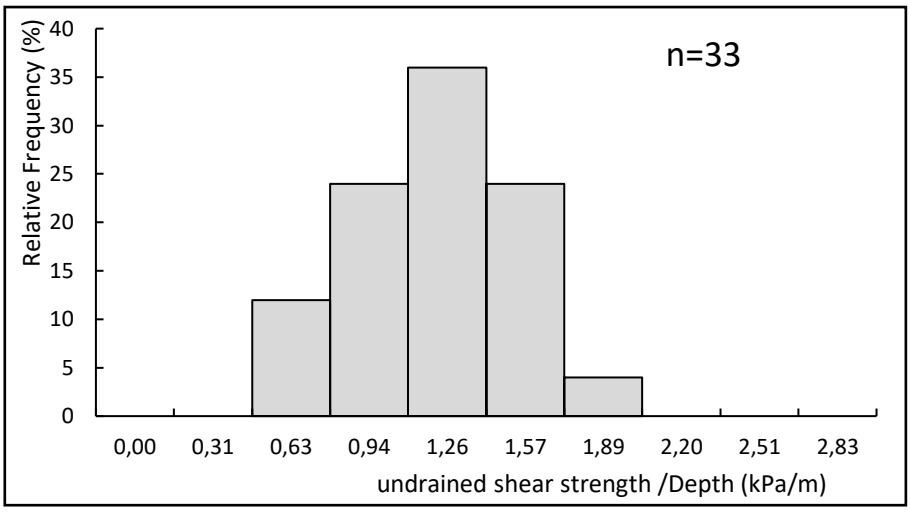


Figure 1.13. *Relative frequency of undrained shear strength/Depth*

1.5 Quantitative Analysis of Variability

In addition to graphical analyses, the variability of a dataset can also be examined quantitatively. Descriptive statistics computed from a dataset (also referred to as sample statistics, since the dataset represents a sample) provide numerical measures of variability. Key features of interest include measures of central tendency, dispersion, data skewness, and the correlation or dependence between data points. The most commonly used statistical measures are presented in this section.

1.5.1 Central Tendency

The most common measure of central tendency in a dataset is the mean (also called the sample mean), The sample mean is computed as follows:

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.8)$$

where: $\hat{\mu}_x$s the sample mean, $x_i$denotes the i-th data value, and $n$is the total number of observations (data points).

For example, the sample mean of the soil density dataset in Table 1.1 is given by $110399.2/64 = 1725$ kg/m^3.

The sample median and the sample mode are other measures of central tendency for a dataset. The median of a sample corresponds to the 50th percentile (i.e., 50% of the cumulative frequencies), whereas the sample mode is the most likely value (the most frequently occurring class).

For instance, the sample median for the soil density dataset is 1710kg/m^3 (Figure 1.8), while the sample mode is 1612,5kg/m^3 (Figure 1.3). The sample mode depends on the bin width used in the frequency plot, and a dataset may have more than one mode.

$$\boldsymbol{sample\ mode = lower\ bound\ of\ the\ modal\ class\ (Cp) + (^{1}/_{2} \times class\ width)} \qquad (1.9)$$

The sample means, medians, and modes for the datasets described above are summarized in Table 1.3. Note that the mean, median, and mode coincide only when the data distribution (frequency plot) is symmetric and unimodal (i.e., it has a single peak) (Baecher & Christian, 2003).

Table 1.3 Statistical analysis of the studied variables

Data	Figure	Mean	Median	Mode	Variance σ^2	Standard Deviation SD	COV	Skewness coefficient ψ
Density (kg/m^3)	Figure 1.2	1725	1710	1612,5	13252,2 (kg/m^3)^2	115,1 (kg/m^3)	0,067	0,31
Flow Rate (10^{-3}m^3/h)	Figure 1.3	15,30(10^{-3} m^3/h)	13,26 (10^{-3} m^3/h)	7,5 (10^{-3} m^3/h)	70,9 (10^{-3} m^3/h)^2	8,42 (10^{-3} m^3/h)	0,55	0,63
Cost	Figure 1.4	2,22	1,74	1,5	3,72	1,93	0,87	2,1
Friction Angle (°)	Figure 1.5	29 (°)	29 (°)	28 (°)	256 (°)^2	16 (°)	0,55	-0,01
Permeability K (cm/s)	Figure 1.6	8,78 $\times 10^{-8}$	1×10^{-8}	$1,25\times10^{-7}$	1,66464E-13	$4,08\times10^{-7}$	4,7	7,1
Log k	Figure 1.7	-7,81	-8	-8,3	0,375769	-0,613	0,078	1,4
undrained shear strength τ (kPa)	Figure 1.8	99	96	60	2756,25	52,5	0,53	0,81
τ/Depth (kPa/m)	Figure 1.9	1,36	1,4	1,41	0,1089	0,33	0,24	-0,4

1.5.2 Dispersion

Dispersion in a dataset is most readily characterized by the sample range (Baecher & Christian, 2003). The range is simply the maximum value in the dataset minus the minimum value. For the soil density data, the range is 480,5kg/m^3(Table 1.1).

The sample variance is a measure of dispersion about the mean value of the dataset. The sample variance is computed as follows:

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2 \qquad (1.10)$$

Where $\hat{\sigma}_x^2$ is the sample variance. The sample variance is the average of the squared deviations of individual data points from the sample mean, and it is always non-negative. For the soil density data, the sample variance is given by:

$\frac{1}{64-1} * 834886$=13252,2 $(kg/m^3)^2$ (tableau 1.1).

The sample standard deviation, $\hat{\sigma}_x$, is the square root of the sample variance, whereas the sample coefficient of variation, $\hat{\delta}_x$(COV), is defined as the standard deviation divided by the mean value:

$$\hat{\delta}_x = \frac{\hat{\sigma}_x}{\hat{\mu}_x} \qquad (1.11)$$

Since the standard deviation has the same units as the mean, the coefficient of variation (COV) is a dimensionless measure of dispersion. For the soil density data, the sample standard deviation and COV are 115,1 and 0,067, respectively.

The statistical measures of dispersion for the different datasets are summarized in Table 1.3. Note the wide range of COV values, with a minimum of 0,067 and a maximum of 4,7. Note also that accounting for the depth-related trend in undrained shear strength (Figure 1.12) reduces the sample COV from 0,53 to 0,24.

1.5.3 Skewness

Since the sample variance is the average of the squared deviations from the sample mean, data values located at the same distance above and below the mean contribute equally. Consequently, the sample variance does not indicate how the data are distributed around the mean (i.e., whether the distribution is symmetric or skewed). The sample skewness coefficient–essentially the average of the cubed deviations from the sample mean–provides a measure of symmetry (or lack of symmetry) for a dataset (Baecher & Christian, 2003).

The skewness coefficient is a dimensionless measure of asymmetry and is given by:

$$\psi = \left[\frac{n}{(n-1)(n-2)}\right]\frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_x)^3}{\hat{\sigma}_x^3} \qquad (1.12)$$

where ψdenotes the sample skewness coefficient. A skewness coefficient equal to zero indicates that the data are symmetrically distributed about the mean value.

A positive skewness coefficient indicates that the distribution is skewed to the right (toward larger values), whereas a negative skewness coefficient indicates that the distribution is skewed to the left (toward smaller values). For the soil density dataset, the sample skewness coefficient is:

$$\frac{64}{63\times62}\times\left(\frac{28880499.8}{115.1^3}\right)=0.31$$

, indicating a right-skewed distribution toward larger values (Figure 1.3).

The skewness coefficients for the other datasets are summarized in Table 1.3. Most datasets are positively skewed. Note that taking the logarithm of permeability reduces the skewness coefficient from 7.1 to 1.4 (Table 1.3).

1.5.4 Correlation or Dependence

Two variables may be related to one another, as illustrated by a scatter plot such as the one shown in Figure 1.12. The sample correlation coefficient provides a measure of the degree of linear dependence between two variables (Baecher & Christian, 2003). It is defined as:

$$\hat{\rho}_{xy}=\frac{\sum_{i=1}^{n}[(x_i-\hat{\mu}_x)(y_i-\hat{\mu}_y)]}{\sqrt{\sum_{i=1}^{n}(x_i-\hat{\mu}_x)^2\sum_{j=1}^{n}(y_j-\hat{\mu}_y)^2}}=\frac{51416.3}{\sqrt{42751.3\times89055.6}}=0.83$$

where $x_i$and $y_i$are observations of two variables. The sample correlation coefficient ranges from $-1{,}0$ to $+1{,}0$. A value of zero indicates that there is no linear relationship between the two variables. A negative value indicates that one variable tends to decrease as the other increases, whereas a positive value indicates that one variable tends to increase as the other increases.

When $\hat{\rho}_{xy}$is close to 1, the relationship between the two variables is strongly linear. For example, the correlation coefficient between undrained shear strength and depth is computed in Table 1.4. The sample correlation coefficient is $\hat{\rho}_{xy}=0{,}83$. This positive value close to one indicates that undrained shear strength tends to increase approximately linearly with increasing depth (Figure 1.12).

Table 1.4 Correlation analysis between undrained shear strength and depth.

Sample ID	τ (kPa) y	Depth (m) x	$(x-\mu_x)^2$ (m^2)	$(y-\mu_y)^2$ (kPa^2)	$(x-\mu_x)(y-\mu_y)$ (kPa m)
1	29,2	21,3	3069,2	4914,0	3883,5
2	32,0	18,2	3422,3	4529,3	3937,1
3	34,6	24,5	2724,8	4186,1	3377,3
4	47,0	40,1	1339,6	2735,3	1914,2
5	48,2	58,3	338,6	2611,2	940,2
6	52,2	33,9	1831,8	2218,4	2015,9
7	53,6	30,7	2116,0	2088,5	2102,2
8	54,3	43,3	1115,6	2025,0	1503,0
9	57,0	27,6	2410,8	1789,3	2076,9
10	60,0	48,9	772,8	1544,5	1092,5
11	63,7	85,9	84,6	1267,4	-327,5
12	63,7	85,9	84,6	1267,4	-327,5
13	67,0	46,4	918,1	1043,3	978,7
14	68,5	104,1	750,8	948,6	-843,9
15	81,4	52,1	605,2	320,4	440,3
16	91,6	125,4	2371,7	59,3	-375,0
17	95,6	94,7	324,0	13,7	-66,6
18	101,7	64,6	146,4	5,8	-29,0
19	101,7	64,6	146,4	5,8	-29,0
20	104,4	76,5	0,0	26,0	-1,0
21	108,5	73,4	10,9	84,6	-30,4
22	110,5	109,8	1095,6	125,4	370,7
23	111,2	92,2	240,3	141,6	184,5
24	112,6	61,5	231,0	176,9	-202,2
25	121,4	82,2	30,3	488,4	121,6
26	128,2	119,2	1806,3	835,2	1228,3
27	144,5	106,6	894,0	2043,0	1351,5
28	154,0	101,0	590,5	2992,1	1329,2
29	158,0	128,6	2693,6	3445,7	3046,5
30	195,3	116,7	1600,0	9216,0	3840,0
31	199,4	122,3	2079,4	10020,0	4564,6
32	209,0	136,7	3600,0	12034,1	6582,0
33	217,0	134,2	3306,3	13853,3	6767,8
$\sum$	3277,0	2531,4	42751,3	89055,6	51416,3
	99,3	76,7			
ρxy	0,83				

1.6 Theoretical Models of Random Variables

A random variable is a mathematical model representing a quantity that varies. A random-variable model describes the possible values the quantity can take and the corresponding probabilities associated with each value.

Just as the frequency plot of a dataset indicates the likelihood of different values occurring (Figure 1.3), a theoretical random-variable model provides a mathematical representation of the information contained in a frequency plot.

Why is a theoretical random-variable model needed to describe a dataset? First, any dataset is finite in size. For example, if another sample of 64 soil density measurements were collected, the resulting frequency plot would differ from that shown in Figure 1.3, and the sample statistics would differ from those summarized in Table 1.3. To obtain the "true" frequencies and statistics, soil density would need to be measured at every location within the soil mass. A random variable therefore serves as a theoretical model of these "true" frequencies and statistics.

Second, most engineering problems involve combinations of uncertain quantities. For instance, piles may undergo excessive displacements when the applied load exceeds the pile capacity. In such cases, variability in both load and pile bearing capacity must be considered. Random-variable models provide a mathematical framework for representing and combining multiple uncertain quantities.

Random variables are generally denoted by uppercase letters; for example, Xmay represent soil density. When a random variable takes a specific value–once it has been observed or measured–it is no longer random and is denoted by a lowercase letter. Thus, $x$represents an observation (realization) of X.

The set of all possible values that Xcan take is called the sample space. For example, the density of a soil must be greater than zero. By definition, the probability of the sample space event, such as $P[X > 0, \text{kg}/m^3]$, is equal to 1,0. An event corresponding to specific values is a subset of the sample space. For example, the event that the soil density is less than 1600 kg/m³ has the corresponding probability $P[X < 1600\ \text{kg}/m^3]$.

The probability distribution of a random variable is a function that describes the likelihood that it takes different values, for example $P[X = x]$. (Ang & Tang, 2007)

1.7 Discrete Random Variables

Discrete random variables can take only discrete values within the sample space. For example, consider the number of projects awarded to a consulting firm next month, denoted X. If the firm has submitted five proposals, then Xmay take one of the following possible values: 0, 1, 2, 3, 4, or 5.

The probability mass function (PMF) of a discrete random variable describes its probability distribution. An example PMF is shown in Figure 1.15. This PMF indicates that the probability of zero successful proposals is $P[X = 0] = 0{,}116$, the probability of one successful proposal is $P[X = 1] = 0{,}312$, and so on.

Note that the sum of the individual probabilities for Xbetween 0 and 5 is equal to 1,0, since this range represents all possible values of X. Probabilities that Xlies within a given range can also be obtained from the PMF. For example, the probability that $X$is greater than 1, $P[X > 1]$, is given by $0{,}336 + 0{,}181 + 0{,}049 + 0{,}005 = 0{,}571$. The PMF of Xcan be expressed mathematically as $P[X = x] = p_X(x)$ (Ang & Tang, 2007).

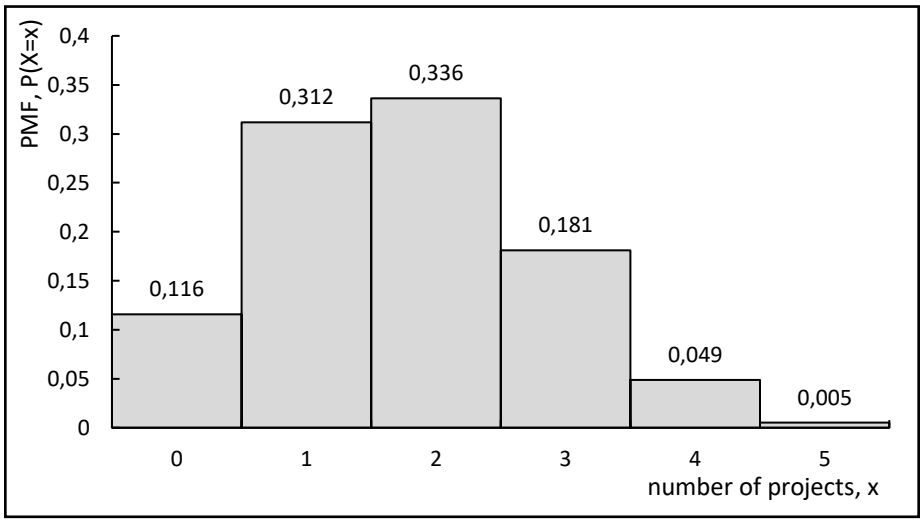


Figure 1.14. *Exemple de la fonction de masse de probabilité PMF.*

The cumulative distribution function (CDF) describes the probability that a random variable takes a value less than or equal to a given value. It is expressed as:

$$F_x(x) = P[X \leq x] = \sum_{xi \leq x} p_x(x_i) \qquad (1.13)$$

In the previous example (Figure 1.15), the probability given by the CDF at 1 is obtained as $F_X(1) = 0{,}116 + 0{,}312 = 0{,}428$.

The resulting CDF for this example is shown in Figure 1.16. Note that the PMF and the CDF convey the same information.

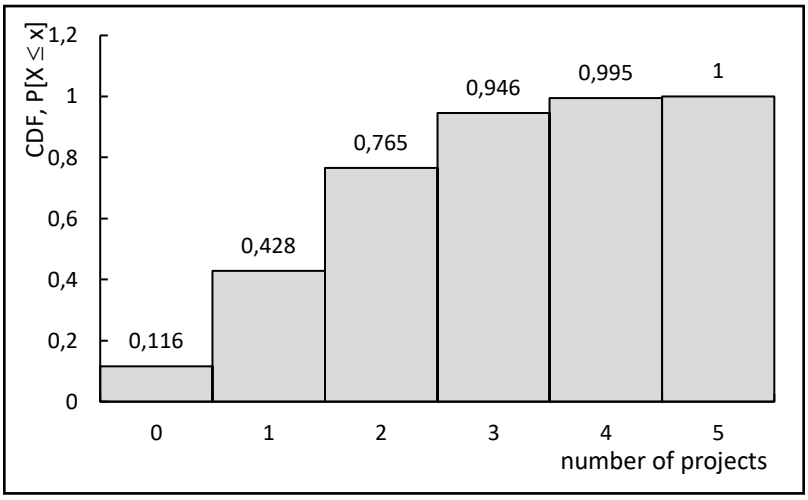


Figure 1.15. *CDF Example*

For a dataset, the PMF and CDF represent theoretical counterparts of the frequency and cumulative frequency plots. The mean value of a discrete random variable is computed as :

$$\mu_X = \sum_i x_i \; p_X(x_i), \qquad (1.14)$$

where $\mu_X$denotes the mean of X. The mean is a weighted average of X, in which each value $x_i$is weighted by its probability of occurrence.

The median is the value of $X$for which the probabilities of values above and below are equal, i.e.,

$$F_X(x_{\text{median}}) = 0{,}5.$$

The mode is the most probable value of X, such that $p_X(x_{\text{mode}})$is maximal.

For the example of the number of projects awarded (Figure 1.15), $\mu_X = 1{,}75$, x_{median}lies between 1 and 2, and $x_{\text{mode}} = 2$.

The variance is computed using the following expression:

$$\sigma_x^2 = \sum (x_i - \mu_x)^2 \, p_x(x_i) \qquad (1.15)$$

where $\sigma_X$denotes the standard deviation of X. The coefficient of variation of X, δ_X(COV), is defined as the ratio of the standard deviation to the mean:

$$\delta_x = \frac{\sigma_x}{|\mu_x|} \qquad (1.16)$$

The skewness coefficient is obtained as follows:

$$\psi_x = \frac{1}{\sigma_x^3} \sum (x_i - \mu_x)^3 p_x(x_i) \qquad (1.17)$$

For the same example of the number of projects, $\sigma_X = 1{,}07$, $\delta_X = 0{,}61$, and $\psi_X = 0{,}28$.

An important tool when working with random variables is the expectation. The expectation of a quantity is its weighted average, where the possible values are weighted by their corresponding probabilities of occurrence.

For example, the expectation of X, denoted $E[X]$, is given by the following expression:

$$E[X] = \sum x_i p_x(x_i) \qquad (1.18)$$

The mean of X, μ_X, is equal to its expectation. The expectation of a function of Xcan be obtained as:

$$E[g(X)] = \sum g(x_i) p_x(x_i) \qquad (1.19)$$

he variance of $X$is equal to the expectation of the squared deviation from the mean:

$$g(X) = (X - \mu_x)^2 \qquad (1.20)$$

Expectation is a useful tool when studying multiple random variables and functions of random variables. As a simple and practical example, consider the random variable describing the number of projects awarded to a consulting firm next month (Figure 1.15). If the revenue from each project is 6000000 DZD, then the expected revenue for next month is obtained as follows:

$$\begin{aligned} E[revenue] &= E[6000000X] \\ &= [0 \times (0{,}116)] + [6000000 \times (0{,}312)] + [12000000 \times 0{,}336] + [18000000 \times 0{,}181] \\ &+ [24000000 \times 0{,}049] + [30000000 \times 0{,}005] = 10488000DZD \end{aligned}$$

We could also evaluate the expected profit. If at least 6000000 in new revenue is required each month to operate, a profit margin of 20% is applied to the next 12000000 of revenue, and a profit margin of 30% is applied to any revenue exceeding 18000000, then the expected profit is computed as follows:

$$\begin{aligned} E[bénéfice] &= [0 \times 0{,}116] + [0 \times 0{,}312] + [1200000 \times 0{,}336] + [2400000 \times 0{,}181] + [4200000 \times 0{,}049] \\ &+ [6000000 \times 0{,}005] = 1073400DZD \end{aligned}$$

Clearly, the firm must secure more projects or reduce its overhead costs. Several of the most common discrete random-variable models are summarized in Table 1.5. The PMF shown in Figure 1.15 is an example of a binomial distribution with $n = 5$(a maximum of five projects can be awarded) and $p = 0{,}35$(the probability of winning a project is assumed to be 35%).

Table 1.5 Common discrete random-variable models (Ang & Tang, 2007)

Distribution	PMF	Mean	Variance	Explanation	Example
Binomial	$p_X(x) = \frac{n!}{x!(n-x)!}\, p^x(1-p)^{n-x},\ x = 0,1,\dots,n$	np	$np(1-p)$	$X$is the number of occurrences "successes" in $n$independent trials, where $p$is the probability of occurrence per trial.	Flood occurrence
Geometric	$p_X(x) = p(1-p)^{x-1},\ x = 1,2,\dots$	$1/p$	$(1-p)/p^2$	$X$is the number of independent trials until the next occurrence "success", where $p$is the probability of occurrence per trial.	Flood return period
Poisson	$p_X(x) = \frac{(\nu t)^x}{x!}\, e^{-\nu t},\ x = 0,1,\dots$	νt	νt	$X$is the number of independent occurrences in a time interval t, where $\nu$is the average occurrence rate.	Earthquake occurrence

1.7.1 Continuous Random Variables

Continuous random variables can take any value within the sample space. Soil density is an example of a continuous random variable, as it may take any value greater than zero.

The probability density function (PDF) of a continuous random variable describes its probability distribution. An example of a PDF is shown in Figure 1.17. The PDF conveys information similar to that provided by a PMF. However, for a continuous random variable, there is an infinite number of possible values in the sample space. Consequently, unlike the discrete case, it is not meaningful to define the probability of the event Xbeing exactly equal to a given value x, since this probability is infinitesimally

small. Instead, probabilities are defined over intervals, and the probability that Xfalls within a very small interval is proportional to the value of the PDF.

In the soil density example, the probability that the density lies within a small interval around 1762 kg/m³ is greater than the probability that it lies within a small interval around 2002 kg/m³ (Figure 1.17). The PDF is denoted mathematically by $f_X(x)$ (Ang & Tang, 2007).

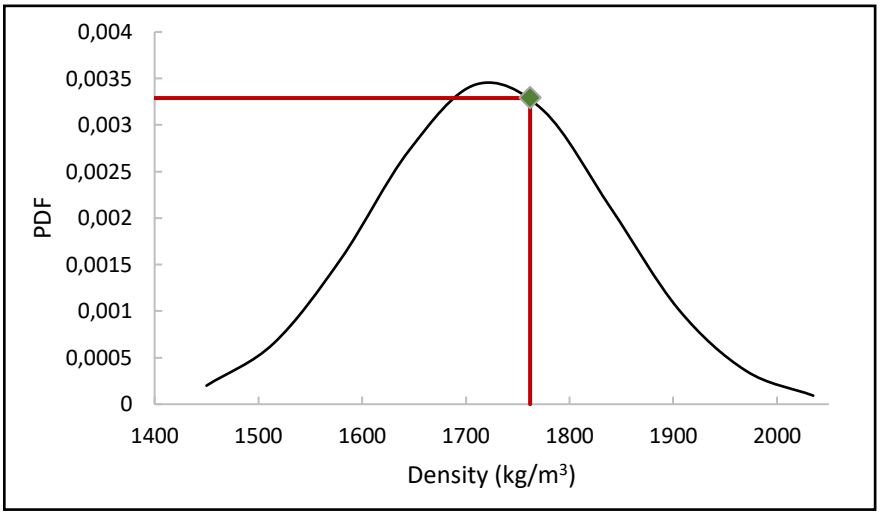


Figure 1.16. *Probability density function for soil density.*

As for a discrete random variable, the cumulative distribution function (CDF) describes the probability that the variable takes a value less than or equal to a given value. It is obtained as follows:

$$F_x(x) = P[X \leq x] = \int_{-\infty}^{x} f_x(\xi) d\xi \qquad (1.21)$$

For example, the CDF evaluated at 1762 kg/m³ for soil density is equal to 0,62 (Figure 1.17). A plot of the CDF for soil density is shown in Figure 1.18.

Since the probability over the sample space is equal to 1.0, the area under the PDF must also be equal to 1,0. Recall that the area under a frequency density plot for a dataset is likewise equal to 1,0. Therefore, theoretical PDFs can be used to model a dataset by superimposing a theoretical PDF on a frequency density plot. For instance, Figure 1.19 shows the theoretical PDF for soil density (Figure 1.17) overlaid on the frequency density plot (Figure 1.7).

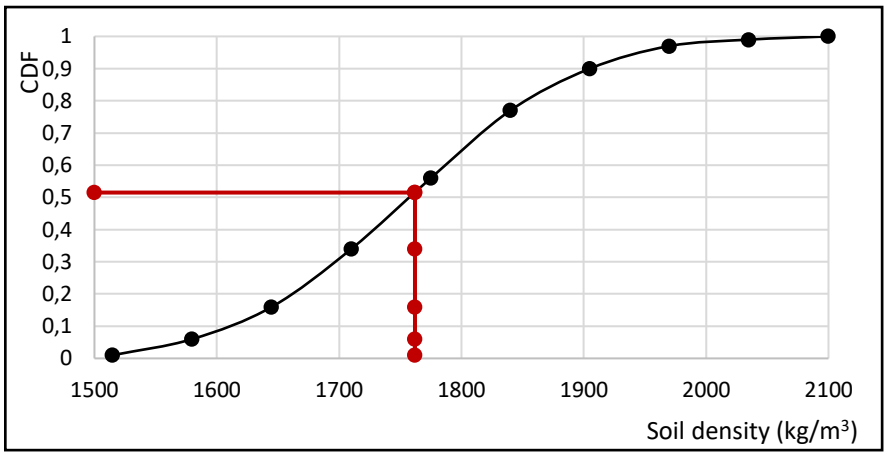


Figure 1.17. *Cumulative distribution function (CDF) for soil density*

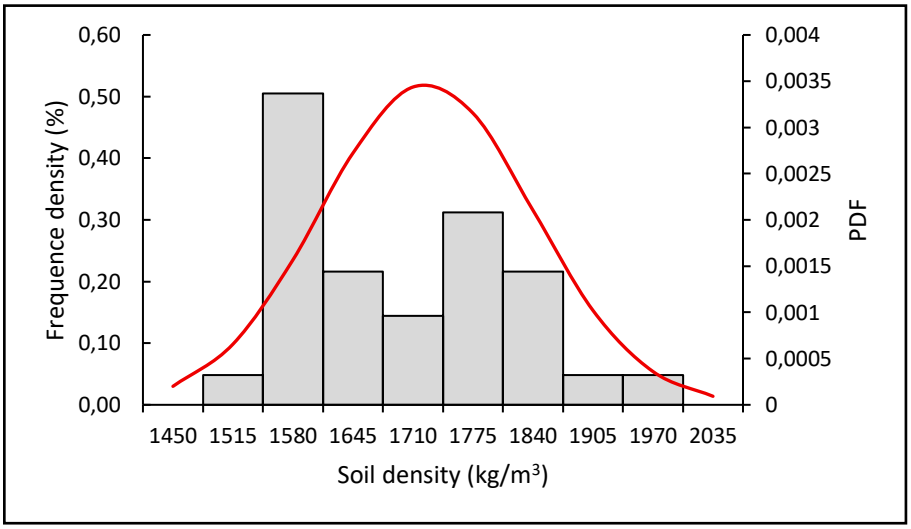


Figure 1.18. *Probability density function and frequency density plot*

The expectation of a continuous random variable is defined in the same way as for a discrete random variable: it is a weighted average, where the values are weighted by their probabilities. However, since there is an infinite number of possible values in the sample space, the summation of values weighted by their probabilities is replaced by an integral:

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x) f_x(x) dx \qquad (1.22)$$

Similarly, the mean, variance, and skewness coefficient are obtained as follows:

$$\mu_x = E[X] = \int_{-\infty}^{+\infty} x f_x(x) dx \qquad (1.23)$$

$$\sigma_x^2 = E[(X - \mu_x)^2] = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f_x(x) dx \qquad (1.24)$$

$$\psi_x = \frac{E[(X-\mu_x)^3]}{\sigma_x^3} = \frac{\int_{-\infty}^{+\infty} (x-\mu_x)^3 f_x(x) dx}{\sigma_x^3} \qquad (1.25)$$

The commonly used models for continuous random variables are summarized in Table 1.6. The normal distribution and the related lognormal distribution are the most frequently used random-variable models.

Table 1.6 Common continuous random-variable models (Ang & Tang, 2007)

Distribution	PDF	Mean	Variance	Explanation	Example
Uniform	$f_X(x) = \frac{1}{b-a}, a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	All values in $[a, b]$ are equally likely.	Test result trend
Triangular	$f_X(x) = \begin{cases} \frac{2}{b-a}\frac{x-a}{u-a}, & a \le x \le u \\ \frac{2}{b-a}\frac{b-x}{b-u}, & u \le x \le b \end{cases}$	$\frac{a+b+u}{3}$	$\frac{a^2+b^2+u^2-ab-au-}{18}$	Bounded distribution with a most likely value u.	Construction cost
Exponential	$f_X(x) = ve^{-vx}, x \ge 0$	$\frac{1}{v}$	$\frac{1}{v^2}$	X represents the time between independent occurrences; v is the occurrence rate.	Aftershock occurrence
Normal	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], -\infty < x < +\infty$	μ	σ^2	X may represent the sum of several random variables (central limit behavior).	Soil shear strength
Lognormal	$f_X(x) = \frac{1}{x\,\sigma_{\ln X}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu_{\ln X}}{\sigma_{\ln X}}\right)\right], x \ge 0$	$\exp\left(\mu_{\ln X}+\frac{1}{2}\sigma^2_{\ln X}\right.$	$[\exp(\sigma^2_{\ln X}) - 1]\exp(2\mu_{\ln X}+\sigma^2_{\ln X})$	X may represent the product of several random variables.	Permeability

The normal distribution (also known as the Gaussian distribution) is the classic bell-shaped curve and occurs frequently in datasets. For example, the undrained shear strength-to-depth data shown in Figure 1.14 are well fitted by a normal distribution (Figure 1.20).

The normal distribution has several important properties. First, it is symmetric (the skewness coefficient $\psi$is zero for a normal PDF). Second, its extreme values decay exponentially. There is a 68% probability that a normally distributed variable lies within $\pm$1standard deviation of the mean, a 95% probability that it lies within ($\mu\pm2\sigma$), and a 99.7% probability that it lies within ($\mu\pm3\sigma$). Therefore, it is very unlikely (less than 1% probability) to observe a value outside $\pm$3standard deviations from the mean.

Finally, any linear transformation of a normally distributed variable is also normally distributed. If $Y = aX + b$and Xfollows a normal distribution, then $Y$is normally distributed with mean $\mu_Y = a\mu_X + b$and standard deviation:

$$\sigma_Y = a\sigma_X \qquad (1.26)$$

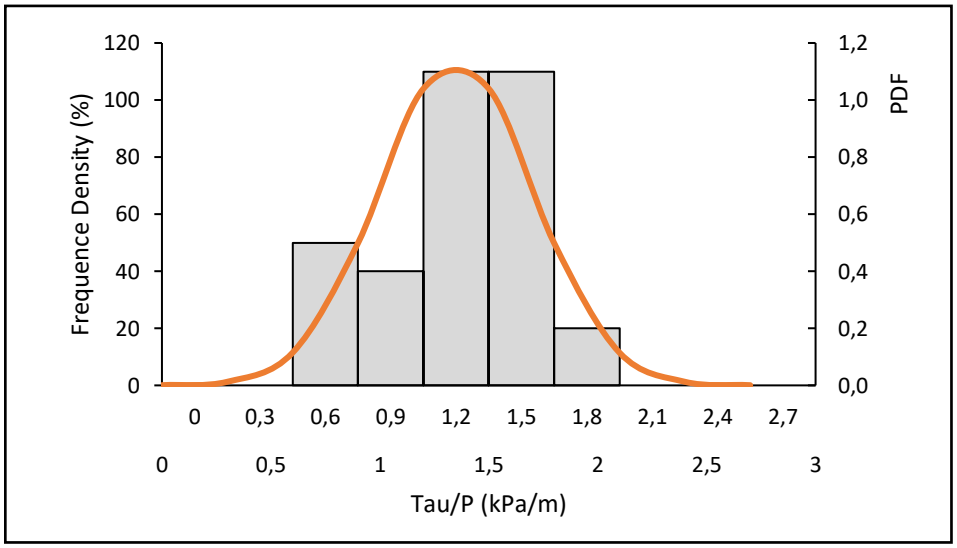


Figure 1.19. *Probability density function for the undrained shear strength-to-depth ratio*

The CDF of a normal distribution (i.e., the integral of the PDF) cannot be expressed in closed analytical form. However, it is widely tabulated and is available in most spreadsheet software packages.

The first step in using these tables is to standardize $X$by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{X-\mu_x}{\sigma_x} \qquad (1.27)$$

where Z is the standardized value of X; it has zero mean and unit variance. The tables then provide the CDF evaluated at $x$through the standardized variable z, such that $F_X(x) = \Phi(z)$, where $\Phi$is called the standard normal distribution function. Standard normal values as a function of $z$are summarized in Table 1.7.

As an example, consider the undrained shear strength-to-depth ratio data in Figure 1.20. The probability that this ratio is less than 1.884 kPa/m is computed as follows:

$$P[X \leq 1.884 \text{ kPa/m}] = F_X(x = 1{,}884) = \Phi\left(\frac{x - \mu_X}{\sigma_X}\right) = \Phi\left(\frac{1{,}884 - 1{,}35}{0{,}35}\right) = \Phi(1{,}53) = 0{,}93699.$$

Similarly, the probability that the ratio exceeds 1.884 kPa/m can be computed using the fact that the total area under the PDF is 1.0:

$$P[X > 1{,}884 \text{ kPa/m}] = 1 - P[X \leq 1{,}884] = 1 - 0{,}93699 = 0{,}063.$$

The probability that the ratio is less than 0.66 kPa/m is computed as:

$$P[X \leq 0{,}66 \text{ kPa/m}] = F_X(x = 0{,}66) = \Phi\left(\frac{x - \mu_X}{\sigma_X}\right) = \Phi\left(\frac{0{,}66 - 1{,}35}{0{,}35}\right) = \Phi(-1{,}97).$$

Since Table 1.7 does not list negative z-values and the normal distribution is symmetric, Φcan be evaluated as:

$\Phi(-1{,}97) = 1 - \Phi(1{,}97) = 1 - 0{,}97558 = 0{,}024$

Table 1.7 $\Phi(z)$ Values (Devore, 2015)

z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
0.00	0.50000	0.50	0.69146	1.00	0.84134	1.50	0.93319	2.00	0.97725	2.50	0.99379	3.00	0.99865
0.01	0.50399	0.51	0.69497	1.01	0.84375	1.51	0.93448	2.01	0.97778	2.51	0.99396	3.01	0.99869
0.02	0.50798	0.52	0.69847	1.02	0.84614	1.52	0.93574	2.02	0.97831	2.52	0.99413	3.02	0.99874
0.03	0.51197	0.53	0.70194	1.03	0.84850	1.53	0.93699	2.03	0.97882	2.53	0.99430	3.03	0.99878
0.04	0.51595	0.54	0.70540	1.04	0.85083	1.54	0.93822	2.04	0.97932	2.54	0.99446	3.04	0.99882
0.05	0.51994	0.55	0.70884	1.05	0.85314	1.55	0.93943	2.05	0.97982	2.55	0.99461	3.05	0.99886
0.06	0.52392	0.56	0.71226	1.06	0.85543	1.56	0.94062	2.06	0.98030	2.56	0.99477	3.06	0.99889
0.07	0.52790	0.57	0.71566	1.07	0.85769	1.57	0.94179	2.07	0.98077	2.57	0.99492	3.07	0.99893
0.08	0.53188	0.58	0.71904	1.08	0.85993	1.58	0.94295	2.08	0.98124	2.58	0.99506	3.08	0.99897
0.09	0.53586	0.59	0.72240	1.09	0.86214	1.59	0.94408	2.09	0.98169	2.59	0.99520	3.09	0.99900
0.10	0.53983	0.60	0.72575	1.10	0.86433	1.60	0.94520	2.10	0.98214	2.60	0.99534	3.10	0.99903
0.11	0.54380	0.61	0.72907	1.11	0.86650	1.61	0.94630	2.11	0.98257	2.61	0.99547	3.11	0.99906
0.12	0.54776	0.62	0.73237	1.12	0.86864	1.62	0.94738	2.12	0.98300	2.62	0.99560	3.12	0.99910
0.13	0.55172	0.63	0.73565	1.13	0.87076	1.63	0.94845	2.13	0.98341	2.63	0.99573	3.13	0.99913
0.14	0.55567	0.64	0.73891	1.14	0.87286	1.64	0.94950	2.14	0.98382	2.64	0.99585	3.14	0.99916
0.15	0.55962	0.65	0.74215	1.15	0.87493	1.65	0.95053	2.15	0.98422	2.65	0.99598	3.15	0.99918
0.16	0.56356	0.66	0.74537	1.16	0.87698	1.66	0.95154	2.16	0.98461	2.66	0.99609	3.16	0.99921
0.17	0.56749	0.67	0.74857	1.17	0.87900	1.67	0.95254	2.17	0.98500	2.67	0.99621	3.17	0.99924
0.18	0.57142	0.68	0.75175	1.18	0.88100	1.68	0.95352	2.18	0.98537	2.68	0.99632	3.18	0.99926
0.19	0.57535	0.69	0.75490	1.19	0.88298	1.69	0.95449	2.19	0.98574	2.69	0.99643	3.19	0.99929
0.20	0.57926	0.70	0.75804	1.20	0.88493	1.70	0.95543	2.20	0.98610	2.70	0.99653	3.20	0.99931
0.21	0.58317	0.71	0.76115	1.21	0.88686	1.71	0.95637	2.21	0.98645	2.71	0.99664	3.21	0.99934
0.22	0.58706	0.72	0.76424	1.22	0.88877	1.72	0.95728	2.22	0.98679	2.72	0.99674	3.22	0.99936
0.23	0.59095	0.73	0.76730	1.23	0.89065	1.73	0.95818	2.23	0.98713	2.73	0.99683	3.23	0.99938
0.24	0.59483	0.74	0.77035	1.24	0.89251	1.74	0.95907	2.24	0.98745	2.74	0.99693	3.24	0.99940
0.25	0.59871	0.75	0.77337	1.25	0.89435	1.75	0.95994	2.25	0.98778	2.75	0.99702	3.25	0.99942
0.26	0.60257	0.76	0.77637	1.26	0.89617	1.76	0.96080	2.26	0.98809	2.76	0.99711	3.26	0.99944
0.27	0.60642	0.77	0.77935	1.27	0.89796	1.77	0.96164	2.27	0.98840	2.77	0.99720	3.27	0.99946
0.28	0.61026	0.78	0.78230	1.28	0.89973	1.78	0.96246	2.28	0.98870	2.78	0.99728	3.28	0.99948
0.29	0.61409	0.79	0.78524	1.29	0.90147	1.79	0.96327	2.29	0.98899	2.79	0.99736	3.29	0.99950
0.30	0.61791	0.80	0.78814	1.30	0.90320	1.80	0.96407	2.30	0.98928	2.80	0.99744	3.30	0.99952
0.31	0.62172	0.81	0.79103	1.31	0.90490	1.81	0.96485	2.31	0.98956	2.81	0.99752	3.31	0.99953
0.32	0.62552	0.82	0.79389	1.32	0.90658	1.82	0.96562	2.32	0.98983	2.82	0.99760	3.32	0.99955
0.33	0.62930	0.83	0.79673	1.33	0.90824	1.83	0.96638	2.33	0.99010	2.83	0.99767	3.33	0.99957
0.34	0.63307	0.84	0.79955	1.34	0.90988	1.84	0.96712	2.34	0.99036	2.84	0.99774	3.34	0.99958
0.35	0.63683	0.85	0.80234	1.35	0.91149	1.85	0.96784	2.35	0.99061	2.85	0.99781	3.35	0.99960
0.36	0.64058	0.86	0.80511	1.36	0.91309	1.86	0.96856	2.36	0.99086	2.86	0.99788	3.36	0.99961
0.37	0.64431	0.87	0.80785	1.37	0.91466	1.87	0.96926	2.37	0.99111	2.87	0.99795	3.37	0.99962
0.38	0.64803	0.88	0.81057	1.38	0.91621	1.88	0.96995	2.38	0.99134	2.88	0.99801	3.38	0.99964
0.39	0.65173	0.89	0.81327	1.39	0.91774	1.89	0.97062	2.39	0.99158	2.89	0.99807	3.39	0.99965
0.40	0.65542	0.90	0.81594	1.40	0.91924	1.90	0.97128	2.40	0.99180	2.90	0.99813	3.40	0.99966
0.41	0.65910	0.91	0.81859	1.41	0.92073	1.91	0.97193	2.41	0.99202	2.91	0.99819	3.41	0.99968
0.42	0.66276	0.92	0.82121	1.42	0.92220	1.92	0.97257	2.42	0.99224	2.92	0.99825	3.42	0.99969
0.43	0.66640	0.93	0.82381	1.43	0.92364	1.93	0.97320	2.43	0.99245	2.93	0.99831	3.43	0.99970
0.44	0.67003	0.94	0.82639	1.44	0.92507	1.94	0.97381	2.44	0.99266	2.94	0.99836	3.44	0.99971
0.45	0.67364	0.95	0.82894	1.45	0.92647	1.95	0.97441	2.45	0.99286	2.95	0.99841	3.45	0.99972
0.46	0.67724	0.96	0.83147	1.46	0.92785	1.96	0.97500	2.46	0.99305	2.96	0.99846	3.46	0.99973
0.47	0.68082	0.97	0.83398	1.47	0.92922	1.97	0.97558	2.47	0.99324	2.97	0.99851	3.47	0.99974
0.48	0.68439	0.98	0.83646	1.48	0.93056	1.98	0.97615	2.48	0.99343	2.98	0.99856	3.48	0.99975
0.49	0.68793	0.99	0.83891	1.49	0.93189	1.99	0.97670	2.49	0.99361	2.99	0.99861	3.49	0.99976

Finally, we can compute the probability that the undrained shear strength is lower than a design value of 12 kPa at a depth of 15 m. Let $Y = 15X$. Since $X$is normally distributed, $Y$is also normally distributed with mean $15(1{,}35) = 20{,}25$kPa and standard deviation $15(0{,}35) = 5{,}25$kPa. The probability that $Y$is less than 12 kPa is therefore computed as:

$$P[Y \leq 12 \text{ kPa}] = \Phi\left(\frac{12 - 20{,}25}{5{,}25}\right) = \Phi(-1{,}57) = 1 - \Phi(1{,}57) = 1 - 0{,}94179 = 0{,}058.$$

The lognormal distribution is related to the normal distribution as follows: if the logarithm of a variable is normally distributed, then the variable itself follows a lognormal distribution. The lognormal distribution is commonly used for three main reasons. First, it arises naturally when several random variables are multiplied; therefore, any process defined as the product of individual random variables tends to be described by a lognormal distribution. Second, the lognormal model is appropriate for quantities that cannot be negative. Since many engineering properties (e.g., strength) are non-negative, the lognormal distribution is often a reasonable choice. Finally, the lognormal distribution is convenient for modeling quantities that span several orders of magnitude, such as permeability.

An example of a lognormal distribution fitted to the permeability dataset is shown in Figure 1.21. Note that this distribution appears symmetric when plotted on a logarithmic scale, but exhibits positive skewness on an arithmetic scale

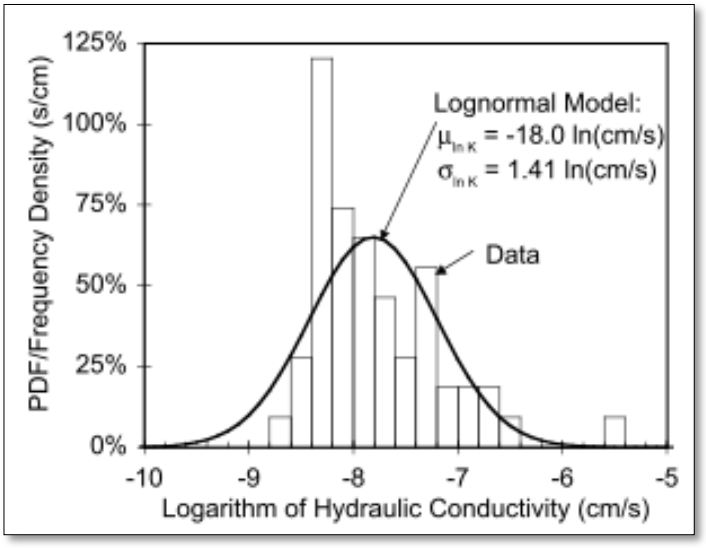


Figure 1.20. *Probability density function for permeability(Ken'ichirou , 1996)*

Since the lognormal distribution is related to the normal distribution, the CDF of a lognormal variable can be computed using the standard normal distribution function. The relationship between the two is as follows:

$$P[X \leq x] = F_X(x) = \Phi\left(\frac{ln(x)-\mu_{lnX}}{\sigma_{lnX}}\right) \qquad (1.28)$$

where Xfollows a lognormal distribution with parameters $\mu_{\ln X}$and $\sigma_{\ln X}$(Table 1.6), which are the mean and standard deviation of $\ln(X)$. For example, the probability that the permeability (Figure 1.21) exceeds 1×10^{-7}cm/s is computed as follows:

$$P\left[X > 1 \times 10^{-7}\frac{cm}{s}\right] = 1 - \Phi\left(\frac{\ln(1 \times 10^{-7}) - (-18)}{1{,}41}\right) = 1 - \Phi(1.33) = 1 - 0{,}908 = 0{,}092$$

1.8 Geotechnical Correlations

Empirical correlations are often used in geotechnical engineering to relate soil properties. For example, the compression index $C_c$may be correlated with the void ratio $e$or the liquid limit w_L; relative density $D_r$may be correlated with SPT results; and shear strength may be correlated with the plasticity index I_p. The objective is to estimate soil parameters required for design using properties that are generally less expensive and easier to obtain. If a strong relationship truly existed between the properties, it would provide a cost-effective means of determining design parameters. In practice, however, empirical correlations are often far from perfect, and the additional implicit uncertainty associated with this approach must be assessed (Phoon & Ching, RISK AND RELIABILITY IN GEOTECHNICAL ENGINEERING, 2015).

Figure 1.22 presents two empirical relationships established from test data. Clearly, one would have greater confidence in using relationship A. However, the following questions arise: Is it reliable to use relationship B? Would reliability improve if the number of indirect tests increased, and by how much? Should only direct tests be used? Could a limited number of direct tests be complemented with less expensive indirect tests?

To address these questions, the probabilistic implications of using an empirical relationship are introduced as follows. Consider the simple case where the empirical relationship is approximately linear and the data scatter is approximately constant about the prediction curve. Linear regression with constant variance can then be used to obtain an estimate of the mean value of Yas a function of x, i.e., $Y(x) = a + bx$, along with an estimate of the variance, i.e., $\mathrm{Var}(Y(x))$. The first represents the predicted value of $Y$for

a given value of x, whereas the second quantifies the error associated with using the empirical correlation for that prediction. The square root of $\mathrm{Var}(Y(x))$is referred to as the calibration error.

If the empirical relationship is established from a limited dataset, an additional estimation error arises due to limited information. However, this error is generally small compared with the calibration error and will be neglected in the examples that follow (Ang & Tang, 2007).

In the following example, the predicted value of a soil parameter and its associated uncertainty will be compared using multiple sources of information, namely: (1) $n$direct tests; (2) $m$indirect tests; and (3) subjective judgment.

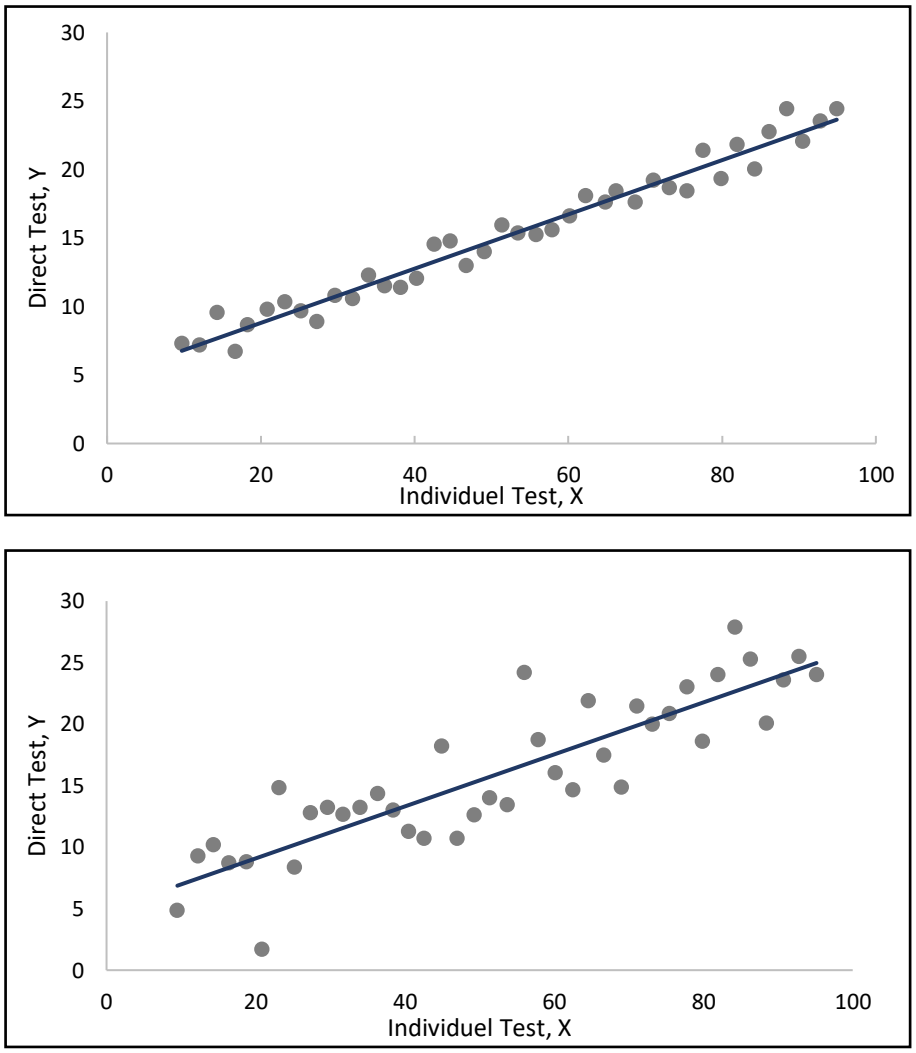


Figure 1.21. *Example of empirical correlations*

To simplify, the tests are assumed to be statistically independent. For direct tests, the estimation error is proportional to the data scatter but inversely proportional to the number of tests n. For indirect tests, the calibration error $\sigma_c$must be added to the data scatter.

Finally, based on subjective judgment, the uncertainty in the soil parameters is described by a distribution whose corresponding mean and standard deviation can be inferred from μ' and σ', respectively. Bayesian probability theory is used to combine the two estimates.

Assuming that the soil samples are statistically independent, the error associated with estimating the mean can be computed as follows (Ang & Tang, 2007) :

1) Direct Tests:

$$\hat{\mu}_X = \frac{1}{n}\sum x_i \qquad (1.29)$$

$$\hat{\sigma}_{\hat{\mu}_X} = \frac{\hat{\sigma}_x}{\sqrt{n}} \qquad (1.30)$$

2) Indirect Tests:

$$\hat{\mu}_X = \frac{1}{m}\sum x_{ci} \qquad (1.31)$$

$$\hat{\sigma}_{\hat{\mu}_X} = \sqrt{\frac{\hat{\sigma}_X^2+\sigma_c^2}{m}} \qquad (1.32)$$

where σ_c denotes the standard deviation of the random calibration error.

3) Subjective judgment

$$\hat{\mu}_X = \mu' \qquad (1.33)$$

$$\hat{\sigma}_{\hat{\mu}_X} = \sigma' \qquad (1.34)$$

The following example illustrates how seven triaxial tests and nine pressuremeter readings values can be combined with subjective judgment to obtain an overall estimate of the mean cohesion of Hong Kong clay, along with a measure of the associated overall estimation error (Figure 1.23) (Ang & Tang, 2007).

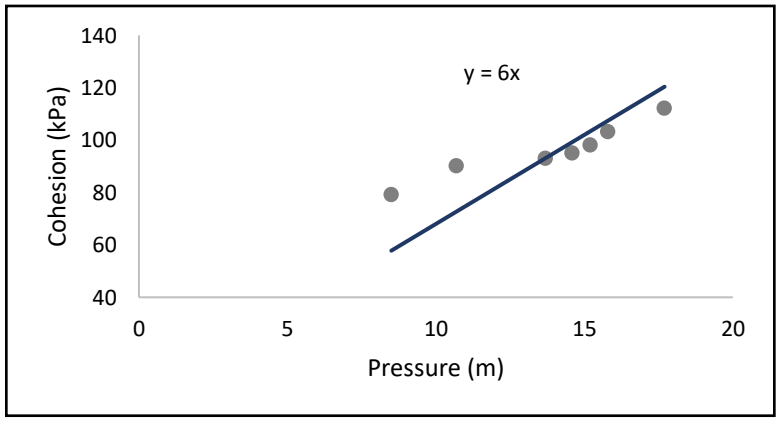


Figure 1.22. *Variation of cohesion as a function of pressure*

The empirical correlation used is:

$c' = 6\ N$ (kPa)

where $N$is the pressure (in m), with a calibration error of 36 kPa based on the data reported by Lumb (1968) for Hong Kong clay (Ang & Tang, 2007).

Example 1 : Estimation of the mean cohesion of the clay (assuming $\hat{\sigma}_X = 10.3$kPa)

1) Prior information (experience): 100 ± 12kPa with 95% probability

$\hat{\mu}_{X1} = 100$

$\hat{\sigma}_{\hat{\mu}_{X1}} = 12$

2) Seven triaxial test: $\{103,90,93,79,112,95,98\}$ kPa

$\hat{\mu}_{X2} = \frac{1}{7}\sum_{i=1}^{7} x_i = 96kPa$

$\hat{\sigma}_{\hat{\mu}_{X2}} = \frac{10.3}{\sqrt{7}} = 3.9$ kPa

3) Nine pressure values: $\{8.5, 13.7,\ 10.7,\ 15.8, 20.4, 21.6, 14.6, 15.2, 17.7\}$m

$\hat{\mu}_X = \frac{1}{9}\sum_{i=1}^{9} 6N$ =92kPa

$\hat{\sigma}_{\hat{\mu}_X} = \sqrt{\frac{\hat{\sigma}_X^2 + \sigma_c^2}{m}} = \sqrt{\frac{10.3^2 + 36^2}{9}} = 12.5kPa$

Combining (1) and (2), we obtain:

$\hat{\mu}_x = \frac{(100 \times 3.9^2) + (96 \times 12^2)}{3.9^2 + 12^2} = 96kPa$

$\hat{\sigma}_{\hat{\mu}_X} = \frac{3.9 \times 12}{\sqrt{3.9^2 + 12^2}} = 3.7kPa$

Combining the result with (3), we obtain:

$\hat{\mu}_X = \frac{96 \times 12.5^2 + 92 \times 3.7^2}{12.5^2 + 3.7^2} = 96kPa$

$\hat{\sigma}_{\hat{\mu}_X} = \frac{12.5 + 3.7}{\sqrt{12.5^2 + 3.7^2}} = 3.5$

The coefficient of variation of the final mean estimate is:

$\frac{\hat{\sigma}_{\hat{\mu}_X}}{\hat{\mu}_X} = 3.7\%$.

Remark:

The estimation error using indirect tests may be smaller than that obtained using direct tests, provided that a larger number of indirect tests is performed and the calibration error is relatively low. When combining estimates from two sources of information, the weighted-mean formula assigns weights inversely proportional to the respective estimation errors. Moreover, the uncertainty of the combined estimate is always lower than the uncertainty of each individual estimate.

1.9 Multiple Random Variables

The following examples are used to illustrate reliability problems involving more than one random variable. In such cases, the degree of correlation between the random variables is an important factor affecting reliability.

1.9.1 Differential Settlement Between Two Footings

Consider two adjacent footings as shown in Figure 1.24. Assume that the settlement of each footing follows a normal distribution with a mean of 5 cm and a standard deviation of 1.3 cm. The allowable differential settlement is 2.5 cm. Determine the probability of unacceptable differential settlement for these footings (Ang & Tang, 2007).

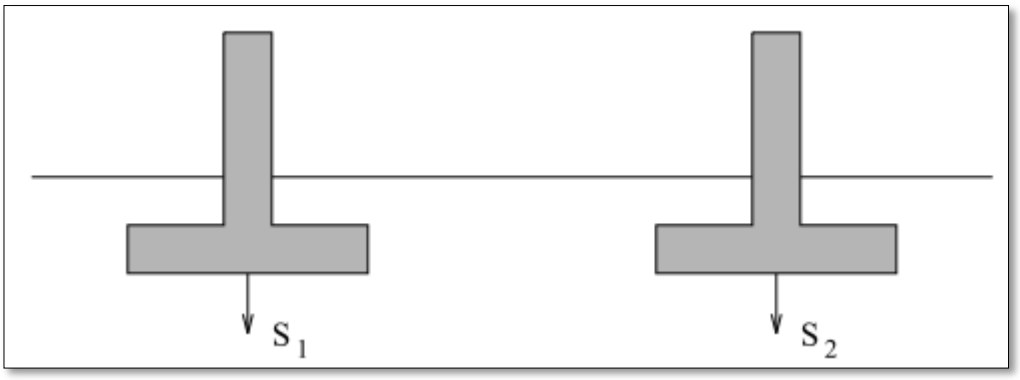


Figure 1.23. *Differential settlement between two footings*

1.9.1.1 Case 1: Statistically independent settlements

Let

$$D = S_1 - S_2$$

where $S_1$and S_2denote the settlements of Footings 1 and 2, respectively.

It can be shown that Dalso follows a normal distribution, with mean and standard deviation given by:

$$\mu_D = \mu_{S_1} - \mu_{S_2} = 5 - 5 = 0$$

$$\sigma_D = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} = \sqrt{1{,}3^2 + 1{,}3^2} = 1{,}8 \text{ cm}$$

Therefore, the probability of unacceptable differential settlement is (Ang & Tang, 2007):

$$P(|\ D\ | > 2{,}5) = P(D > 2{,}5) + P(D < -2{,}5) = 2P(D > 2{,}5)$$

$$= 2\left[1 - \Phi\left(\frac{2{,}5 - 0}{1{,}8}\right)\right] = 2(1 - 0{,}91774) = 0{,}16$$

1.9.1.2 Case 2: Correlated settlements

In practice, the loads and soil properties affecting the settlements of adjacent footings are often similar. Assume the correlation coefficient is $\rho = 0{,}8$. The variable Dremains normally distributed with zero mean, but its standard deviation becomes:

$$\sigma_D = \sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2 - 2\rho\sigma_{S_1}\sigma_{S_2}}$$
$$= \sqrt{1{,}3^2 + 1{,}3^2 - 2(0{,}8)(1{,}3)(1{,}3)} = 0{,}82 \text{ cm}$$

Thus, the probability of unacceptable differential settlement is:

$$P(|\ D\ | > 2{,}5) = 2\left[1 - \Phi\left(\frac{2{,}5 - 0}{0{,}82}\right)\right] = 0{,}002$$

On the other hand, if the settlements of the two footings are identical, i.e., perfectly correlated with $\rho = 1$, then:

$$\sigma_D = \sqrt{1{,}3^2 + 1{,}3^2 - 2(1)(1{,}3)(1{,}3)} = 0$$

and therefore, the probability of unacceptable differential settlement is zero.

Figure 1.24 illustrates how the probability of unacceptable differential settlement decreases as the correlation increases. In other words, correlation improves performance reliability; neglecting the correlation effect may be overly conservative.

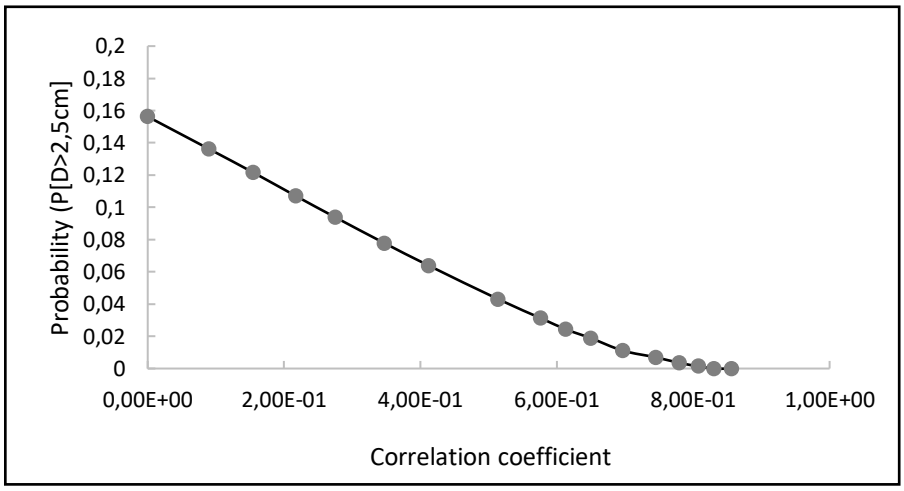


Figure 1.24. *Variation of probability as a function of the correlation coefficient*

1.9.1.3 Extension to Multiple Footings

Consider a footing grid as shown in Figure 1.26. Determining the probability of unacceptable settlement or differential settlement using an analytical approach such as the one presented above becomes very cumbersome. A more practical procedure is the Monte Carlo simulation method (Ang & Tang, 2007)

This method relies on simulating possible scenarios of settlement values for each footing based on the probability distributions of the random variables involved, and then inferring the probability of various events from the outcomes observed over all simulation runs.

In the illustrated example, three different models are investigated in which a compressible zone may be present (in Models 2 and 3) within an otherwise stiff medium. The compressible zone is treated as an anomaly A, whose occurrence beneath adjacent footings may be correlated (as in Model 3).

The probabilities of various settlement performance indicators (e.g., maximum settlement exceeding the allowable settlement, or maximum differential settlement between adjacent footings exceeding the allowable differential settlement) are presented in Figure 1.26 for the three models.

In general, the potential presence of anomalies degrades performance by increasing the likelihood of unsatisfactory behaviour. However, correlation in anomaly occurrence beneath adjacent footings improves performance. Using the Monte Carlo simulation procedure, the fraction of footings experiencing excessive settlement can also be readily estimated, as illustrated (Ang & Tang, 2007).

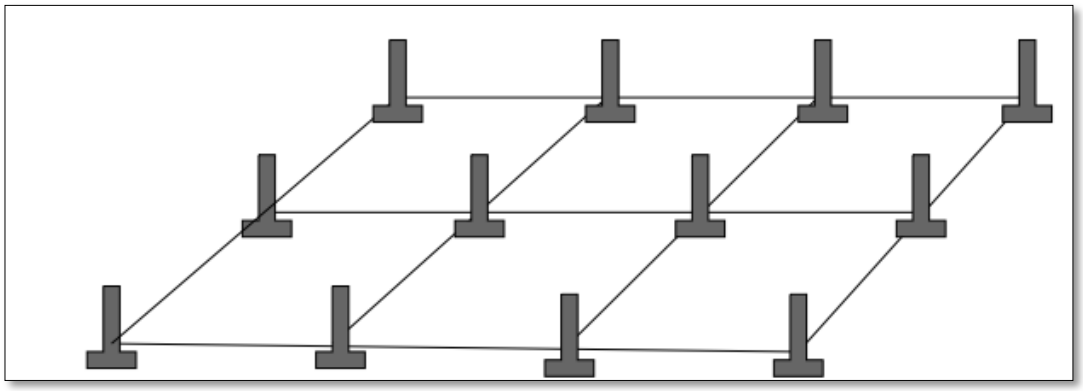


Figure 1.25. *Footing grid in which the settlement of each footing is denoted by S_i (Ang & Tang, 2007)*

Model 1: $\boldsymbol{S} \sim \boldsymbol{\mathcal{N}}(\mathbf{2,5},\ \mathbf{0,76})$cm; correlation coefficient $\boldsymbol{\rho} = \mathbf{0,6}$.

Model 2: Probability of occurrence of a compressible zone $\boldsymbol{A}$: $\boldsymbol{P}(\boldsymbol{A}) = \mathbf{0,01}$; settlement in the compressible zone $\boldsymbol{S_A} \sim \boldsymbol{\mathcal{N}}(\mathbf{5},\ \mathbf{1,5})$cm.

Model 3: Transition probability $\boldsymbol{P}(\boldsymbol{A} \rightarrow \boldsymbol{A}) = \mathbf{0,02.}$

Allowable maximum settlement: $\boldsymbol{S_{max}} = \mathbf{5}$cm,

Allowable maximum differential settlement: $\boldsymbol{\Delta S_{max}} = \mathbf{3,8}$cm.

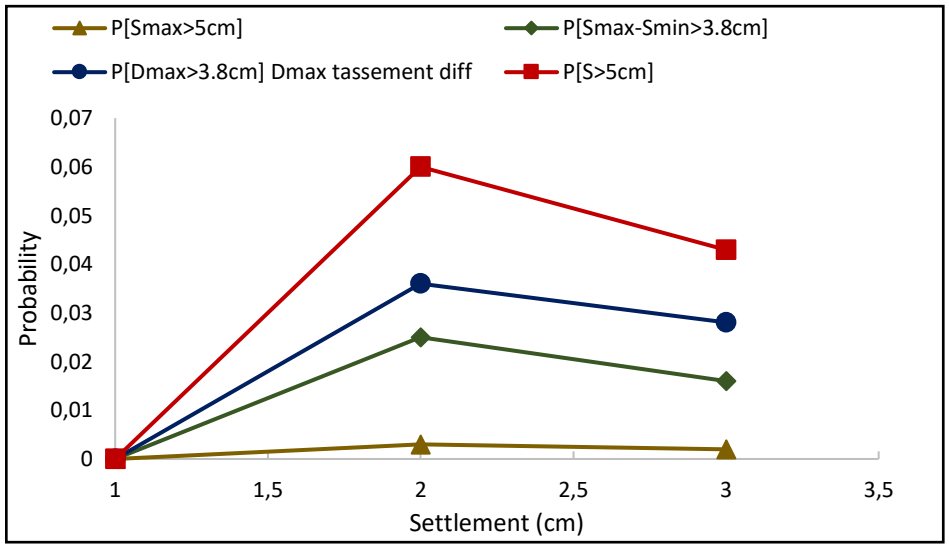


Figure 1.26. *Probability variation as a function of settlement for the different cases (Ang & Tang, 2007)*

1.9.2 Consolidation Settlement

In this example, a first-order uncertainty analysis will be applied to the settlement problem. Note that the relative contribution to the overall uncertainty depends both on the uncertainty (variability) of the individual variables and on their sensitivity factors.

The settlement is expressed as follows (Ang & Tang, 2007):

$$S = N\left(\frac{C_c}{1+e_0}\right) H log_{10}\left(\frac{p_0+\Delta p}{p_0}\right) \qquad (1.35)$$

where:

$N$is the model error factor, $C_c$is the compression index, $p_0$denotes the effective stress at point B, and $\Delta p$is the increase in stress at the same point.

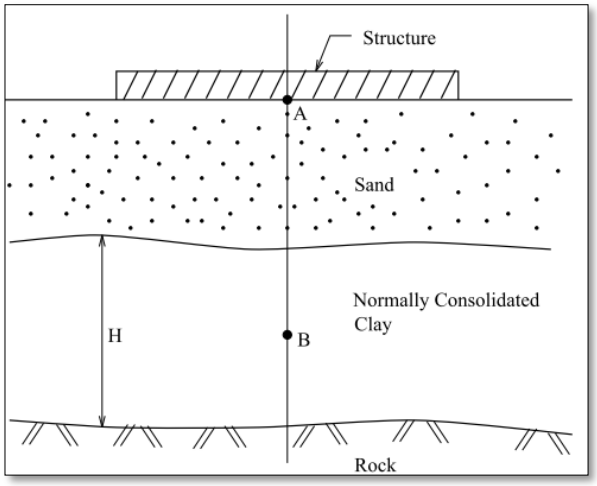


Figure 1.27. *Consolidation Settlement for a clay (Ang & Tang, 2007)*

Table 1.8 summarizes the statistical values of the different parameters.

Table 1.8 Statistical values of the different parameters (Ang & Tang, 2007)

Variable	Mean	Standard Deviation σ	COV δ
N	1	0,1	0,1
Cc	0,396	0,099	0,25
e_0	1,19	0,179	0,15
H	4,27m	8,40	0,05
p_0	178kPa	0,186	0,05
Δp	24kPa	0,1	0,2

The first step in uncertainty calculation is as follows:

If $Y = g(X_1, X_2, \dots, X_m)$, the estimates of the mean, μ_Y, and the coefficient of variation, δ_Y, of $Y$are obtained as follows:

$$\mu_Y = g(\mu_{X1}, \mu_{X2}, \dots, \mu_{Xm}) \qquad (1.36)$$

$$\delta_Y^2 = \sum_{j=1}^{m} \left(\frac{\partial g}{\partial X_j} \frac{\mu_{Xj}}{\mu_Y} \right)_{\mu}^2 \times \delta_j^2 = \sum_{j=1}^{m} S_j^2 \delta_j^2 \qquad (1.37)$$

In this case, $\mu_S = 1.6$.

Defining

$$s_j = \left(\frac{\partial S}{\partial X_j} \right) \left(\frac{\mu_{X_j}}{\mu_S} \right), \qquad (1.38)$$

the terms involved in estimating the uncertainty in Scan be computed as follows:

Table 1.9 Uncertainty calculation for S (Ang & Tang, 2007)

Variable	Mean μ_{x_i}	COV δ_j	S_j	$S_j^2\delta_j^2$	%
N	1	0,1	1	0,01	8,4
Cc	0,396	0,25	1	0,0625	52,4
e_0	1,19	0,15	-0,55	0,0068	5,7
H	4,27m	0,05	1	0,0025	2,1
p_0	178kPa	0,05	-0,94	0,0022	1,8
Δp	24kPa	0,2	0,94	0,0353	29,6

Give: δ_S =0,345,

1.9.3 Multiple Failure Modes (Retaining Wall)

Three failure modes can be identified in Figure 1.29, namely:

(i) overturning of the wall,

(ii) horizontal sliding of the wall,

(iii) bearing capacity failure of the foundation.

The system probability of failure is therefore the probability that at least one of these failure modes occurs.

Due to correlation between the failure modes, the system failure probability cannot, in practice, be determined exactly and can only be evaluated in terms of bounds. The First-Order Reliability Method (FORM) is first used to determine the probability of failure of each individual mode.

In the conventional approach, one can only compute the factor of safety for each mode; however, within a probabilistic framework, the failure probabilities of the individual modes can be combined to estimate the overall probability of failure of the system.

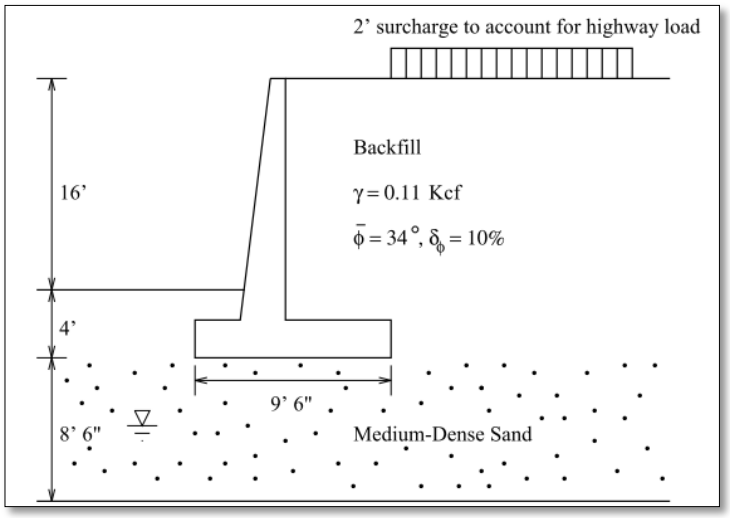


Figure 1.28. *Reliability of a Retaining Wall (Ang & Tang, 2007)*

Potential failure modes

Overturning:

$$g_1(X) = 112.5 - 195.1\, tan^{\,2}\left(45 - \frac{\varphi}{2}\right) \qquad (1.39)$$

Base sliding:

$$g_2(X) = 20.14\, tan\ \delta - 26.6\, tan^{\,2}\left(45 - \frac{\varphi}{2}\right) \qquad (1.40)$$

Bearing capacity failure of the foundation soil: negligible in this example.

For the overturning failure mode:

$p_{F1} = 0.3 \times 10^{-7}$.

For the sliding failure mode:

$p_{F2} = 0.01044$.

The first-order bounds on the failure probability are:

$$0{,}01044 \leq p_F \leq 0.01044 + 0.3 \times 10^{-7}.$$

The result shows that the probability of failure is approximately 0,01044. The first-order bound is sufficient in this case because there is a dominant failure mode, namely horizontal sliding.

1.10 Random functions

1.10.1 Regionalized variables

Geotechnical quantities vary in space and can be regarded as regionalized variables exhibiting a high degree of irregularity. They can only be represented graphically in an approximate manner. The behaviour of regionalized variables includes a random component, which requires a probabilistic interpretation (Matheron G. , 1966). The qualitative or structural characteristics of regionalized variables include spatial localization, continuity, anisotropy, and transition phenomena.

1.10.2 Localization

A regionalized variable does not take values everywhere, but within a well-defined spatial domain referred to as the geometric field. For a grade variable, for example, the geometric field Vcorresponds to the geological formation, or possibly a portion of it.

A regionalized variable is sometimes defined as a point function $Z(s)$. More often, interest lies in average values of the variable over a small domain v, referred to as the support. For a grade variable, the support corresponds to the volume of the sampled material. The support must be defined precisely, including its shape, dimensions, and orientation. Changing the support $v$or the field Vmodifies the apparent

characteristics of regionalization. For instance, a grade will appear more dispersed when defined over a smaller support or over a larger field (Matheron G. , 1966).

1.10.3 Continuity

A second essential characteristic is the degree of continuity of the regionalized variable in its spatial variation. For a given sampling network, the estimation will be more accurate when the variable is more continuous. If continuity is not ensured, one refers to a nugget effect (Matheron G. , 1966).

1.10.4 Anisotropy

Spatial directions are not always equivalent. A regionalized variable may vary slowly or smoothly along one direction while exhibiting much faster or more irregular variations along a perpendicular direction. Such behaviour is often linked to zonal structures (faults, underground galleries, etc.) (Matheron G. , 1966).

1.10.5 Transition phenomenon

In a sedimentary deposit, stratification may induce transition effects through the presence of discontinuity networks within the geometric field of regionalization. Values may remain (almost) constant within a given layer, but exhibit abrupt changes when passing from one layer to another.

1.10.6 Definition of random functions

A random function is defined by a probability law expressed over a space of functions. Any function generated according to this probability law within that function space is called a realization of the random function. The random function $Z(s)$can also be viewed as a vector random variable with an infinite number of components (the numerical values taken by $Z(s)$at each spatial location).

Thus, a regionalized variable is a realization of a random function. The regionalized variable is denoted z_s, while the random function is denoted $Z(s)$, where $s$represents the spatial position.

1.10.7 Stationarity

A random function is said to be stationary if its spatial law is invariant under translation, i.e., if the values $Z(s_1), \dots, Z(s_k)$taken at $k$locations $s_1, s_2, \dots, s_k$follow the same probability distribution as $Z(s_1 + h), \dots, Z(s_k + h)$for any translation vector h (Matheron G. , 1966).

This means that the phenomenon repeats itself indefinitely in space. Three types of stationarity are commonly used in geostatistics (Guillot, 2004): strict stationarity of order 1, second-order (weak) stationarity, and intrinsic stationarity.

1.10.7.1 Strict stationarity (order 1)

A random function $Z(s)$is first-order (strictly) stationary if its expectation (mean) is the same at every spatial location (Guillot, 2004):

$$\exists\, m \in \mathbb{R}\ such\ that\ E[Z(s)] = m, \forall s. \qquad (1.41)$$

1.10.7.2 Second-order (weak) stationarity

A random function is second-order stationary if the expectation and the covariance are invariant under translation, and the variance is assumed constant (Floch, 2018):

$$\exists m \in \mathbb{R}\ /E[Z(s)] = m \qquad \forall m \qquad (1.42)$$

$$E[Z(s)] = m(s) \forall s \qquad (1.43)$$

The variance is constant (Floch, 2018):

$$E[(Z(s) - m)^2] = \sigma^2 \qquad (1.44)$$

which implies that (Floch, 2018):

$$m(s + h) = m(s) = m \quad \forall s \qquad (1.45)$$

The covariance depends only on the spatial lag (Floch, 2018):

$$\exists C / Cov[Z(s + h), Z(x)] = E[(Z(s + h) - m)(Z(s) - m)] = C(h) \qquad (1.46)$$

When the stationarity assumptions are satisfied, the statistical description of $Z$is considerably simplified. All second-order moments can then be derived from the mean $m$and the covariance function C (Guillot, 2004):

$$C(h) \ = \ C(s - (s + h)) \qquad (1.47)$$

A covariance function is defined for all pairs of locations $s$and $(s+h)$. When $Z$is stationary, the covariance function reduces to a function of a single spatial argument, namely the lag vector h (Guillot, 2004):

$$C(s,(s+h)) = C(s-(s+h)) = C(h), \quad (1.48)$$

This covariance function Calso does not depend on the direction of the lag vector h, but only on its magnitude ‖ h ‖ (Guillot, 2004):

$$C(h) = C(|h|) \quad (1.49)$$

1.10.7.3 Intrinsic Stationarity

The intrinsic stationarity assumption is stated as follows (Matheron G. , 1971) (Floch, 2018):

$$E[(Z(s+h)-Z(s))^2] = 0 \quad (1.50)$$

A new function, called the variogram, can then be defined (Floch, 2018):

$$\gamma(s,s+h) = \gamma(h) = \frac{1}{2}Var(Z(s+h)-Z(s)) = \frac{1}{2}E[Z(s+h)-Z(s)]^2 \quad (1.51)$$

Second-order stationarity implies intrinsic stationarity, but the converse is not true. A random function may allow the variogram to be defined even when the covariance and autocorrelation functions cannot be defined (Floch, 2018).

1.11 Covariance

The covariance function makes it possible to interpret the relationship between all pairs of points. If we consider two locations $s_i$and s_j, the covariance can be defined as (Floch, 2018):

$$Cov\big[Z(s_i),Z(s_j)\big] = E\big[(Z(s_i)-m)\big(Z\big(s_j\big)-m\big)\big] \quad (1.52)$$

According to the second-order stationarity criterion, the covariance depends only on the separation (lag) between the two locations $s_i$and s_j, denoted h. The covariance function can therefore be written as (Floch, 2018):

$$C(h) = Cov[Z(s+h),Z(s)] = E[(Z(s+h)-m)(Z(s)-m)] \quad (1.53)$$

This function represents how the covariance between observations varies as a function of the separation distance (lag) h. When $h = 0$, the covariance is equal to the variance (Floch, 2018) :

$$C(0) = E[(Z(s)-m)^2] = \sigma^2 \quad (1.54)$$

he properties of the covariance function are symmetry and positive semi-definiteness (Floch, 2018):

$$C(-h) = C(h) \quad (1.55)$$

$$|C(h)| \leq C(0) \quad (1.56)$$

$$Var[\sum_{i=0}^{n} \lambda_i Z(s_i)] = \sum_{i,j=0}^{n} \lambda_i \lambda_j C(s_i - s_j) \quad (1.57)$$

1.11.1 Estimation of the Covariance Function

The covariance function is estimated from $n$pairs of points (Floch, 2018):

$$\hat{C}(h) = \frac{1}{n(h)} \sum_{i=1}^{n} (Z(s_i) - m)\,(Z(s+h) - m) \quad (1.58)$$

1.11.2 Operations on Variance and Covariance

For: $Z = f(X,Y) = aX + bY$ et $W = cX + dY$:

$$VAR(Z) = a^2 VAR(X) + b^2 VAR(Y) + 2abCOV(X,Y) \quad (1.59)$$

$$COV(Z,X) = aVAR(X) + bCOV(X,Y) \quad (1.60)$$

$$COV(Z,W) = acVAR(X) + bdVAR(Y) + adCOV(X,Y) + bcCOV(X,Y) \quad (1.61)$$

In the presence of two functions (random variables) X and Y:

$$COV(X,Y) = E[(X - E(X))(Y - E(Y))] \quad (1.62)$$

$$COV(X,Y) = E(XY) - E(X)E(Y) \quad (1.63)$$

$$COV\ (X,Y) = COV(Y,X) \quad (1.64)$$

« C » is a constant: $COV(CX,Y) = CCOV(X,Y)$ (1.65)

« C » is a constant: $COV(X + C,Y) = COV(X,Y)$ (1.66)

$$COV(X + Y,Z) = COV(X,Z) + COV(Y,Z) \quad (1.67)$$

$$COV(\sum_{i=1}^{n} X_i\ , \sum_{i=0}^{n} Y_i) = \sum\sum COV(X_i, Y_j) \quad (1.68)$$

1.12 Autocorrelation Function

The autocorrelation function is defined as a function of the lag $h$through the ratio $C(h)/C(0)$. Its value lies between $-1$and $+1$. The following relationships can be derived when second-order stationarity holds (Floch, 2018):

$$\gamma(h) = C(0) - C(h) \qquad (1.69)$$

$$\gamma(h) = \sigma^2(1 - \rho(h)) \qquad (1.70)$$

1.13 Spatial Distribution

1.13.1 Symbol plot

Spatial distribution can be examined by producing a symbol plot, in which a symbol (e.g., a cross, star, or circle) is displayed at each sampling location, with its size proportional to the observed value Z_i. Visual inspection of a symbol plot allows a qualitative detection of spatial variations in the data, often referred to as a trend or drift (Guillot, 2004).

Figure 1.1 shows an example of a symbol plot for rainfall observations from 467 meteorological stations in Switzerland recorded on May 8, 1986 (Diggle & Ribeiro, 2007). The radii of the circles are proportional to the recorded rainfall. The results indicate a heterogeneous spatial distribution with an irregular trend.

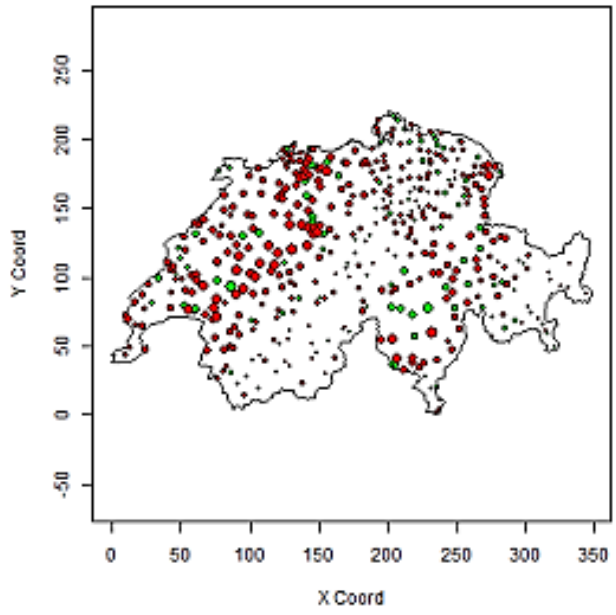


Figure 1.29. *Rainfall study in Switzerland (green: sampled observations; red: validation data) (Diggle & Ribeiro, 2007) (Floch, 2018)*

1.13.2 Variogram Cloud

The variogram cloud is used to analyze dependence between data points. It consists of a scatter of points formed by pairs

$$\left[(s_j - s_i), \frac{1}{2}(Z(s_j) - Z(s_i))^2\right] \qquad (1.71)$$

It can be displayed either as a scatter plot or as a boxplot. Figure 1.2 shows an example of a variogram cloud for rainfall measurements recorded in Switzerland (Diggle & Ribeiro, 2007) (Floch, 2018).

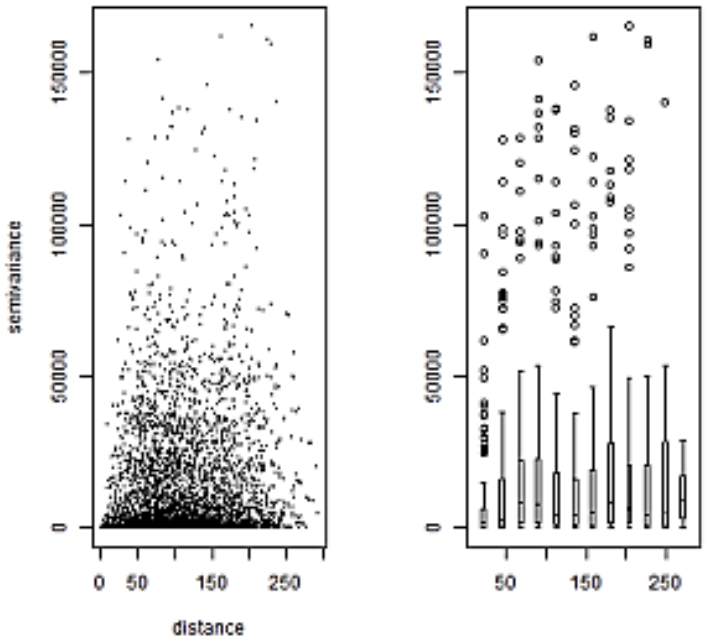


Figure 1.30. *Variogram cloud of rainfall observations in Switzerland (8 May 1986) (Floch, 2018)*

1.14 Fitting a Function

The empirical covariance function does not necessarily satisfy the positive definiteness property, and the shape of these curves is often irregular. This requires defining a theoretical covariance function that ensures both positive definiteness and a sufficiently regular behaviour.

1.14.1 Parametric Families of Covariance Functions

Several theoretical models have been proposed that ensure both positive definiteness and a well-defined standard behaviour (Figures 1.3 and 1.4) (Guillot, 2004).

a. The pure nugget model:

$$\rho(h) = I_0(h) \qquad (1.72)$$

This structure is attributable to measurement error.

The exponential variogram model:

$$\rho(h) = exp(-h/a) \qquad (1.73)$$

In the **exponential model**, the covariance decreases linearly at the origin and decays very rapidly for large values of h, without ever becoming exactly zero.

b. The Gaussian model:

$$\rho(h) = exp(-\frac{h}{a})^2 \qquad (1.74)$$

In this case, the decay near the origin is very slow (all derivatives at $h = 0$are zero), which produces a very smooth behavior.

These last two models depend on a scale parameter a, which controls the strength and range of spatial dependence: the larger $a$is, the farther spatial dependence extends. The three models correspond to a random function with unit variance, since $\rho(0) = 1$.

A covariance model with variance $\sigma^2$is obtained by multiplying a standardized correlation model by σ^2:

$$C(h) = \sigma^2 \rho(h).$$

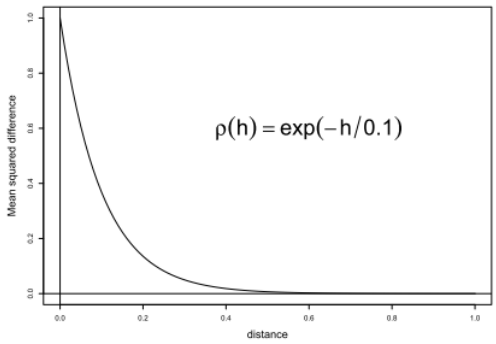
A simple calculation shows that $\sigma^2$is the limit of the variogram as $h \to +\infty$; this value is called the sill of the variogram.

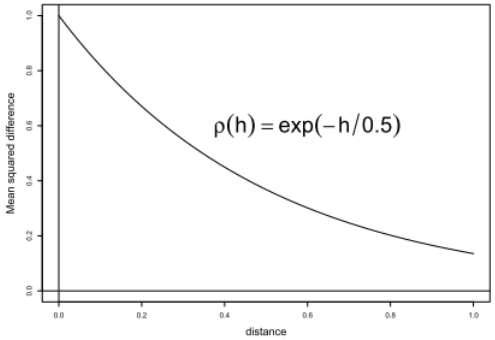
Finally, if $C_1$and $C_2$are two covariance functions, then for any positive real numbers $a_1$and a_2, the linear combination $a_1C_1 + a_2C_2$is also a covariance function, corresponding to the covariance of $\sqrt{a_1}\,Z_1 + \sqrt{a_2}\,Z_2$(Guillot, 2004).

The most commonly used model is (Guillot, 2004):

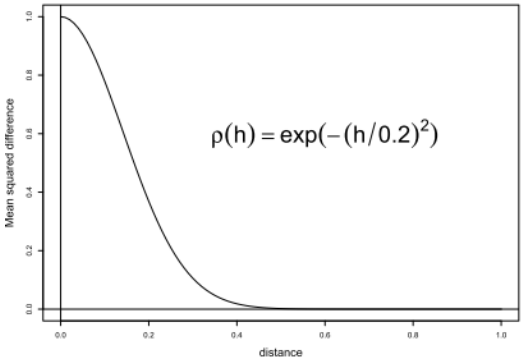
$$C(h) = \sigma^2 II_{pep}(h) + C'(h) \qquad (1.75)$$

where $C'(h)$is a covariance function that is continuous at the origin (e.g., exponential or Gaussian). This corresponds to decomposing the studied variable into a spatially structured component and a spatially unstructured component.

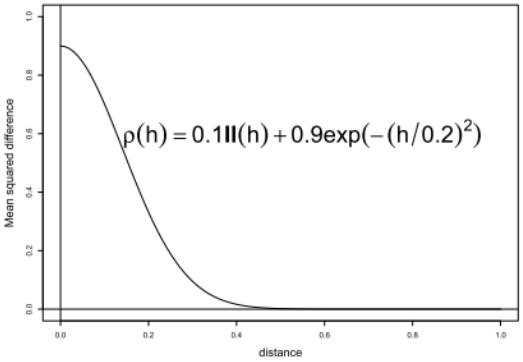




(a) Exponential model with two different ranges

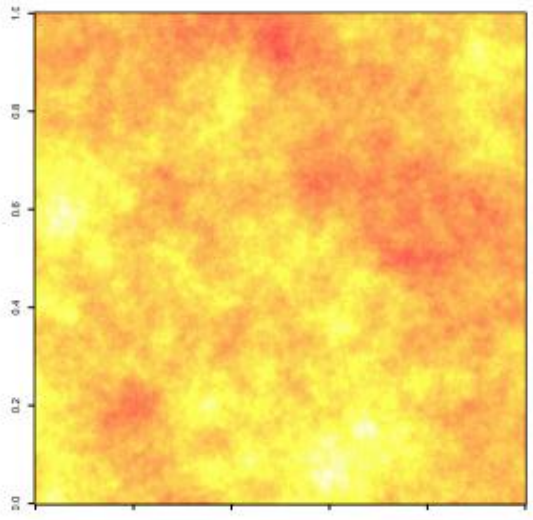


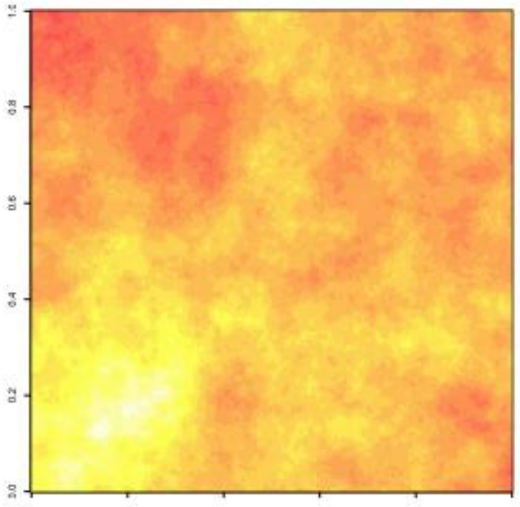
(b) Gaussian



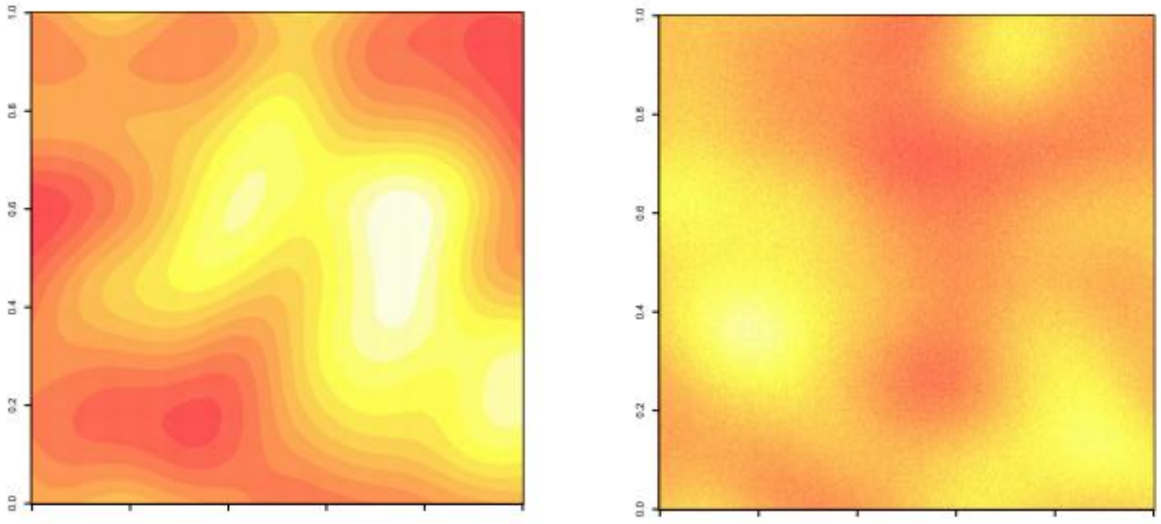
© Gaussian with nugget e effect

Figure 1.31. *Exponential model with two different ranges (Guillot, 2004)*





a. Exponential model with two different ranges



b. Gaussian c. Gaussian with a nugget effect

Figure 1.32. *Realizations of random functions (Guillot, 2004)*

1.14.2 Fitting Methods

To estimate the parameters of the theoretical covariance model, one first selects a simple parametric family—most often positive linear combinations of basic structures (Guillot, 2004):

$$C(h) = \sigma^2 II_{pep}(h) + C'(h) \qquad (1.76)$$

The two or three real parameters involved are then estimated using one of the following approaches.

1.14.2.1 Manual fitting

This is the simplest approach: the model parameters are chosen to obtain the best visual fit, with particular attention (by decreasing order of importance) (Guillot, 2004).

- To the behaviour at the origin,
- The behaviour at short distances (linear versus parabolic),
- And the distance at which the structure stabilizes

1.14.2.2 Least squares

The unknown parameter $\theta$is determined by minimizing the sum of squared errors (Guillot, 2004).

1.15 Conclusion

This first chapter has established the conceptual foundations required for any rigorous geostatistical approach. Introducing random functions made it possible to move beyond the classical deterministic framework by explicitly accounting for the spatial uncertainty inherent to geotechnical properties. In the context of geomaterials, this formalization is essential: mechanical parameters (c', φ', γ, k, m_v, etc.)are no longer treated as constants, but as realizations of a spatially correlated random field.

The stationarity assumption—whether second-order or intrinsic—was presented as a structuring mathematical framework that makes the problem tractable while preserving a consistent physical interpretation. Although simplifying, this assumption forms the basis of spatial estimation and simulation.

The study of covariance and the variogram highlighted the central role of spatial dependence structure. The variogram, in particular, is the fundamental tool for characterizing spatial continuity: it quantifies the loss of correlation with distance and enables the identification of key structural parameters such as range, sill, and nugget effect. In geotechnical applications, these parameters have direct physical meaning related to depositional processes, consolidation history, or weathering.

Finally, fitting a theoretical variogram model was introduced as a critical step linking observation to modeling. Moving from the experimental variogram to an admissible function condition the mathematical validity of subsequent steps, particularly kriging. This operation requires both statistical rigor and geological judgment in order to avoid overfitting or artificial interpretations of variability.

Overall, this chapter establishes the theoretical framework that will enable, in the following chapters, the development of advanced variographic analysis, kriging theory, and numerical applications. It constitutes the conceptual bridge between classical statistics and advanced spatial modeling, and provides an indispensable foundation for any probabilistic and reliable approach in geotechnical engineering and geomaterials science.

1.16 Exercises

The comprehension questions and numerical applications provided at the end of this chapter are intended to reinforce the fundamental notions related to random variables and the statistical analysis of variability. They are designed to assess students' understanding of the theoretical concepts, to develop proficiency with quantitative tools, and to prepare students for the introduction of spatial dependence and variographic analysis in the following chapters. The progression adopted—from knowledge recall to critical analysis—aims to foster a thorough and well-structured understanding of the fundamentals of geostatistics.

1.16.1 Comprehension Questions

1. Define a discrete random variable and a continuous random variable.
2. Explain the difference between a classical random variable and a regionalized variable.
3. Why is the concept of a regionalized variable particularly suitable for modeling geotechnical parameters?
4. Let Xbe a random variable representing soil density. Provide the mathematical definitions of the :

– Expectation,

– The variance,

– The standard deviation.

5. Explain the usefulness of the coefficient of variation in geotechnical engineering.
6. Why is variance alone insufficient to characterize a spatial phenomenon?
7. How does the natural variability of soil justify the introduction of a probabilistic approach in geotechnical engineering, and why may a purely deterministic approach be insufficient?
8. Why is it sometimes necessary to transform data (e.g., logarithmic transformation of permeability) before statistical analysis?
9. Explain why two samples taken at nearby depths may exhibit similar values of a geotechnical parameter. Does this observation challenge the assumption of data independence?
10. Consider two soils having the same mean shear strength but different standard deviations. Which soil exhibits greater variability, and what practical consequences may this have for the design of a structure?

1.16.2 Numerical Applications

1. Exercise 1: Basic descriptive statistics

Consider the undrained shear strengths (kPa): 65, 72, 80, 75, 90, 85, 78, 88.

1. Compute the mean.
2. Compute the variance.
3. Compute the standard deviation.
4. Compute the coefficient of variation.
5. Interpret the level of dispersion obtained.

2. Exercise 2: Histogram construction

Given the soil densities (kg/m³): 1580, 1600, 1620, 1650, 1670, 1690, 1700, 1720, 1750, 1780, 1800, 1820.

1. Determine the number of classes using Sturges' rule.
2. Determine the class width.
3. Construct the frequency table.
4. Identify the modal class.
5. Comment on the symmetry of the distribution.

3. Exercise 3: Logarithmic transformation

Permeability measurements (cm/s): 1.2×10^{-8}; 5.0×10^{-9}; 3.5×10^{-7}; 7.8×10^{-8}; 2.4×10^{-7}.

1. Compute the base-10 logarithm of the values.
2. Compute the mean and standard deviation before transformation.
3. Compute the mean and standard deviation after transformation.
4. Compare the dispersions.
5. Explain the benefit of the transformation.

4. Exercise 4: Geotechnical correlations

Given the pairs (depth z, undrained shear strength C_u): (2 m, 45 kPa), (4 m, 60 kPa), (6 m, 78 kPa), (8 m, 95 kPa).

1. Compute the correlation coefficient.
2. Provide a physical interpretation of the result.
3. What does strong spatial correlation mean in geotechnical engineering?

5. Exercise 5: Comparison of two soils

Two sites exhibit the following characteristics:

Site	Mean Cu (kPa)	Standard deviation (kPa)
A	100	10
B	100	30

1. Compute the coefficient of variation for each site.
2. Which site has the higher relative uncertainty?
3. Which site is potentially more critical for design?
4. Explain why the mean alone is insufficient to compare these two soils.

1.16.3 Critical Analysis

A geotechnical engineer has eight undrained shear strength tests $C_u$performed on a site intended for a building foundation: 82, 95, 88, 110, 76, 102, 90, 97 (kPa). The computed statistics are:

- Mean = 92.5 kPa,
- Standard deviation = 10.8 kPa,
- Coefficient of variation ≈ 11.7%.

The engineer decides to use the mean value (92.5 kPa) as the representative design value for shallow foundations.

1. Is this decision sufficient from a statistical standpoint? Justify your answer.
2. What important information is ignored when using only the mean?
3. How can data dispersion influence structural safety?
4. If the tests come from different depths, what additional limitation arises in this analysis?
5. How does this example demonstrate the need to go beyond descriptive statistics toward spatial dependence analysis?

*The detailed solutions to the exercises in this chapter are provided in **Appendix A***

Chapter 2.

Variogram Analysis

Variogram Analysis

After establishing the theoretical foundations related to regionalized variables and the concepts of stationarity and covariance, it becomes necessary to quantitatively characterize the spatial dependence structure of the studied phenomena. This characterization constitutes a central step in the geostatistical workflow.

The variogram is the fundamental tool used to describe the spatial organization of variability. It measures how dissimilarity between two observations evolves as a function of their separation distance and, when relevant, their direction. Variogram analysis enables the identification of key structural parameters such as the correlation range, the nugget effect, the sill, and possible anisotropies—elements that are essential for understanding the spatial behaviour of geotechnical parameters.

This chapter is devoted to constructing the experimental variogram, examining its theoretical properties, and fitting admissible models. This step is decisive, since the quality of variogram modelling directly controls the reliability of subsequent kriging estimates.

2.1 Introduction

Variogram analysis is the core step of any geostatistical study. Following the introduction in the previous chapter of random variables, stationarity, and covariance, it is now necessary to explicitly characterize the spatial dependence structure of the data. The variogram provides the fundamental framework for this characterization.

In geotechnical engineering, the physical and mechanical properties of soils typically exhibit spatial continuity: nearby points tend to have more similar values than points farther apart. This continuity—although affected by the natural heterogeneity of geological formations—forms the basis of spatial inference. The variogram quantifies this dependence by measuring the average variability between observations separated by a given lag vector (Chilès & Delfiner, 2012).

Formally, the variogram is defined as one half of the expected squared increment of a second-order stationary random function. It expresses how dissimilarity between two points evolves with distance and, potentially, direction. It therefore provides a compact representation of the spatial structure of the phenomenon under study (Cressie & Wikle, 2011).

In modern applications, the variogram plays a decisive role because it directly governs the quality of kriging estimates. Indeed, kriging weights are computed from the adopted variogram model; poor

modelling of the spatial structure may therefore lead to biased estimates or an underestimation of uncertainty (Diggle & Ribeiro, 2007; Chilès & Delfiner, 2012).

Recent developments in spatial statistics have further enriched variographic analysis by incorporating: the treatment of geometric and zonal anisotropy,

- The modelling of nested structures,
- The investigation of nugget effects associated with measurement error or micro-scale variability,
- The use of robust estimators for the experimental variogram,
- And the integration of non-stationary models in heterogeneous contexts (Banerjee et al., 2014).

Within geotechnical engineering, variogram analysis not only improves the spatial interpolation of soil parameters, but also helps to better understand the spatial organization of heterogeneities and to optimize sampling strategies.

This chapter focuses on:

- The definition of the experimental variogram,
- The relationship between the variogram and covariance,
- The fundamental properties of the variogram,
- The main theoretical models (spherical, exponential, Gaussian, etc.),
- The fitting of a theoretical model to the experimental variogram,
- And the analysis of nugget effect, range, and anisotropy.

These elements provide the essential foundation for the rigorous implementation of the kriging methods developed in the following chapter.

2.2 Experimental Variogram

Geotechnical soil parameters often exhibit substantial variability. Even within the same site, different values and interpretations may be obtained. It is generally observed that parameters measured at locations close to one another tend, on average, to be more similar than values measured farther apart.

For example, consider water content measurements collected at different locations. Although these values vary from point to point, they are not independent of their spatial location. The difference in water content Zbetween two points $x$and $x + h$tends to be smaller when the separation distance is small (Figure 2.1).

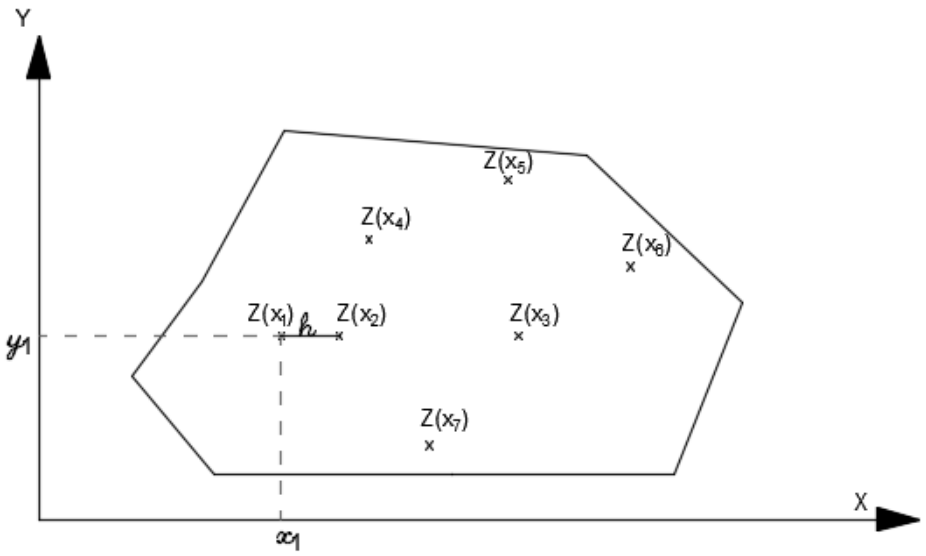


Figure 2.1. Explanatory diagram

From a statistical standpoint, the data exhibit spatial correlation. This correlation is stronger when points are closer together and when the phenomenon is continuous and smoothly varying. To quantify the degree of spatial correlation–or, more precisely, the degradation of this correlation with distance–geostatistics relies on the variogram function $\gamma(h)$. This function provides, as a function of the separation distance $h$between two locations, the mean value of $\frac{1}{2}[Z(x+h)-Z(x)]^2$. Figure 2.2 illustrates the principle of the method (Bourgine, 1996).(Bourgine, 1996)

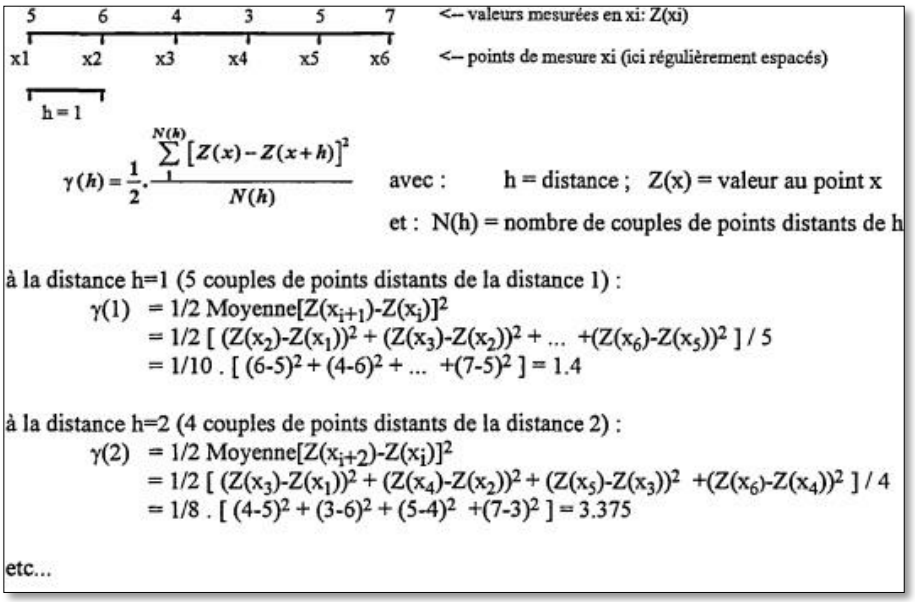
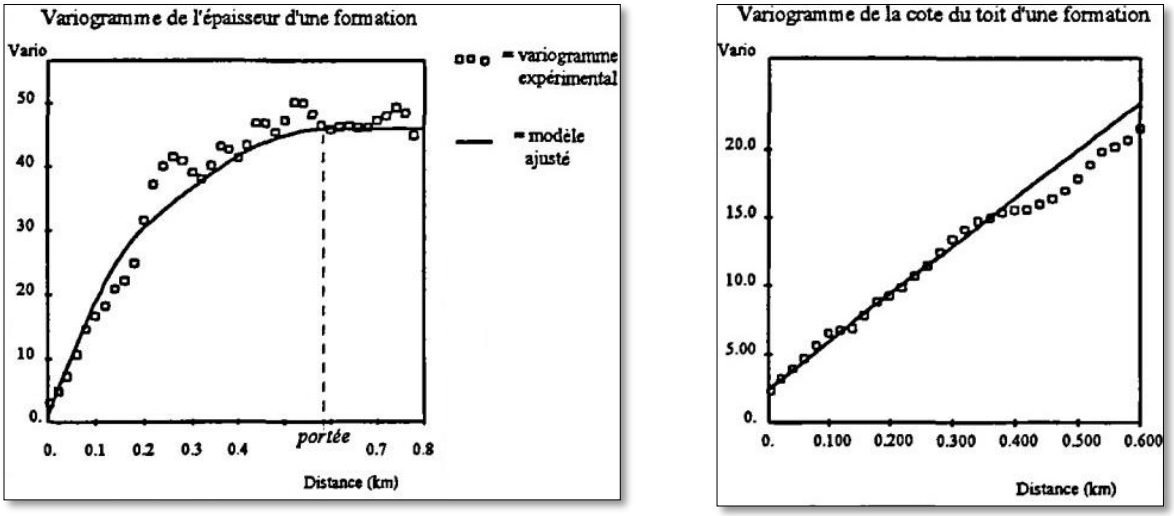


Figure 2.2. Principle of variogram computation (Bourgine, 1996).

The variogram is a key tool for analysing natural phenomena. The shape and behaviour of the experimental variogram provide a concise, synthetic view of the main characteristics of the phenomenon under study, including spatial continuity, the range (correlation distance) under stationarity conditions (Figure 2.3a), anisotropy, nested variability scales, and the absence of a finite range (Figure 2.3b), which may indicate drift or a regional trend and thus a non-stationary process. The variogram also helps define an appropriate support size and contributes to optimizing a measurement network, whether for additional site investigation or monitoring purposes (Bourgine, 1996).



(a) Stationary variogram with a finite range (b) Non-stationary variogram without a finite range

Figure 2.3. Typical examples of variograms (Bourgine, 1996)

The experimental variogram is computed from the observed data and is estimated from the set of point pairs considered:

$$\gamma(h) = \frac{1}{2N}\sum_{i=1}^{N}[Z(x_i) - Z(x_i + h)]^2 \qquad (2.1)$$

where $\boldsymbol{N(h)}$is the number of data pairs separated by the lag $\boldsymbol{h}$.

2.3 where $N(h)$is the number of data pairs separated by the lag h

Equation (2.2) and Figure 2.4 illustrate the relationship between the variogram and the covariance function.

$$\gamma(h) = \frac{1}{2}Var\left(Z(x) - Z(x+h)\right) = \frac{1}{2}\left[Var\left(Z(x)\right) + Var\left(Z(x+h)\right) - 2Cov\big(Z(x), Z(x+h)\big)\right] = \sigma^2 - Cov\big(Z(x), Z(x+h)\big) \qquad (2.2)$$

$$\gamma(h) = \sigma^2 - C(h) \qquad (2.3)$$

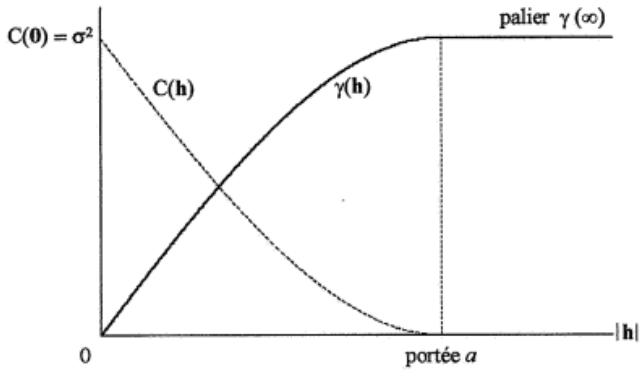


Figure 2.4. Relationship between the variogram and the covariance function (Emery, 2001).

This relationship is fundamental and is routinely used in geostatistics. When the covariance function $C(h)$is well defined, then:

$$C(h) = E[Z(x) - Z(x+h)] - m^2 \quad (2.4)$$

$$\gamma(h) = C(0) - C(h) \quad (2.5)$$

For variograms with a sill, the variogram and the covariance function provide the same information about the spatial behaviour. However, the variogram has two advantages over the covariogram (Marcotte, 2011): the variogram remains defined even when no sill exists, and its expression does not involve the constant mean m, so there is no need to estimate m, unlike when computing the covariogram directly.

2.4 Variable Transformation

A strongly asymmetric data distribution, characterized by a few very large positive values compared with many much smaller ones, can significantly affect the computation of the experimental variogram and lead to instability. To address this issue, a logarithmic transformation ("log-transformation") may be applied. For the random variable Z, one defines an associated random variable Y (Emery, 2001):

$$Y = ln\left(1 + \frac{Z}{\mu}\right) \quad (2.6)$$

where μ denotes the arithmetic mean of the data for the variable Z. This transformation mitigates the effect of very high values without accentuating the differences among low values. If Y follows a bi-Gaussian distribution, the following relationship may be established between the variogram of Z and that of Y (Emery, 2001):

$$\gamma_Z(h) = [(m_Z + \mu)^2 + VAR(Z)]\left[1 - e^{\left[-\frac{\sigma^2 \gamma_Y(h)}{VAR(Y)}\right]}\right] \quad (2.7)$$

with:

$$m_Z = E(Z) \quad (2.8)$$

$$\sigma^2 = ln\left[1 + \frac{VAR(Z)}{(\mu + m_Z)^2}\right] \qquad (2.9)$$

2.5 Theoretical Variogram Models

The experimental variogram is fitted using a theoretical model. Several models may be proposed, provided that the following conditions are satisfied (Deverly, 1984)

$$\gamma(h) = \frac{1}{2}VAR[Z(x+h) - Z(x)] \Rightarrow \begin{cases} \gamma(h) = \gamma(-h) \\ \gamma(0) = 0 \\ \gamma(h) \geq 0 \end{cases} \qquad (2.10)$$

Figure 2.5 illustrates the general shape of a variogram. The variogram increases with the lag distance h: the farther apart two samples are, the less correlated they tend to be. The more or less rapid increase of the variogram near the origin (for small h) reflects a higher or lower degree of regularity in the mineralization (or, more generally, in the spatial continuity of the phenomenon). The behavior at large distances is related to the overall amplitude of possible fluctuations (Deverly, 1984).

In most cases, the variogram reaches a sill, which corresponds to the variance of the random function $Z(x)$. Structures that generate variograms with a sill are referred to as transition phenomena (Figure 2.6). The distance beyond which the sill is reached is called the variogram range, denoted a. The range represents the zone of influence of the samples: for $h < a$, samples are correlated; for $h \geq a$, they are no longer correlated.

At very short distances, as $h$decreases toward 0, $\gamma(h)$may tend toward a non-zero value C_0, referred to as the nugget effect. This indicates the presence of a discontinuity at the origin (Deverly, 1984).

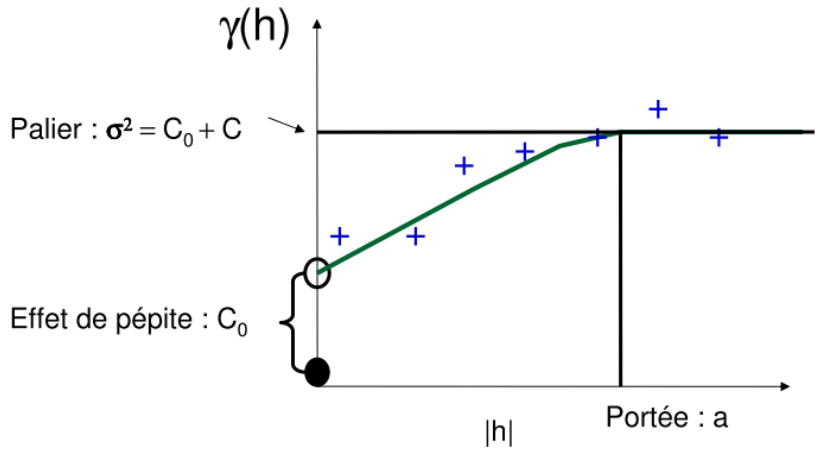
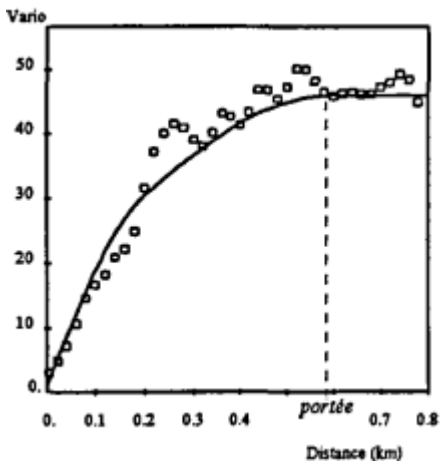
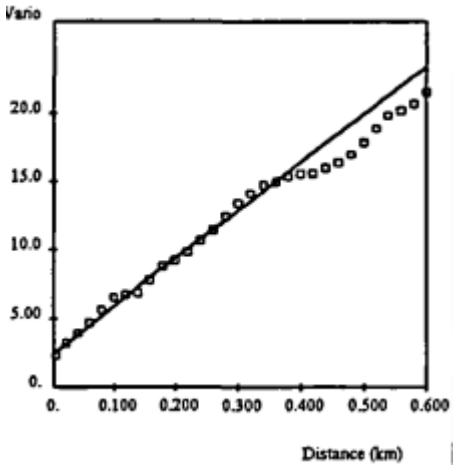


Figure 2.5. General shape of a variogram (Marcotte, 2011)



a. Stationary variogram with sill and finite range



b. Non-stationary variogram without sill or finite range

Figure 2.6. Typical variogram patterns (Bourgine, 1996)

2.5.1 Properties of the Variogram

Variograms with a sill are characterized by three parameters: the range, the nugget effect, and the sill.

2.5.1.1 The Range a

The range a is the distance beyond which two observations no longer resemble each other, on average; they are no longer linearly related (zero covariance). At this distance, the value of the variogram is equal to the variance of the random variable (Marcotte, 2011).

2.5.1.2 The Sill

The variogram increases until it reaches a limiting value, called the sill, after which it becomes flat. However, it may also increase without bound in the presence of drift. This sill is equal to the variance of the random function Z(x):

$$\gamma(\infty) = VAR[Z(x)] = C(0) \qquad (2.11)$$

Consequently, the covariance function C(h) is well defined:

$$C(h) = C(0) - \gamma(h) \qquad (2.12)$$

The experimental variance of the data σ^2 can be expressed as follows (Deverly, 1984):

$$\sigma^2 = \frac{1}{n}\sum_{n=1}^{n}\left[Z(x_i) - \frac{1}{n}\sum_{j=1}^{n} Z(x_j)\right]^2 = \frac{1}{2n^2}\sum_i \sum_j \left[Z(x_i) - Z(x_j)\right]^2 \qquad (2.13)$$

One may then write:

$$\sigma^2 = C_0 + C \qquad (2.14)$$

Sometimes variograms do not exhibit a sill. This corresponds to cases in which the covariance and the variance do not exist.

2.5.2 The Nugget Effect

The nugget effect, denoted C_0, represents very short-scale variability and reflects the presence of a discontinuity at the origin. It may be due to: (i) micro-scale structures that are not observable at the sampling scale; and (ii) measurement errors (sampling and analytical errors).

It should be emphasized that:

$$\lim_{h \to 0} \gamma(h) = C_0 \qquad (2.15)$$

$$\gamma(0) = 0 \qquad (2.16)$$

The extreme case is the pure nugget effect, for which $\gamma(h) = C_0$for all $h > 0$, with $\gamma(0) = 0$by definition. This situation characterizes a complete absence of spatial correlation.

2.6 Theoretical Variogram Models

Once the experimental variogram has been computed, a corresponding mathematical model must be selected. This model should be both operational and simple to use; this step is referred to as fitting the experimental variogram. The two main characteristics of a variogram are (Emery, 2001):

- Its behaviour at the origin, which reflects the degree of regularity of the regionalization,
- The presence or absence of a sill. The presence of a sill is synonymous with second-order stationarity; in that case, a covariance function exists and can be derived from the variogram through:

$$C(h) = \gamma(\infty) - \gamma(h) \qquad (2.17)$$

The proposed models are isotropic, depending only on $|h|$. They may be classified into three categories:

- models with a sill and a specific behavior at the origin (discontinuous, linear, or parabolic);
- models with a sill and a hole effect (e.g., cardinal-sine, Bessel Jmodel)
- models without a sill (e.g., power, linear, and logarithmic models) .

2.6.1 Sill and Transition Models

Each of these models is associated with a corresponding covariogram model (Equation (2.12)). They may be classified according to their behaviour at the origin as discontinuous, linear, or parabolic (Emery, 2001).

2.6.1.1 Discontinuous Behaviour at the Origin (Nugget-Effect Model with Sill C)

This model indicates the absence of spatial structuring, arising either from measurement errors or from the presence of a microstructure that is not experimentally detectable. It is illustrated in Figure 2.7.

$$\gamma(h) = \begin{cases} 0 & pour \quad h = 0 \\ C & pour\ h > 0 \end{cases} \qquad (2.18)$$

2.6.1.2 Linear Behaviour at the Origin

A variogram exhibiting linear behaviour at the origin increases proportionally to distance for small lags, reflecting a lower degree of spatial continuity than that of a model with parabolic behaviour. This type of behaviour is generally associated with the spherical and exponential models and indicates structured variability without immediate discontinuity near the origin (Figure 2.8).

A. Spherical model with range a and sill C:

$$\gamma(h) = \begin{cases} C\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3}\right) & pour\ 0 \le h \le a \\ C = \sigma^2 & pour\ h \ge a \end{cases} \qquad (2.19)$$

B. Exponential model:

Unlike the spherical model, which reaches its sill at $h = a$, the exponential model reaches its sill only asymptotically. However, a practical range equal to $3a$may be defined, for which the variogram reaches 95% of its sill value.

$$\gamma(h) = C\left(1 - e^{-h/a}\right) \qquad \forall h \ge 0 \qquad (2.20)$$

2.6.1.3 Parabolic behaviour at the origin

A variogram exhibiting parabolic behaviour at the origin increases proportionally to the square of the distance for small lags, reflecting strong regularity and a high degree of spatial continuity in the phenomenon under study. This behaviour is particularly characteristic of the Gaussian model and indicates the absence of local discontinuities, together with a notably smooth spatial structure near the origin (Figure 2.9).

A. Cubic model with range a and sill C

$$\gamma(h) = \begin{cases} C\left(7\frac{h^2}{a^2} - \frac{35}{4}\frac{h^3}{a^3} + \frac{7}{2}\frac{h^5}{a^5} - \frac{3}{4}\frac{h^7}{a^7}\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \qquad (2.21)$$

B. Gaussian model with range a and sill C

The sill is reached asymptotically, and the practical range may be taken as $a\sqrt{3}$.

$$\gamma(h) = C(1 - e^{-h^2/a^2}) \qquad (2.22)$$

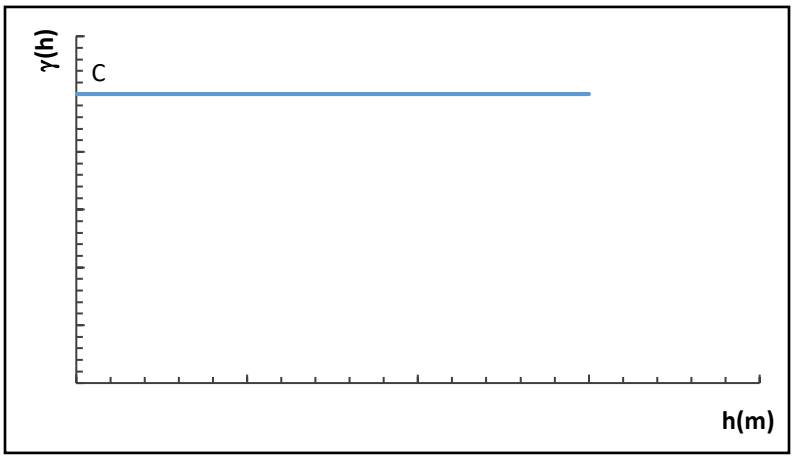
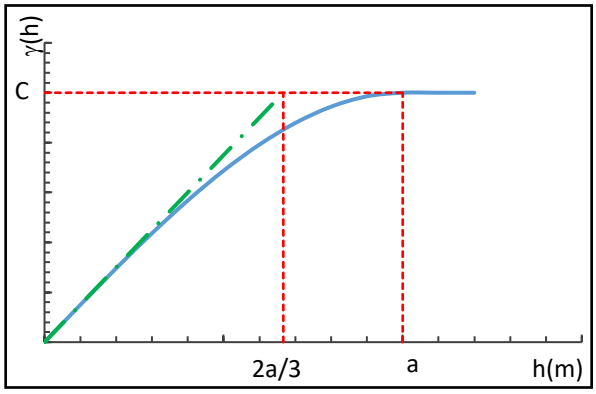
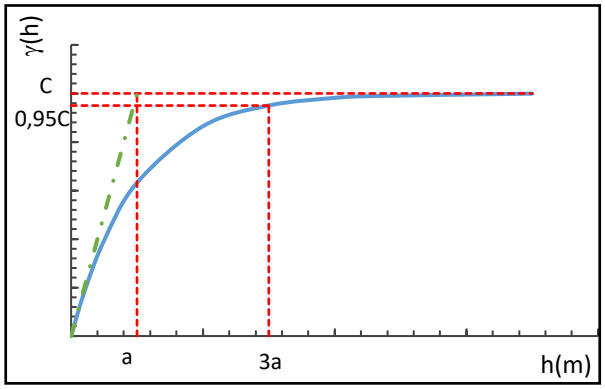


Figure 2.7. Discontinuity at the origin (pure nugget-effect model)

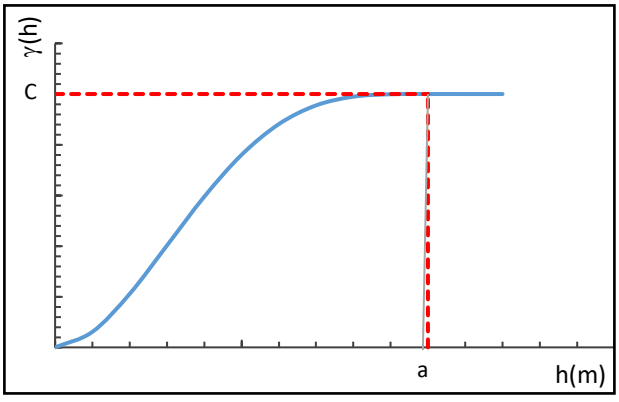


a. Spherical model

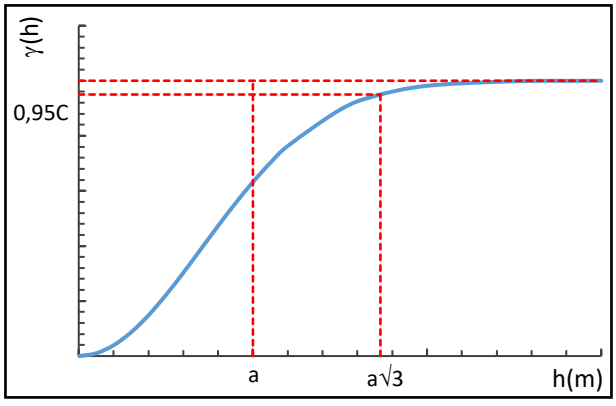


b. Exponential model

Figure 2.8. Linear behaviour at the origin



a. Cubic model



b. Gaussian model

Figure 2.9. Parabolic behaviour at the origin

2.6.1.4 Other Transitional Models

Other transitional models with a sill, which are less commonly used, are also available (Figure 2.10) (Emery, 2001):

A. Triangular Model

The associated covariance is equal, up to a multiplicative factor, to the geometric covariogram of a segment of length a,

$$\gamma(h) = \begin{cases} C\frac{h^{\alpha}}{a} & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \qquad (2.23)$$

with: $0 < \alpha \le 1$

B. Quadratic model

$$\gamma(h) = \begin{cases} C\left(2\frac{h}{a} - \frac{h^2}{a^2}\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \qquad (2.24)$$

C. Circular model

$$\gamma(h) = \begin{cases} C\left[\frac{2}{\pi}\frac{h}{a}\sqrt{1-\frac{h^2}{a^2}} + \frac{2}{\pi}arcsin\left(\frac{h}{a}\right)\right] & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \qquad (2.25)$$

D. Pentaspherical model

$$\gamma(h) = \begin{cases} C\left(\frac{15}{8}\frac{h}{a} - \frac{5}{4}\frac{h^3}{a^3} + \frac{3}{8}\frac{h^5}{a^5}\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \qquad (2.26)$$

E. Stable model

The corresponding practical range for this model is $a\sqrt[\alpha]{3}$. The exponential and Gaussian models are special cases of this model, corresponding to the parameter values $\alpha = 1$and $\alpha = 2$, respectively.

$$\gamma(h) = C\left[1 - exp\left(-\frac{h^{\alpha}}{a^{\alpha}}\right)\right] \qquad avec\ 0 < \alpha \le 2 \qquad (2.27)$$

F. Gamma model

This model is defined by a sill C, a scale factor a, and a parameter α. Its practical range is given by $a\left(\sqrt[\alpha]{20}-1\right)$. The gamma model with $\alpha = 1$corresponds to a hyperbolic model.

$$\gamma(h) = C\left[1 - \frac{1}{\left(1+\frac{h}{a}\right)^{\alpha}}\right] \qquad avec\ \alpha > 0 \qquad (2.28)$$

G. Cauchy model

This model is defined by a sill C, a scale factor a, and a parameter α. Its practical range is given by $a\sqrt{\sqrt[\alpha]{20}-1}$.

$$\gamma(h) = C\left[1 - \frac{1}{\left(1+\frac{h^2}{a^2}\right)^{\alpha}}\right] \qquad avec\ \alpha > 0 \qquad (2.29)$$

H. Modified Bessel model

his model is defined by a sill C, a scale factor a, and a parameter α.

$$\gamma(h) = C\left[1 - \frac{1}{2^{\alpha-1}\Gamma(\alpha)}\left(\frac{h}{a}\right)^{\alpha} K_{\alpha}\left(\frac{h}{a}\right)\right] \qquad avec\ \alpha > 0 \qquad (2.30)$$

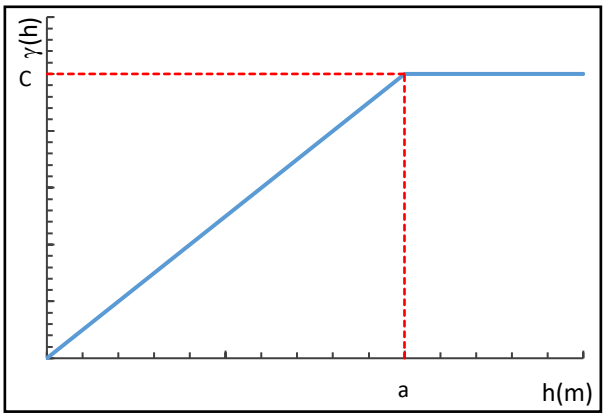
where:

Γ is the Euler gamma function, which interpolates the factorial function;

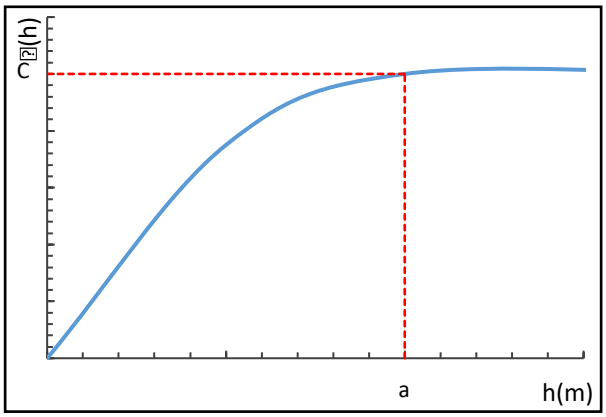
K_{α} is the modified Bessel function of the second kind of order α, defined as follows:

$$K_{\alpha}(u) = \frac{\pi}{2sin\ (\alpha\pi)}\left[\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(-\alpha+k+1}\left(\frac{u}{2}\right)^{2k-\alpha} - \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(\alpha+k+1)}\left(\frac{u}{2}\right)^{2k+\alpha}\right] \qquad (2.31)$$

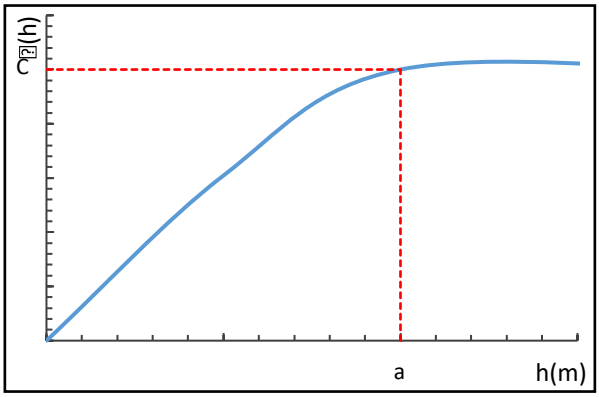
For $\alpha = \frac{1}{2}$, this reduces to the exponential model.



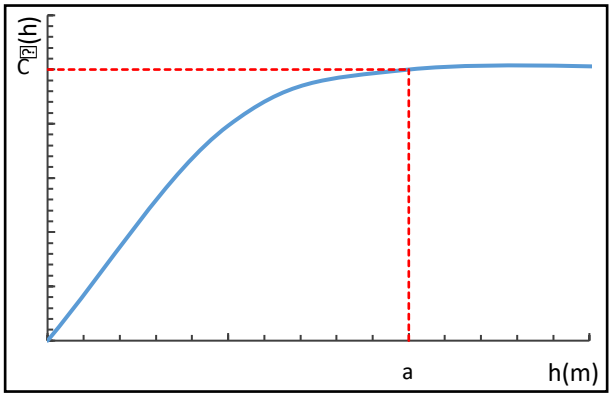
a- Triangular model



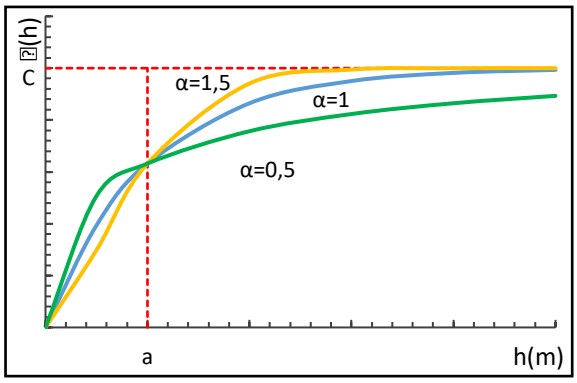
b- Quadratic model



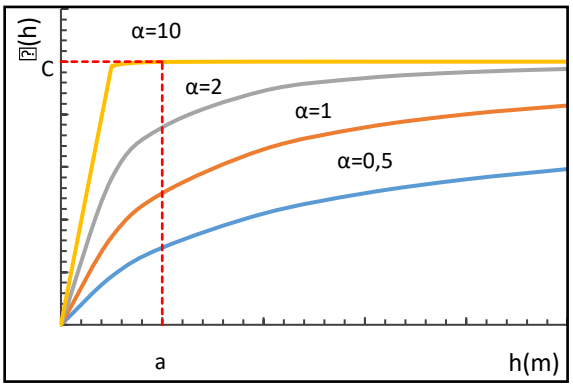
c- Circular model



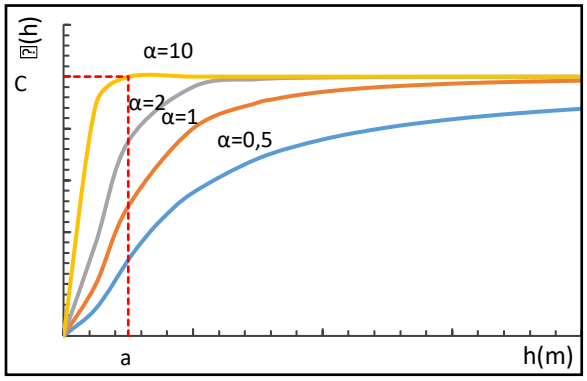
d- Pentaspherical model



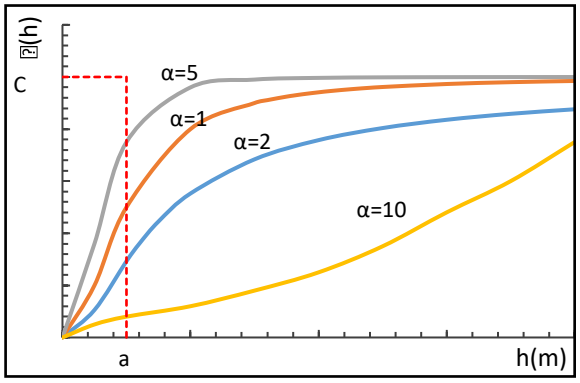
e- Stable model



f- Gamma model



g- Cauchy model



h- Bessel model

Figure 2.10. Transitional models with a sill

2.6.2 Models without a sill

These models correspond to strictly non-stationary random functions (Figure 2.11) (Emery, 2001).

2.6.2.1 Power model

The variogram for this model is given by:

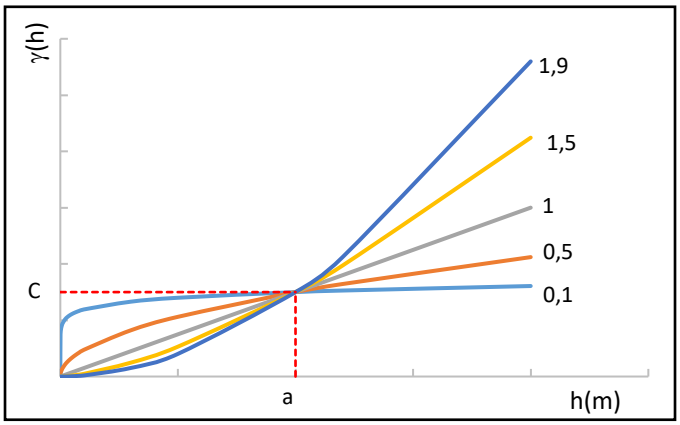
$$\gamma(h) = Ch^{\theta} \qquad avec\ 0 < \theta < 2 \qquad (2.32)$$

The parameter $\theta$is related to the degree of continuity of the regionalized variable. In the limit as $\theta \to 0$, the variogram approaches a pure nugget effect. Conversely, as θincreases and approaches 2, the behavior at the origin becomes close to parabolic, indicating that the random function $Z(s)$becomes increasingly smooth (Emery, 2001).

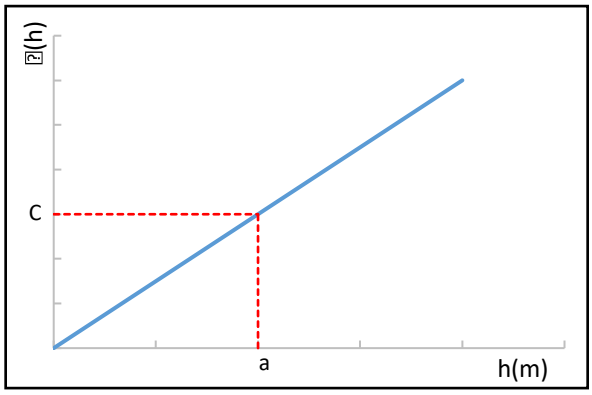
2.6.2.2 Linear model

This is a special case of the power model with exponent equal to 1 (Emery, 2001).

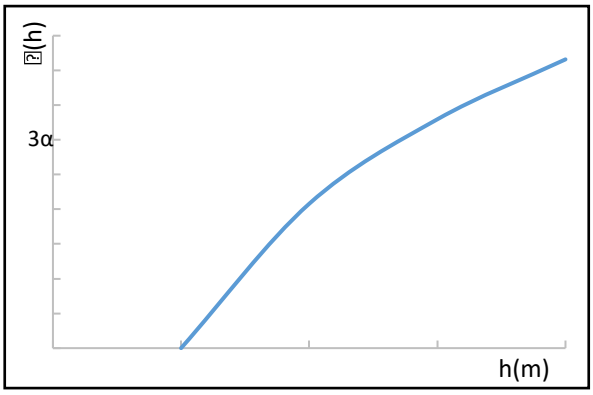
$$\gamma(h) = Ch \qquad (2.33)$$



a- Power model



b- Linear model



c- Logarithmic model

Figure 2.11. Models without a sill

2.6.2.3 Logarithmic model

he logarithmic model is quite particular, because it tends to $-\infty$as $h \to 0$, which contradicts the variogram definition. In fact, it does not refer to a random function, but rather to a random distribution. However, averages computed over non-point supports become random functions. This means that the logarithmic model is only applicable to describe non-point variables (physically, no measurement is truly pointwise, as it is always performed over a support of nonzero volume (Emery, 2001).

$$\gamma(h) = 3\alpha ln\ (h) \qquad (2.34)$$

where $\alpha$is a positive scalar called the absolute dispersion.

2.6.3 Hole-effect models

A hole effect occurs when the variogram is not monotonic but exhibits one or more oscillations. These oscillations often have a physical interpretation, such as a damped periodic phenomenon (Emery, 2001).

2.6.3.1 Periodic or semi-periodic models with parabolic behaviour at the origin

Several models exhibit periodic or semi-periodic behaviour with a parabolic form near the origin, including the cosine model, damped cosine model, cardinal-sine model, and the Bessel J model. These models are illustrated in Figure 2.12 (Emery, 2001).

A. Cosine model

This variogram oscillates indefinitely; it has neither a finite range nor a practical range. The associated random function is a perfect sinusoid with period and amplitude $2C$, with a random phase(Emery, 2001):

$$\gamma(h) = C[1 - cos\ (2\pi h/a] \qquad (2.35)$$

B. Damped cosine model

The parameters involved are a, b, α, and the sill C (Emery, 2001):

$$\gamma(h) = C\left[1 - \left(e^{-bh^{\alpha}} cos\ (2\pi \frac{h}{a}\right)\right] \qquad avec\ 0 < \alpha \leq 2 \qquad (2.36)$$

C. Cardinal-sine model (sinc model)

The parameters involved are $a$and the sill C (Emery, 2001):

$$\gamma(h) = C\left[1 - \frac{sin(h/a)}{h/a}\right] \qquad (2.37)$$

For this variogram, the practical range is $20.37a$, and the half pseudo-period is $4.49a$, at which distance the variogram equals $1.21C$. The ratio of this value to the sill, which measures the amplitude of the hole effect, is the maximum that can be obtained with an isotropic three-dimensional model (Emery, 2001).

D. Bessel J model

The parameters involved are a, α, and the sill C (Emery, 2001):

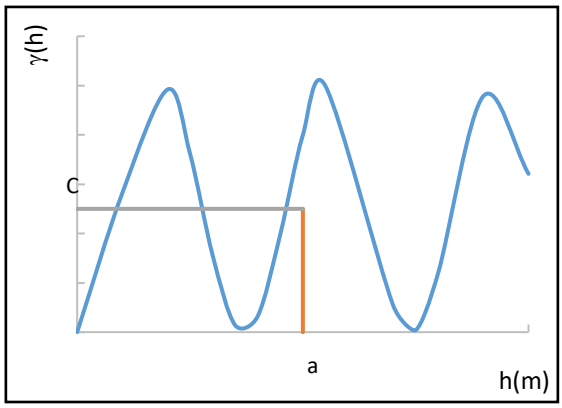
$$\gamma(h) = C\left[1 - \left(\frac{h}{2a}\right)^{-\alpha} \Gamma(\alpha + 1) J_{\alpha}\left(\frac{h}{a}\right)\right] \qquad (2.38)$$

where:

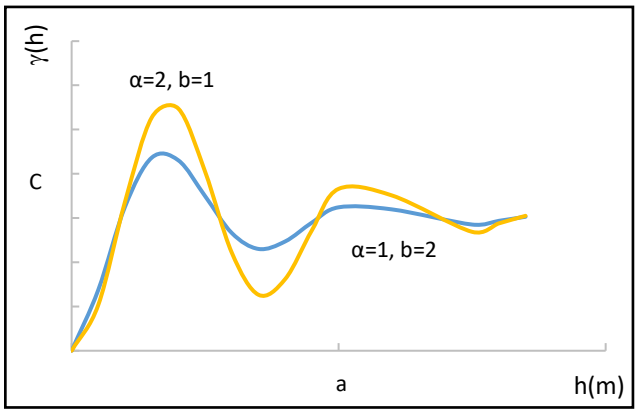
- Γ is the Euler gamma function, which interpolates the factorial function;
  - J_{α} is the Bessel function of the first kind of order α, defined as follows:

$$J_{\alpha}(u) = \left(\frac{u}{2}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(\alpha+k+1)} \left(\frac{u}{2}\right)^{2k} \qquad (2.39)$$

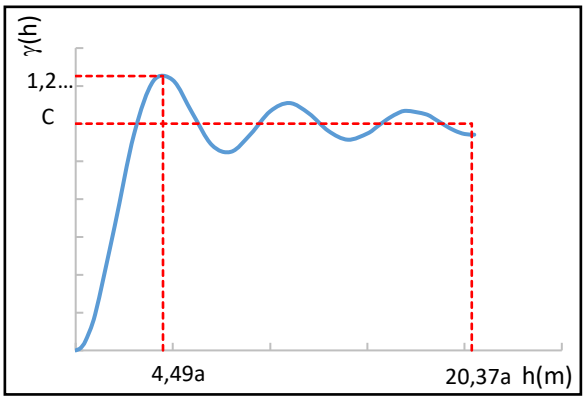
The cosine and cardinal-sine models constitute special cases of this model, corresponding to the parameter values $\alpha = -0.5$and $\alpha = 0.5$, respectively (Emery, 2001).



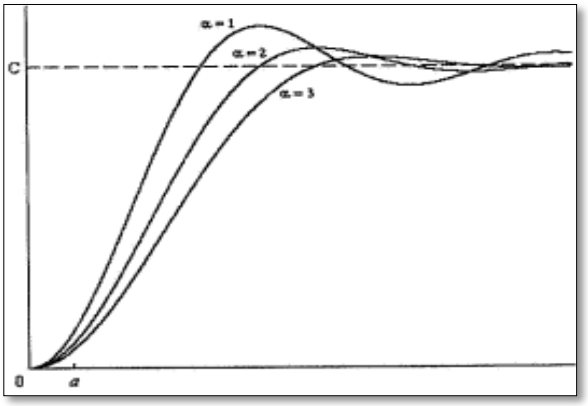
a. Cosine model



b. Damped cosine model



c. Cardinal Sin model



d. J-Bessel model

Figure 2.12. Periodic or semi-periodic models with parabolic behaviour at the origin (hole effect)

2.6.3.2 Periodic or Semi-Periodic Models with Linear Behaviour at the Origin

A. Truncated Polynomial Models

These models exhibit a range $a$and a sill C (Emery, 2001):

$$\gamma(h) = \begin{cases} C\left(\frac{5}{2}\frac{h}{a} - \frac{5}{2}\left(\frac{h}{a}\right)^3 + \left(\frac{h}{a}\right)^5\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \quad (2.40)$$

$$\gamma(h) = \begin{cases} C\left(\frac{35}{12}\frac{h}{a} - \frac{35}{8}\left(\frac{h}{a}\right)^3 + \frac{7}{2}\left(\frac{h}{a}\right)^5 - \frac{25}{24}\left(\frac{h}{a}\right)^7\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \quad (2.41)$$

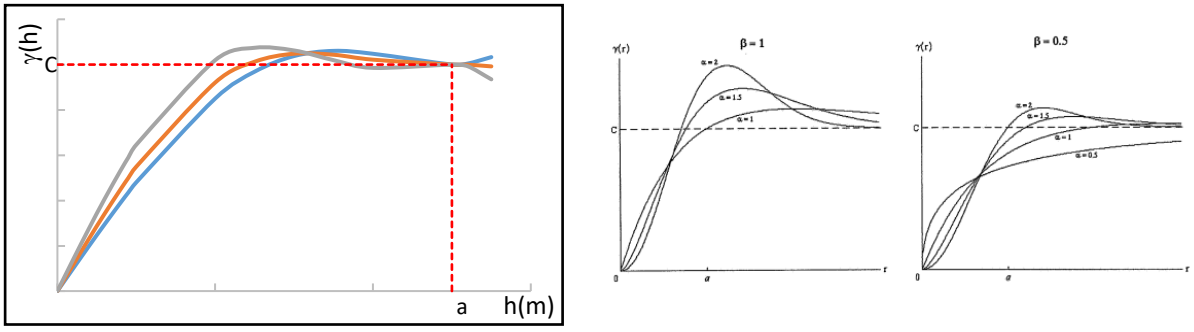
$$\gamma(h) = \begin{cases} C\left(\frac{7}{2}\frac{h}{a} - 7\left(\frac{h}{a}\right)^3 + 7\left(\frac{h}{a}\right)^5 - \frac{5}{2}\left(\frac{h}{a}\right)^7\right) & pour\ 0 \le h \le a \\ C & pour\ h \ge a \end{cases} \quad (2.42)$$

B. Generalized Stable Model

The parameters defining this variogram are a, α, β, and the sill C(Figure 2.13) (Emery, 2001):

$$\gamma(h) = C\left[1 - \left[[1 - \beta(h/a)^{\alpha}]exp[-(h/a)^{\alpha}]\right]\right] \quad (2.43)$$

The hole effect becomes more pronounced as αapproaches 2 and βapproaches α. When $\beta = 0$, the model reduces to the stable model. The parameter αcontrols the behavior at the origin (through h^{α}), whereas βgoverns the amplitude of the hole effect for a fixed value of α. The parameter $a$is a scale factor related to the practical range of the model (Emery, 2001).



a. Truncated polynomial models b. Generalized stable models

Figure 2.13. Periodic or semi-periodic models with linear behaviour at the origin

2.7 Variographic Analysis: Fitting a Model to an Experimental Variogram

The experimental variogram cannot be used directly. First, it is defined only for a limited set of lag distances and is therefore incomplete. Second, it is necessarily non-negative. The aim is therefore to determine a theoretical variogram model that best fits the experimental variogram. This stage, referred to as structural analysis or variographic analysis, constitutes a critical step in any geostatistical study, as an incorrect interpretation may lead to erroneous results (Chile's & Delfiner, 2012).

The theoretical models presented above may be combined to account for different scales of variability (nested structures) and to incorporate short-range variability through the nugget effect.

Various types of anisotropy may also be encountered, making it difficult to model the spatial variation properly; such situations must therefore be reduced to an isotropic case for modelling purposes.

2.7.1 Nested Structures

Natural phenomena frequently exhibit several scales of variability acting simultaneously. The theoretical variogram may therefore result from the superposition of several elementary structures, known as nested structures. For example, if the data are considered at the meter scale, the variogram may incorporate variability acting at the millimeter and centimeter scales, whereas kilometer-scale variation may not be reflected. A hierarchy therefore exists among these structures, which are embedded within one another, hence the term “nested” (Chile's & Delfiner, 2012).

The variographic analysis of a nested structure commonly relies on a widely used model, namely the linear model of regionalization, in which the variogram $\gamma(h)$is expressed as the sum of basic variograms (Deverly, 1984)(Figure 2.14):

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) + \cdots + \gamma_s(h) \qquad (2.44)$$

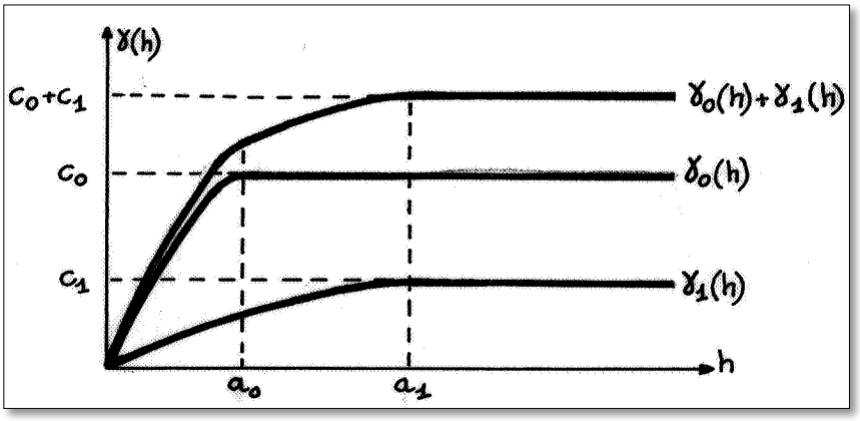


Figure 2.14. Nested Structures (Deverly, 1984)

2.7.2 Nugget Effect

The nugget effect represents a discontinuity in the variogram, arising either from measurement errors or from the presence of a microstructure (Chile's & Delfiner, 2012).

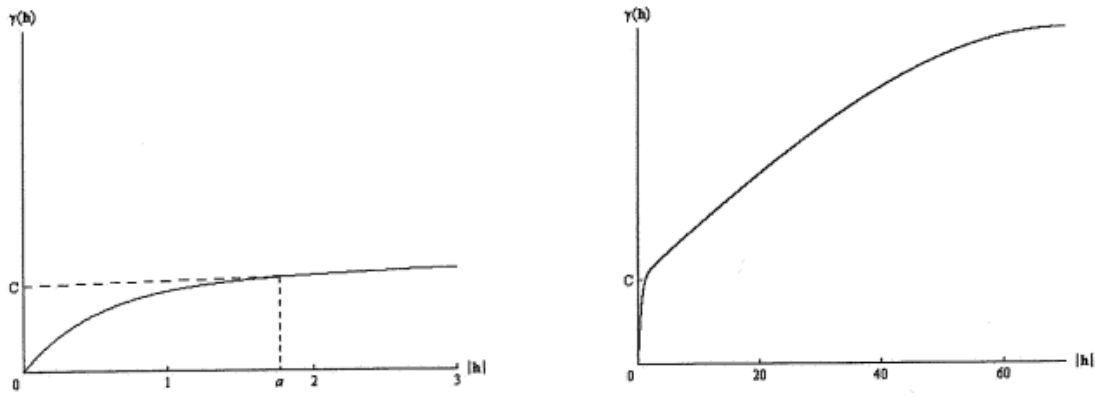
2.7.2.1 Measurement Errors

The measurement of geotechnical parameters, $Z(s)$, is associated with measurement errors, $\varepsilon(s)$. These errors are assumed to be uncorrelated with one another and with the values of $Z(s)$, and they are characterized by a zero mean and a variance denoted by C_0. Thus, one may write:

$$\gamma_{Z+\epsilon}(h) = \begin{cases} \gamma_Z(h) + C_0 & pour\ h \neq 0 \\ 0 & pour\ h = 0 \end{cases} \qquad (2.45)$$

2.7.2.2 Presence of a Microstructure

In the linear model of regionalization, the concept of scale is fundamental. Consider, for instance, a nested structure composed of two components: the first has a centimetric range $a$and a sill C, while the second has a kilometric range (Figure 2.15). In this case, the variogram $\gamma(h)$exhibits, near the origin, a region of increase that reaches the sill Cvery rapidly (within only a few centimeters). At the kilometer scale, this increase becomes indistinguishable from a discontinuity at the origin, i.e., a nugget effect of amplitude C (Chile's & Delfiner, 2012).



a- Microscopic scale b- Macroscopic scale

Figure 2.15. Nugget effect (Chile's & Delfiner, 2012)

2.7.3 Anisotropy

Anisotropy arises when a formation exhibits different degrees of variability depending on direction. In such cases, the behavior of the variogram varies with spatial direction. Since theoretical variogram models are generally formulated for the isotropic case, it is therefore necessary to consider the transformations that allow anisotropic behavior to be represented.

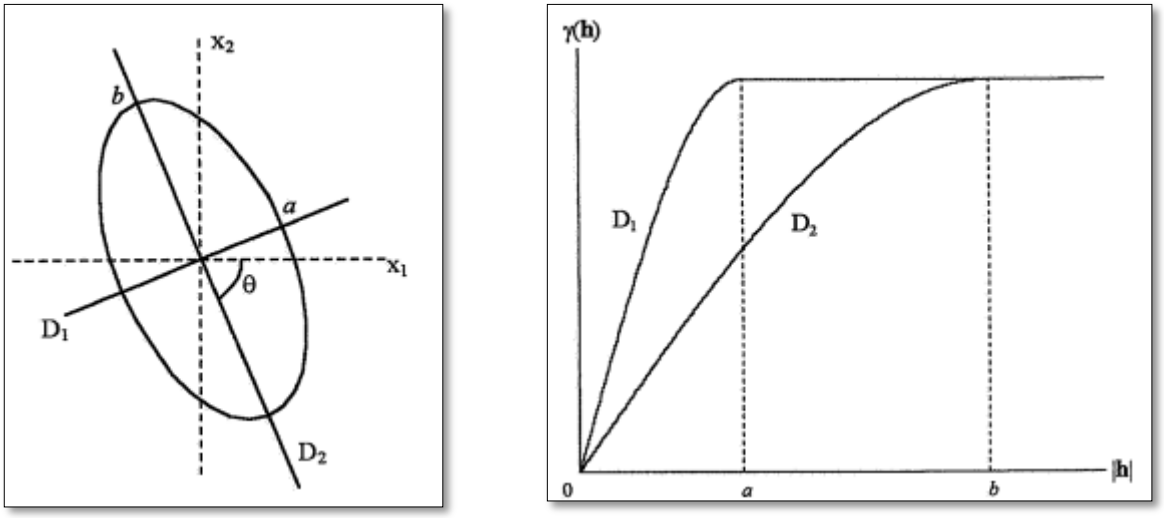
In practice, anisotropy is identified by comparing experimental variograms computed along several directions; for example, in the two-dimensional case, along directions oriented at $0°$, $45°$, $90°$, and $135°$relative to the x-axis. This analysis is often complemented by constructing a variogram map, i.e., a contour map of the experimental variogram values as a function of the vector h(therefore of both its magnitude and direction). Under isotropic conditions, the experimental variograms in different directions coincide, and the variogram map takes the form of circles, or spheres in the three-dimensional case. The variogram then depends only on the length of the vector h.

Otherwise, the phenomenon is anisotropic. Several types of anisotropy can be distinguished, in particular geometric anisotropy and zonal anisotropy (Deverly, 1984) (Emery, 2001) (Chile's & Delfiner, 2012).

2.7.3.1 Geometric Anisotropy

Anisotropy is said to be geometric when the variogram map consists of concentric ellipses (Figure 2.16a). The directional variograms then have the same sill in all directions, but different ranges from one direction to another (Figure 2.16b). A simple linear transformation of the coordinates is then sufficient to reduce the problem to an isotropic case.

Geometric anisotropy is defined by two parameters: the anisotropy ratio and the anisotropy angle. The anisotropy ratio is given by a/b, where $a$denotes the minor axis and $b$the major axis. It measures the intensity of the anisotropy, which becomes more pronounced as this ratio moves away from 1. The anisotropy angle $\theta$is the angle between the major axis (the direction of b) and the x-axis, and it defines the orientation of the anisotropy (Deverly, 1984) (Emery, 2001) (Chile's & Delfiner, 2012).



a. Range ellipse

b. Variograms along the principal directions of anisotropy

Figure 2.16. Example of geometric anisotropy with anisotropy ratio $a/b$and angle θ (Emery, 2001)

The equivalent range can be calculated using the following equation (Marcotte, 2011):

$$\left(\frac{a_\theta cos\theta}{b}\right)^2 + \left(\frac{a_\theta sin\theta}{a}\right)^2 = 1 \qquad (2.46)$$

$$a_\theta = \frac{ab}{\sqrt{(acos\theta)^2 + (bsin\theta)^2}} \qquad (2.47)$$

Accordingly, $\gamma(h,\theta)$may be evaluated either by using $a_\theta$or by adjusting the distance $h$to account for anisotropy (Marcotte, 2011).

$$\gamma(h_\theta, \theta) = \gamma(h_g) \qquad (2.48)$$

$$h_g = \sqrt{(h_\theta cos\theta)^2 + \left(\frac{b}{a} h_\theta sin\theta\right)^2} \qquad (2.49)$$

2.7.3.2 Zonal Anisotropy

The variograms exhibit different sills and ranges depending on the directions along which they are computed. Zonal anisotropy may, for example, be modelled using nested structures, each characterized by

its own variability and anisotropy. The treatment of anisotropy should ultimately lead to the definition of an equivalent isotropic variogram (Figure 2.17).

A typical example of zonal anisotropy is hydraulic head, which is inherently anisotropic. Maximum variations are observed in the direction of flow, whereas, in the direction perpendicular to the flow, the hydraulic head remains constant.

Modelling zonal anisotropy is generally quite delicate and requires substantial experience. The simplest representation of zonal anisotropy consists in combining one or several isotropic components with a component exhibiting geometric anisotropy for which $b$is infinite (Marcotte, 2011).

$$\gamma_{zonal}(h,\theta) = \gamma_{isotrope}(h) + \gamma_p(hsin\theta) \qquad (2.50)$$

where the subscript p denotes the anisotropic model corresponding to the direction of minimum range.

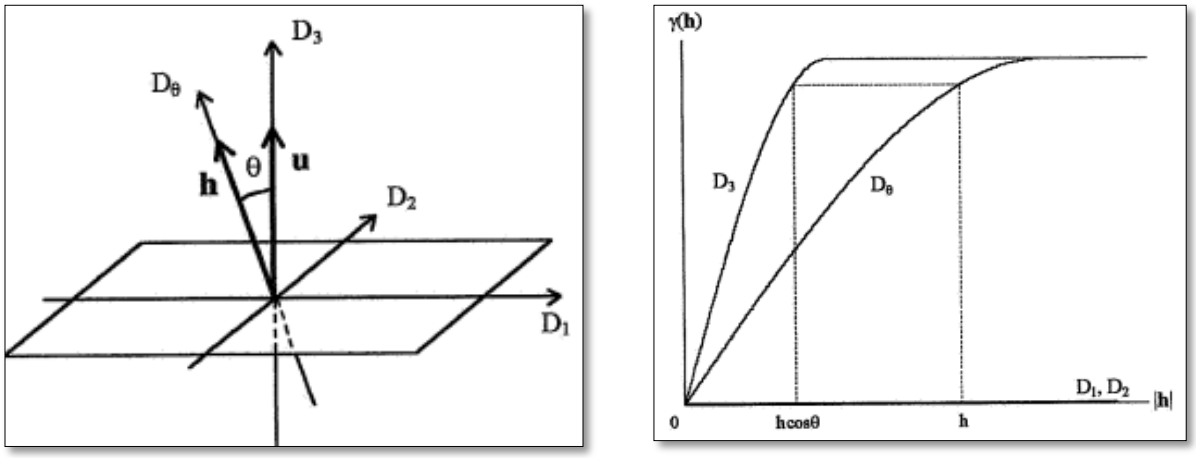


Figure 2.17. Example of vertical zonal anisotropy (Marcotte, 2011).

2.8 Mean and Regularized Variograms

The average value of the variogram $\gamma(h = x_1 - x_2)$, when the point $x_1$moves over the volume $V_1$and the point $x_2$over the volume V_2, is called the mean variogram and is denoted by $\bar{\gamma}(V_1, V_2)$. It is defined as:

$$\bar{\gamma}(V_1, V_2) = \frac{1}{V_1 V_2}\int_{V_1}\int_{V_2} \gamma(x_1 - x_2)dx_1 dx_2 \qquad (2.51)$$

The regularized variogram $\gamma_V(h)$describes the variability observed over supports V. It is expressed in terms of the point variogram as follows (Matheron & Blondel, 1962) (Deverly, 1984):

$$\gamma_V(h) \approx \gamma(h) - \bar{\gamma}(V,V) \qquad (2.52)$$

2.9 Calculation of Estimation and Dispersion Variances

Estimation and dispersion variances can be expressed as functions of the variogram.

2.9.1 Estimation Variance

The aim is to establish the results that provide a measure of the accuracy of estimates obtained by any (linear) estimation method.

Let $Z_v$be a random variable to be estimated (here, the subscript $v$indicates that the estimation concerns a block; if $v = 0$, the estimation refers to a point (Matheron & Blondel, 1962) (Deverly, 1984) (Chile's & Delfiner, 2012):

$$Z_v^* = \sum_{i=1}^{n} \lambda_i Z_i \qquad (2.53)$$

where:

Z_i: observed value at point x_i;

Z_v^*: estimator of Z_v.

The estimation error is defined as follows (Chile's & Delfiner, 2012):

$$e = Z_v - Z_v^* \qquad (2.54)$$

The variance of this error is the estimation variance(Chile's & Delfiner, 2012):

$$Var\ (e) = VAR(Z_v) + Var(Z_v^*) - 2Cov(Z_v, Z_v^*) \qquad (2.55)$$

Substituting the expression of Z_v^* in terms of the Z_i yields(Chile's & Delfiner, 2012):

$$\sigma_e^2 = Var(Z_v) + \sum_i \sum_j \lambda_i \lambda_j Cov\big(Z_i, Z_j\big) - 2\sum_i \lambda_i Cov(Z_i, Z_v) \qquad (2.56)$$

This expression can be rewritten in terms of the variogram (Deverly, 1984):

$$\sigma_e^2 = 2\sum_i \lambda_i \bar{\gamma}(x_i, v) - \bar{\gamma}(v, v) - \sum_i \sum_j \lambda_i \lambda_j \gamma\big(x_i, x_j\big) \qquad (2.57)$$

The estimation variance may therefore be computed either from the covariogram or from the variogram. In the case of models without a sill, only the variogram can be used, provided that $\sum \lambda_i = 1$ (Chile's & Delfiner, 2012).

The estimation variance thus depends on (Matheron & Blondel, 1962) (Deverly, 1984):

- the variability of the parameters, through the variogram;
- the geometry of the samples (their volume, shape, and, where appropriate, the relative arrangement of the elementary units constituting v), through the term $\bar{\gamma}(v, v)$;
- the geometry of the volume to be estimated, through the term $\bar{\gamma}(V, V)$;
- the relative geometry of the sample volume $v$with respect to the volume V, through the term $\bar{\gamma}(V, v)$; this term depends on the number and spatial arrangement, relative to volume V, of the units constituting volume v.
- The variance of the estimation error is therefore a measure of the accuracy of the estimate. One may seek to choose the $\lambda_i$so that $\sigma_e^2$is minimized. This is precisely the principle underlying kriging.,

2.9.2 Remarks on Variogram Computation and Fitting

The computation and fitting of a variogram require the verification of a number of conditions (Marcotte, 2011) (Chile's & Delfiner, 2012):

- The experimental variogram should be computed using the maximum possible number of defined pairs.
- For each point of the experimental variogram, a minimum of 30 pairs should be used ($N(h) \geq 30$). If this is not possible for certain lag classes, less weight should be assigned to those points. If the number of pairs is very small (≤ 10), the point should be discarded altogether.
- The first points of a variogram are the most informative (small h), since these values have the greatest influence on geostatistical calculations.
- When $h$exceeds approximately $d_{\max}/2$, the corresponding variogram values are generally disregarded, where $d_{\max}$denotes the size of the phenomenon under study in the direction considered.
- The simplest possible models that adequately reproduce the experimental values should be selected.

2.10 Conclusion

Variogram analysis constitutes a decisive step in any geostatistical study. It enables the quantitative characterization of the spatial dependence structure of a phenomenon and describes the continuity or discontinuity of the properties under investigation through a consistent mathematical model.

Although the experimental variogram provides an initial representation of spatial variability, it remains a discrete estimate affected by uncertainty. It cannot therefore be used directly in estimation procedures. Fitting an admissible theoretical model that satisfies the conditions of mathematical validity is thus an essential phase, referred to as structural analysis or variographic analysis. This step directly governs the quality of subsequent kriging estimates, as well as the evaluation of the associated uncertainties.

The study of the fundamental variogram parameters (nugget effect, range, and sill) makes it possible to interpret physically the structure of the medium under investigation. Likewise, identifying possible anisotropies, whether geometric or zonal, contributes to a more faithful representation of the spatial organization of heterogeneities.

Variogram modelling is therefore not merely a technical step, but a genuine stage of scientific interpretation of the phenomenon under study. An incorrect identification of the spatial structure may lead to biased estimates and an underestimation of uncertainty, whereas a properly fitted model ensures the consistency and robustness of the results.

The concepts established in this chapter provide the direct foundation for the implementation of the optimal estimation methods developed in the following chapter, in particular kriging, whose formulation is entirely based on the adopted variogram model.

2.11 Exercises

The exercises proposed in this section are intended to reinforce understanding of the principles of variographic analysis and to develop mastery of its methodological aspects. They are designed to familiarize the student with the computation of the experimental variogram, the identification of its characteristic parameters, and the physical interpretation of the results obtained.

The proposed progression addresses, in sequence, computational aspects, structural analysis, and critical reflection on the choice of theoretical models. This approach provides direct preparation for the implementation of kriging-based estimation methods developed in the following chapter.

2.11.1 Comprehension Question

1. Define the experimental variogram and give its mathematical expression.
2. What is the physical meaning of the vector $h$in variographic analysis?
3. What is the value of the variogram for two samples taken side by side ($h \to 0$)in the case of a perfectly continuous phenomenon, and then in the case of a phenomenon exhibiting a nugget effect? Provide a physical interpretation of these two situations.
4. Explain why the variogram is an increasing function of distance for a spatially correlated phenomenon.
5. Why cannot the experimental variogram be used directly in kriging?
6. Explain the difference between the spherical, exponential, and Gaussian models in terms of behaviour at the origin and continuity of the phenomenon.
7. How can the presence of anisotropy be identified from directional variograms?
8. What is the difference between geometric anisotropy and zonal anisotropy?
9. Why can a poor interpretation of the variogram lead to biased estimates in kriging?

2.11.2 Numerical Applications

1. Exercise N°01: Computation of an Experimental Variogram

Consider measurements of a parameter Zalong a profile (in m):

X (m)	0,10	20	30
Z (x)	12,15	14	18

Compute the experimental variogram $\gamma(h)$for $h = 10$m and $h = 20$m using:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2$$

Plot the resulting experimental variogram (points $h, \gamma(h)$) and comment on its evolution.

2. Exercice 2 : Construction by Distance Classes

Six points are available on a site (1D coordinates in m), with the following values:

X (m)	0,5	12	18
Z (x)	10,11	9	13

A discretization into distance classes of width 10 m is adopted:

Class 1: $(0,10]$

Class 2: $(10,20]$

Class 3: $(20,30]$

Class 4: $(30,40]$

1. Compute $\gamma(h)$for each class, taking $h$as the class center.

2. Indicate, for each class, the number of pairs $N(h)$.

3. Conclude on the influence of the number of pairs on the stability of the variogram.

3. Exercise 3: Estimation of Model Parameters from a Variogram

An experimental (omnidirectional) variogram provides the following values:

$$\gamma(10) = 2{,}0$$
$$\gamma(20) = 5{,}0$$
$$\gamma(30) = 7{,}5$$
$$\gamma(40) = 8{,}5$$
$$\gamma(50) = 9{,}0$$
$$\gamma(60) = 9{,}1$$

1. Assume that the sill is reached beyond 50-60 m.

2. Determine a reasonable estimate of the nugget effect C_0, the sill $C_0 + C$, and the range a.

3. Propose a compatible spherical or exponential model and justify your choice.

4. Exercise 4 : Directional Variograms and Anisotropy

Experimental variograms computed along two orthogonal directions on a 2D site are available:

Direction $0°$: estimated range $a_0 = 60$m, sill ≈ 12

Direction $90°$: estimated range $a_{90} = 25$m, sill ≈ 12

1. Identify the type of anisotropy (geometric or zonal).

2. Determine the anisotropy ratio.

3. Propose a simple geometric transformation to reduce the problem to an isotropic case.

5. Exercise 5: Model Selection and Continuity at the Origin

Consider two theoretical models that may be fitted to an experimental variogram:

- Model A: Gaussian model with a small nugget effect
- Model B: Spherical model with a non-negligible nugget effect

In a geotechnical context where measurements come from in situ tests exhibiting measurement variability and micro-heterogeneity, discuss which of the two models is the more realistic. Support your answer by referring to the behaviour of the variogram at the origin, the nugget effect, and the expected impact on kriging estimates.

2.11.3 Critical Analysis

A geotechnical study has 22 values of a parameter Z(for example, C_u) obtained from boreholes distributed over a site. The analyst computes an omnidirectional experimental variogram and quickly fits a spherical model “by eye.” A high nugget effect and a short range are obtained. On this basis, ordinary kriging is performed and a very smoothed estimation map is produced. Cross-validation indicates a mean error close to zero, but large local errors are observed in certain areas.

You are asked to assess this approach critically.

1. Identify the preliminary checks that should have been carried out before the variographic analysis (nature of the data, outliers, possible transformation, presence of trend).
2. Explain how a high nugget effect may arise from different causes (measurement error, micro-variability, poor discretization of distance classes, mixing of populations), and indicate how these causes may be distinguished.
3. Discuss the influence of selecting the model “by eye” (model type, range, sill, nugget) on kriging weights, the smoothing effect, and the estimation variance.
4. Propose a procedure for validating and improving the variogram model (directional variograms, anisotropy analysis, comparison of several models, cross-validation criteria, sensitivity to parameters).
5. Conclude on the technical risks of a poor interpretation of the variogram in a geotechnical context (poor localization of weak zones, under- or overestimation of uncertainty, inappropriate site investigation or design decisions).

*The detailed solutions to the exercises in this chapter are provided in **Appendix B***

Chapter 3. Kriging Theory

Kriging Theory

The characterization of spatial structure through variographic analysis naturally leads to the central issue of geostatistics: the optimal estimation of a variable at an unsampled location. Kriging theory constitutes the methodological culmination of this approach, as it uses the variogram model to produce rigorously founded estimates.

Kriging is based on the formalism of random functions and on the search for an unbiased linear estimator with minimum variance. It simultaneously incorporates the geometric configuration of the data and the spatial correlation structure previously identified. This dual consideration distinguishes kriging from purely deterministic interpolation methods.

This chapter presents the mathematical foundations of kriging by successively introducing its assumptions, its different formulations, particularly simple kriging and ordinary kriging, and the properties of the resulting estimator. Particular attention is given to the decisive role of the variogram model in the quality of the estimates and in the assessment of the associated uncertainty

3.1 Introduction

After establishing the theoretical foundations of regionalized variables and analysing the spatial structure of phenomena through the variogram, it is appropriate to address the central problem of geostatistics: the optimal estimation of a variable at an unsampled point from a finite set of observations. This problem finds its formal formulation in kriging theory.

The term kriging was introduced by Georges Matheron in tribute to the pioneering work of the South African mining engineer Danie G. Krige in the field of mineral resource estimation (Matheron G. , 1966). Developed within the framework of random function theory, kriging is now regarded as one of the most rigorous and accomplished spatial interpolation methods (Chile's & Delfiner, 2012).

Unlike deterministic interpolation methods, kriging is based on the interpretation of the regionalized variable as a realization of a random function whose spatial dependence structure has been modelled by means of the covariance function or the variogram. The estimation is then formulated within a probabilistic framework aimed at determining an unbiased linear estimator with minimum variance (Cressie & Wikle, 2011). The variogram model established in the previous chapter plays a decisive role, since it directly governs the calculation of the weights assigned to neighbouring data.

Kriging offers major advantages. On the one hand, it simultaneously incorporates geometric information (the number and configuration of measurement points) and the structural information contained in the variogram model. On the other hand, it provides a quantitative measure of the uncertainty associated with the estimate through the kriging variance, thereby making it possible to assess the precision of the results (Emery, 2001) (Chile's & Delfiner, 2012). This ability to quantify uncertainty is of fundamental importance in geotechnical engineering, where risk control is essential.

Depending on the assumptions adopted regarding the mean of the phenomenon under study, different kriging formulations may be distinguished. Simple kriging assumes a known and constant mean, whereas ordinary kriging considers an unknown but locally constant mean (Wackernagel, 2003). These two approaches form the basis of classical kriging theory and will be developed in this chapter.

The objective is to establish the systems of equations that allow the computation of the estimation weights, to analyse the mathematical properties of the resulting estimator, particularly in terms of unbiasedness and minimum variance, and to examine the practical implications in terms of precision and smoothing effect.

Thus, kriging theory represents the logical culmination of the geostatistical approach: once spatial variability has been modelled, this structure can be exploited to produce optimal, consistent, and quantitatively justified estimates of geotechnical parameters at any point within the domain under study.

3.2 Computation of Kriging Weights

Assume that $Z(s)$is intrinsically stationary, that its variogram $\gamma(h)$is known, but that its mean $m$is unknown. Let the available data be $Z = [Z(s_1), \dots, Z(s_i), \dots, Z(s_N)]$. The objective is to predict the value of Zat an unsampled location and thus estimate $Z(s_0)$. The kriging estimator is defined as a linear combination of the observations (Guillot, 2004):

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \qquad (3.1)$$

Such an estimator is statistically satisfactory if it is unbiased and if the variance of the estimation error is small. The objective is therefore to determine weights $\lambda_i$that ensure zero bias while minimizing the variance. This can be written as follows (Guillot, 2004):

$$E\left[\hat{Z}_{s_0} - Z_{s_0}\right] = 0 \qquad (3.2)$$

$$Var\left[\hat{Z}_{s_0} - Z_{s_0}\right] = min\left\{Var\left[\sum_i \lambda_i Z(s_i) - Z_{s_0}\right], \lambda \in \mathbb{R}^n\right\} \qquad (3.3)$$

3.2.1 Simple Kriging

If $Z$is assumed to have zero mean, the bias is written as follows (Guillot, 2004):

$$E\left(Z_{s_0}^{k_s}\right) = E\left(\sum_i \lambda_i Z_{s_i}\right) = \sum_i \lambda_i E\left(Z_{s_i}\right) = 0 \qquad (3.4)$$

Condition (2.57) is therefore automatically satisfied. The second condition requires that the weights $\lambda_i$be the solution of the following matrix system:

$$C\lambda = C_0 \qquad (3.5)$$

where Cdenotes the covariance matrix between the random variables at the measurement points, that is:

$$C = \left(C(s_i - s_j)\right)_{i,j} \qquad (3.6)$$

Denoting:

$$\lambda^{ks} = C^{-1}C_0 \qquad (3.7)$$

The solution to our problem is written as follows:

$$Z^{ks}(s_0) = (Z_1, \dots, Z_n)\lambda^{ks} \qquad (3.8)$$

3.2.2 Ordinary Kriging

Suppose that the objective is to estimate a block $v$centered at point x_0. Let $Z_v$denote the true (unknown) value of this block, and $Z_v^*$the corresponding estimator. The estimator is linear (Equation (2.51))(Guillot, 2004):

$$Z_v^* = \sum_{i=1}^n \lambda_i Z_i \qquad (3.9)$$

where the $Z_i$denote the random variables corresponding to the sampling points. The objective is to minimize:

$$\sigma_e^2 = Var(Z_v - Z_v^*) = Var(Z_v) + Var(Z_v^*) - 2Cov(Z_v, Z_v^*) \qquad (3.10)$$

Substituting the expression of the estimator into this equation yields Equation (2.54):

$$\sigma_e^2 = Var(Z_v) + \sum_i \sum_j \lambda_i \lambda_j Cov\left(Z_i, Z_j\right) - 2\sum_i \lambda_i Cov(Z_i, Z_v) \qquad (3.11)$$

For the estimator to be unbiased, it is necessary that (Guillot, 2004):

$$\sum_i \lambda_i = 1 \qquad (3.12)$$

Indeed, in this case (Floch, 2018):

$$E(Z_v^*) = \sum_i \lambda_i E(Z_i) = \sum_i \lambda_i m = m \qquad (3.13)$$

We therefore have a quadratic (and thus convex) minimization problem under an equality constraint, which is solved using the Lagrange multiplier method. The Lagrangian is then defined as (Floch, 2018):

$$L(\lambda) = \sigma_e^2 + 2\mu(\sum_{i=1}^n \lambda_i - 1) = Var(Z_v) + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Cov\left(Z_i, Z_j\right) - 2\sum_{i=1}^n \lambda_i Cov(Z_v, Z_i) + 2\mu(\sum_{i=1}^n \lambda_i - 1) \qquad (3.14)$$

where μ is the Lagrange multiplier. The minimum is attained when all partial derivatives with respect to each λ_i and with respect to μ vanish. This leads to the ordinary kriging system. (Guillot, 2004):

$$\sum_{j=1}^n \lambda_j Cov\left(Z_i, Z_j\right) + \mu = Cov(Z_v, Z_i) \qquad \forall i = 1 \dots n \qquad (3.15)$$

$$\sum_{j=1}^{n} \lambda_j = 1 \qquad (3.16)$$

The minimum estimation variance, called the kriging variance, is obtained by substituting the kriging equations into the general expression for the estimation variance (Floch, 2018):

$$\sigma_k^2 = Var(Z_v) - \sum_{i=1}^{n} \lambda_i Cov(Z_v, Z_i) - \mu \qquad (3.17)$$

This kriging variance does not depend on the observed values; it depends only on the variogram and on the configuration of the points used for estimation relative to the point (or block) to be estimated (Chile's & Delfiner, 2012)

3.2.2.1 Ordinary Kriging System Written in Terms of the Variogram

Since the estimation variance can also be expressed directly in terms of the variogram, the kriging system may likewise be written in terms of the variogram, noting that:

$$C(h) = \sigma^2 - \gamma(h) \qquad (3.18)$$

and that:

$$\sum_{i=1}^{n} \lambda_i = 1 \qquad (3.19)$$

Then:

$$\sum_{j=1}^{n} \lambda_j \gamma(x_i, x_j) - \mu = \bar{\gamma}(v, x_i) \qquad \forall i = 1 \dots n \qquad (3.20)$$

$$\sum_{j=1}^{n} \lambda_j = 1 \qquad (3.21)$$

$$\sigma_k^2 = \sum_{i=1}^{n} \lambda_i \bar{\gamma}(v, x_i) - \bar{\gamma}(v, v) - \mu \qquad (3.22)$$

3.2.2.2 Matrix Formulation of Ordinary Kriging

It is useful to express the ordinary kriging system and the ordinary kriging variance in matrix form:

$$K_0 \lambda_0 = k_0 \qquad (3.23)$$

$$\sigma_{k_0}^2 = \sigma_v^2 - \lambda'_0 k_0 \qquad (3.24)$$

Where:

$$K_0 = \begin{bmatrix} \sigma^2 & Cov(Z1, Z2) & Cov\,(Z1, Zn) & 1 \\ Cov\,(Z2, Z1) & \sigma^2 & Cov\,(Z2, Zn) & 1 \\ Cov\,(Zn, Z1) & Cov\,(Zn, Z2) & \sigma^2 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \qquad (3.25)$$

$$K_0 = \begin{bmatrix} Cov\,(Z1, Zv) \\ Cov\,(Z2, Zv) \\ Cov\,(Zn, Zv) \\ 1 \end{bmatrix} \qquad (3.26)$$

$$\lambda_0 = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \\ \lambda_n \\ \mu \end{bmatrix} \qquad (3.27)$$

All terms $\mathrm{Cov}(Z_i, Z_v)$and $\mathrm{Var}(Z_v)$are obtained directly from the variogram by numerical integration (Chile's & Delfiner, 2012).

Considering that:

$$vZ_v = \frac{1}{v}\int_v Z(x)dx \qquad (3.28)$$

where $x$denotes a location within the block v. By exploiting the linearity of the expectation operator, it can be shown that:

$$Cov(Z_i, Z_v) = \frac{1}{v}\int_v Cov(Z_i, Z(x))dx \qquad avec\ x \in V \qquad (3.29)$$

and

$$Var(Z_v) = \frac{1}{v^2}\int_V \int_V Cov(Z(x), Z(y))dxdy \qquad x, y \in V \qquad (3.30)$$

3.3 Examples of Ordinary Kriging

The following cases are presented solely to provide some intuition regarding the behaviour of kriging. A spherical variogram with a finite range a is assumed.

3.3.1 Estimation of a Point from Another Point Located at a Distance h

$$(Si\ h>a, \sigma_{ko}^2 = 2\sigma^2\ )\lambda_1 = 1, \sigma_{ko}^2 = 2(\sigma^2 - C(h)) = 2\gamma(h) \qquad (3.31)$$

3.3.2 Estimation of a Point Located at $x_0$from Two Points Located at $x_1$and x_2

$$\lambda_1 = \frac{\sigma^2 + C(x0,x1) - C(x1,x2) - C(x0,x2)}{2(\sigma^2 - C(x1,x2))}\ ;\ \lambda_2 = \frac{\sigma^2 + C(x0,x2) - C(x1,x2) - C(x0,x1)}{2(\sigma^2 - C(x1,x2))} \qquad (3.32)$$

3.3.3 Estimation of a Point from nPoints in the Presence of a Pure Nugget-Effect Variogram

$$\lambda_i = \frac{1}{n}, et\ \sigma_{ko}^2 = \frac{(n+1)}{n}\sigma^2 \qquad (3.33)$$

3.4 Properties of Kriging

The main properties and characteristics associated with kriging are as follows:

- It is linear, unbiased, and of minimum variance by construction.
- It is an exact interpolator: if a known point is estimated, the known value is recovered.
- It exhibits a screening effect: the nearest points receive the greatest weights. This screening effect varies depending on the configuration of the data and on the variogram model used for kriging. The larger the nugget effect, the weaker the screening effect.
- It takes into account the size of the domain to be estimated and the relative positions of the points.
- Through the use of the variogram, it accounts for the continuity of the phenomenon under study (nugget effect, anisotropy, etc.).
- It generally produces a smoothing effect, i.e., the estimates are less variable than the true values (point or block) being estimated, while remaining almost conditionally unbiased. This means that when a cutoff grade is applied to estimated values, the predicted grade is approximately recovered. This is a very important property in mining. It implies that the estimator used is smoother than the value it seeks to estimate, which is indeed the case for kriging.
- It is transitive: if a value observed at a given point coincides with the kriged value at that point, then the kriged values at other points are not modified by including this new point in the kriging process. The kriging variances, however, are reduced. Likewise, if a certain number of points are kriged and the kriged values are then used as if they were new data, subsequent kriging estimates remain unchanged (except for the kriging variance) (Matheron & Blondel, 1962).

3.5 Exact Interpolator

Examples of 1D kriging interpolation using different variogram models (Figures 3.1, 3.2, 3.3, and 3.4) (Rivoirard, 2003):

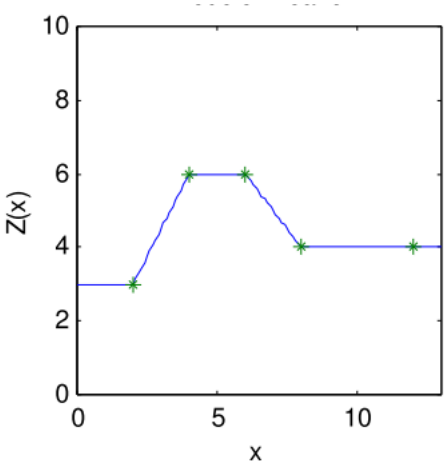


Figure 3.1. Linear model (Rivoirard, 2003)

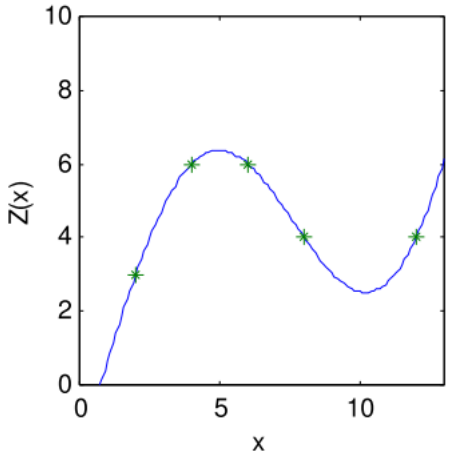


Figure 3.2. Gaussian model (a=10) (Rivoirard, 2003)

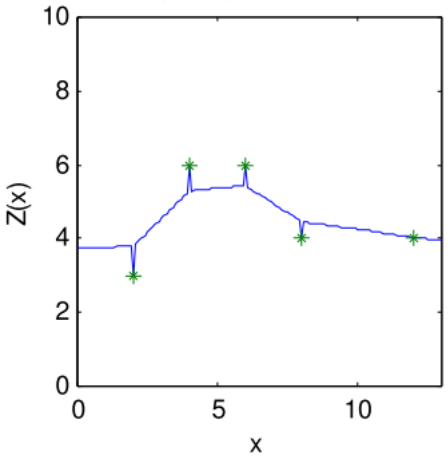


Figure 3.3. Spheric model (C_0=25% et a=10) (Rivoirard, 2003)

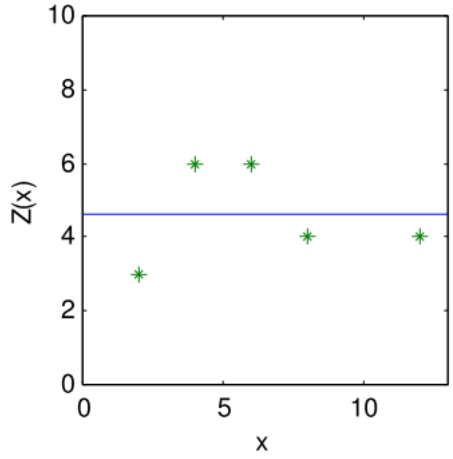


Figure 3.4. Pure nugget effect (Rivoirard, 2003)

At the sampling points, kriging returns the sampled value. To avoid discontinuities in maps, it is therefore recommended not to krige exactly at a sampling point. In practice, one ensures that there is at least an "epsilon" distance between the point to be kriged and the sampling point. Since the nugget effect often represents measurement error, it is justified to depart slightly from the observed values.

3.6 Screening effect

3.6.1 Extreme case: linear model in 1D

- It decreases as the nugget effect increases (there is no screening effect in the presence of a pure nugget effect) (Figure 3.5).
- It makes it possible to restrict kriging systems to neighbouring observations (moving neighbourhoods) (Chile's & Delfiner, 2012)

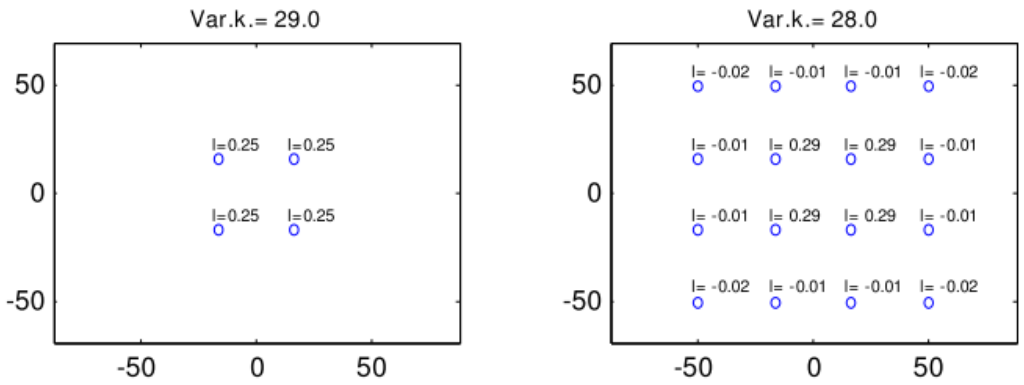


Figure 3.5. Spheric variogram C=100, a=100 et C_0=0 (Chile's & Delfiner, 2012)

3.7 Influence of the Field Size

As the size of the estimated domain increases, the weights tend to become equal, and the estimation variance first decreases and then increases if the domain to be estimated becomes larger than the domain containing the data (extrapolation) (Figure 3.6) (Chile's & Delfiner, 2012).

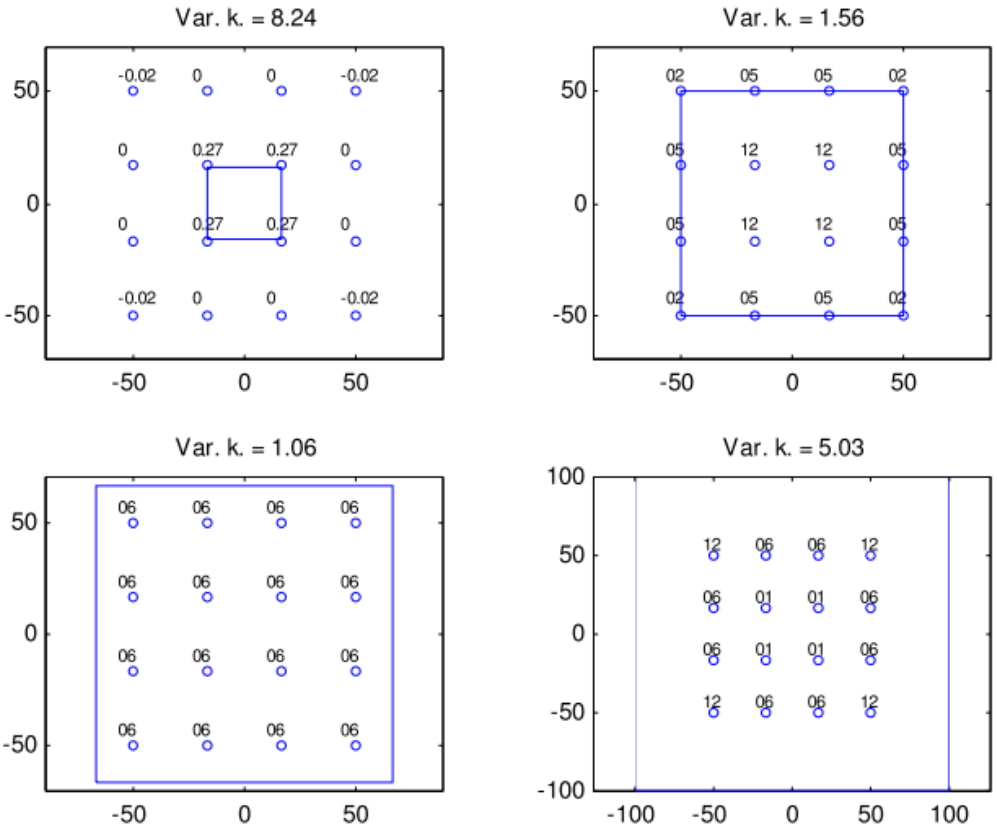


Figure 3.6. Influence of the Field Size (Chile's & Delfiner, 2012)

3.8 Relative Positions of the Points

Unlike inverse-distance methods, the relative positions of the points are of fundamental importance. Each point is automatically weighted according to its "zone of influence." (By contrast, inverse-distance weighting would assign a weight of $1/3$to each point in both cases.) A spherical variogram is assumed throughout, with $a = 100$, $C = 100$, and $C_0 = 0$(Figure 3.7)(Chile's & Delfiner, 2012).

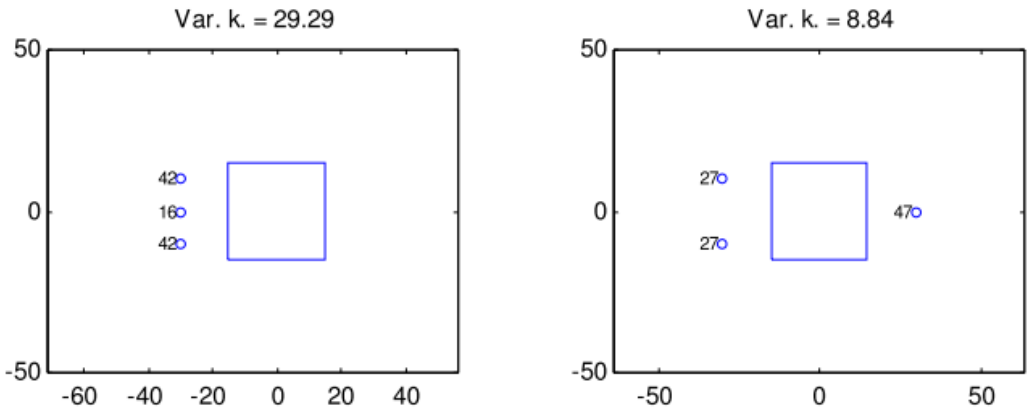


Figure 3.7. Influence of the Relative Positions of the Points (Chile's & Delfiner, 2012)

3.9 Influence of the Nugget Effect and the Range

The larger the nugget effect (relative to a fixed sill), the greater the estimation variance. Conversely, the larger the range, the smaller the estimation variance (Figure 3.8) (Chile's & Delfiner, 2012).

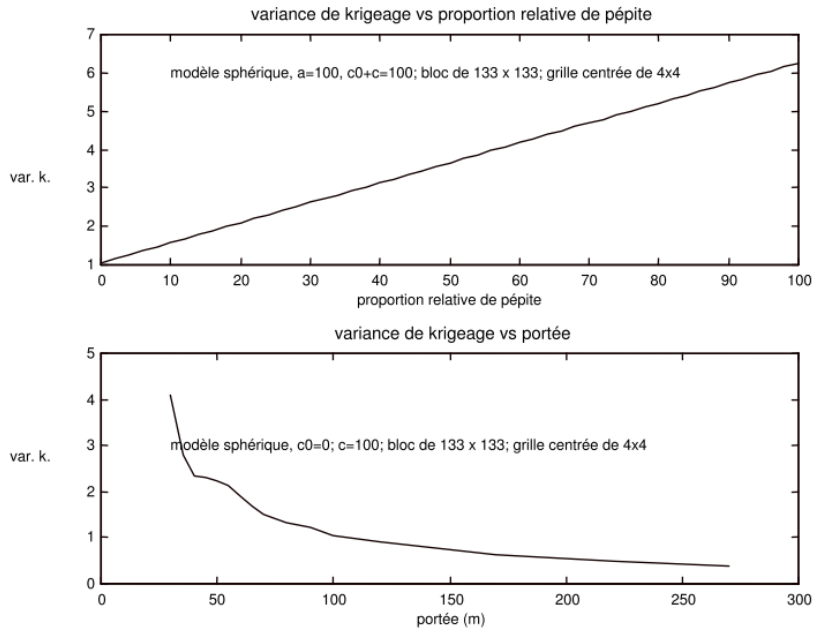


Figure 3.8. Influence of the Nugget Effect and the Range (Chile's & Delfiner, 2012)

3.10 Influence of Anisotropy

Sampling must be adapted by increasing the sampling density in the direction of the shortest range. In this example, a spherical model is assumed, with $C_0 = 0$, $C = 100$, $a_x = 200$, and $a_y = 50$. The three examples shown in Figure 3.9 correspond to the same sampling density (one sample per area of 33×33units). For the same sampling cost, much more accurate estimates can therefore be obtained if the sampling strategy is adjusted to account for anisotropy (Marcotte, 2011).

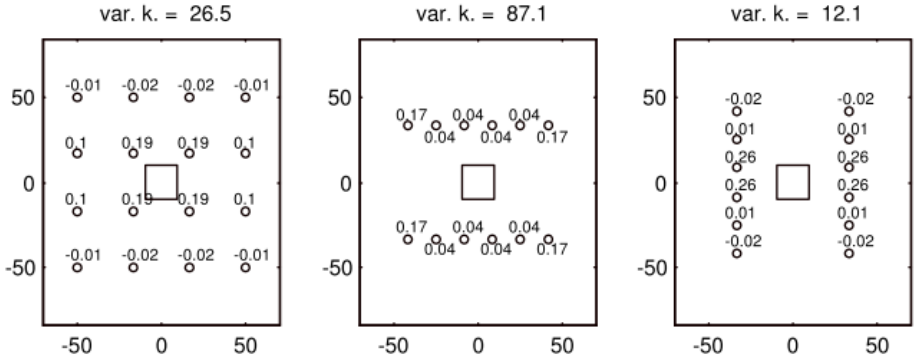
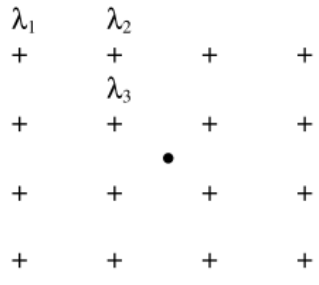
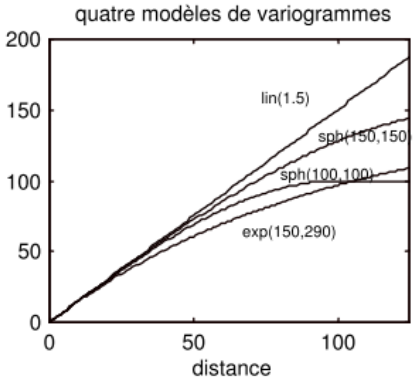


Figure 3.9. Influence of Anisotropy (Marcotte, 2011)

3.11 Influence of Model Selection

Model selection has only a limited influence on kriging results, provided that each model yields an equivalent fit at short distances. In Figure 3.10, the field measures $100\,\text{m} \times 100\text{m}$, and the points are spaced 33,3m apart. The point to be estimated is located at the midle of the grid. The theoretical models provide approximately the same values for distances ranging from 0 to 25 m, whereas the central points, which receive the highest weights, are located 24 m from the point to be estimated (Matheron & Blondel, 1962)





Sphérique:	C=100 a=100m λ_1=-.02 λ_2=-.01 λ_3= .29 σ_k^2= 28.0	Sphérique:	C=150 a=150m λ_1=-.01 λ_2=-.01 λ_3= .29 σ_k^2= 27.8
Exponentiel:	C=150 a=290m λ_1=-.01 λ_2=-.01 λ_3= .28 σ_k^2= 28.2	Linéaire:	Pente=1.5 λ_1=-.01 λ_2=-.01 λ_3= .28 σ_k^2= 27.6

Figure 3.10. Influence of Model Selection (Matheron & Blondel, 1962)

3.12 Smoothing Effect

It follows directly from the ordinary kriging equations that (Rivoirard, 2003):

$$Var\ (Z_v) = Var(Z_v^*) + \sigma_{ko}^2 + 2\mu \qquad (3.34)$$

For a fixed v, the term $\mathrm{Var}(Z_v)$does not depend on location, whereas the terms $\mathrm{Var}(Z_v^*)$, σ_{k0}^2, and μdo depend on the block considered and on the available samples. Under normal conditions, $\sigma_{k0}^2 + 2\mu > 0$, which explains the smoothing effect mentioned above (KRIGE, 1981).

Example:

Consider a square block of size 10x10 estimated from its four corners. The variogram is spherical, with a sill equal to 1 and a range of 20. The estimation is performed by ordinary kriging (with equal weights of 0,25) (Matheron G. , 1971).

The following quantities can be calculated:

$$\begin{aligned} \mathrm{Var}(Z_v)&=0.6278 \\ \sigma_{ko}^2 &= 0.1311 \end{aligned} \qquad (3.35)$$

Moreover,

$$\mathrm{Var}(Z_v^*)=1/16*(4*1+8*0.3125+4*0.1161)=0.4353 \qquad (3.36)$$

By substituting into the ordinary kriging equations, one obtains:

$$\mu = 0.0307 \qquad (3.37)$$

Thus, it follows that:

$$0.4353+0.1311+2*0.0307=0.6278 \text{ et } \mathrm{Var}(Z_v^*)<\mathrm{Var}(Z_v) \quad (3.38)$$

3.13 Conditional Bias

Consider the true block grade $Z_v$and its estimate, Z_v^*. Assume that the conditional expectation of $Z_v$given $Z_v^*$is linear, which is the case when both variables follow a bivariate Gaussian distribution. One then has (KRIGE, 1981):

$$E\left[\frac{Z_v}{Z_v^*}\right] = a + bZ_v^* \quad (3.39)$$

Where:

$$et\ a = (1\text{-}b)mb = \frac{Cov(Z_v,Z_v^*)}{Var\ (Z_v^*)} \quad (3.40)$$

By construction, one has:

$$Var\ (Z_v^*) + \mu = Cov\ (Z_v, Z_v^*) \Rightarrow b = 1 + \frac{\mu}{Var(Z_v^*)}, a = \frac{-\mu}{Var(Z_v^*)} \quad (3.41)$$

and therefore:

$$E\left[\frac{Z_v}{Z_v^*}\right] = Z_v^* + \frac{\mu}{Var(Z_v^*)}(Z_v^* - m) \quad (3.42)$$

This indicates that kriging exhibits conditional bias. This bias becomes very small when the estimate is accurate (small kriging variance, a Lagrange multiplier close to zero, and a large $\mathrm{Var}(Z_v^*)$).

In general, the Lagrange multiplier is slightly negative, which implies that the regression slope is less than 1. Consequently, when kriged values are used directly, high values tend to be slightly overestimated, whereas low values tend to be underestimated.

Note: for the polygonal (nearest-neighbour) estimator, one has:

$$b = \frac{Cov(Z_v,Z_v^*)}{Var(Z_v^*)} = \frac{Cov(Z_v,Z_i)}{\sigma^2} < 1 \quad (3.43)$$

This estimator exhibits a conditional bias that becomes more pronounced as the point used is farther from the block to be estimated (Matheron & Blondel, 1962).

3.13.1 Smoothing and Conditional Bias

As previously shown, from Equation (3.33) one has:

$$b = \frac{Cov(Z_v,Z_v^*)}{Var\ (Z_v^*)} \quad (3.44)$$

This may be rewritten as:

$$b = \frac{\rho\sigma_v\sigma_v^*}{Var\ (Z_v^*)} = \frac{\rho\sigma_v}{\sigma_v^*} \quad (3.45)$$

ρ is the correlation coefficient between $Z_v$and Z_v^*, and it is necessarily less than or equal to 1, while $\sigma_v^* = \mathrm{Var}(Z_v^*)^{0.5}$. For $b = 1$, it is therefore necessary that $\sigma_v^* \leq \sigma_v$. One may thus conclude that if an estimator is

more variable than the quantity it seeks to estimate, then it necessarily exhibits conditional bias (the regression slope will be less than 1). This is the case, for example, for the polygonal (nearest-neighbor) estimator, for which the variance of the estimated values is equal to the variance of the point data. The smoothing of the estimator (a characteristic property of kriging) is therefore an essential prerequisite for the absence of conditional bias. (Marcotte, 2011).

3.14 Practical Aspects of Kriging

3.14.1 Kriging Grid

Kriging is frequently performed over a regular grid of points or blocks.

In the case of point kriging, the objective is generally to produce a map of the variable under investigation.

The kriging grid must therefore be sufficiently dense to ensure that the resulting map reflects the kriging estimator itself, rather than artifacts introduced by the specific contouring or interpolation procedure used to draw iso-lines.

A poorly discretized grid may lead to visual smoothing or artificial structures unrelated to the spatial model (KRIGE, 1981).

3.14.2 Neighbourhood Used for Kriging

In practical applications, kriging is typically carried out using moving neighbourhoods. The main considerations are as follows:

- Moving search neighbourhoods are generally adopted.
- A sufficient number of data points should be included (typically > 10; in some cases, up to 50–100, depending on data density and spatial structure).
- The search area must be large enough to guarantee a minimum number of data points for the kriging system.
- When anisotropy is present, an elliptical search neighbourhood aligned with the direction of maximum spatial continuity is recommended.
- However, a circular search neighbourhood may be adequate if the number of retained data points is sufficiently increased.
- A quadrant-based search strategy ensures a more uniform spatial distribution of neighbouring data. It is common practice to impose at least two or three data points per quadrant to avoid directional bias (Matheron & Blondel, 1962).

Example

Consider a circular search neighbourhood with a maximum of two data points allowed per quadrant:

- Points 3 and 11 are rejected because they fall outside the search radius.
- Point 8 is rejected because two other data points within the same quadrant are closer to the estimation location.

This strategy improves numerical stability and reduces clustering effects, thereby enhancing the robustness of the kriging estimator.

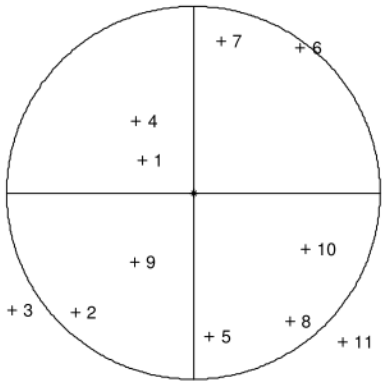


Figure 3.11. Neighbourhood in Kriging (Matheron & Blondel, 1962)

3.15 Cross-Validation

An effective procedure for validating both the variogram model and the search neighbourhood adopted for kriging consists of performing cross-validation.

The principle is to remove each observation in turn and to estimate its value using the remaining neighbouring data. For each location, two values are therefore obtained:

- the observed (true) value, and
- the kriging estimate obtained without using that observation.

These paired values can then be compared to assess whether:

- the variogram model provides unbiased and reliable predictions,
- the selected search neighborhood is appropriate,
- and the kriging variance is consistent with the actual prediction errors.

More formally, let Z_i^* denote the kriging estimate at location i obtained after removing the observed value Z_i. Let $\sigma_{k,i}^2$be the corresponding kriging variance.

The following quantities are defined:

- Estimation error (residual):

$$e_i = Z_i - Z_i^*$$

- Standardized residual:

$$n_i = \frac{e_i}{\sigma_{k,i}}$$

An adequate variogram model and neighbourhood configuration should yield:

- a mean of residuals closes to zero (unbiasedness),
- a small variance of residuals,
- and standardized residuals with mean close to zero and variance close to one.

These indicators provide quantitative diagnostics for assessing the consistency between the theoretical spatial model and the actual predictive performance of kriging (KRIGE, 1981):

$$et \sum_i n_i \approx 0 \sum_i e_i \approx 0 \qquad (3.46)$$

$$ou \sum_i {e_i}^0 \, min \sum_i |e_i| min \qquad (3.47)$$

$$\left(\frac{1}{n}\sum_i n_i^2\right)^{0,5} \approx 1 \qquad (3.48)$$

In addition, it is essential to examine:

- the histogram of the residuals e_i,
- the histogram of the standardized residuals n_i,
- as well as their spatial distribution.

This complementary analysis serves two main purposes:

- Detection of extreme values (outliers):

To verify whether the previously computed global statistics (mean error, variance of standardized residuals) are unduly influenced by one or two anomalous observations.

- Assessment of spatial homogeneity:

To ensure that the residuals are spatially homogeneous and do not exhibit structured patterns, clustering, or directional bias.

If residuals display spatial organization, this may indicate:

- inadequacy of the variogram model,
- inappropriate modelling of anisotropy,
- insufficient neighbourhood configuration,
- or the presence of an unmodeled deterministic trend.

Such diagnostic verification is fundamental in applied geostatistics, particularly in geotechnical engineering contexts where localized model deficiencies may significantly affect reliability assessments (Matheron & Blondel, 1962).

3.15.1 Illustration of Cross-Validation

The following figures present the results of simulations performed on 1,600 points (40 × 40 grid) using variable sampling intervals (represented along the horizontal axis of the graphs).

- On the vertical axis:

  In the upper figure, the following quantities are displayed:

  1. the mean squared kriging errors obtained through cross-validation,
  2. and the mean kriging variance.

- In the lower figure, the mean standardized kriging errors (normalized by the kriging variance) are shown.

All kriging estimations were carried out using 50 neighbouring data points.

The true model used for the simulations is:

- a spherical variogram model with range $a = 10$and nugget effect $C_0 = 0$for the first three figures,
- a pure nugget effect model for the last figure.
- In all cases, the variance of the simulated data is equal to 1 (Chile's & Delfiner, 2012).

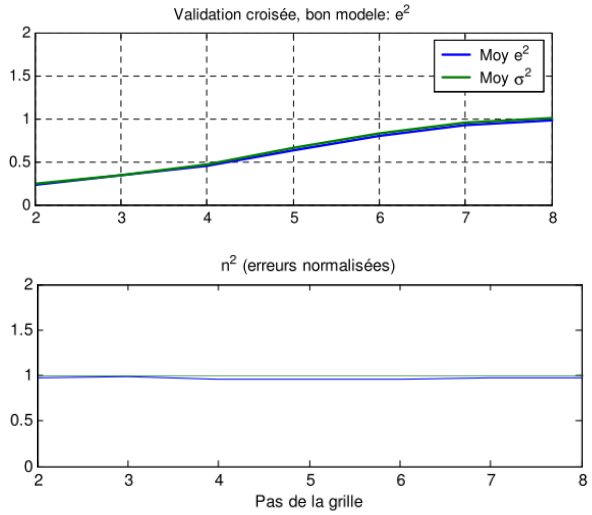


Figure 3.12. Kriging performed using the correct model (Chile's & Delfiner, 2012)

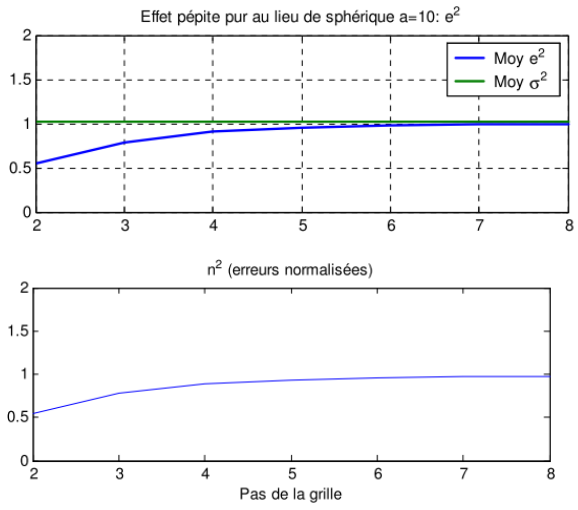


Figure 3.13. Model with a pure nugget effect instead of the true model (Chile's & Delfiner, 2012)

It can be observed that:

- The kriging variance accurately predicts the increase in precision associated with a denser sampling grid.
- The standardized errors exhibit a variance equal to 1, as theoretically expected.

Furthermore:

- For widely spaced grids (sampling interval between 6 and 8), the spatial structure is weakly resolved, and the kriging variance provides a reasonably accurate prediction of the actual estimation precision.
- For densely spaced grids (sampling interval between 2 and 4), the kriging variance exceeds the empirical variance of the errors (conservative or pessimistic behaviour). This leads to a variance of the standardized errors lower than 1 (Chile's & Delfiner, 2012).

This behaviour highlights the sensitivity of kriging diagnostics to the adequacy of the assumed spatial model and the interaction between sampling density and structural continuity.

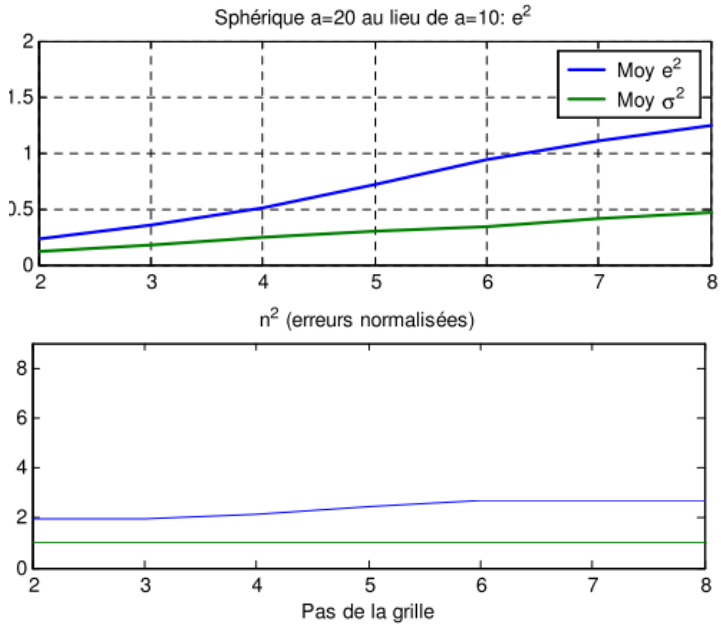


Figure 3.14. Spherical model with $a = 20$instead of the true value $a = 10$ (Chile's & Delfiner, 2012)

It can be observed that the kriging variance underestimates the variance of the prediction errors (optimistic behaviour); consequently, the variance of the standardized errors is greater than 1.

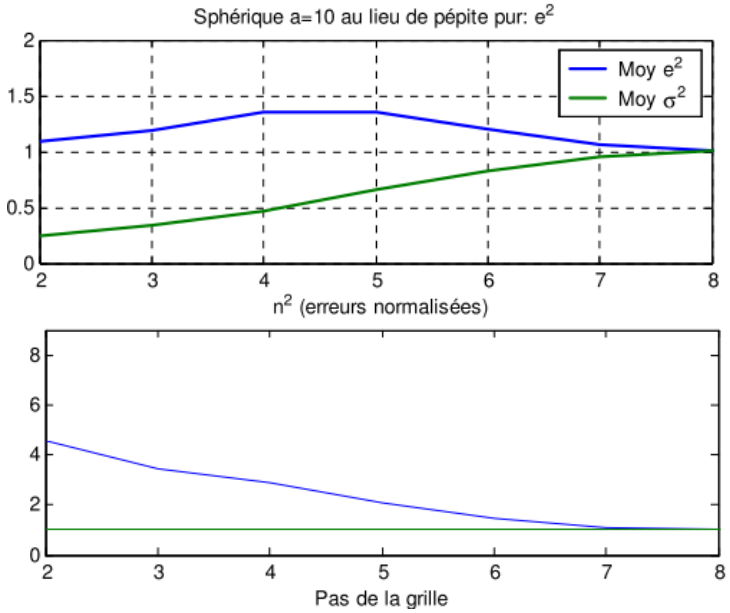


Figure 3.15. Spherical model with a = 10 compared to the pure nugget effect model (Chile's & Delfiner, 2012)

It can be observed that:

- The variance of the prediction errors first increases and then decreases as a function of the grid spacing.
- Although a grid with spacing 4 contains four times more samples than a grid with spacing 8, the variance of the errors is higher. When an incorrect model is specified, the additional information provided by the denser sampling is not properly exploited. It may therefore be hazardous to adopt an overly optimistic variogram model.
- The deterioration in precision can be explained by the strong screening effect of kriging weights when data points are located close to the estimation location (dense grid). In such cases, the estimate effectively becomes an average of only a few nearby points, instead of incorporating the full set of 50 neighbours as in more widely spaced grids, where correlations (and therefore screening effects) are weaker.
- The variance of the standardized errors is significantly greater than 1, indicating that the adopted model is excessively optimistic (Chile's & Delfiner, 2012)

3.15.2 Additional Validation Measures

The experimental variance of the grades (or variable of interest),

$$(\hat{\sigma}^2 = \frac{1}{n}\sum(Z_i - \bar{Z})^2) \qquad (3.49)$$

should be equal to the dispersion variance of a point within the deposit (Chile's & Delfiner, 2012).

The kriging smoothing relationship naturally provides an additional model validation tool. Once the variogram model is fixed, block variances can be computed for different block sizes. One may also perform block kriging for various block dimensions and calculate:

- the experimental variance of the kriged block values $\hat{\sigma}^2_{Z^*_v}$,
- the mean of the Lagrange multipliers $\bar{\mu}$,
- and the mean kriging variance $\bar{\sigma}^2_{k0}$.

The following relationship should then hold (Chile's & Delfiner, 2012):

$$\hat{\sigma}^{2*}_{Z_v} \approx D^2\left(\frac{Z_v}{G}\right) - \overline{\sigma}_{ko}^{\ 2} - 2\overline{\mu} \qquad (3.50)$$

This relation expresses the consistency between theoretical dispersion variance and empirical variability of kriged estimates, and constitutes a powerful diagnostic criterion.

3.16 Conclusion

Kriging theory represents the natural culmination of the geostatistical approach. After characterizing the spatial structure of a phenomenon through the variogram, kriging allows this information to be exploited to produce optimal estimates within a rigorous probabilistic framework.

Based on the search for a linear, unbiased estimator with minimum variance, kriging differs fundamentally from classical interpolation methods by its ability to simultaneously incorporate both the geometric configuration of data and the spatial correlation structure. However, the quality of the resulting estimates depends directly on the adequacy of the adopted variogram model, underscoring the necessity of a carefully conducted structural analysis.

The formulations examined (simple kriging and ordinary kriging) highlight the decisive role of assumptions concerning the mean of the random function. In practice, ordinary kriging is the most commonly used approach due to its flexibility and adaptability to real-world conditions.

Beyond point estimation, kriging provides an explicit measure of uncertainty through the estimation variance. This feature is of major importance in geotechnical engineering, where the reliability of parameters directly governs structural safety and performance.

Kriging is therefore not merely an interpolation technique; it constitutes a comprehensive analytical and decision-support tool that integrates estimation and uncertainty quantification. It also lays the foundation for more advanced geostatistical methods, including multivariate approaches and stochastic simulation techniques, which extend and enrich the theoretical framework developed in this chapter.

3.17 Exercises

The exercises proposed in this section are intended to consolidate the understanding of the fundamental principles of kriging and to ensure mastery of its methodological implementation.

They are designed to enable students to:

- apply the mathematical formulations of simple kriging and ordinary kriging,
- interpret the estimation weights,
- analyse the kriging variance as an indicator of uncertainty.

The adopted progression leads from the formulation of the kriging system to a critical analysis of the obtained results, highlighting the decisive influence of the variogram model on estimation quality.

This approach prepares students for a rigorous use of the numerical tools presented in the following chapter.

3.17.1 Comprehension Questions (Kriging Theory)

1. Define kriging and state its principal objective in geostatistics.
2. What is the difference between a deterministic interpolation method and kriging?
3. On which fundamental assumption concerning the regionalized variable is kriging based?
4. Explain the meaning of “linear unbiased estimator of minimum variance.”
5. What is the difference between simple kriging and ordinary kriging?
6. Why is the variogram model indispensable for solving the kriging system?
7. What does the kriging variance represent physically? Does it depend on the measured values or only on their spatial configuration?
8. Why do kriging weights depend on the spatial structure rather than directly on the observed values?
9. Under what conditions can kriging produce a smoothing effect of extreme values? Explain the mechanism.
10. Why can kriging be considered a method that simultaneously integrates estimation and uncertainty quantification, unlike classical interpolation techniques?

3.17.2 Numerical Applications

1. Exercise 1: Principle of kriging

Consider an unsampled location $x_0$surrounded by three measurement points x_1, x_2, x_3.

1. Write the general expression of the kriging estimator.
2. What do the weights $\lambda_i$represent?

3. What condition must these weights satisfy in the case of ordinary kriging?

4. Why does this condition guarantee unbiasedness?

2. Exercise 2: Simple Kriging

Assume that the mean of the phenomenon is known and equal to

$$\mu = 10$$

A single measurement point is available:

$Z(x_1) = 14$

The variogram is defined as follows:

$$\gamma(x_0, x_1) = 2$$
$$\gamma(x_1, x_1) = 0$$

1. Write the simple kriging system.

2. Compute the weight λ_1.

3. Compute the estimate $Z^*(x_0)$.

4. How should the result be interpreted if $x_0$is very far from x_1?

3. Exercise 3: Ordinary Kriging between two points

Two data points are available:

$$Z(x_1) = 12$$
$$Z(x_2) = 18$$

Given:

$$\gamma(x_1, x_2) = 4$$
$$\gamma(x_0, x_1) = 2$$
$$\gamma(x_0, x_2) = 2$$

1. Write the ordinary kriging system.

2. Solve the system to determine $\lambda_1$and λ_2.

3. Compute the estimate at location x_0.

4. Comment on the symmetric case.

4. Exercise 4: Kriging Variance

Consider a spherical variogram model with:

Nugget effect $C_0 = 1$

Structured contribution $C = 5$

Range $a = 30$ m

A location $x_0$lies at a distance of 10 m from a single measurement point.

1. Compute $\gamma(10)$using the spherical model.
2. Write the expression of the kriging variance in this simple case.
3. Interpret how the kriging variance evolves as distance increases.
4. Does the kriging variance depend on the measured values?

3.17.3 Critical Analysis of Kriging

An engineer performs ordinary kriging on a site characterized by a strong nugget effect. The resulting map is strongly smoothed, and extreme values observed in the field are attenuated.

1. Explain why kriging produces a smoothing effect.
2. What is the role of the nugget effect in this phenomenon?
3. Why does kriging not necessarily reproduce the maximum observed values?
4. In what situation would a geostatistical simulation be more appropriate than kriging?

*The detailed solutions to the exercises in this chapter are provided in **Appendix C***

Chapter 4.

Software and Applications

Software and Applications

The operational implementation of geostatistical methods requires the use of appropriate numerical tools capable of efficiently processing spatial data and solving the systems of equations associated with estimation. The theoretical developments related to variogram analysis and kriging thus reach their full practical relevance through the use of specialized software.

Modern computational tools not only automate calculations but also enable:

- visualization of the spatial structure of phenomena,
- testing and comparison of alternative variogram models,
- kriging-based estimation,
- and quantitative assessment of associated uncertainty.

They constitute an essential decision-support framework in geotechnical engineering, where natural soil variability must be incorporated in a rational and scientifically consistent manner.

This chapter presents the principal functionalities of geostatistical software, the most commonly used computational environments, and illustrative applications in geotechnical engineering. It establishes the link between the mathematical formalism developed in the previous chapters and professional engineering practice, while highlighting both the capabilities and the limitations of these tools.

4.1 Introduction

The theoretical foundations of geostatistics (regionalized variables, variogram analysis, and kriging theory) fully reveal their relevance when confronted with real field data. In geotechnical engineering, site investigation campaigns generate heterogeneous, spatially irregular, and often limited datasets. Transforming these discrete measurements into a continuous, coherent, and operational representation of the subsurface constitutes a major challenge.

The development of numerical tools has profoundly transformed geostatistical practice. Specialized software now enables the operational implementation of mathematically demanding methods such as kriging estimation and conditional simulation, whose computational complexity would be difficult to manage without algorithmic support (Chile's & Delfiner, 2012).

Beyond mere automation of calculations, these tools provide a comprehensive spatial analysis platform, including:

- statistical data exploration,
- directional variogram modeling,
- detection and modeling of anisotropy,
- point and block estimation,
- cross-validation procedures,
- and uncertainty mapping (Cressie & Wikle, 2011).

The increasing integration of geostatistics within GIS environments and scientific programming languages such as R and Python further reflects its central role in modern spatial analysis (Pebesma, 2004) (Gräler, Pebesma, & Heuvelink, 2016).

In the geotechnical context, these tools extend far beyond the production of interpolated maps. They constitute a decision-support framework that enables:

- optimization of borehole and sampling layouts,
- assessment of the spatial variability of mechanical parameters,
- explicit incorporation of uncertainty into reliability analyses,
- and improved management of geotechnical risk (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

However, computational power cannot replace scientific rigor. The quality of results depends directly on the relevance of the adopted variogram model and on a clear understanding of the assumptions underlying the applied methods. As emphasized by (Chile's & Delfiner, 2012), computational tools represent an extension of geostatistical reasoning, not a substitute for it.

The objective of this chapter is therefore twofold: to present the principal software and computational environments used in geostatistics, and to illustrate, through practical applications in geotechnical engineering, the operational implementation of the concepts developed in the preceding chapters. The aim is to establish a clear link between mathematical formalism and professional engineering practice.

4.2 Main Functions of Geostatistical Software

Geostatistical software packages constitute integrated platforms designed to implement the full sequence of the geostatistical workflow, from data exploration to estimation and simulation.

Their architecture is generally structured around a coherent methodological chain:

Exploratory Data Analysis → Variogram Modelling → Estimation → Validation → Simulation

This structured approach reflects the theoretical framework established by classical linear geostatistics (Matheron G. , 1971) (Chile's & Delfiner, 2012).

4.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) constitutes an essential preliminary step in the geostatistical workflow. Its objective is to characterize the statistical distribution of the data and to verify the assumptions required for subsequent modelling.

Typical functionalities provided by geostatistical software include:

- computation of descriptive statistics (mean, variance, standard deviation, skewness, kurtosis, coefficient of variation);
- histograms and probability density functions;
- cumulative distribution curves;
- data transformations (logarithmic transformation, normal-score transformation);
- detection of outliers;
- spatial location maps of the data.

This phase enables assessment of the normality of distributions, a condition often required for certain kriging methods or Gaussian simulation approaches (Cressie N. A., 1993) (Cressie & Wikle, 2011).

In geotechnical engineering, this step is particularly critical for variables such as hydraulic conductivity or shear strength parameters, whose distributions are frequently skewed and exhibit significant variability (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

4.2.2 Variogram Analysis

Variogram analysis constitutes the core of spatial modelling in geostatistics. Software tools enable computation of the experimental variogram from pairs of observations separated by a distance vector h.

The principal functionalities include:

- definition of distance classes (lag distance);
- computation of omnidirectional and directional variograms;
- detection and modelling of anisotropy;
- visualization through variogram maps;
- fitting of admissible theoretical models (spherical, exponential, Gaussian, Matérn);
- automatic estimation of model parameters (range, sill, nugget effect).

The fitting of a valid theoretical model is essential, since the associated covariance function must be positive definite to guarantee the existence and stability of the kriging estimator (Chile's & Delfiner, 2012) (Diggle & Ribeiro, 2007).

Modern software environments also provide parametric estimation techniques based on maximum likelihood or weighted least squares, offering statistically consistent parameter inference within a rigorous framework (Cressie & Wikle, 2011).

4.2.3 Kriging Estimation

Geostatistical software implements the principal kriging variants:

- simple kriging (known mean);
- ordinary kriging (unknown but locally constant mean);
- universal kriging (deterministic trend component);
- co-kriging (correlated variables);
- block kriging.

The computation relies on solving a linear system derived from the minimization of estimation variance under the unbiasedness constraint (Matheron G. , 1971) (Cressie N. A., 1993).

The outputs typically include:

- the estimated value;
- the kriging variance;
- the kriging weights assigned to each data point;
- continuous estimation maps.

The ability to provide a quantitative measure of uncertainty constitutes a major advantage over deterministic interpolation methods (Chile's & Delfiner, 2012).

4.2.4 Cross-Validation and Diagnostic Assessment

Cross-validation is used to evaluate the consistency between the variogram model and the observed data. The standard procedure consists of sequentially removing each data point and estimating it using the remaining observations.

The diagnostic indicators generally provided include:

- mean error (bias);
- mean squared error;
- standardized error statistics;
- ratio between experimental error variance and theoretical kriging variance.

This step is essential to prevent overfitting of the variogram model and to ensure predictive reliability (Diggle & Ribeiro, 2007).

4.2.5 Geostatistical Simulation

Beyond mean estimation, modern software packages provide **conditional simulation methods**, enabling the generation of equiprobable realizations of the spatial random field.

The principal approaches include:

- sequential Gaussian simulation (SGS);
- indicator simulation;
- multi-Gaussian simulation.

These techniques allow representation of the full spectrum of possible spatial variability beyond the kriging mean. They are particularly valuable for probabilistic risk assessment in geotechnical engineering, where uncertainty propagation plays a critical role (Chile's & Delfiner, 2012) (Cressie & Wikle, 2011).

4.2.6 Integration with Geographic Information Systems (GIS)

Geostatistical software is increasingly integrated within GIS environments, enabling:

- overlay of multiple information layers;
- multi-criteria spatial analysis;
- management of complex spatial datasets.

Open-source scientific environments such as R (notably the *gstat* package) and Python promote scientific reproducibility and analytical transparency (Pebesma, 2004) (Gräler, Pebesma, & Heuvelink, 2016).

4.3 Main Software Used in Geostatistics

The development of geostatistical methods has been accompanied by the emergence of specialized software tools designed for their operational implementation. These tools differ in terms of:

- level of specialization;
- application domain (mining, environmental sciences, geotechnical engineering);
- software architecture (proprietary or open source);
- and degree of integration with GIS platforms.

Three broad categories can be distinguished:

1. Historical and academic libraries;
2. Specialized professional software;
3. Open-source scientific environments.

4.3.1 Historical and Academic Libraries

4.3.1.1 GSLIB (Geostatistical Software Library)

GSLIB is one of the earliest structured libraries dedicated to geostatistics. Developed by Deutsch and Journel (1998), it comprises a collection of algorithms implementing classical kriging and simulation methods.

Main functionalities:

- computation of experimental variograms;
- fitting of theoretical models;
- simple and ordinary kriging;
- sequential Gaussian simulation;
- co-kriging.

GSLIB has played a fundamental role in disseminating geostatistical methods and remains a key pedagogical reference (Deutsch & Journel, 1998). However, its text-based interface may limit its accessibility for beginner students.

4.3.2 Specialized Professional Software

4.3.2.1 Isatis.neo (Geovariances)

Isatis.neo is an industrial-grade software widely used in mining, environmental, and geotechnical applications. It provides a comprehensive graphical interface enabling:

- exploratory data analysis;
- directional variogram modelling;
- simple, ordinary, and universal kriging;
- co-kriging;
- conditional simulation;
- management of large-scale databases.

It also integrates advanced modules for multivariate analysis and simulation, consistent with modern developments in geostatistics (Chile's & Delfiner, 2012).

4.3.2.2 ArcGIS Geostatistical Analyst

The Geostatistical Analyst extension of ArcGIS enables direct integration of geostatistical methods within a GIS environment.

Functionalities:

- kriging and co-kriging;
- interactive variogram analysis;

- cross-validation;
- production of estimation and uncertainty maps.

Its principal advantage lies in full spatial integration with other geographic information layers (Cressie & Wikle, 2011).

4.3.2.3 Surfer (Golden Software)

Surfer is primarily a mapping software package but incorporates several geostatistical interpolation methods, including:

- ordinary kriging;
- universal kriging;
- standard variogram models.

It is particularly used for rapid production of thematic maps in engineering applications.

4.3.3 Open-Source Software and Scientific Environments

4.3.3.1 R and gstat package

The R programming language has become one of the most widely used environments in spatial statistics. The gstat package, developed by Pebesma (2004), provides functionalities for:

- variogram computation;
- kriging;
- simulation;
- spatio-temporal analysis.

The work of Gräler et al. (2016) further extended these capabilities to spatio-temporal modeling.

Advantages:

- free and open access;
- scientific reproducibility;
- scripting flexibility;
- large and active scientific community.

4.3.3.2 Python (PyKrige, scikit-gstat)

Python is increasingly used in scientific and engineering applications. Libraries such as:

- PyKrige;
- scikit-gstat

enable implementation of kriging and variogram analysis within a modern scientific computing environment.

These tools facilitate integration with broader data-processing workflows, including machine learning pipelines and advanced numerical analysis frameworks.

4.3.3.3 SGeMS (Stanford Geostatistical Modeling Software)

SGeMS is an open-source software primarily oriented toward geostatistical simulation. It is particularly suited for:

- conditional simulation;
- multivariate analysis;
- mining applications.

4.3.4 Criteria for Software Selection in Geotechnical Engineering

The choice of software depends on:

- the volume of available data;
- the complexity of the spatial structure;
- the required level of precision;
- budget constraints;
- the need for scientific reproducibility.

In academic contexts, open-source environments (R, Python) are often preferred due to their methodological transparency and reproducibility (Pebesma, 2004). In industrial practice, specialized software such as Isatis.neo or ArcGIS is more commonly adopted.

4.4 Applications in Geotechnical Engineering

The application of geostatistics in geotechnical engineering addresses a central issue: characterizing a naturally heterogeneous medium from a limited number of discrete observations. Soils exhibit intrinsic spatial variability resulting from geological formation and evolution processes. This variability directly influences the mechanical behaviour of structures and must be incorporated quantitatively into engineering analyses (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

Geostatistical methods provide a rigorous framework for modelling this variability and explicitly integrating uncertainty into design assessments.

4.4.1 Mapping of Soil Mechanical Parameters

One of the most direct applications of geostatistics in geotechnical engineering concerns the spatial interpolation of soil mechanical and physical parameters, such as:

- cohesion $\boldsymbol{c}$;
- internal friction angle $\boldsymbol{\phi}$;

- pressuremeter modulus;
- undrained shear strength;
- hydraulic conductivity;
- bulk density.

Because data from boreholes are discrete, kriging enables the production of a continuous spatial estimate from these point observations (Chile's & Delfiner, 2012). Unlike deterministic interpolation methods, kriging incorporates the spatial correlation structure identified through the variogram.

Estimation maps allow:

- visualization of weak or heterogeneous zones;
- identification of spatial gradients;
- guidance of design decisions.
- Simultaneously, kriging variance maps provide a quantitative measure of spatial uncertainty, which is particularly relevant for risk management during the design phase (Cressie N. A., 1993).

4.4.2 Optimization of Site Investigation Campaigns

Geostatistics also enables optimization of borehole placement. The variogram range indicates the distance beyond which observations become spatially independent (Chile's & Delfiner, 2012).

This information allows:

- determination of optimal borehole spacing;
- avoidance of costly oversampling;
- identification of zones requiring additional investigation.

Such an approach contributes to cost reduction while maintaining an adequate level of reliability. Spatial analysis thus supports rational planning of site investigation campaigns (Cressie & Wikle, 2011).

4.4.3 Probabilistic Analysis and Structural Reliability

Spatial variability of geotechnical parameters constitutes a major source of uncertainty in stability and design analyses. Geostatistical methods enable explicit integration of this variability within probabilistic frameworks.

The coupling of:

- kriging;
- conditional simulation;
- reliability methods (FORM, Monte Carlo);

allows evaluation of failure probability while accounting for the spatial structure of soil properties (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

Geostatistical simulation, in particular, generates multiple equiprobable realizations of the spatial parameter field, providing a more realistic representation of variability than mean estimation alone (Chile's & Delfiner, 2012).

4.4.4 Three-Dimensional Subsurface Modelling

Modern geostatistical software enables three-dimensional modeling of geological formations. This capability is particularly useful for:

- deep foundations;
- tunnels;
- excavations;
- dams.

Three-dimensional representation improves understanding of spatial organization of layers and discontinuities and facilitates integration into numerical models such as finite element analyses.

Spatio-temporal approaches further allow analysis of parameter evolution over time (Cressie & Wikle, 2011).

4.4.5 Geotechnical Risk Management

In geotechnical engineering, explicit consideration of uncertainty is essential for ensuring structural safety. Geostatistical methods enable:

- identification of zones with high uncertainty;
- estimation of parameter confidence intervals;
- adjustment of safety margins.

A probabilistic approach based on spatial variability characterization provides improved risk control compared with purely deterministic methods (Phoon & Kulhawy, Characterization of geotechnical variability, 1999).

4.5 Limitations and Precautions in Use

Although geostatistical software provides powerful tools for spatial data analysis and interpolation, their use requires thorough understanding of underlying theoretical assumptions. The quality of results depends less on tool sophistication than on the relevance of the adopted model and the rigor of the analysis.

4.5.1 Dependence on the Variogram Model

Kriging relies entirely on the selected variogram model. Incorrect identification of spatial structure (range, sill, nugget effect, anisotropy) may lead to biased estimates or underestimation of uncertainty (Chile's & Delfiner, 2012).

Because the experimental variogram is itself an estimate derived from limited data, its interpretation involves a degree of subjectivity. Fitting a theoretical model must therefore be accompanied by rigorous validation (Diggle & Ribeiro, 2007).

4.5.2 Stationarity Assumptions

Classical kriging methods generally assume second-order or intrinsic stationarity. However, in geotechnical contexts, formations may exhibit pronounced spatial trends linked to geological processes (Cressie N. A., 1993).

When these assumptions are violated, estimates may become inconsistent. In such cases, one should:

- explicitly identify and model the trend (universal kriging);
- or work within locally stationary domains.

Verification of assumptions is a fundamental step in the analysis.

4.5.3 Smoothing Effect of Kriging

Kriging produces a mean estimate that minimizes variance under the unbiasedness constraint. This property induces a smoothing effect, reducing local variability relative to observed values (Chile's & Delfiner, 2012).

Consequently:

- extreme values are attenuated;
- the variance of the estimated field is lower than the true variance.

For analyses requiring realistic variability representation (e.g., probabilistic assessments), conditional simulation may be preferable (Cressie & Wikle, 2011).

4.5.4 Sensitivity to Outliers and Sampling Density

Outliers may strongly influence the variogram and subsequent estimates. Thorough exploratory analysis is indispensable prior to modelling (Wackernagel, 2003).

Moreover, result quality depends on the density and spatial distribution of observations. Excessive clustering in one area and lack of data elsewhere may generate unreliable estimates in poorly sampled zones.

Kriging variance maps help identify such areas of high uncertainty.

4.5.5 Illusion of Numerical Precision

Software produces numerically precise outputs (multi-decimal estimates, smooth continuous maps). However, apparent numerical precision must not be confused with true physical accuracy.

As emphasized by Cressie (2015), geostatistical estimation remains conditional upon model assumptions. The engineer must maintain a critical perspective and interpret results within their geological and technical context.

4.5.6 Computational Limitations

For large datasets, solving the kriging system may become computationally demanding, since it involves inversion of matrices whose dimension increases with the number of data points (Diggle & Ribeiro, 2007).

Local neighbourhood strategies or approximation techniques can be used to reduce computational complexity

4.6 Conclusion

Geostatistical software now represents a crucial link between the mathematical formalism of geostatistics and its operational implementation in geotechnical engineering. It translates theoretical concepts, variogram analysis, kriging, conditional simulation, into practical decision-support tools.

The integration of exploratory analysis, variogram modelling, estimation, and validation functions provides a complete methodological framework for rational exploitation of site investigation data. In particular, the simultaneous production of estimation and uncertainty maps constitutes a fundamental contribution in contexts where natural soil variability directly influences structural reliability.

Nevertheless, software performance does not eliminate the need for deep understanding of theoretical assumptions. Result quality depends closely on variogram model adequacy, data representativeness, and critical interpretation of numerical outputs. The engineer must therefore avoid confusing numerical precision with physical accuracy.

Ultimately, geostatistical software does not replace scientific reasoning; it operationalizes it. Mastery of these tools, combined with solid theoretical understanding, enables engineers to address contemporary geotechnical challenges through an approach that explicitly integrates variability and uncertainty, thereby contributing to more reliable, rational, and controlled engineering practice.

4.7 Exercises

4.7.1 Comprehension Questions

1. Why is the use of specialized software indispensable for the practical implementation of geostatistical methods?
2. What are the main functionalities expected from a geostatistical software package?
3. Why is exploratory data analysis an essential preliminary step before any variogram modelling?
4. How does the choice of computational parameters (lag size, angular tolerance, distance classes) influence the quality of the experimental variogram obtained using software?
5. What is the difference between point kriging and block kriging in geotechnical applications?
6. Why is cross-validation an important step when using geostatistical software?
7. What are the risks associated with "automatic" use of software without critical assessment of the results?
8. In which situations is geostatistical simulation preferable to kriging for modelling a geotechnical site?
9. How does integrating geostatistical software with GIS (Geographic Information Systems) improve the spatial analysis of soil parameters?
10. Why must the engineer maintain a critical perspective on numerical outputs produced by geostatistical software, despite their apparent computational precision?

4.7.2 Critical Analysis: Use of R Software in Geostatistics

A student uses R (packages gstat and sp) to model the undrained shear strength C_u of a site based on 35 boreholes. The workflow is as follows:

- data import and rapid visualization;
- automatic computation of the experimental variogram using default parameters;
- automatic fitting of an exponential model using the function fit.variogram;
- ordinary kriging on a regular grid;
- production of an estimation map and a kriging variance map;
- no in-depth exploratory analysis and no trend verification.

Questions

1. Why can using default parameters for variogram computation be problematic?
2. What preliminary checks should have been performed before fitting the variogram model?
3. Why does automatic fitting (fit.variogram) not eliminate the need for critical interpretation of the obtained model?

4. How could the presence of a trend (e.g., variation of $\boldsymbol{C_u}$with depth) bias the variogram computed under the stationarity assumption?
5. Why must the kriging variance map be analyzed jointly with the estimation map?
6. What are the risks of a purely visual interpretation of the interpolated map produced in R?
7. Propose a rigorous methodological workflow to improve this study in R.

The detailed solutions to the exercises in this chapter are provided in ***Appendix D***

General Conclusion

Geostatistics has become a fundamental tool for the analysis and modeling of spatially distributed phenomena, particularly in geotechnical engineering, where soil heterogeneity constitutes a major source of uncertainty.

This course manual has progressively introduced the theoretical and methodological foundations required for a rigorous application of geostatistical methods. The first chapter established the essential statistical bases for understanding regionalized variables and the concept of spatial dependence. The second chapter developed variogram analysis, a crucial step for quantitatively characterizing the correlation structure of the studied medium. The third chapter presented kriging theory, which represents the culmination of the geostatistical approach by providing an optimal linear estimator that explicitly incorporates spatial structure. Finally, the last chapter connected theory to practice through the presentation of software tools and geotechnical applications.

One of the major contributions of geostatistics lies in its ability to explicitly integrate spatial uncertainty into engineering analyses. Unlike purely deterministic approaches, it provides not only an estimate of parameters at any location within the domain of interest, but also a quantitative measure of the associated precision. This probabilistic dimension is essential in a context where structural safety depends directly on the reliability of the adopted parameters.

However, geostatistics should not be regarded as an automatic tool. The quality of results depends closely on the relevance of the underlying assumptions, the rigor of variogram analysis, the representativeness of available data, and the critical interpretation of outputs. Mastery of software tools cannot replace a thorough understanding of the theoretical concepts on which the methods are based.

In geotechnical engineering, where decisions directly affect the safety of people and infrastructure, geostatistics must be viewed as a decision-support framework that enhances understanding of natural soil variability and improves risk management.

This course manual aims to provide Master 1 students with the necessary foundations to understand, apply, and critically assess geostatistical methods. It also opens the way toward more advanced developments, such as stochastic simulation, multivariate approaches, spatio-temporal models, and integration with structural reliability methods.

Geostatistics thus fully aligns with modern engineering practice based on uncertainty quantification, rational decision-making, and optimized site investigation strategies, contributing to a safer and more scientifically grounded geotechnical engineering discipline.

Bibliographical References

Ang, A. H.-S., & Tang, W. H. (2007). *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering* (éd. 2éme édition). Hoboken, New Jersey (USA): John Wiley & Sons Ltd.

Bacconet, C., & Azzoue, R. S. (1991). Géostatistique et géotechnique. *Journnées de la géostatistique* (pp. 63-76). Fontainebleau: Cahier de géostatistique, Fascicule 1.

Baecher, G. B., & Christian, J. T. (2003). *Reliability and Statistics in Geotechnical Engineering.* The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd.

Banerjee, S., Carlin, B. P., & Gelfa, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data.* New York: Taylor & Francis Group. doi:https://doi.org/10.1201/b17115

Bourgine, B. (1996). *Potentiel d'application de la géostatistique en géoingénierie.* France: BRGM R 39049.

Chile's, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty* (éd. 2nd Edition). Hoboken, New Jersey: John Wiley & Sons, Inc.

Cressie, N. A. (1993). *Statistics for Spatial Data.* New York: John Wiley & Sons, Inc. doi:10.1002/9781119115151

Cressie, N., & Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* Hoboken, NJ: A JOHN WILEY & SONS, INC.

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

Deverly, F. (1984). *Thèse de doctorat : "Echantillonnage et géostatistique".* Paris: Ecole nationnale supérieure de Paris.

Devore, J. L. (2015). *Probability and Statistics for Engineering and the Sciences* (éd. 9 éme édition). Boston, MA: Cengage Learning.

Diggle, P. J., & Ribeiro, P. J. (2007). *Model Based Geostatistics.* New York: Springer.

Emery, X. (2001). *Géostatistique lionéaire.* (E. d. Minas", Éd.) Montpellier.

Floch, J. M. (2018). *Insee Méthodes Manuel d'analyse spatiale "Théorie et mise en oeuvre pratique avec R"* (Vol. 131). (I. n. économiques, Éd.) Montrouge: INSEE Eurostat 2018. Récupéré sur https://ec.europa.eu/eurostat/documents/3859598/9462709/INSEE-ESTAT-SPATIAL-ANA-18-EN.pdf

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. *The R Journal, 08*(01), 204 - 218. doi:10.32614/RJ-2016-014

Guillot, G. (2004). *Introduction à la géostatistique.* Paris: Institut National Agronomique de Paris-Grignon.

Ken'ichirou , K. (1996). Lognormal Distribution Model for Unsaturated Soil Hydraulic Properties. (A. G. (Agu), Éd.) *Water Resources Research*. doi:https://doi.org/10.1029/96WR01776

KRIGE, D. (1981). *Lognormal-de Wijsian Geostatistics for Ore Evaluation.* (D. G. S. Baker, Éd.) Johannesburg: the South African Institute of Mining and Metallurgy. Récupéré sur ISBN 0-620-03006-2

Marcotte, D. (2011). *Cours de géostatistique.* Montréal: École Polytechnique - Département CGM.

Matheron, G. (1966). Présentation des variables régionalisées. *Journal de la société statistique de Paris*, 263-275.

Matheron, G. (1971). *The Theory of Regionalized Variables and Its Applications* (Vol. 05). (É. d. Paris, Éd.) Paris: Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu.

Matheron, G., & Blondel, F. (1962). *Traité de géostatistique appliquée. Tome I.* Paris: Technip.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences, 30*(07), 683-691. doi:https://doi.org/10.1016/j.cageo.2004.03.012

Phoon, K. K., & Ching, J. (2015). *RISK AND RELIABILITY IN GEOTECHNICAL ENGINEERING.* (T. &. Group, Éd.) N W: CRC Press.

Phoon, K. K., & Kulhawy, F. H. (1999). Characterization of geotechnical variability. *Canadian Geotechnical Journal, 36*(04), 612-624. doi:https://doi.org/10.1139/t99-038

Rivoirard, J. (2003). *COURS DE GEOSTATISTIQUE MULTIVARIABLE.* (C. d. géostatistique, Éd.) Paris: école des mines de Paris.

Wackernagel, H. (2003). *Multivariate Geostatistics : An Introduction with Applications* (éd. 03). (S.-V. B. Heidelberg, Éd.) Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-662-05294-5

Appendices

Appendix A: Solutions to Exercises

Chapter 01: Theoretical Foundations of Geostatistics

A.1. Solutions to Comprehension Questions

1. Discrete and Continuous Random Variables

A discrete random variable takes values in a finite or countable set. A continuous random variable can take any value within a real interval. In geotechnical engineering, measured physical parameters (strength, hydraulic conductivity, bulk density) are generally modelled as continuous variables.

2. Classical Variable and Regionalized Variable

A classical random variable is defined without explicit reference to spatial location. A regionalized variable depends on a spatial position $x$and is written $Z(x)$. This concept is particularly suited to geotechnics because soil properties vary spatially.

3. Expectation, Variance, and Standard Deviation

Let Xbe a continuous random variable:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\ f_X(x)\ dx$$

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$$

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

4. Coefficient of Variation

$$CV = \frac{\sigma}{\mu}$$

It measures relative dispersion independently of measurement units.

5. Limitation of Variance

Variance describes global dispersion but does not provide information about spatial structure. Two datasets may have the same variance yet exhibit very different spatial organizations.

6. Probabilistic Approach

Natural soil variability introduces an irreducible uncertainty. A purely deterministic approach ignores this uncertainty and may lead to biased risk assessment.

7. Logarithmic Transformation

It stabilizes variance and reduces skewness when data follow a lognormal distribution (a frequent case for hydraulic conductivity).

8. Proximity of Samples

Nearby samples may show similar values due to spatial continuity, which challenges the independence assumption.

9. Comparison of Two Soils

The soil with the larger standard deviation exhibits greater variability and may be more critical for design.

A.2. Solutions: Numerical Applications

1. Exercise 1

Data: 65, 72, 80, 75, 90, 85, 78, 88

$$\mu = 79.1 \text{ kPa}$$
$$\sigma = 8.2 \text{ kPa}$$
$$CV \approx 10.4\%$$

Interpretation: moderate dispersion.

2. Exercise 2

Given $n = 12$

Sturges' rule:

$$N = 1 + 3.3\log(12) \approx 5$$

Class width:

$$I = \frac{1820 - 1580}{5} = 48$$

Interpretation: slightly symmetric distribution.

3. Exercise 3

Logarithmic transformation: values become more clustered and the standard deviation decreases relatively, resulting in improved statistical stability.

4. Exercise 4

Correlation coefficient:

$$r \approx 0.99$$

Interpretation: very strong linear correlation between depth and strength.

5. Exercise 5

$$CV_A = 10\%, CV_B = 30\%$$

Interpretation: Site B exhibits greater variability, implying increased uncertainty and risk.

A.3. Solutions: Critical Analysis

1. The mean alone is insufficient for characterization.
2. Dispersion directly affects the probability of underestimation.
3. If tests originate from different depths, the independence assumption is invalid.
4. It becomes necessary to introduce spatial structure through variogram analysis.

Appendix B: Solutions to Exercises Chapter 02: Variogram Analysis

B.1. Solutions to Comprehension Questions

1. Definition of the Experimental Variogram

The experimental variogram measures the average dissimilarity between observations separated by a lag vector $\boldsymbol{h}$. It is defined as:

$$\boldsymbol{\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i + h) - Z(x_i)]^2}$$

where $\boldsymbol{N(h)}$is the number of data pairs separated by distance $\boldsymbol{h}$.

2. Meaning of the Lag Vector h

The vector $\boldsymbol{h}$represents the separation distance (and possibly direction) between two spatial locations. It defines the spatial scale at which variability is analyzed.

3. Variogram Value for $h \to 0$

For a perfectly continuous phenomenon:

$$\boldsymbol{\gamma(0) = 0}$$

If a nugget effect exists:

$$\boldsymbol{\gamma(0) = C_0 > 0}$$

This reflects microscale variability or measurement errors.

4. Increasing Behavior of the Variogram

For a spatially correlated phenomenon, dissimilarity increases with distance until reaching the sill, corresponding to the total variance.

5. Limitations of the Experimental Variogram

The experimental variogram:

is defined only for sampled distances,

is discrete,

may be irregular,

does not guarantee mathematical admissibility (positive definiteness condition).

A theoretical admissible model must therefore be fitted.

6. Difference Between Spherical, Exponential, and Gaussian Models

1. Spherical: linear growth at the origin; reaches sill at finite range.

2. Exponential: rapid growth; sill reached asymptotically.

3. Gaussian: parabolic behaviour at origin; strong local continuity.

4. Behaviour near the origin indicates the regularity of the phenomenon.

7. Identification of Anisotropy

Directional variograms are computed.

- Same sill, different ranges → geometric anisotropy.
- Different sills → zonal anisotropy.

8. Geometric vs. Zonal Anisotropy

Geometric anisotropy:

- identical sills,
- different ranges.

Zonal anisotropy:

- different sills,
- different structural contributions by direction.

9. Impact of Poor Model Interpretation

A poorly fitted model alters kriging weights and may lead to:

- excessive smoothing,
- underestimation of uncertainty,
- misrepresentation of critical zones.

Estimation quality directly depends on the variogram model.

10. Variogram as Spatial Signature

The variogram summarizes the spatial dependence structure. It describes correlation scale, variability intensity, and local continuity. It is the mathematical representation of spatial organization.

B.2. Numerical Applications

1. Exercise 1: Experimental Variogram Calculation

Data:

$\boldsymbol{x} = \{\mathbf{0}, \mathbf{10}, \mathbf{20}, \mathbf{30}\}$m

$$\boldsymbol{Z(x)} = \{\mathbf{12}, \mathbf{15}, \mathbf{14}, \mathbf{18}\}$$

For $\boldsymbol{h} = \mathbf{10}$m

Pairs: (0,10), (10,20), (20,30)

$$(\mathbf{15} - \mathbf{12})^{\mathbf{2}} = \mathbf{9}, (\mathbf{14} - \mathbf{15})^{\mathbf{2}} = \mathbf{1}, (\mathbf{18} - \mathbf{14})^{\mathbf{2}} = \mathbf{16}$$

Sum = 26, $\boldsymbol{N(10)} = \mathbf{3}$

$$\boldsymbol{\gamma(10)} = \frac{\mathbf{26}}{\mathbf{2} \times \mathbf{3}} = \mathbf{4.33}$$

For $\boldsymbol{h} = \mathbf{20}$m

Pairs: (0,20), (10,30)

$$(\mathbf{14} - \mathbf{12})^{\mathbf{2}} = \mathbf{4}, (\mathbf{18} - \mathbf{15})^{\mathbf{2}} = \mathbf{9}$$

Sum = 13, $\boldsymbol{N(20)} = \mathbf{2}$

$$\boldsymbol{\gamma(20)} = \frac{\mathbf{13}}{\mathbf{2} \times \mathbf{2}} = \mathbf{3.25}$$

Comment: small samples may produce non-monotonic behaviour; the global trend should be considered.

2. Exercise 2: Distance Classes

Data:

$\boldsymbol{x} = \mathbf{0}, \mathbf{5}, \mathbf{12}, \mathbf{18}, \mathbf{27}, \mathbf{35}$m

$$\boldsymbol{Z} = \mathbf{10}, \mathbf{11}, \mathbf{9}, \mathbf{13}, \mathbf{12}, \mathbf{14}$$

Classes: (0,10], (10,20], (20,30], (30,40]

Results

$$\boldsymbol{\gamma(5)} = \mathbf{2.60}$$
$$\boldsymbol{\gamma(15)} = \mathbf{2.40}$$
$$\boldsymbol{\gamma(25)} = \mathbf{4.875}$$
$$\boldsymbol{\gamma(35)} = \mathbf{8.00}$$

The last class is unstable due to only one pair.

3. Exercise 3: Parameter Estimation

Experimental values stabilize near 9.0–9.1

Estimated sill:

$$C_0 + C \approx 9.1$$

Estimated range:

$$a \approx 55 \text{ m}$$

Possible nugget:

$$C_0 \approx 0.5–1$$

A spherical model is consistent with finite stabilization.

4. Exercise 4: Directional Variograms

Direction 0°: $a_0 = 60$m, sill ≈12

Direction 90°: $a_{90} = 25$m, sill ≈12

Geometric anisotropy.

Anisotropy ratio:

$$\frac{60}{25} = 2.4$$

Coordinate rescaling by 2.4 restores isotropy.

5. Exercise 5: Model Selection

Model A: Gaussian, low nugget

Model B: Spherical, non-negligible nugget

In geotechnics, a non-zero nugget is realistic due to measurement errors and micro-heterogeneity. Model B is generally more appropriate.

A higher nugget reduces short-range correlation, increases smoothing, and raises kriging variance locally.

B.3. Critical Analysis

1. Required Preliminary Checks

- Outlier control
- Unit consistency
- Exploratory analysis
- Possible transformation
- Trend detection
- Definition of stationary domain

2. Causes of High Nugget

- Measurement errors
- Micro-variability
- Inadequate lag choice
- Mixed populations

Diagnosis: recompute variograms, separate facies, compare directional results.

3. Influence of Visual Fitting

Range, sill, and nugget directly influence kriging weights and final maps.

4. Improvement Strategy

- Adequate lag selection
- Directional analysis
- Anisotropy modelling
- Multiple model testing
- Weighted least squares fitting
- Cross-validation

5. Technical Risks

- Misidentification of weak zones
- Under/overestimation of uncertainty
- Inappropriate site investigation decisions
- Potentially unsafe or overconservative designs

These corrections emphasize that variogram analysis is the structural backbone of kriging and reliability-oriented geotechnical modelling.

Appendix C: Solutions to Exercises Chapter 03: Kriging Theory

C.1. Questions de compréhension

1. Définition du krigeage

Kriging is a geostatistical estimation method used to predict the value of a regionalized variable at an unsampled location based on neighboring observations, while accounting for their spatial configuration and the correlation structure described by the variogram.

2. Difference from Deterministic Interpolation

Deterministic interpolation methods (e.g., inverse distance weighting) rely solely on geometric considerations.

Kriging additionally incorporates the spatial structure of the phenomenon through the variogram model and provides an explicit measure of estimation uncertainty.

3. Fundamental Assumption

Kriging assumes that the regionalized variable can be modeled as a realization of a random function whose spatial dependence structure is described by the variogram (or equivalently, the covariance function).

4. Linear Unbiased Minimum-Variance Estimator

The kriging estimator is expressed as:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

It is:

linear: a weighted combination of observed values;

unbiased: the expected estimation error is zero;

minimum variance: among all linear unbiased estimators, it minimizes the estimation error variance.

5. Difference Between Simple and Ordinary Kriging

Simple kriging: the mean $\mu$is assumed known.

Ordinary kriging: the mean is unknown but assumed locally constant. The following constraint is imposed:

$$\sum_{i=1}^{n} \lambda_i = 1$$

This constraint ensures unbiasedness when the mean is constant but unknown.

6. Role of the Variogram Model

The variogram model provides the semivariances (or covariances) required to construct and solve the kriging system. Without a valid admissible model, the system cannot be solved correctly.

7. Meaning of the Kriging Variance

The kriging variance quantifies the uncertainty associated with the estimate at a given location. It depends only on:

- the spatial configuration of the data;
- the adopted variogram model.
- It does not depend on the observed values themselves.

8. Independence of Weights from Observed Values

The kriging weights $\lambda_i$are determined exclusively by spatial structure and relative distances between points. They do not depend directly on the numerical values of observations, ensuring estimator objectivity.

9. Smoothing Effect

Kriging produces a weighted average; extreme values are attenuated because the estimate tends toward a local mean.

A significant nugget effect reduces short-range correlation and enhances smoothing.

10. Estimation and Uncertainty Quantification

Kriging simultaneously provides:

- a point estimate $Z^*(x_0)$;
- an estimation variance σ_k^2.
- It therefore constitutes a decision-support method that explicitly integrates uncertainty, unlike classical deterministic interpolators.

C.2. Numerical Applications

1. Exercise 1 – Principle of Kriging

The general estimator is:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

The weights $\lambda_i$represent the relative influence of each observation on the estimate at x_0.

In ordinary kriging:

$$\sum_{i=1}^{n} \lambda_i = 1$$

This condition guarantees unbiasedness under the assumption of a locally constant but unknown mean.

2. Exercise 2 : Simple Kriging

Given:

$$\mu = 10, Z(x_1) = 14, \gamma(x_0, x_1) = 2$$

For simple kriging with one data point:

$$Z^*(x_0) = \mu + \lambda_1[Z(x_1) - \mu]$$

The weight is expressed using covariance:

$$\lambda_1 = \frac{C(x_0, x_1)}{C(x_1, x_1)}$$

As distance increases, covariance decreases:

$$x_0 \rightarrow \text{far from } x_1 \Rightarrow C(x_0, x_1) \rightarrow 0 \Rightarrow \lambda_1 \rightarrow 0$$

Thus:

$$Z^*(x_0) \rightarrow \mu = 10$$

Interpretation: far from data, the estimate converges toward the mean.

3. Exercise 3 : Ordinary Kriging with Two Points

Given:

$$Z_1 = 12, Z_2 = 18, \gamma_{12} = 4, \gamma_{01} = 2, \gamma_{02} = 2$$

Ordinary kriging system:

$$\lambda_1\gamma_{11} + \lambda_2\gamma_{12} + \mu = \gamma_{01}$$
$$\lambda_1\gamma_{21} + \lambda_2\gamma_{22} + \mu = \gamma_{02}$$
$$\lambda_1 + \lambda_2 = 1$$

Since:

$$\gamma_{11} = \gamma_{22} = 0$$

The system reduces to:

$$4\lambda_2 + \mu = 2$$
$$4\lambda_1 + \mu = 2$$
$$\lambda_1 + \lambda_2 = 1$$

By symmetry:

$$\lambda_1 = \lambda_2 = 0.5$$

Estimate:

$$Z^*(x_0) = 0.5(12) + 0.5(18) = 15$$

Interpretation: symmetric configuration leads to the arithmetic mean.

4. Exercise 4 : Kriging Variance (Spherical Model)

Spherical variogram:

$$\gamma(h) = C_0 + C\left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right] \text{ for } h \le a$$

Given:

$$C_0 = 1, C = 5, a = 30, h = 10$$

Compute:

$$1.5\frac{10}{30} = 0.5$$

$$0.5\left(\frac{10}{30}\right)^3 \approx 0.0185$$

Thus:

$$\gamma(10) = 1 + 5(0.5 - 0.0185) = 1 + 5(0.4815) = 1 + 2.4075 \approx 3.41$$

Interpretation: the kriging variance increases with distance from known points and depends solely on spatial configuration and the variogram model.

C.3. Critical Analysis of Kriging

1. Kriging smooths because it produces a weighted average.
2. A high nugget effect reduces local correlation and increases smoothing.
3. Extreme values are attenuated because the estimate tends toward a local mean.
4. Geostatistical simulation is preferable when realistic reproduction of local variability and extreme values is required.

Appendix D: Solutions to Comprehension Questions

Chapter 04 Software and Applications

D.1. Comprehension Questions

1. Necessity of Specialized Software

Geostatistical methods require computing experimental variograms, fitting admissible models, and solving potentially large linear systems. For real datasets, these operations are impractical to perform manually. Specialized software automates the computations while providing visualization and spatial diagnostics that are essential for interpretation.

2. Expected Functionalities of Geostatistical Software

A geostatistical software package should enable:

- exploratory data analysis (descriptive statistics, histograms);
- computation of the experimental variogram;
- fitting of theoretical variogram models;
- kriging estimation (point and block kriging);
- geostatistical simulation;
- production of estimation maps and kriging variance maps.

3. Importance of Exploratory Data Analysis

EDA is required to identify:

- outliers;
- skewness and distribution shape;
- the need for transformations (e.g., logarithmic);
- potential spatial trends.

Without this step, the variogram and subsequent estimates may be biased or physically inconsistent.

4. Influence of Computational Parameters

Choices such as lag size, number of bins, and angular tolerance control the number of pairs in each lag class. Poor parameter selection can produce noisy or unstable experimental variograms, leading to incorrect model fitting and degraded kriging performance.

- Point Kriging vs. Block Kriging
- Point kriging estimates the value at a specific location.
- Block kriging estimates the average value over an area or volume.

In geotechnical engineering, block kriging is often more relevant because it better represents the average behavior of the soil volume interacting with a structure.

5. Role of Cross-Validation

Cross-validation removes each observation in turn and compares the observed value with its estimate. It is used to check:

- absence of global bias;
- coherence of the variogram model;
- predictive quality of kriging.

6. Risks of Automatic Use

Using software without critical analysis may lead to:

- an inadequate variogram model;
- incorrect estimation of range or sill;
- underestimation of uncertainty;
- erroneous interpretation of produced maps.

7. Simulation versus Kriging

- Kriging yields a smoothed estimate (conditional mean).
- Geostatistical simulation reproduces natural variability and extreme values.

Simulation is preferable when local variability must be represented or when probabilistic/risk analyses are required.

8. Integration with GIS

GIS integration enables:

- consistent handling of spatial coordinates;
- overlay with other layers (geology, topography);

- improved visualization and interpretation of results.

9. Need for Critical Engineering Judgment

Software executes computations based on user inputs but cannot validate the scientific relevance of assumptions. The engineer remains responsible for selecting appropriate models, verifying assumptions, and ensuring that interpretations are technically and geologically consistent.

D.2. Critical Analysis: Software Use and Applications (R-*gstat*)

1. Default Variogram Parameters

Using default settings (bin width, maximum distance, angular tolerance) may result in:

- too few pairs in some bins;
- noisy or unstable variograms;
- poor representation of short- or long-range structures.

Parameters must be adapted to data density and spatial configuration.

2. Required Preliminary Checks

Before variogram modeling, one should:

- analyze the data distribution (histogram, skewness);
- detect outliers;
- verify unit consistency;
- consider transformations (log for lognormal variables);
- investigate trends (depth effect, slope, facies).

Otherwise, the variogram may capture artifacts rather than true spatial structure.

3. Limits of Automatic Fitting (fit.variogram)

Automatic fitting minimizes a mathematical criterion (often least squares), but:

- it does not ensure geological plausibility;
- it may yield unrealistic parameters (excessive range, nugget issues before constraints);
- it does not replace expert interpretation.

The fitted model must be assessed visually and physically.

4. Effect of an Unmodeled Trend

If a trend exists (e.g., $C_u$increasing with depth), computing the variogram under stationarity may:

- inflate variability at large distances;
- mask the true short-range structure;

- lead to an inappropriate variogram model.

Trend removal (detrending) or universal kriging should then be considered.

5. Joint Interpretation of Estimation and Variance Maps

- The estimation map shows predicted values.
- The kriging variance map shows spatial uncertainty.

A low estimated value associated with high variance must be interpreted cautiously; decisions should incorporate both maps.

6. Risks of Purely Visual Interpretation

Interpolated maps may create an illusion of continuity and precision even when:

- data are sparse;
- kriging variance is high;
- the variogram is poorly fitted.

Graphical smoothness does not guarantee scientific reliability.

7. Rigorous Methodological Workflow in R

A robust approach should include:

- complete EDA (summary, histograms, boxplots);
- trend assessment and possible detrending;
- directional variograms;
- comparative fitting of several models (spherical, exponential, Gaussian);
- cross-validation (krige.cv);
- residual and standardized error diagnostics;
- joint interpretation of estimation and variance outputs.

R and geostatistical packages provide powerful tools, but they do not replace theoretical understanding. The software performs computations; the engineer remains responsible for model choice, assumption validity, technical interpretation, and the engineering decisions that follow.