CHAPTER 1
INTRODUCTORY CHAPTER

1.1 Introduction
The assumption that data observations are independent is frequently used in mathematical statistics. Data observations *close* together in time or space, however, should not always be considered independent. Temporal and spatial data often possess some form of correlation between observations, and to assume that each observation is taken independently of any other would be statistically incorrect.

As an illustration, consider the following example which has been taken and adapted from Cressie (1991 : 13-15).

Example 1.1
Suppose that that the random variables $Z(1)$, $Z(2)$,…, $Z(9)$ are independent and identically distributed from a normal distribution with unknown mean μ and known variance σ^2, and that they are located in two-dimensional Euclidean space as follows:

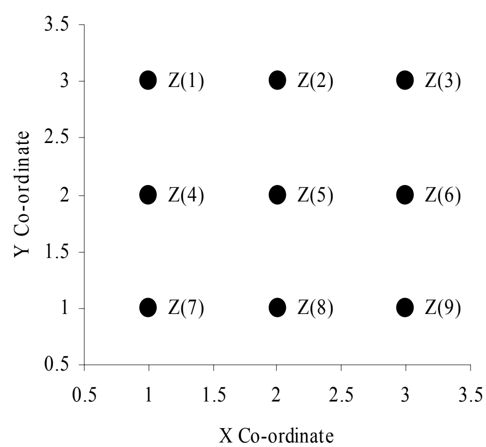

Figure 1.1 The location of the nine random variables $Z(1)$, $Z(2)$, …, $Z(9)$ in two-dimensional Euclidean space.

The minimum-variance unbiased estimator of μ is

$$\bar{Z} = \frac{1}{9}\sum_{i=1}^{9} Z(i)$$

with

$$\bar{Z} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{9}\right) \tag{1.1}$$

and therefore, the 2-sided 95% confidence interval for μ is

$$\left(\bar{Z} - (1.96)\frac{\sigma}{3}, \bar{Z} + (1.96)\frac{\sigma}{3}\right). \tag{1.2}$$

Now, instead of independent data, assume that the nine random variables are positively correlated such that

$$\operatorname{cov}(Z(i), Z(j)) = \begin{cases} \sigma^2 & \text{for } d(i,j) = 0 \\ \rho\sigma^2 & \text{for } d(i,j) = 1 \\ \rho^2\sigma^2 & \text{for } d(i,j) = \sqrt{2} \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

where $d(i, j)$ is the Euclidean distance between the random variables $Z(i)$ and $Z(j)$, and $0 < \rho < 1$.

In matrix notation, $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where:

$$\mathbf{Z} = \begin{pmatrix} Z(1) \\ . \\ . \\ . \\ Z(9) \end{pmatrix}_{9\times 1} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu \\ . \\ . \\ . \\ \mu \end{pmatrix}_{9\times 1}$$

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & \rho & \rho^2 & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho^2 & \rho & \rho^2 & 0 & 0 & 0 \\ 0 & \rho & 1 & 0 & \rho^2 & \rho & 0 & 0 & 0 \\ \rho & \rho^2 & 0 & 1 & \rho & 0 & \rho & \rho^2 & 0 \\ \rho^2 & \rho & \rho^3 & \rho & 1 & \rho & \rho^2 & \rho & \rho^2 \\ 0 & \rho^2 & \rho & 0 & \rho & 1 & 0 & \rho^2 & \rho \\ 0 & 0 & 0 & \rho & \rho^2 & 0 & 1 & \rho & 0 \\ 0 & 0 & 0 & \rho^2 & \rho & \rho^2 & \rho & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho^2 & \rho & 0 & \rho & 1 \end{pmatrix}_{9\times 9}$$

By defining

$$\mathbf{a} = \begin{pmatrix} \frac{1}{9} & \frac{1}{9} & . & . & . & \frac{1}{9} \end{pmatrix}_{1\times 9}$$

the unbiased estimator of μ is now written as

$$\overline{Z} = \mathbf{aZ}$$

with

$$\overline{Z} \sim \mathrm{N}\left(\mathbf{a\mu}\,,\, \mathbf{a\Sigma a}^T\right).$$

It is easily shown that in this case

$$\mathrm{var}\left(\overline{Z}\right) = \mathbf{a\Sigma a}^T = \frac{\sigma^2}{9}\left[1 + \frac{24}{9}\rho + \frac{16}{9}\rho^2\right]. \tag{1.4}$$

Accordingly, the 2-sided 95% confidence interval for μ is

$$\left(\overline{Z} - (1.96)\frac{\sigma}{3}\sqrt{\psi}\,,\, \overline{Z} + (1.96)\frac{\sigma}{3}\sqrt{\psi}\right) \tag{1.5}$$

where

$$\psi = 1 + \frac{24}{9}\rho + \frac{16}{9}\rho^2 .$$

The confidence interval defined by Equation 1.5 is wider than the confidence interval defined by Equation 1.2. Such a result is to be expected, as nine *dependent* random variables certainly contain less information about the overall mean than do nine *independent* random variables.

By rewriting Equation 1.4 as

$$\operatorname{var}(\overline{Z}) = \frac{\sigma^2}{9^{\#}}$$

where

$$9^{\#} = 9\left(1 + \frac{24}{9}\rho + \frac{16}{9}\rho^2\right)^{-1}$$

so as to imitate the form of the variance of the unbiased estimator in Equation 1.1, and assuming, say, ρ = 0.2, it is calculated in Equation 1.6 that the nine correlated random variables of this example provide the same amount of information about the overall mean μ as would be provided by 5.61 independent random variables[†].

For $\rho = 0.2$

$$9^{\#} = 9\left(1 + \frac{24}{9}(0.2) + \frac{16}{9}(0.2)^2\right)^{-1} = 5.61 \tag{1.6}$$

Figure 1.2 on the next page displays the form of the relationship between ρ and $9^{\#}$ for $\rho \in (0,1)$.

[†] In other words, the nine correlated random variables will have the same sized confidence interval for μ (defined by Equation 1.5) as would 5.61 independent random variables (whose confidence interval is defined by Equation 1.2).

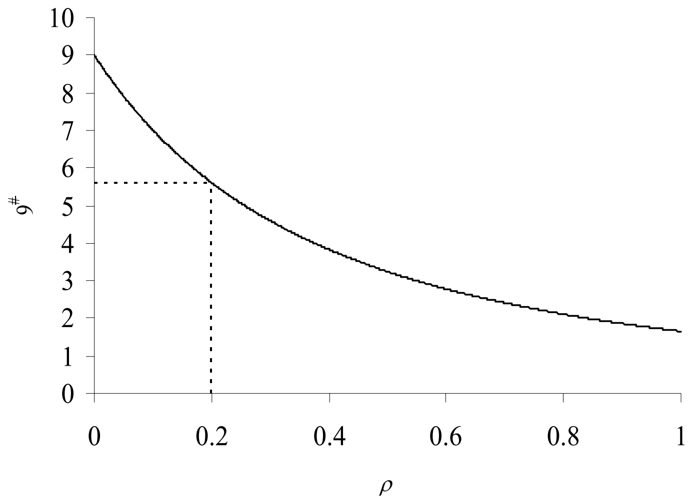

Figure 1.2 The relationship between $\rho \in (0,1)$ and $9^{\#}$. The dotted line clearly shows that for $\rho = 0.2$ the nine correlated random variables provide the same precision in estimation of the overall mean μ as would be provided by 5.61 independent random variables.

This example underlines the importance of acknowledging dependence between observations whenever such correlation exists. The use of inappropriate assumptions in any statistical analysis can lead to incorrect solutions (such as incorrect confidence intervals as in Example 1.1) and result in poor decision-making with potentially serious repercussions.

Data can often contain some form of spatial dependence or spatial correlation, and in such circumstances, it is of paramount importance that the spatial location of the data observations be considered. A particular area of statistics, entitled *spatial statistics*, encompasses a considerable number of models and techniques that take into account the spatial location of the data and allows for the existence of spatial correlations. *Geostatistics* comprises a particular branch of spatial statistics.

1.2 An Elementary Description of Geostatistics

Olea (1990 : 43) defines geostatistics as:

> The study of phenomena that fluctuate in space. Originally the field started with the objective to improve forecasts of ore grade and reserves in mines, but the mathematical generality of the approach led to applications of geostatistics to other problems in the earth sciences. Although not all geostatistical methods are probabilistic in nature, the most important accomplishments have been in estimation and forecasting, extending the probabilistic methods of stochastic processes and time series analysis to the spatial domain.

A more concise definition is supplied in Journel & Huijbregts (1978 : 1): "'Geostatistics is the application of the formalism of random functions to the reconnaissance and estimation of natural phenomena.'"†

For an elementary description of what geostatistics entails, consider a spatial region **Ω**‡ within which a particular attribute (such as gold grade) can be measured. Suppose further that values of this attribute are observed at n locations within **Ω**. These n observed data values "… may be continuous or discrete, they may be spatial aggregations or observations at points in space, their spatial locations may be regular or irregular, and those locations may be from a spatial continuum or a discrete set." (Cressie, 1991 : 8). Briefly, geostatistics is often regarded as encompassing a number of techniques that…

1. make use of both the value and spatial location of the n observed attribute values to estimate and describe the spatial relationship existent between all locations within the spatial region **Ω**,

and then

2. use both the value and spatial location of the n observed attribute values, together with the spatial relationship determined in the previous step, to predict the values of the attribute at any location within **Ω**.

---

† According to Journel & Huijbregts (1978 : 1) this definition is cited from Matheron (1962-63).

‡ Of either two- or three-dimensions.

Note 1.1 The predictions are generated by considering some linear or non-linear combination of the n observed attribute values, depending on the type of geostatistical method used.

This description of geostatistics is very basic and is merely intended to provide an initial impression of what geostatistics is all about. In the chapters that follow, the theory and techniques of geostatistics will be studied in greater detail, providing a comprehensive coverage of the subject. By the end of the dissertation, it will become evident that geostatistics clearly has a great deal to offer to the analysis of spatial data.

Note 1.2 In geostatistics, the words prediction and estimation generally mean the same thing. So, to geostatistically predict the value of an attribute at a particular spatial location is the same as geostatistically estimating the value of an attribute at a particular spatial location. Olea (1990 : 84) backs up this statement by quoting that, "In geostatistics, the term [prediction] is used … interchangeably with … estimation."

1.3 The History and Development of Geostatistics

(A large part of Section 1.3 has been researched from Matheron & Kleingeld (1987), Cressie (1990), Wackernagel (1995 : 1) and Krige (2000)).

Geostatistics had its origins in the late 1940's and the 1950's with the work of two South Africans, Herbert Sichel and Danie Krige (see, for example, Sichel, 1947 and Krige, 1951). These two individuals were particularly interested in the application of statistics in ore evaluation and mining. Their work aroused the interest of the Frenchman Georges Matheron, and in the late 1950's and early 1960's he developed the theoretical foundations of geostatistics (Matheron 1963b). Over the following years, many new contributions were made and the subject of geostatistics grew into a highly specialized discipline. Matheron & Kleingeld (1987), Cressie (1990) and Krige (2000) provide a more detailed account of the beginnings and evolution of geostatistics.

Note 1.3 Round about the same time G. Matheron was introducing the name geostatistics and developing its theoretical foundations, similar work was been carried out in meteorology by L. S. Gandin and in forestry by B. Matern (according to Matheron & Kleingeld, 1987 : 9 ; Cressie, 1990 and Dowd, 2000 : 14). In his article describing the origins of kriging, Cressie (1990) not only acknowledges D. G. Krige and G. Matheron for their significant contributions to spatial prediction, but also accentuates in some detail other important works (with respect to spatial prediction) that took place outside of the mining environment.

"Geostatistics … is particularly [well] suited to the study of natural phenomena" (Journel & Huijbregts, 1978 : 2), and while geostatistics was originally developed for use in the mining industry (Wackernagel, 1995 : 1 and Olea, 1990 : 43), it is nowadays frequently used in such fields as agriculture, forestry, atmospheric science and mining. Provided that geostatistics is used with understanding, it can prove a powerful tool in many spatial statistical problems.

"From the very beginning geostatistics was a problem (i.e. industry) driven discipline …." (Dowd, 2000 : 15). The acceptance of geostatistics thus always hinged mainly on the usefulness and reliability of the solutions it could provide (Journel & Huijbregts, 1978 : 3). As it turns out, geostatistics *has* indeed proved itself in the various industries - Armstrong (1998 : 1) states, "Over the past 30 years, geostatistics has proved its superiority as a method for estimating reserves in most types of mines (precious metals, iron ore, base metals etc.)." In addition, numerous case studies (e.g. Sabourin, 1975 and Rendu, 1979) have been carried out which also document the usefulness of geostatistics. The case studies just mentioned compare the results of certain geostatistical estimation techniques with actual mining figures, and both conclude that geostatistics provides impressive estimates.

It is evident from the quantity of geostatistical articles and books that continue to be published and the number of geostatistical conferences that are planned for the future that the boundaries of geostatistical knowledge are still expanding. New techniques are regularly being introduced, for example Arik (2002), and new areas of application are also regularly being identified. The recent articles by Farias et al. (2000) and Santos et al. (2000) confirm that geostatistical analysis can prove beneficial in various areas other than mining. These two articles describe, respectively, the application of geostatistics in crop farming and in bird migration patterns. There is no doubt that geostatistical analysis continues as an essential branch of spatial data analysis.

1.4 Plan and Purpose for this Dissertation

The theory of geostatistics is often divided into the two categories of linear geostatistics and non-linear geostatistics. It is the intention of this dissertation to clearly and concisely present the theory of linear geostatistics. Linear geostatistics comprises geostatistical methods whose spatial predictions are based on linear combinations of the available data. Linear geostatistics is far less complex than its non-linear counterpart and its theory is also more widely accessible (Journel & Huijbregts, 1978 : 555-556 and Wackernagel, 1995 : 2). With the exception of lognormal kriging which is introduced in Chapter 7, the theory presented in this dissertation falls almost entirely within the category of linear geostatistics.

Note 1.4 Readers interested in non-linear geostatistics are referred to Rivoirard (1994).

On completion of the theoretical part of this dissertation, some of the models and techniques studied will be applied to spatial data sets. This will offer valuable insight into the application of geostatistical theory to practical situations. Tests will be carried out to

1. establish how various decisions regarding semi-variograms can affect geostatistical estimates,
2. compare geostatistical estimates with inverse distance weighting and polygonal estimates,

and

3. ascertain the theoretical necessity of additive regionalized variables.

In 1986 and 1987 several journal articles and letters contesting the validity of geostatistics were published in *Mathematical Geology*, and a fiery debate ensued. Chapter 9 of this dissertation will summarize these articles and letters and use the research in the previous chapters to uncover whether the criticisms posed against geostatistics during this debate are legitimate or not.

Through these studies the dissertation aims to examine critically the geostatistical literature from both a statistical and an applied perspective. It is hoped that this assessment of the theoretical soundness, the performance and the merits of geostatistics will provide a detailed insight into the subject and add something meaningful to the area of geostatistical understanding. It is also hoped that this dissertation will be considered of value to both the individual already familiar with geostatistics and the individual wishing to research the subject.

1.5 Justification for this Study

As has already been emphasized in this introductory chapter, geostatistics is in use in many industries and has regularly performed well. However, geostatistics is in no way a perfect science. As is the case with any other statistical estimation and prediction technique, geostatistics cannot be expected to provide *exact* estimates.

Companies requiring spatial predictions are constantly on the lookout for improved and more reliable techniques and methodologies since their profitable existence is often dependent on these predictions. Journel (1986 : 120) remarks that currently geostatistics "… provides the most complete set of tools yet available…." It is therefore sensible that the theories and techniques of geostatistics be constantly scrutinized and tested (as is done in this dissertation) and that attempts to improve upon these theories and techniques are always encouraged.

1.6 A Preview of the Chapters to Follow

In view of the fact that

1. geostatistics was originally developed for use in the mining industry,
2. geostatistics is still widely used in the mining industry, and
3. most geostatistical books are aimed at mining engineers, geologists, and other individuals working in mineral extraction,

the dissertation will begin in Chapter 2 with the presentation of selected information and fundamental concepts related to geostatistics and the mining industry. Certain mining related terms will be used throughout this dissertation, and these are defined in Chapter 2. In addition, areas of application of geostatistics in the mining industry and the statistical concept of distributions in the earth sciences will be discussed. Of particular importance is the concept of support and the support effect, which is introduced in this chapter. Support and the support effect is applicable not only in the mining industry, but in all geostatistical analyses. Also included in Chapter 2 is a brief overview of the South African mining industry.

Classical statistical theory can be effectively applied in any geostatistical analysis, and many geostatistical texts thus begin by introducing some classical statistical theory and techniques. Accordingly, Chapter 3 presents some classical statistical techniques and considerations that can be successfully applied in geostatistical analyses. Obtaining an unbiased estimate of the distribution of samples and blocks is discussed, and the theory of the two- and three-parameter lognormal distributions is presented. The graphical technique of grade-tonnage curves is also introduced.

The actual theoretical study of geostatistics begins in Chapter 4. In this chapter the probabilistic model of geostatistics is introduced, together with the important concepts of random functions and regionalized variables. Three important assumptions that are often required by the probabilistic model of geostatistics (viz. second-order stationarity, intrinsic stationarity and ergodicity) are also discussed. Finally, three alternative approaches to spatial estimation are introduced.

Chapter 5 discusses in great detail the theory of covariance functions and variograms. The requirements of valid covariance and variogram functions are put forward and various other properties of covariance and semi-variogram functions are discussed. Chapter 5 also describes the concepts of isotropy and anisotropy, and presents the most commonly used semi-variogram models in geostatistics. The estimation of covariance and semi-variogram functions, and the fitting of valid models to these estimates, is also discussed, as is the theory of robust and relative semi-variograms.

In Chapter 6, the concept of support plays a very important role. In this chapter the concepts of estimation and extension variance, dispersion variance and regularization are introduced. Additionally, the average point support covariance $\overline{C}$ and the average point support semi-variogram $\bar{\gamma}$ are defined, and various methods of calculating $\overline{C}$ and $\bar{\gamma}$ are discussed.

Chapter 7 presents the theory of the geostatistical estimation techniques called kriging. The well known geostatistical estimators of ordinary and simple kriging are introduced followed by detailed discussions of kriging techniques in general. Included in these discussions are the calculation and interpretation of kriging weights, the concept of a search neighbourhood and kriging as an exact interpolator. Also included is a study of the concept of kriging variance, the effect of smoothing in geostatistical estimation and the use of cross validation. These discussions are followed by the theoretical presentation of universal and lognormal kriging, and a section covering the issues of kriging with a relative (or proportional) semi-variogram, and testing the available data for the presence of trend.

Chapter 8 is a lengthy chapter which focuses on the application of geostatistical theory to spatial data sets. The responsibilities of a geostatistician are discussed and a general warning is given against ‘black box’ geostatistics. Following this, the tests mentioned earlier in Section 1.4 are carried out. These tests provide insight into the application of geostatistical theory to spatial data sets and highlight the effects that various decisions taken by the geostatistician can have on the final geostatistical results. Furthermore, the tests also help evaluate the performance of geostatistics under various assumptions and conditions.

Chapter 9 deals with the journal articles and letters (contesting the validity of geostatistics) mentioned earlier in Section 1.4. A summary of these articles and letters is presented, followed by an attempt to address certain of the concerns and statements raised. The chapter is concluded with some closing remarks concerning these publications.

Chapter 10 brings the dissertation to conclusion. A review of what was studied in the preceding nine chapters is presented, and the author’s concluding remarks on the dissertation and geostatistics in general is given.

Three appendices are included after Chapter 10. Appendix A contains various statistical and geostatistical tables and charts which assist in certain statistical and geostatistical calculations. Appendix B presents some basic classical statistical theory, techniques and considerations which have not been included in Chapter 3. Most of Appendix B should be well known to someone with a statistical background, but has been included here for completeness. Finally, Appendix C provides the mathematical details of the concepts of positive and negative definiteness introduced in Chapter 5.

A detailed list of references follows the three appendices and concludes the dissertation. The required format of this reference list was researched from Campbell, Ballou & Slade (1990), Faculty of Science (1995), and Central Queensland University (2004). A bibliography is not included in this dissertation, as all bibliographic works happen to be included in the list of references.

1.7 The Spatial Data used in this Dissertation

As far as the spatial data used in this dissertation is concerned, the main data set is a South African Wits type gold mine data set consisting 2321 point support samples. At each point sample location, the gold-ore thickness (measured in centimeters) and the gold-ore grade (measure in grams per ton) have been measured. This data is based on real-life South African gold mining data and has kindly been provided by Dr. Christina Dohm[†]. The other data sets used in this dissertation have been simulated by the author himself, using either *S+SpatialStats* (2001) or *Geostokos Toolkit* (c.2003). These two geostatistical software packages allow for geostatistical analysis of spatial data, and are used throughout the dissertation. *Geostokos Toolkit* (c.2003) was kindly provided for this study by Dr. Isobel Clark.

[†] Dohm, C. E. (2003, pers. comm., 4 July) and Dohm, C. E. (2005, pers. comm., 30 March).