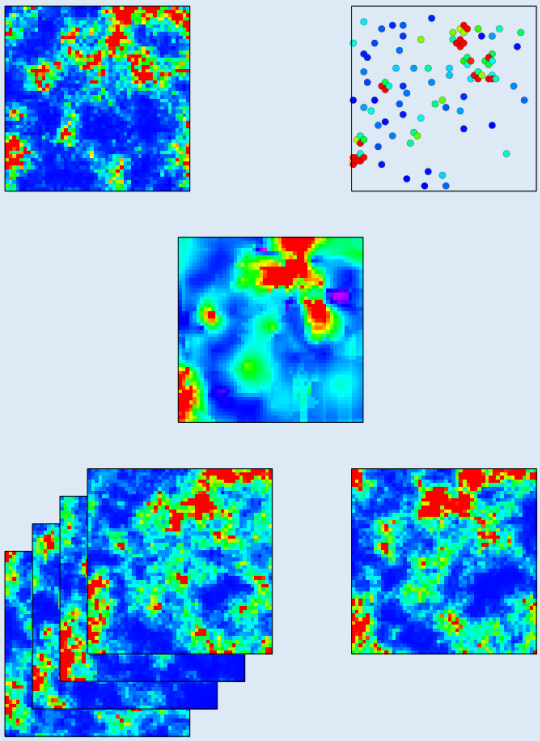
Citation in Applied Geostatistics

2021 Edition



Clayton V. Deutsch

Citation in Applied Geostatistics

Professor Clayton V. Deutsch
Ph.D., PEng, FCAE

Deutsch, C. V. (2021). *Citation in Applied Geostatistics.*

October 30, 2021

Contents

Chapter 1

Getting Started

After more than 20 years, 50 cohorts and 650 students through the program it is time for some formalized notes. The content and exercises have stabilized in the last five years. A lecture plan for the first four or five days:

Four Days	**Five Days**
• Overview • Notation, mathematics • Stationarity, composites, outliers	• Overview • Notation, mathematics • Stationarity, composites, outliers
• Probability • Statistics • Declustering	• Probability • Statistics • Declustering
• Bayes Law • Coordinate rotation, anisotropy • Grids and block size	• Bayes Law • Review
• Bivariate distributions • Covariance and correlation • Principal component analysis	• Coordinate rotation, anisotropy • Grids and block size • Bivariate distributions
	• Covariance and correlation • Principal component analysis • Review

This notes package is not exhaustive - more like a roadmap for the content being covered. Some references are given and lecture content is summarized.

Those additional references, relevant Geostatistics Lessons (cited, but see geostatisticslessons.com), papers and books will be required [35, 13, 28, 37, 9, 38].

The content is delivered in four or five days per week for four weeks. Historically, the Edmonton cohort has been four days per week and most others have been five days of instruction per week. The roadmap presented here is for four four-day weeks. Some redistribution of content, see Table above, additional review and demonstrations complete the five day version. Each Chapter is a week, each section is a day and each subsection is a lecture (8:30 to 9:45, 10-11, 11:05 to noon). There are exercises and self study in the afternoon. Participants are encouraged to work together, but everyone must submit their work individually. Virtual work is more complicated, but worth the engagement. Any question or concern that you have is surely shared by other participants in the course.

The Citation in the first few years were in response to specific demand from Chile and South Africa. Marcelo Arancibia from Maptek South America championed the formal Citation as a credential for industry. His fingerprint on the Citation is not to be under estimated. The first formal Citation through the Faculty of Extension at the University of Alberta was in Chile in 2002. Figure 1.1 shows the participants to date. There is no way to acknowledge all of the people that have had a meaningful impact on the Citation, yet I can mention a few. Graeme Lyall was there at the start. Oy Leuangthong contributed to the material and teaching in the early days. Jeff Boisvert contributes to teaching as time went on. Eric Gonzalez was there from the start and translated and helped more times than can be counted in Chile. Claudia Monreal is the modern face of the Citation in South America and has translated and helped countless students. Many teaching assistants have helped including Chad Neufeld, John Manchuk, Brandon Wilde, Jared Deutsch, Ryan Barnett, Miguel Cuba, Diogo Silva, Felipe Pinto, Ben Harding and Oktay Erten.

The project is an important part of the Citation program. This is an unscripted application of geostatistical tools with data provided by the participant (or the instructor). The results are proprietary to the participant, which facilitates integrating the project into ongoing work. An observation is that the participant will not be happy with the project at the deadline, but they should hold their nose and submit anyway. The project is a demonstration of independent application of geostatistics and consists of, approximately, a 20 page PDF document representing 100+ hours of independent work. Do the best possible and submit on time.

The project should have some background, but do not copy in large parts of a geology or previous resource report. Just enough background for the gestatistics to make sense. There should be a clear problem statement and

Year	Edmonton	Chile/SA	Denver	Other	Number	Total
1999	1					1
2000	2					2
2001	3					3
2002	2	10				12
2003	4	12				16
2004	8	8				16
2005	6	10				16
2006	19	10		**Brazil**	8	37
2007	13	19	11			43
2008	11	13		**Brazil**	12	36
2009	9	7				16
2010	16	17	6			39
2011	9	11		**RSA**	16	36
2012	12	20	14			46
2013	24	16		**RSA**	12	52
2014	22	11				33
2015	11	11	13			35
2016	16	9				25
2017	10	9	10	**Perth**	10	39
2018	7	9		**Mexico**	11	27
2019	11	13	10	**Mexico**	6	40
2020	17	12		**Edinburgh**	9	38
2021	17		20	**RSA**	29	46
Total	250	227	84		113	674

Figure 1.1: Participants at the preparation of this edition.

workflow. Some ideas for a good project: (1) a good kriged model for one variable in one rock type with exploratory data analysis (EDA), declustering, variograms, change of support, validation and model checking - this is good for beginners. (2) a drill hole spacing study with quantified uncertainty, (3) a multivariate geostatistical study that includes machine learning and geometallurgy, or (4) a geological domain uncertainty study. Details are provided during the course.

Software is essential for modern geostatistics and there are many commercial and other alternatives. The approach advocated in the course is to use pygeostat http://www.ccgalberta.com/pygeostat/welcome.html and GSLIB/CCG executables. *Any* software can be used, but participants will likely find that their commercial software has not implemented everything that is used in the class. It is recommended to use the provided software. An alternative that implements everything that I feel needs to be implemented is the RMS Platform (resourcemodelingsolutions.com/).

1.1 Introduction

1.1.1 Overview of Geostatistics

Geostatistics is a philosophical approach and a toolkit that applies statistical and numerical analysis principles to rock properties within some spatial domain. In most circumstances, less than one trillionth of the rock is sampled from the domain. Moreover, there is geological variability at all scales. The inevitable conclusion of sparse sampling and variability at all scales is that there is uncertainty. Our job is to quantify this uncertainty, communicate it and make the best decision possible. The focus of this course is the fundamental principles of: (1) geological heterogeneity modeling, (2) uncertainty assessment, and (3) decision-making and resource reporting.

Historically, science involved (1) extensive data collection and physical experimentation, then (2) deduction of laws consistent with the data. Now, many aspects of science follow a more inductive approach concerned with (1) understanding and quantifying physical laws, and (2) numerical modeling for inference. We now accept that uncertainty cannot be removed. Rev. Thomas Bayes (1702-1761) first used probability inductively and established a mathematical basis for probability inference.

Figure 1.2 sets the stage for a hypothetical question. The kidney shaped area is a geological domain under consideration. The red dots are drill holes that intersected ore. The white dots are drill holes that intersect waste. The black square is an unsampled location. The blue squiggly line represents seismic

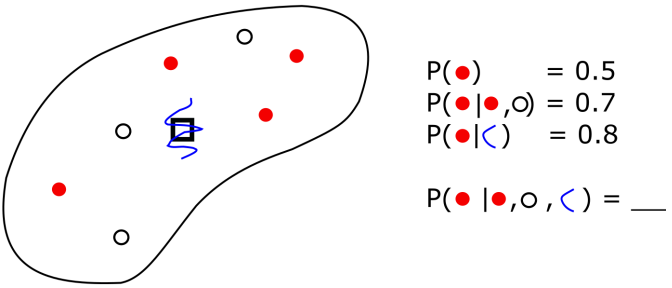


Figure 1.2: Illustration of the challenge of integrating probabilistic information.

data. The overall probability of ore is 0.5. The probability of ore at the unsampled location conditional to the drill holes is 0.7. The probability of ore at the unsampled location conditional to the seismic is 0.8. Intuitively (since assumptions would be required for calculation) what is the probability of ore at the unsampled location conditional to both the drill holes and the seismic? This is discussed in the class.

Danie Krige and Herb Sichel studied reserve estimation problems in South Africa from the 1950's establishing the problem [41]. Professor Georges Matheron (1930-2000) built the major concepts of the theory for estimating resources he named Geostatistics [45]. The Traité de géostatistique appliquée (Editions Technip, France, 1962-63) defines the fundamental tools of linear geostatistics: variography, variances of estimation and dispersion, and kriging.

The author considers that there have been four paradigms of geological modeling, see Figure 1.3 for a graphical illustration. First (I), hand drawn maps and sections - epitomized by an image of the Turin papyrus, which is considered the oldest geological map. Second (II), computer implementation of what we do by hand with the machine in a faster and, hopefully, more objective manner - epitomized by an image of Danie Krige, who (with Georges Matheron) championed the use of modern technology for resource estimation. Third (III), the quantification of uncertainty that is something we could never do by conventional hand techniques - epitomized by an image of Andre Journel's world-beating enthusiasm, teaching and mentoring. Finally, fourth (IV), the active management of uncertainty where we can change the uncertainty by our actions and optimize our decisions - epitomized by an image of Markowitz's efficient frontier, which is a result of kriging-like equations and expresses the concept that we can trade value for a reduction in uncertainty. These four paradigms help put what we do in context and prepare ourselves for the future that will include intensive automation, machine learning and artificial intelligence.

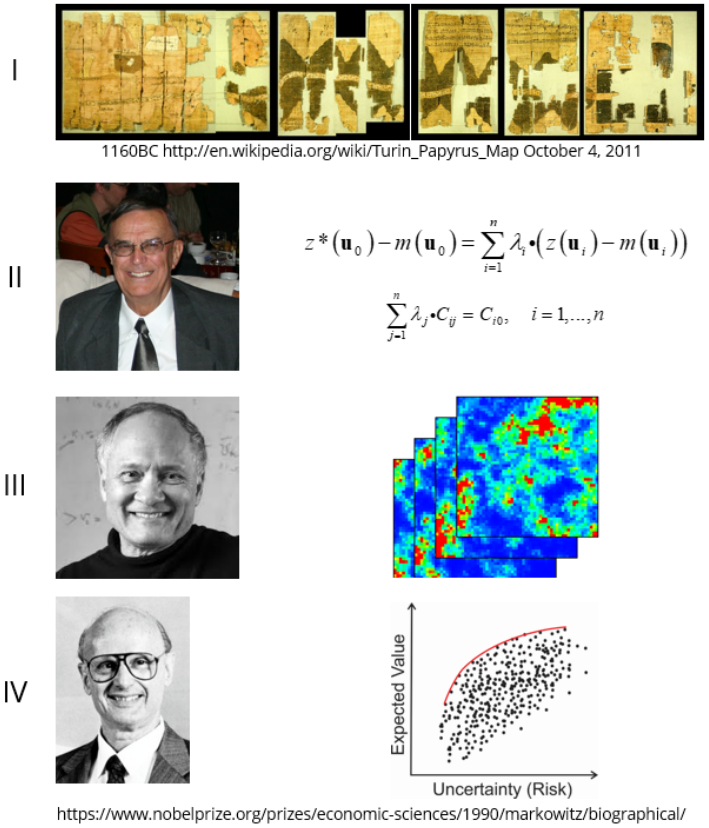


Figure 1.3: Illustration of the four paradigms of geological modeling. The photos are of Danie Krige (taken by the author), Andre Journel (from Stanford web page), and Harry Markowitz (reference given).

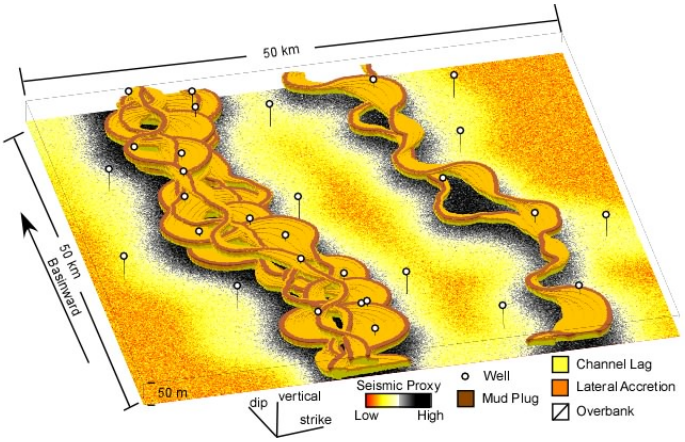


Figure 1.4: Illustration of a fluvial geological environment (Figure 3.43 from [51]). The data are related together in complex ways due to the underlying geological features and not the mechanism of choosing their locations.

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis, Geostatistics is a branch of applied statistics that places emphasis on the: (1) geological context of the data, (2) spatial relationship between the data, and (3) data of different volumetric support and precision.

There is no way to make geological data independent and identically distributed; the underlying processes impart spatial structure that is important for us to quantify and transfer into our final predictions. Figure 1.4 provides a sketch of this concept. Data that are collected a large distance apart in the basinward direction are related through the geological processes. The mechanism of choosing the data locations is less relevant than the underlying geological processes.

There is a single true distribution of properties in each deposit at any instance in geological time. The true distribution is the result of a complex succession of physical, chemical, and biological processes. Although some of these depositional and diagenetic processes may be understood, we do not completely understand all of the processes and have no access to the initial and boundary conditions in sufficient detail to predict the unique true distribution. We adopt a numerical geostatistical model because we have no alternative.

Geostatistics is useful for: (1) putting geology into numbers, (2) estimation, (3) quantifying uncertainty, (4) sample design, and (5) simulation / risk analysis. Geostatistics does not work well as a black box, replace the need for

common sense and good judgment, save time or make modeling any easier. Nevertheless, geostatistics is the best approach at present for resource modeling. Enough validation tests with k-fold validation and reconciliation to production data have been performed.

1.1.2 Notation and Some Mathematics

Upper case letters (X, $Z, \ldots$) are often reserved for random variables (RVs). These variables are not entirely random, but they are not certain; a probability distribution is used to quantify what we know and do not know about them. Lower case letters (x, $z, \ldots$) often represent outcomes of random variables - perhaps measured data or perhaps simulated outcomes. Bold font ($\mathbf{u}$, $\mathbf{h}$,) is often reserved for vectors with some component in coordinate directions. The symbol $\in$ means that something belongs to. The symbol $\forall$ means that something is true for all possibilities. The summation $\sum$ and product $\prod$ symbols are commonly used for adding or multiplying a list of numbers.

The values of a regionalized variable over a domain A could be considered a random function and denoted as follows:

$$\{Z(\mathbf{u}), \mathbf{u} \in A\}$$

We subscript a list of numbers with characters $i, j, \ldots$ or α, $\beta, \ldots$. The letter n or N is often used to denote the number of data. The letter L is often used to denote the number of realizations. The letter K is often used to denote the number of rock types or the number of variables. Multiple realizations of a multivariate block model could be denoted:

$$\{z_{k,l}(\mathbf{u}_i); i = 1, \ldots, N; k = 1, \ldots, K; l = 1, \ldots, L\}$$

My intention is to define additional notation on first usage.

Derivatives are essential for optimization. They are the slope of a function with respect to a variable. When the derivative is zero, the function is a minimum or maximum. They are rarely derived from first principals. They can be looked up in mathematical handbooks. The derivation of $f(x) = x^2$ will be given in class and the generalization will be explained. The derivatiave of $f(x) = x^n$ with respect to x is nx^{n-1} and we use this in geostatistics to minimize error variance. It is interesting to reflect on how efficient derivatives are for optimization considering the brute force alternative of exploring the entire space. In presence of 30 variables, conventional optimization would require computing the variable values that minimize 30 derivatives. If we were to try that in a brute force manner by discretizing each variable by 10 values, then we would have to evaluate the function10^{30} times, which is inconceivable.

Integration is used to solve the area or the average problem. The integration of $f(x) = x$ will be given in class and the generalization will be explained. The integration of $f(x) = x^n$ between the bounds of a and b by infinitesmal intervals dx is given by:

$$\int_a^b x^n dx = \frac{1}{n+1} x^{n+1} \Big|_a^b = \frac{1}{n+1} \left(b^{n+1} - a^{n+1}\right)$$

In general, the solution to the integration of different functions is found in the literature. It is good if the geostatistician understands integration as it relates to expected values. The fundamental theorem of calculus lays out the link between derivatives and integration.

In geostatistics, we often deal with linear sums such as $Y = a_1 z_1 + a_2 z_2 + \ldots + a_n z_n = \sum_{i=1}^n a_i z_i$. The quadratic form of a sum is a double sum: $\sum_{i=1}^n \sum_{j=1}^n a_i z_i \; a_j z_j$. The differentiation of a sum by a particular coefficient is important in our build up to kriging.

$$\frac{\partial \left(\sum_{i=1}^n \sum_{j=1}^n a_i z_i \; a_j z_j \right)}{\partial a_i} = 2 z_i \sum_{j=1}^n a_j z_j$$

This is part of the first exercise. These basic mathematical concepts are not used on a day-to-day basis by a geostatistician or resource estimator; however, they form the basis of the algorithms we use and will help the rest of the course make sense.

Matrix notation conveniently organizes numbers into rows and columns. Matrices of the same size could be added or subtracted. Matrix multiplication is based on rows multiplied by columns: $(mxn)(nxr) = mxr$ matrix. The order matters in matrix multiplication. The transpose of a matrix switches the rows and columns. Aside from simplicity of notation, we will only use matrices in the Citation to summarize dual kriging. Some additional linear algebra will be required for that, but the essence of the idea should be clear without extensive background.

1.1.3 Stationarity

All statistical analysis requires a decision of how to pool the data for statistical analysis. The data and the study area are divided into reasonable subsets. It is unlikely that we would consider all of the data and the entire study area as one domain. The basis to subset the study area / data is necessarily subjective and not a hypothesis because there is no reference to test against. We must:

1. Divide the volume of interest into reasonable domains

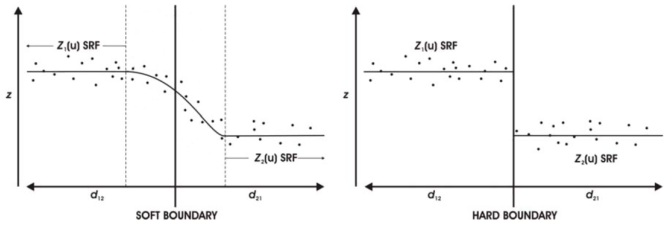


Figure 1.5: Illustration of a soft and hard boundary between two rock types. The distance axes is the distance into 2 from 1 (to the right) and into 1 from 2 (to the left). The variable is on the ordinate axis.

2. Choose whether or not to model a gradational trend in each domain
3. Establish the nature of contacts between domains - See Figure 1.5 for two schematic contact plots.
4. Determine representative statistics (perhaps with a trend)
5. Assess the uncertainty in critical statistics (usually the mean)

Each of these steps will be discussed to some extent in the course (this book). Not all of these will be covered in detail in this section.

The first aspect of stationarity is our decision of how to group the data for (geo)statistical analysis considering: (1) depositional or alteration zones of different quality/spatial behaviour, (2) rock types within the zones, (3) spatially coherent. There is often a compromise geological precision and stable statistics. The geological meaning of the subsets is important. These subsets are sometimes called geological units, rock types or domains. The precise terminology is context dependent.

The second aspect of stationarity is our decision of how the statistical parameters vary in space. When a parameter can be assumed reasonably constant over the domain, then we refer to it as stationary. When we consider the parameter to be locally varying, then we refer to it as non-stationary. Some sketches could be drawn for the mean and variance in a 1-D setting.

A more formal definition of stationarity is sometimes preferred. A formal definition comes from time series and is paraphrased as *a Stationary process is one where the unconditional joint probability distribution does not change when shifted in time.* This connotes a location independence of statistical parameters and not the initial part of stationarity, that is, the choice of a volume and the data within the volume to work with. Some specific forms of stationarity that are encountered include:

- Strict/strong stationarity - the entire probability distribution or spatial law of the variable is invariant under translation.

- Weak stationarity - 1st and 2nd moments (mean, variance, covariance, ...) are invariant under translation. First and second order stationarity are commonly assumed.
- Intrinsic stationarity - increments of the regionalized variable separated by specific lag ($\mathbf{h}$) vectors are stationary. The variable could have a first order trend, but the increments are stationary. Higher order increments could be considered for more complex trends.
- Quasi stationarity - the regionalized variable is assumed to follow constant statistical parameters within spatial windows or, commonly, search neighbourhoods.
- Trend stationarity - the regionalized variable is stationary (strong or weak) after removal of a trend. This is commonly assumed.

Increasingly in modern geostatistics we are explicitly defining or describing stationarity with auxilliary variable(s) like trend models and locally varying anisotropy direction/magnitude models.

This course is not principally concerned with data collection, sampling theory, database integrity, but these issues must be mentioned. Standard best practices should be followed in all aspects of data collection, preparation and assaying. In general, geostatistical tools have no ability to detect problem data: (1) errors appear like short scale geological microstructure, (2) biases can be detected between data sources, but the truth cannot be discerned from geostatistical analysis. Special geostatistical analysis is required for nonisotopic data or data that is not sampled at the same locations.

The term compositing refers to the procedure of combining adjacent values into longer down-hole intervals. The grade of each new interval is calculated on the basis of the weighted average of the original sample grades. These are weighted by length and possibly by specific gravity and core recovery. Compositing typically leads to an average representing the entire thickness of the zone or some regular length interval. There are special considerations for partial lengths at the end of drill holes and for partial lengths at the boundaries between domains.

The main reasons to composite are to (1) focus on a the scale of relevance for mining resource estimation, (2) filter high frequency variations to show a much better variogram, (3) mitigate the string effect in kriging to improve estimation [14], and (4) to be compliant with virtually all software that assumes the data represent the same volume. Length weighting is not a good idea in estimation - it would only be correct if the variogram is a pure nugget effect. Regular length or bench composites are common. The length should be small enough to permit resolution of the final simulated grid spacing.

Outliers and extreme values are a concern in deposits with highly skewed distributions [47]. Clearly, errors in the data should be corrected and samples

that are clearly erroneous should be rejected. We cannot blindly follow the advice in statistics books since extreme values are justifiably removed in many cases for robust statistics, but these extreme values represent a significant fraction of the metal in many deposits. We cannot blindly use all data. The goal is to ensure: (1) no conditional (local) bias in the volumes around the outliers, and (2) no unwarranted dependence on few samples. An outlier should be considered in a spatial context - a high value among other high values is of lesser concern. Cross validation could help with this. Trying to isolate the high values into their own stationary domain mitigates the problem.

An advice is to *follow local customs*. This is not an excuse to be lazy, but a practical recommendation to make the resource estimate consistent with past and similar resource estimates. We also consider probability plots; looking for inflection points. Another helpful measure is the Tukey fence [59] where the upper threshold is considered:

$$z_{\text{limit}} = z_{0.75} + k\,(z_{0.75} - z_{0.25})$$

where k is 1.5 or 3.0 depending on the degree of outlier being considered and $z_{0.75}/z_{0.25}$ are quarties of the distribution (discussed below). Metal at risk, that is, the fraction of the metal based on a small number of extreme values could be considered.

The simulation approach in the Lesson [10] is neat, but a significant amount of professional work. This would be considered in advanced projects or when significant geostatistical expertise is available.

1.1.4 Exercise W1-1

The objective of this exercise is to review some mathematical principles and to become familiar with some notation. Please write out by hand and show all important steps. Photograph/scan the pages and submit a PDF for marking.

1. Consider the following function $f(x, y) = (ax + by)(x + y)$. Calculate the derivative of this function with respect to x, that is, calculate $\partial f(x, y)/\partial x$. Also calculate the derivative of the function with respect to y, that is: $\partial f(x, y)/\partial y$.
2. Calculate the following integral:

$$\int_0^5 \left(\frac{1}{2}x^2 + x^3 - \frac{1}{4}x^5\right) dx$$

3. Consider the three matrices below:

$$\mathbf{A} = \begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 2 & 3 \end{bmatrix}$$

what is the result of **AB** and **CB**?

4. For the summation below, calculate the derivative $\partial f(\lambda_i, i = 1, \ldots, 4)/\partial \lambda_k$ with respect to λ_k where k is an index between 1 and 4. You may want to compute all four and see if they are any different.

$$f(\lambda_i, i = 1, \ldots, 4) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j$$

5. (optional) Expand the following system of equations from matrix form to summation form. In addition, solve this linear system of equations for λ_1 and λ_2 symbolically where $n = 2$.

$$\begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1) & \cdots & C(\mathbf{u}_1 - \mathbf{u}_n) \\ \vdots & & \vdots \\ C(\mathbf{u}_n - \mathbf{u}_1) & \cdots & C(\mathbf{u}_n - \mathbf{u}_n) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n) \end{bmatrix} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_0) \\ \vdots \\ C(\mathbf{u}_1 - \mathbf{u}_n) \end{bmatrix} =$$

1.2 Statistics

1.2.1 Probability

There are many variable types encountered in geostatistics. The main distinction is continuous and categorical variables. Categorical variables could be nominal, ordinal, binary or a count. In most cases we consider them to be a nominal description or an indicator of a stationary domain. Continuous variables are mostly grades, thicknesses and variables that average linearly, but variables that average non-linearly are being encountered increasingly in geometallurgy and more complete rock characterization, see Figure 1.6. A power law transform is sometimes condidered to linearize variables [12]. Compositional data will be addressed with multivariate.

For practical purposes, the meaning of probability is clear. It is the proportion of times something would occur under similar circumstances. Categorical or discrete variables can take one of K states. Uncertainty is represented as the probability of each state: $p_k, k = 1, \ldots, K$. Probability is proportion of times each state will occur under similar circumstances. Must be nonnegative and sum to one. Perfect knowledge would be one p value equal to 1 and all others zero. Complete uncertainty would be $p_k = 1/K$ for all categories. Prior proportions often tell us that some states are more likely. Virtually all categorical variable distributions are non-parametric, that is, specified by proportions derived from data. Benford's Law and the Binomial distribution are two parametric categorical variable distributions that are sometimes used.

Benford's Law is an interesting categorical variable parametric distribution

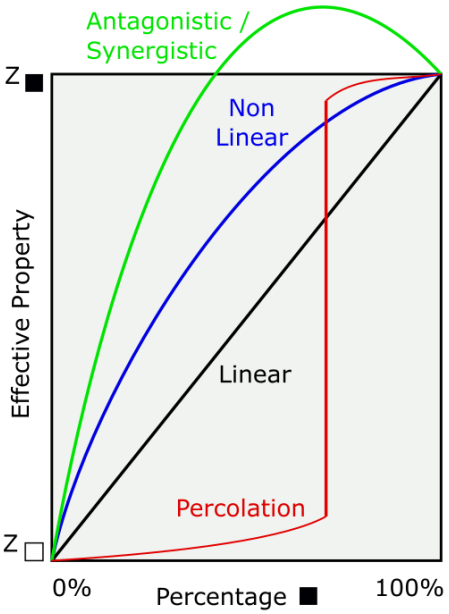


Figure 1.6: Illustration of the nature of averaging of different continuous variables.

for the probability of non-zero leading digit in natural numbers spanning several orders of magnitude:

$$p(d) = Log_{10}(1 + 1/d) \qquad d = 1, \ldots, 9$$

This could be used to detect fraudulent data. The other categorical variable distribution used in Mining and Petroleum is the Binomial distribution for probability of k successes in n independent trials where p is the probability of one trial:

$$f(k, n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

As mentioned, these are uncommon. Non-parametric and, often, non-stationary distributions are used for categorical variable distributions.

The universal approach to represent uncertainty in a continuous variable is the cumulative distribution function (CDF). Replace the unknown truth z_{true} by a random variable Z. The cumulative distribution function (CDF) of an RV is defined as:

$$F(z) = \text{Prob}\{Z \leq z\}$$

The CDF $F(z)$ is non decreasing within $[0, 1]$. $F(z_{min}) = 0$ and $F(z_{max}) = 1$. The CDF is flexible in representing great uncertainty ($F(z)$ spread over a large range of z) and certainty ($F(z)$ a step function at the certain z value).

There are three ways to infer a CDF:

1. Degree of belief - for global parameters that are inaccessible
2. Proportions from data - global probabilities and distributions
3. Mathematical model - local conditional parameters

The probability of Z occurring in an interval from a to b (where $b > a$) is the difference in the cdf values evaluated at points b and a. The probability density function (pdf) is the derivative of the cdf, if it exists:

$$f(z) = F'(z) = \lim_{dz \to 0} \frac{F(z + dz) - F(z)}{dz} \geq 0$$

The CDF and PDF contain equivalent information. It may be convenient to consider one in mathematical calculations / computer implementations and another in visualization or for other calculations.

A parametric distribution is one where $F(z)$ or $f(z)$ is given by an equation with some parameters. There may or may not be a generating mechanism for the probabilities. Most earth science distributions are non-parametric and inferred by proportions from data.

The generating mechanism for the Gaussian distribution is summarized in central limit theorem (sum of many iid $\rightarrow$ Gaussian), as shown on Figure 1.7.

$$y = G^{-1}(F(z)) \text{ or } Y = \frac{Z - m}{\sigma} \quad ; g(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}$$

The multivariate Gaussian distribution is used extensively in geostatistics because it is simply parameterized and tractible.

The lognormal distribution is closely related to the Gaussian or normal distribution and is also common in geostatistics. $Z \rightsquigarrow \text{logN}(m, \sigma)$ when $Y = ln(z)$ is $N(\alpha, \beta^2)$. The parameters are related by:

$$\alpha = ln(m) - \beta^2/2 \qquad \beta^2 = ln\left(1 + \frac{\sigma^2}{m^2}\right)$$

$$m = e^{\alpha + \beta^2/2} \qquad \sigma^2 = m^2\left(e^{\beta^2} - 1\right)$$

Considering a bivariate distribution that we have not yet talked about, the correlation in Gaussian and lognormal units is also related:

$$\rho_Z = \frac{m^2}{\sigma^2}\left(e^{\beta^2 \rho_Y} - 1\right)$$

This relationship is used in back transforming the variogram of log-grades, which could be much more stable than the variogram of the original units.

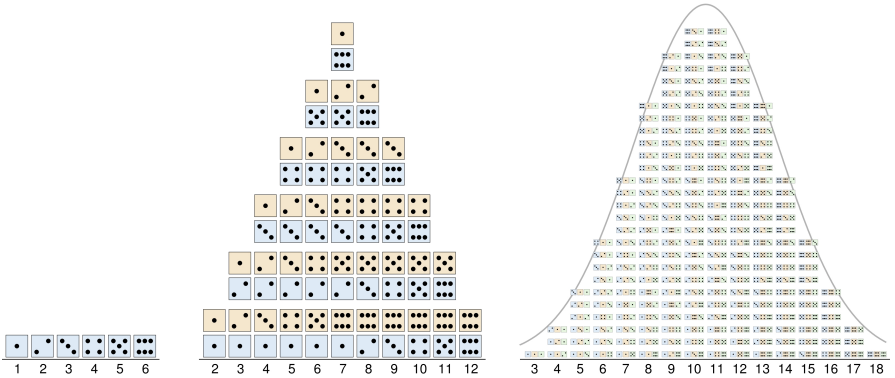


Figure 1.7: Illustration of the generating mechanism underlying the Gaussian distribution - the sum of independent and identically distributed random variables.

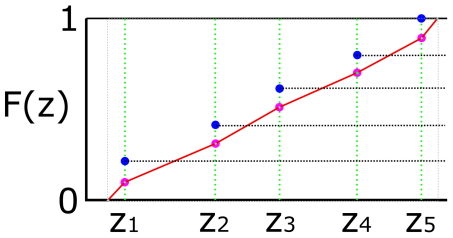


Figure 1.8: Illustration of a non parametric distribution.

The uniform distribution will be considered in the exercise. It is simple and amenable to an exercise to make participants more familiar with the mathematics, notation and probabilistic concepts. It is also important since all probabilties for a given CDF are uniformly distributed. This is true regardless of the variable or the distribution. Fixing the distribution as a marginal distribution, then conditioning leads to the beta distribution.

A Nonparametric distribution comes from the data themselves. There are often enough data. Geological data do not follow a convenient parametric model. It is possible to transform any univariate distribution to any other distribution if required for mathematical calculations.

The procedure to build a non-parametric distribution: (1) sort the data in ascending order, (2) compute cumulative weights to be less than or equal to, (3) average with the value from the preceeding data, and (4) interpolate between the points to fill in $F(z) \forall z$. See Figure 1.8 for an illustration.

1.2.2 Statistics

Statistics could refer to a large field of study or to summary measures that characterize probability distributions. This section focuses on the latter.

The p-quantile of the distribution $F(z_p)$ is the value z_p for which $F(z_p) = p$. Thus, the quantile can be expressed in an inverse form of the CDF $q(p) = F-1(p)$. The probability value lends meaning to a particular number. The P10, P50 and P90 are important quantiles used to summarize a distribution. The three quartiles, that is, $z_{0.25}$, $z_{0.5}$ and $z_{0.75}$ divide the distribution into four equal parts. The interquartile range (IQR) is another measure of dispersion of variation of a distribution. The difference between the mean and the Median, m-$z_{0.5}$, is a measure of skewness.

A Q-Q plot compares multiple univariate distributions to a reference distribution. Q-Q plot is a plot of matching quantiles a straight line implies that the two distributions have the same shape. A shift up or down indicates a difference in the center of the distribution. A slope different from 1 indicates a difference in the variance. A non-linear shape indicates a difference in the shape of the distribution.

A Q-Q plot is *not* for paired values. A scatterplot is used for that. Some valid applications of a Q-Q plot include (1) different measurements (DDH versus BH, log versus core, ...), (2) checking the histogram of a 3-D model (representative data distribution versus the 3-D model), (3) comparing distributions of a variable within different rock types, (4) comparing distributions within different drillholes, and (5) comparing different models. An example Q-Q plot is shown on Figure **??**.

The expected value is a statistical operator that is the probability weighted average of the random variable: Expected value of a constant is the constant. The expected value is a probability weighted average:

$$E\{\cdot\} = \int_{z_{min}}^{z_{max}} \cdot f(z)dz$$

The most important expected value is the mean:

$$m = E\{Z\} = \int_{z_{min}}^{z_{max}} zf(z)dz$$

In theory, the expected value is an integral (probability weighted average); in practice, the expected value is estimated by a weighted average:

$$\hat{m} = \sum_{i=1}^{n} w_i z_i$$

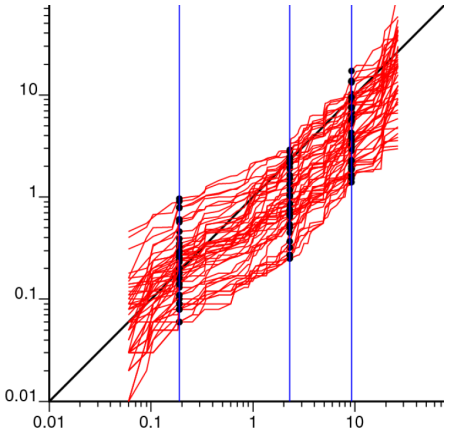


Figure 1.9: An example Q-Q plot comparing the distributions from 100 realizations against the reference distribution for the stationary domain. Note that the result do not look particularly good - the distributions appear systematically biased on average.

The mean is also known as the first moment, that is, the center of mass of the probability density function. The variance is a second order moment defined as:

$$\sigma^2 = E\{[Z - m]^2\} = E\{Z^2\} - m^2$$

The variance is a measure of the spread of the data from the mean. The standard deviation is the square root of the variance. It also measures data variability from the mean. The dimensionless coefficient of variance (CV) is the ratio of the standard deviation over the mean (σ/m). This works for positive variables. Conventional wisdom indicates that a variable with a $CV < 0.4$ is not that variable and potentially straightforward to model. A variable with $CV > 2.0$ is quite variable and the domain should be considered for subdivision. These are not fixed values, but general guidelines.

The mean is the correct number to replace a probability distribution when the variable under consideration averages linearly. Consider mass percentage measurements of 9, 1, 1 and 1 for four units of rock of equal mass. If the piles were combined, the mass percentage would be 3 and not a quantile or a more robust estimator. This is true for mass fractions, volume fractions, rock type proportions and other variables that average linearly. The correct effective value is not the median, mode, Exceptions are when the variable does not average linearly such as color, rate constants (permeability, work index, etc.).

Many geostatistical calculations require continuous variables to exactly follow the standard normal distribution, that is, $N(0, 1)$ normal with a mean of zero and a variance of one. The normal score transform is used to transform non-

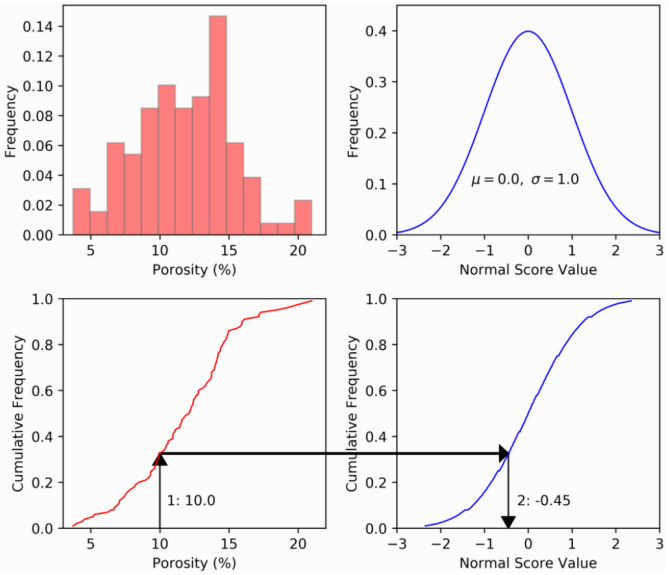


Figure 1.10: Illustration of the normal score transform (see Lesson). A value of 10 is transformed.

normal data following any distribution to a Gaussian or normal distribution. The steps: (1) determine the representative Z-variable distribution $F_Z(z)$, (2) transform each z data by matching quantiles, $y = G^{-1}(F_Z(z))$ where $G(y)$ is the CDF of the $N(0,1)$ distribution. Figure 1.10 shows this for one z value of 10.0. This is taken from the Lesson [52] on the normal score transform.

Regarding the normal score transform, a representative distribution $F_Z(z)$ is the most important consideration. Spikes of constant values coming from detection limit or round off are also a concern. The ties must be broken to avoid a spike in Gaussian units that is not allowed. Entirely random despiking leads to a too-high nugget effect in the final variogram. A combination of a local average and random despiking has been shown to work well. Another option is to isolate the barren/unmineralized material into its own domain/rock type.

1.2.3 Declustering

The data are considered trustworthy (although every measurement is associated with some error), but they are almost certainly collected in a preferential manner. Correcting the statistics for this will be done with declustering and debiasing. Although clear when the mathematics are reviewed, it is impor-

tant to note that declustering does **not** change the data values. The data remain unchanged. The weight applied to the data may vary. The representative distribution for transformation and other calculations may come from another source, but the data do not change.

Data are rarely collected to be statistically representative: Interesting (best) areas are delineated more completely, Samples taken preferentially from good quality rock. These data collection practices should not be changed; they lead to the best economics and the greatest number of data in portions of the study area that are the most important. There is a need, however, to adjust the histograms and summary statistics to be representative of the entire volume of interest

Declustering techniques assign each datum a weight based on closeness to surrounding data: $w_i, i = 1, \ldots, n$ histogram and cumulative histogram use $w_i, i = 1, \ldots, n$ instead of $1/n$. Debiasing uses a secondary variable or trend to establish representative histogram or proportions

Historical mapping algorithms correct for preferential sampling: no need for declustering in inverse distance or OK. There is a need for representative proportions and histograms in modern geostatistics: Checking models, Global resource assessment, and as an input to simulation. Simulation does not correct for preferential sampling even though kriging is used inside simulation. Cokriging with a secondary data does not correct the distribution: correlation with the rank order of the data is used the conditional distributions are not used directly

Declustering weights are taken proportional to the areas or volumes of interest. Weights are very sensitive to the border. Despite the apparent criticism of nearest neighbor declustering, the technique is being used increasingly and perhaps even more than the competing techniques discussed below.

Cell Declustering is considered to be more robust in 3-D and when the limits are poorly defined, see Lesson [18] and Figure 1.11:

1. divide the volume of interest into a grid of cells $l = 1, \ldots, L$
2. count the occupied cells Lo and the number in each cell $n_l, l = 1, \ldots, L_o$
3. weight inversely by number in cell (standardize by L_o)

Fixing the cell size and changing the origin often leads to different declustering weights. To avoid this artifact a number of different origin locations are considered for the same cell size and the weights are averaged for each origin offset, see Figure 1.12. It is highly unlikely, but if a data falls on a cell boundary it is randomly moved to an adjacent possible cell.

The cell size should be the spacing of the data in the sparsely sampled areas.

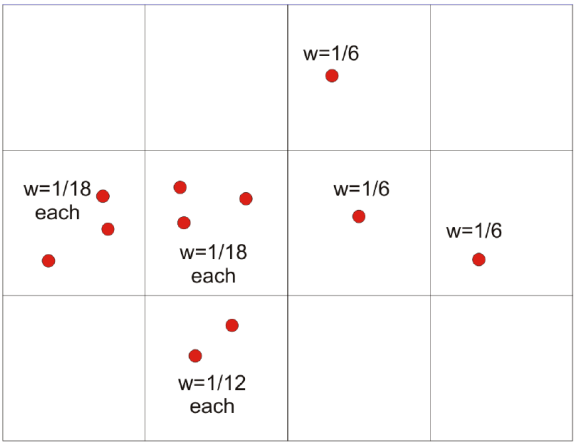


Figure 1.11: Illustration of the concept of cell declustering. A grid of cells is placed over the data, the number of data within each cell counted and the weight is inversely proportional to the number of data in the cell.

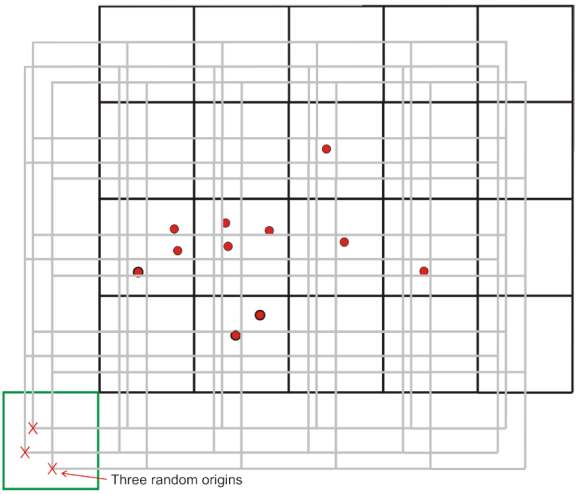


Figure 1.12: Illustration of how the origin of the cell declustering grid network is moved to stabilize the results.

The declustered mean is often plotted versus the cell size for a range of cell sizes a diagnostic plot that may help get the cell size. The long range features of the diagnostic plot are often influenced by trends and non-stationary features.

Declustering by Conventional Estimation amounts to perform an estimation of grid block values within the applicable RT limits and within reasonable distance to the data. Consider inverse distance or kriging. Accumulate the weight given to each data in the estimation procedure. Clustered data will receive less weight overall. Sensitive to the search, variogram, May not work well if the string effect of kriging is a problem. Results of simulation are perfectly consistent with kriging a nice property.

A rhetorical question: what do we do when there are too few data or the data are not representative? Nothing, unless there is some secondary information. An empirical calibration approach is sometimes referred to as *debiasing* or *soft data declustering*: (1) map a secondary variable X at all locations that could be geophysical, structural or some kind of trend model, (2) develop a bivariate relationship between X and the Y variables, and (3) generate a distribution of Y by combining conditional distributions. The secondary variable is something geometric, geophysical or geological.

There is no recipe for correct application. Our goal is to go beyond a limited sample to the underlying true population. It is essential to decluster facies proportions as well as the distributions of continuous variables. Future geostatistical analysis will depend heavily on simple statistics inferred early in the modeling efforts.

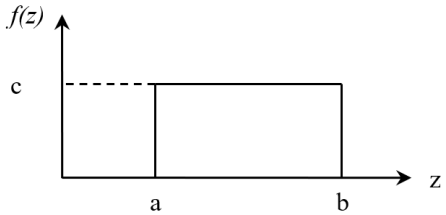
The volume of influence method is gaining ascendancy - fewer artifacts than thought. The sensitivity to the border actually helps infer a representative distribution. Trend models somewhat alleviate the need for declustering, but a good trend model will be built considering declustering weights.

1.2.4 Exercise W1-2

The objective of this exercise is to review some probabilistic concepts and to continue becoming familiar with notation. Please write out by hand and show all important steps. Photograph/scan the pages and submit a PDF for marking.

Consider the uniform distribution specified below:

1. Write the definition and equation for the cumulative distribution function (cdf) of the uniform distribution.



2. What is the value of c that makes $f(z)$ a licit probability distribution? Write your answer in terms of a and b.
3. What is the expected value (or mean) of the variable Z in terms of a, b, and c? Work out the integral.
4. What is the variance of the variable Z in terms a, b, and c? Work out the expected value of Z^2 and solve for the variance using $\sigma^2 = EZ^2 - [EZ]^2$.
5. What is the width of the 90% probability interval expressed in terms of a and b? Use the results of Question 1 and solve for the 5th and 95th percentiles.

1.3 More Prerequisites

1.3.1 Bayes' Law

The arithmetic of probability intrigued Thomas Bayes. He thought deeply on the subject and formulated the fundamental principles that are still used today. It was clear to Bayes that probability is *the proportion of times something would happen in similar circumstances*, probability could not be negative and the sum of probabilities would be one in a closed set. The definition of a conditional probability and Bayes Law are fundamental to probability and statistics:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(A|B) = P(A) \cdot \frac{1}{P(B)} \cdot P(B|A)$$

Conditional probability is the prior probability modified by the evidence. The rarity of the evidence and the relevance of the evidence (likelihood) enter the equation. Additional details and an example to geostatistical mapping are contained in a Lesson [21].

The example on Figure 1.13 is developed in the class. The definition of conditional probability is used extensively in geostatistics. The meaning of

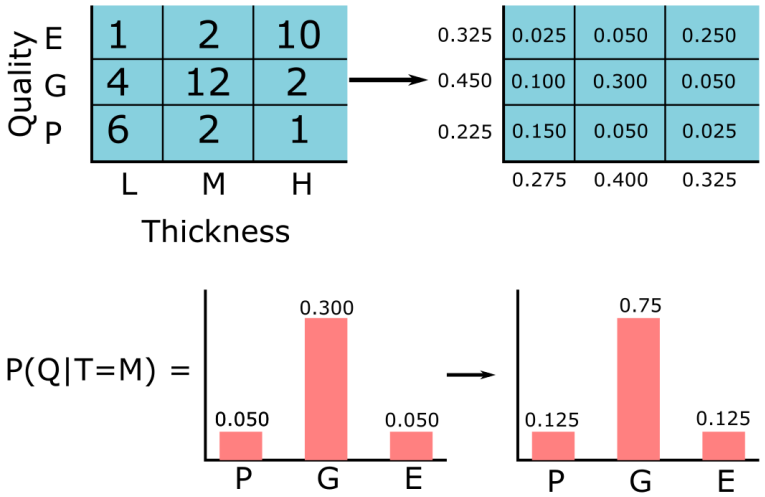


Figure 1.13: Illustration of the definition of conditional probability.

this is revealed in the requirement to standardize probabilities once evidence is available.

1.3.2 Coordinate Rotation and Anisotropy

Coordinates and anisotropy are important topics in geostatistics. The fundamentals of this will also set the stage for principal component analysis (PCA) coming later in the week.

The spatial coordinates of a deposit form a metric space with a measure of distance d where: $d(x, y) \geq 0$, $d(x, y) = 0 \iff x = y$, and $d(x, y) + d(y, z) \geq d(x, z)$. Measures of distance include:

- Chebyshev distance - maximum along any coordinate
- Manhattan distance - measured along axes at right angles
- Euclidean distance - ordinary straight line

$$h = \sqrt{\left(\frac{h_{maj}}{a_{maj}}\right)^2 + \left(\frac{h_{min}}{a_{min}}\right)^2 + \left(\frac{h_{ter}}{a_{ter}}\right)^2}$$

- Minkowski Distance - p exponent ($p_M = 1, p_E = 2$)

Shortest path distances, LVA, Dijkstra algorithm ... and fast marching methods provide an avenue for improved geological modeling.

Real world coordinates are often UTM Easting (X), UTM Northing (Y) and

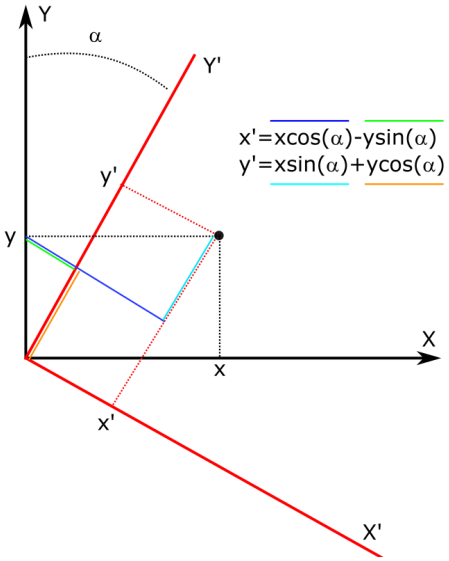


Figure 1.14: Illustration for the rotation around one axis

meters above sea level (Z). These coordinates lead to a natural left hand rule. Imagine your left hand with the thumb sticking upward (Z), the middle finger pointing to your right (X) and your pointing figure pointing away from you (Y). This is the starting point for coordinate rotation.

New coordinate systems should never be invented on the fly. Modern software allows for rotated grid models without changing the actual coordinates.

Aeronautics definition of anisotropy with roll, pitch and yaw is very intuitive, but not used in geology.

Coordinate rotation: left hand rule. The Lesson [26] is helpful. Consider a succession of three 2-D rotations (see Figure 1.14 for elementary rotation:

1. Clockwise positive around vertical (angle one)
2. Down negative around what was X (angle two)
3. Polite positive around what was Y (angle three)

Every software is different. Three angles of rotation specify the major, minor and tertiary directions. The standardized Euclidean distance enters virtually all calculations.

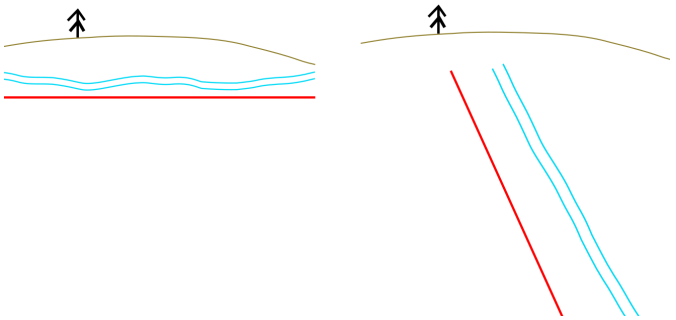


Figure 1.15: Two illustrations of typical tabular deposits that would be considered within the framework of simple tabuluar deposits.

1.3.3 Grids for Geologicical Modeling

Cartesian coordinates and regular 3-D grids are fairly straightforward to understand. Stratigraphic-like grids for tabular deposits are important to capture undulations and thickness variations. Generalized tetrahydra grids will become more popular as software libraries are established.

Considering a regular Cartesian grid, an important consideration is the grid size. There are many considerations for this including (1) the supporting data for predictions - one quarter the data spacing is a reasonable *natural* grid size recommendation (smaller grid sizes will not be estimated more reliably and larger grid sizes would lose information), (2) the desire to respect geological boundaries - smaller blocks or subblocks are desired to capture surfaces that would not be represented by coarse blocks, and (3) engineering considerations where the block size is influenced by pit or stope design considerations. A high resolution grid is recommended even if the block estimates are no better than a lower resolution grid - at least the result will respect geological boundaries and support engineering calculations.

Tabular Deposits

Simple tabular deposits are commonly encountered. There are stratigraphic deposits, lateritic weathered deposits, simple veins, shear zones and many other situations where a tabular geometry is encountered. A first consideration in the case of tabular deposits is to establish a reference plane. The top or hangingwall and bottom or footwall are established relative to the reference plane. Figure 1.15 shows two example reference surfaces; one flat and one steeply dipping.

Given arbitrarily oriented data that intersect a tabular structure, we could consider total least squares plane fitting $\pi : ax + by + cz = d$ to determine the reference plane. The Eigenvector corresponding to the minimum Eigenvalue of $A^T A$ defined the normal vector:

$$A = \begin{bmatrix} x_1 - \overline{x} & y_1 - \overline{y} & z_1 - \overline{z} \\ x_2 - \overline{x} & y_2 - \overline{y} & z_2 - \overline{z} \\ \vdots & \vdots & \vdots \\ x_n - \overline{x} & y_n - \overline{y} & z_n - \overline{z} \end{bmatrix}$$

This is like PCA (defined below) where the major axes of continuity are calculated through a cloud of points. The $A^T A$ matrix could be considered a covariance matrix (defined below). In any case, by default (horizontal or vertical), manual specification or automatic calculation a reference plane must be defined to model tabular deposits.

The commonly accepted approach is to model one position and multiple thickness values relative to the reference plane. Modelling multiple positions would often lead to crossing surfaces or unreasonably large thicknesses. The most continuous surface would be chosen for the position variable, then thicknesses relative to that would be modeled.

Figure 1.16 illustrates the overall approach. Eight drill holes are shown at the top where three do not intersect the structure - the drillholes with dashed lines are the ones that do not intersect the structure (Figure 1.16.a). The red line is the reference plane and the black dots represent the top and base of the structure of interest. Figure 1.16.b illustrates the modeling of the top structure (chosen as the position variable) by the blue line. There may be multiple realizations for the top structure. The drill holes that do not intersect the structure are not considered in modeling of the top position. Figure 1.16.c shows the thickness modeled as the green surface below the blue datum of the reference position. Once again, the drill holes that do not intersect the structure are ignored. Figure 1.16.d shows the modeled position and thickness put back together relative to the reference plane. Finally, Figure 1.16.e shows that the extent of the position and thickness are clipped to account for holes and for limits within the plane of continuity.

Missing values should be ignored, then clipped by boundary modeling. The exact nature of the pinch out or termination may not be exactly reproduced, but this is of little consequence in general given the thin and laterally extensive nature of these deposits.

Modeling continuous properties within tabular deposits defined by a top and bottom (perhaps multiple realizations of the top and bottom are available) required the use of stratigraphic coordinates. Figure 1.17 shows four common cases for stratigraphic coordinates: proportional, erosion - base conforming,

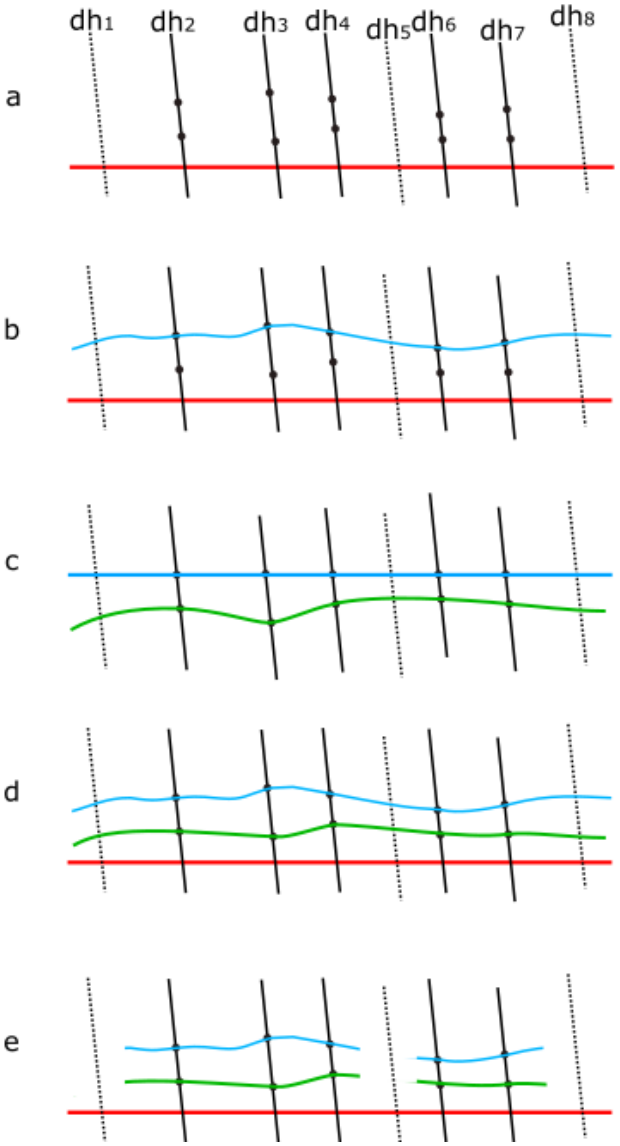


Figure 1.16: Illustration of how we model the position, then thickness, then assemble everything in original units.

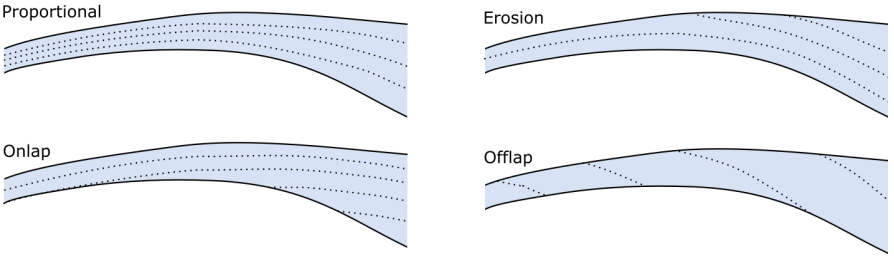


Figure 1.17: Four common cases for stratigraphic coordinates.

onlap - top conforming, and offlap - arbitrary surface conforming.

The coordinate perpendicular to the reference plane (Z in the equation below) is modified so that a flattened or relative coordinate is considered.

$$z_{rel}(x,y) = \frac{z(x,y) - z_{base}(x,y)}{z_{top}(x,y) - z_{base}(x,y)} \cdot T$$

Where x and y represent coordinates define the reference plane of continuity. If the transform is proportional, then the top and bottom are taken locally. If the transform is erosion, then the top is replaced by the bottom translated upward by a constant. If the transform is onlap, then the bottom is replaced by the top translated downward by a constant. The top and base are defined arbitrarily in the case of offlap. This transformation can be reversed at any time.

The 3-D models of tabular deposits should not be shown in flattned space. Highly deviated drill holes relative to the plane of continuity are particularly problematic; some form of geometric data imputation must be considered.

1.3.4 Exercise W1-3

The objective of this exercise is to become familiar with the use of declustering to infer a representative probability distribution. The declustering software in GSLIB or any other software could be used.

1. Consider the Au variable in the skarn2d.dat data file. Decluster the data set using cell declustering. Cell declustering is widely used because it is robust and not very sensitive to edge effects.
   Plot a location map of the sampling locations and propose a reasonable cell declustering cell size based on the data spacing in sparsely sampled areas. Also plot a naive (no declustering) histogram of the variable of interest.

Run cell declustering for a range of cell sizes: explain your choice of parameters in your assignment. Plot the declustered mean versus cell size, then on the basis of this plot and data spacing, choose a cell size, and justify your choice.
Plot a histogram of the variable using the declustering weights and comment on differences with the unweighted histogram.

2. One method to construct a trend model for potential soft data declustering (debiasing) and global mean inference is global kriging. Use global kriging with a long range and large nugget effect for creating a trend model. Visualize the trend model and calculate a histogram of the trend model for the global mean. Compare with the previous question.
3. Consider the Au variable in the Misima.dat data file. Have a quick look at the data and run declustering following the principles of Question 1.

1.4 Bivariate Statistics

1.4.1 Bivariate Distributions

Geostatistical modeling considered a very high dimensional distribution of variables and locations. This high dimensional distribution is almost always understood through models of the bivariate distributions. Figure 1.18 illustrates the magnitude of the distributions we are dealing with.

The bivariate CDF tells us everything about how two variables are related:

$$F_{X,Y}(x,y) = \text{Prob}\{X \leq x; Y \leq y\}$$

This is a contour map in the space of the two variables X and Y. The bivariate PDF contains equivalent information:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

Like the univariate case, the PDF integrates to the CDF. It may be convenient to consider one or the other for visualization or mathematical purposes. Figure 1.19 shows a bivariate with marginal, joint and conditional distributions.

In many cases of presenting bivariate distributions it matter which variable is shown on the X abscissa axis and which is shown on the Y ordinate axis. If one of the variables could be considered an independent or data variable, then it is shown on the X. By convention, the dependent variable is shown on the Y axis. The only conditional distribution that matters is the one of $Y|X = x$. Connecting the expected values of the conditional distributions of

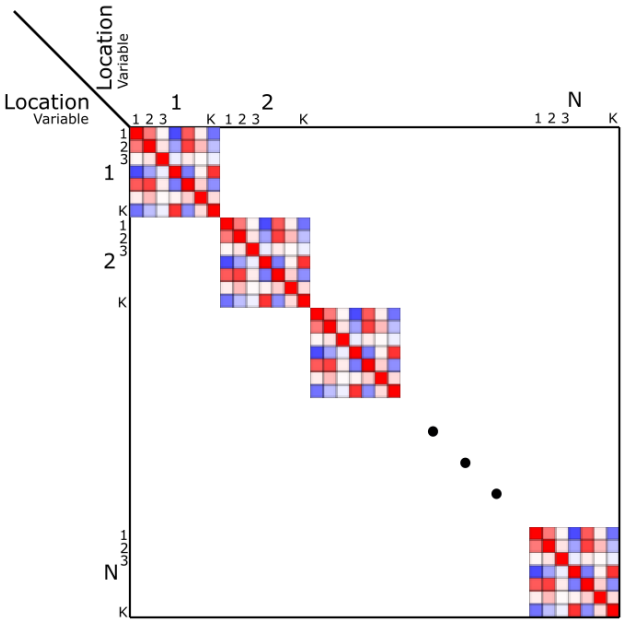


Figure 1.18: Schematic illustration of the high dimension of geostatistical problems: K variables at N locations.

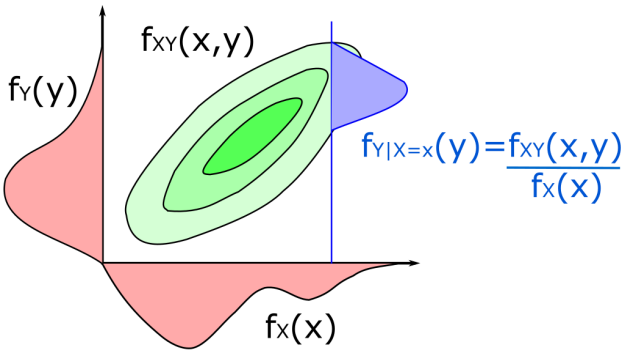


Figure 1.19: Bivariate distribution with marginal and conditional.

$Y|X = x$ would generate a conditional expectation or regression curve. Given knowledge of $X = x$, then this curve would give the best estimate of Y.

The bivariate Gaussian probability distribution is sometimes used. There are few parametric distributions. The Gaussian PDf is given by:

$$g_{XY}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}[x^2 - 2\rho xy + y^2]\right\}$$

$$E\{Y|X = x\} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

$$Var\{Y|X = x\} = \sigma_Y^2(1 - \rho^2)$$

Some properties of this distribution are seen again in the multivariate distribution: (1) conditional expectations are linar functions of conditioning data, and (2) conditional variances are data-value independent.

1.4.2 Covariance and Correlation

Univariate distributions are often summarized by their first two moments: the mean and variance. A bivariate distribution is summarized by the covariance or the standarided correlation coefficient. The covariance is a two-point measure used throughout statistics and geostatistics. The covariance between random variables X and Y is written:

$$\begin{aligned} Cov\{X, Y\} = C_{XY} &= E\{[X - m_X][Y - m_Y]\} \\ &= E\{XY\} - m_X m_Y \end{aligned}$$

The expected value is a double integral over X and Y. In practice it is calculated as a sum over available pairs. The covariance is a measure of the closeness to a perfect linear relationship. If the linear relationship is a direct one, then the covariance will be positive and if the relationship is an inverse one, then the covaraince will be negative. Figure 1.20 illustrates this scematically.

The covariance has the units of X multiplied by Y. It is common to consider the correlation coefficient, that is, the standardized covariance that falls within -1 and +1:

$$\text{correlation } \rho = C_{XY}/(\sigma_X \cdot \sigma_Y) \quad \in [-1, +1]$$

If the bivariate data were bivariate Gaussian, then the covariance or correlation with the variances would completely define the distribution. In many practical cases, there are complex features in the data that make the covariance an incomplete measure.

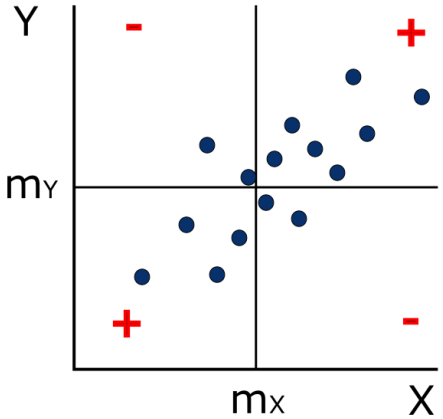


Figure 1.20: Illustration of how a direct relationship leads to a positive covariance and an inverse relationship makes the covariance negative.

1. Outliers can enhance an otherwise poor correlation or reduce an otherwise good correlation
2. Non-linear relationships would be captured to some extent by the correlation coefficient, but the conditional distributions would have less variance than implied.
3. Constraints from mineralogy or fractional measurements (total copper and acid soluble copper, for example) are not fully captured by the correlation coefficient.
4. The proportional effect or heteroscedasticity are not captured by the correlation coefficient.

The correlation coefficient described above is sometimes called the Pearson correlation coefficient - named after Karl Pearson a famous statistician. The correlation coefficient is sensitive to extreme values (outliers) and an alternative was proposed by Charles Spearman. The Spearman correlation coefficient is also known as the rank correlation coefficient. The idea is to rank transform the both variables from 1 to n - preserving the order, then compute the Pearson correlation coefficient on the rank order values and not the original data values. The Spearman correlation coefficient could be used to establish whether extreme values are destabilizing the correlation.

The correlation coefficient is not the coefficient of determination R^2 although it would be R if the distribution were bivariate Gaussian. it is better to reserve the use of R^2 for a measure of goodness in prediction. This will be covered later.

Uncorrelated does not equal independence - there could be some dependence

caused by one or more of the considerations mentioned above.

Correlation does not imply causation. This is a mainstay of modern statistics. There are many examples including the correlation between crime rate and number of places of worship per square kilometer. They may be related, but through something else such as population density. Using variables without cause and effect is still possible for prediction if the results are stable; just not used for explanation.

1.4.3 Principal Component Analysis

Principal Component Analysis (PCA) is well established and very interestingin modern data science and geostatistics. It is nice to see something novel and interesting in the first week - aside from an overview and review of statistics.

PCA is a classic dimension reduction and decorrelation technique that was developed by Pearson [48] and Hotelling [33] and adapted to geostatistics in the 70s and 80s. PCA transforms multivariate data that are correlated in an arbitrary manner into orthogonal linear combinations of the original variables, that is, factors that are all uncorrelated. It is useful to imagine rotation in the high dimension space of the data. The position of the data in the high dimensional space of the data does not change, but the new coordinates of the PCA make the data appear uncorrelated. The direction where the data show the most variance is the first principal component and they are ordered in decreasing variance. The Lesson [4] by Ryan Barnett is a good place for an overview.

The reasons to consider PCA are threefold.

1. Understand underlying factors - there are times when understanding the principal components - the independent directions that explain the variability in the data, explains aspects of the nature of the multiple variables. This is sometimes possible with data in geostatistics.
2. Reduce dimension - the principle components are ordered in decending order of the variance they explain. At a certain point, perhaps when 99% of the variance has been explained, the remaining pricipal components can be dropped from the analysis.
3. Decorrelation - the principal components are uncorrelated with each other and it is possible to simulate them independently, then the correlation structure is reintroduced in the reverse rotation.

Consider K variables that are standardized (centered and scaled to have a variance of one). This could be done by a simple standardization (subtract the

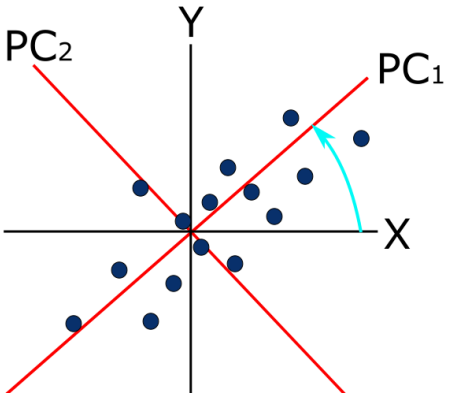


Figure 1.21: Illustration of the rotation to make two variables (labeled X and Y here) to appear uncorrelated.

mean and divide by the standard deviation) or by a normal score transform. We would often consider an normal score transform since the shape of the distribution and outliers are managed. The only reason to consider a simple standardization is if kriging or some other linear technique is being considered for the PCA factors.

From an intuitive or graphical perspective, PCA amounts to rotate the K dimensional coordinate system to make the variables uncorrelated. The values represented in the rotated space are referred to as principal components. This is done by the spectral decomposition of the KxK covariance matrix of the variables. The principal components are uncorrelated, but do not have unit variance. They are ordered so that the first has the greatest variability and the last has the least. Details of this are given in the fourth week (last Chapter), but participants are asked to go through the steps here to gain an understanding of covariance and multivariate data.

The concept of PCA linked to coordinate rotation is powerful. Think about leaving the multivariate data at their locations and rotating the coordinates so that the data appear uncorrelated (see Figure 1.14 for an illustration). Rotation in a high dimension is difficult to imagine, but the analogy is correct.

The mathematics of PCA is based on the covariance (same as correlation since the variables are standard) matrix of the original data: $\mathbf{\Sigma}$. A spectral decomposition of this matrix yields the directions of the PC factors (the eigenvectors) and the variances in those directions (the eigenvalues):

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathbf{T}}$$

Where $\mathbf{V}$ is the matrix of eigenvectors and $\mathbf{D}$ is the diagnonal eigenvalue ma-

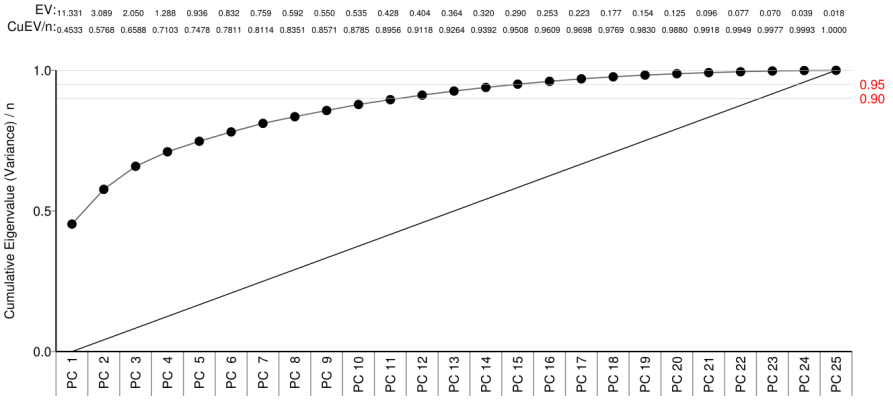


Figure 1.22: Example plot of the cumulative variance explained versus the PC number for 25 variables.

trix. The PC factors are computed by multiplying the data values (locations) by the **V** rotation matrix.

The sum of the variances of the PC factors (the eigenvalues) is equal to the sum of the variances of the original variables, that is, K. A plot of the cumulative variance explained versus the principal component number informs on whether or not dimension reduction is viable, see Figure 1.22 for an example. Note that 11 PCs would explain 90% of the variance and 15 PCs would explain 95% of the variance. Also note in this example that if 15 PCs were chosen for geostatistical modeling, the 15 are combinations of all variables (they are 15 vectors in a 25 dimensional space) and the back transformation of the 15 would lead to 25 predicted variables.

The loading plot is a way to understand the directions of each principal component in a high dimensional space. Simply looking at the components of the Eigenvectors and Eigenvalues may be enough for some, but most of us need a visual display. PCA leads to unique results, but a vector expressed in a forward direction is mathematically (nearly) equivalent to a vector in the reverse direction. The loading plot shows the importance of each principal component and the contribution of each variable to each principal component.

Regarding the exercise, a normal score transform of the principal components may impart some residual non-zero correlation between the transformed variables. This would not happen with multivariate Gaussian data, but real data are rarely truly multivariate Gaussian.

1.4.4 Exercise W1-4

The objectives of this exercise are to learn how to normal score transform data, calculate a correlation matrix and apply principal component analysis to determine orthogonal combinations of variables which account for a large amount of variation. Consider the Ni, Fe, SiO2 and MgO variables in nilat.dat.

1. Plot cross plots between all variables of interest.
2. Normal score transform the data. Note that you do not need to use declustering weights for this exercise, but you may optionally also complete this exercise with declustering weights to see if there are any differences in the correlations.
3. Calculate a correlation matrix for the normal score transformed data. Also plot the normal score bivariate distributions and comment on any changes to the bivariate from Question 1.
4. Run principal component analysis on the normal score transformed variables to construct orthogonal linear combinations of the normal score data.
5. See how the linear combinations are constructed and visualize the loading of each variable on to the principal components to see which variables explain the greatest amount of variance in the data set.
6. Normal score transform the principal components and repeat Question 3. Compare the results and comment on any differences.

Chapter 2

Variograms and Kriging

This week presents much of classic geostatistics. Variograms, volume variance relations (change of support) and kriging are fundamental to spatial modeling in presence of sparse data. A lecture plan for the second week:

Four Days	**Five Days**
• The variogram • Calculation • Variogram volumes	• The variogram • Calculation • Variogram volumes
• Robust variogram estimators • Variogram interpretation • Variogram modeling	• Robust variogram estimators • Variogram interpretation • Variogram modeling
• Volume variance relations • Additivity of variance • Change of shape	• Volume variance relations • Additivity of variance • Change of shape
• Linear estimation • Simple kriging • Properties of kriging	• Review and Demonstration • Linear Estimation • Simple kriging
	• Ordinary kriging • Properties of kriging • Review and Demonstration

2.1 Variograms I

2.1.1 The Variogram

The variogram is the essential tool in geostatistics that quantifies the spatial arrangement of a regionalized variable. Figure 2.1 illustrates the concept with three images of 2500 locations where 140 locations have the same measured values, yet the spatial continuity are quite different. The upper image is quite random and the lower image is the quite continuous. The line charts to the left illustrate the concept of variograms: the horizontal axes are distance and the vertical axes are variability. The variability increases very quickly on the upper image. Note that the distance where the variability reaches a maximum (the range) is the same on all three images.

Two point statistics are used throughout geostatistics. The covariance and correlation coefficient were presented in the previous chapter.

$$\begin{aligned} Cov\{X,Y\} = C_{XY} &= E\{[X - m_X][Y - m_Y]\} \\ &= E\{XY\} - m_X m_Y \end{aligned}$$

$$\text{correlation } \rho = C_{XY}/(\sigma_X \cdot \sigma_Y)$$

The variogram is closely related to these measures. The variogram gained popularity in early geostatistics because it was considered a more robust alternative in presence of non-stationarity.

Central to the variogram is the concept of a vector lag distance $\mathbf{h}$ that is a distance in a particular direction. Experimental pair separated approximately the vector distance are assembled (see Figure 2.2 for a sketch that shows pairs following a specified lag and tolerance - note that some data are used more than once and some data are never used) and the variogram is calculated as the expected value of the squared difference in the values separated by the lag:

$$\text{variogram } 2\gamma(\mathbf{h}) = E\{[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})]^2\}$$

The variogram is calculated for a set of lag distances to obtain a continuous function. Figure 2.3 shows a geometric interpretation of the variogram - the expected squared distance from the 1:1 line.

Under stationarity the variogram, covariance and correlation coefficient are

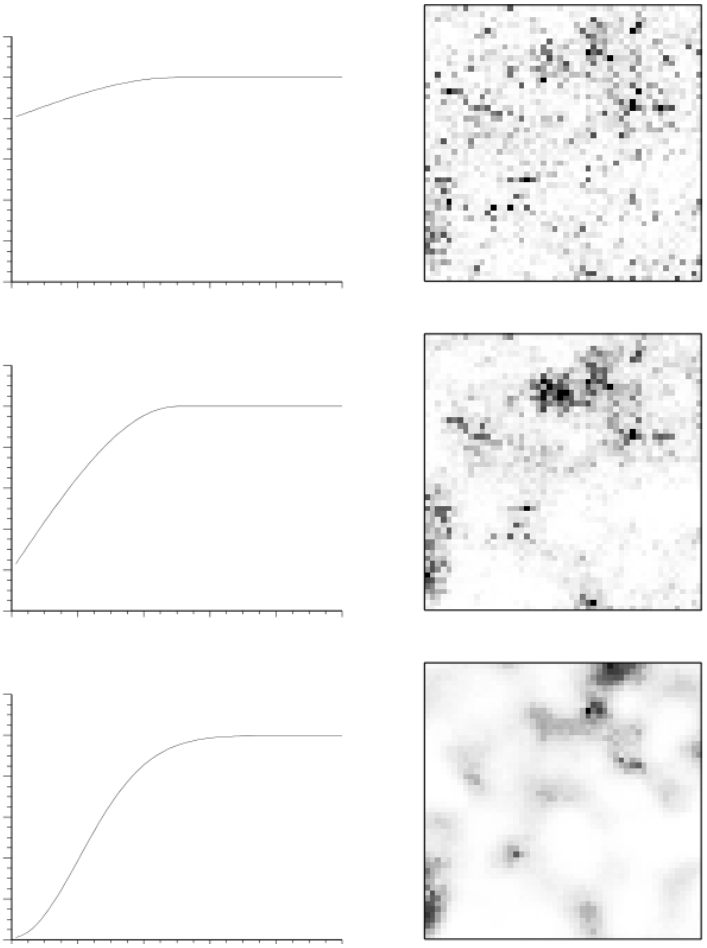


Figure 2.1: Three images of 2500 locations where 140 locations have the same measured values, yet the spatial continuity (sketches to the left) are quite different.

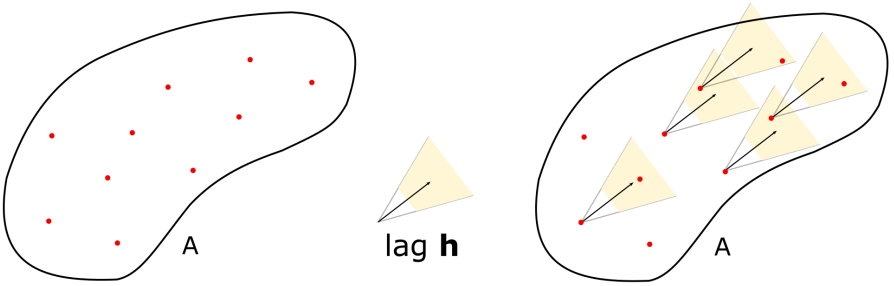


Figure 2.2: Illustration of scanning a lag vector (and tolerance) over a set of data to find pairs that satisfy the lag constraints.

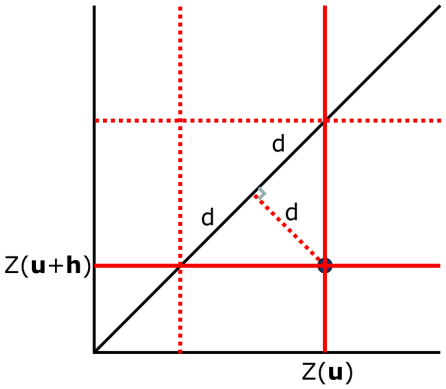


Figure 2.3: Illustration that the variogram is the squared distance from the 1:1 line.

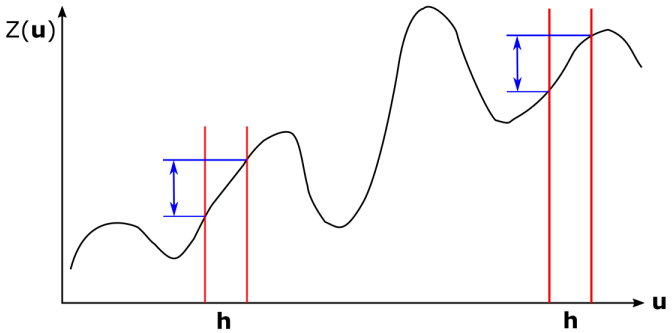


Figure 2.4: Concept of why variogram persisted instead of covariance.

equivalent tools for characterized two-point correlation:

$$\begin{aligned}
2\gamma(\mathbf{h}) &= E\{Z(\mathbf{u})^2 - 2Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h}) + Z(\mathbf{u}+\mathbf{h})^2\} \\
&= E\{Z(\mathbf{u})^2\} - m^2 - 2\left(E\{Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})\} - m^2\right) \\
&+ E\{Z(\mathbf{u}+\mathbf{h})^2\} - m^2 \\
&= \sigma^2 - 2C(\mathbf{h}) + \sigma^2 \\
&= 2\sigma^2 - 2C(\mathbf{h})
\end{aligned}$$

The stationarity of the variogram lag over the domain allows experimental pairs of data to be pooled to calculate experimental variogram values. This assumption of stationarity is another form of second order stationarity that includes directions and magnitudes of continuity.

Each h-scatterplot is summarized by an average squared difference and the values are plotted together. Three h-scatterplots at different distances are plotted to the right. The h-scatterplots may be useful to view in presence of very few data; normally there are many pairs and there is little that can be detected

2.1.2 Variogram Calculation

Variography involves (1) calculation, (2) interpretation, and (3) modeling. The three overlap, but it is reasonable to learn about the steps in this order. An important assumption of stationarity is embedded in the concept of the variogram, that is, the spatial statistics depend on the lag separation vector, but are independent of location. The expected variability between two locations separated by 120m in the North-South direction is the same for all positions within the domain. This assumptions makes it possible to calculation variograms from the available data. The lag vector with some tolerance

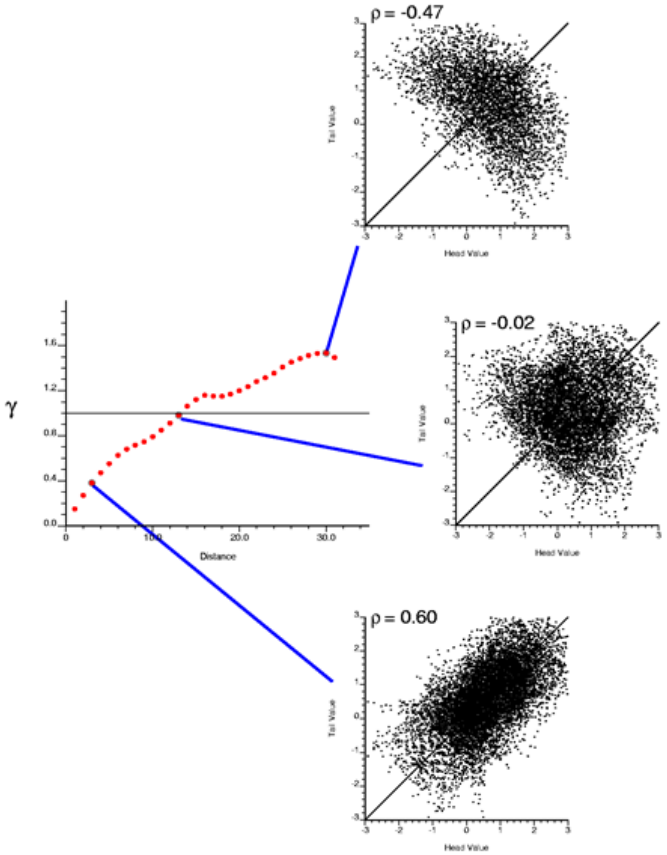


Figure 2.5: Illustration of three different lag scatterplots - from the bottom (1) a value below the sill - positively correlated, (2) a value at the sill - uncorrelated, and (3) a value above the sill - negatively correlated.

is scanned over all pairs of data and the expected variability is calculated directly. The concept that the lag vectors *float* and are *anchored* is a key concept in variography.

Another key concept that arises from the previous lecture is that of the sill. The scalar variance of the data entering variogram calculation is important. When the variogram is below that value, the pairs of values are positively correlated. When the variogram is above that value, the pairs are negatively correlated. Although variogram interpretation comes later, there is inevitable overlap in the steps and this is fundamental for interpreting the variogram.

Calculating variograms for regularly spaced data is straightforward. Irregularly spaced data is more challenging because of the tolerance parameters required. The basic parameters are covered in a Lesson [24]. Another Lesson covers details for tabular deposits [25].

Step One: Choosing Directions. Geological variables are anisotropic. An initial challenge in variogram calculation is to determine the three principal directions of continuity. Section 1.3.2 explained the concept of anisotropy, but our understanding of anisotropy often comes from variogram calculation. Considerations in choosing the principal directions for variogram calculation include:

1. Start with an omnidirectional variogram, that is, for all directions combined together considering the tabular nature of the deposit - no point combining the vertical with the areal in a stratigraphic deposit of 100:1 horizontal to vertical anisotropy. A well crafted omnidirectional variogram (perhaps in the plane of greatest continuity) often yields the most well-behaved variogram.
2. Visualizing the data within the context of solid conceptual geological model may provide the most important information. The variograms may not be the best, but having variograms highly consistent with the geological conceptual model is essential. Reviewing the data in 3-D and in 2-D maps and sections is very useful. Visualizing data at scattered locations in 3-D is difficult. Often, we would build a *neutral* model by kriging or with inverse distance and visualize that model - thresholding high and low values to highlight important structures.
3. Choose directions based on orientation of geological unit. There is no guarantee that the short scale variogram structure would follow the larger scale orientation and size of the geological domain, but that is reasonable. This is often self-fulfilling since the variogram range would always look longer in directions where the domain is larger and there is more possibility of finding pairs.
4. Consider multiple directions before choosing a set of 3 perpendicular directions (major horizontal direction and two perpendicular to major direction). A trial-and-error approach is susceptible to the data config-

uration and random chance.

The creation of a neutral model as mentioned in step 2 can be important. The variogram or anisotropic parameters should not have an inordinate influence on the final model. In presence of a tabular deposit, then the interpolator should be isotropic in the plane of the structure and with a reasonable anisotropy relative to the direction of least continuity., Ordinary kriging with a variogram showing 30% nugget (to filter high frequency variations) and a range one half of the domain size (to show long range structure) works well. Global inverse distance with appropriate anisotropy parameters works well.

Step Two: Choosing Lag Paramters. The unit lag separation distance should coincide with data spacing (the close data spacing if there is any flexibility). This would be chosen differently in each principal direction determined in the previous step. The lag tolerance is typically chosen to be one half of the unit lag separation distance except when (1) the data are very regularly spaced and a smaller tolerance can be considered, or (2) when there are few data and erratic variograms and a tolerance more than one half of the unit lag may be chosen to provide more stable variograms.

Multiple lag distances and lag tolerance parameters may be required when there are multiple nested grids of data spacing. For example, there may be a larger data spacing over the entire domain and a smaller data spacing in a smaller area of greater interest. Then, some lags at the small data spacing should be considered and some lags at the larger data spacing should be considered. A different lag tolerance would be chosen for the small and large lag spacing. The experimental variogram values from both spacings would be combined in variogram interpretation and modeling.

The number of *multiples* of the unit lag must also be chosen. This depends on the direction. In general, less than 10 lags are always relevant - larger distances have little input to local prediction (although highly irregular data spacing may override this). In addition to this rule of 10, the variogram is only valid for a distance one half of the field size. So, we choose the number of lags accordingly.

Another practical detail in variogram calculation is that we would start with the most informed direction - often the down-hole or vertical variogram, then move to directions that are less well informed.

Step Three: Choosing Direction Tolerance. The angle tolerance for directions is required - if the plane of continuity being considered is reasonably isotropic, then this could be as large as 25 degrees. In the case of tabular deposits, the bandwidth parameters perpendicular to the plane of greatest continuity is important. Flattening the vertical coordinate is essential before variogram calculation, then the bandwidth perpendicular to the plane of

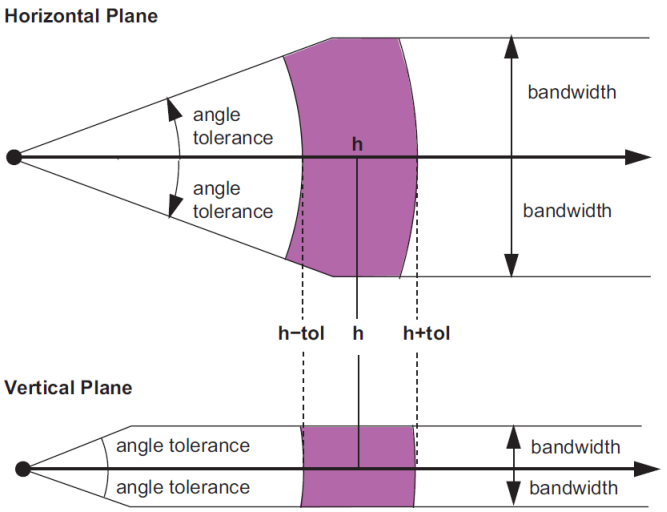


Figure 2.6: Illustration of the parameters used for variogram tolerance specification. The plane of greatest continuity may not be horizontal, but it is labelled as such in the figure.

continuity could consider the data spacing (a few multiples) or the vertical variogram range (keep with 1/3 of the vertical variogram range). In general, we would restrict as much as possible.

The tolerance parameters in variogram calculation must be chosen carefully. If the tolerance is too large, then the variogram may be stable, but it is not precise and not representative of the underlying spatial distribution. If the tolerance is too small, then the variogram may be precise, but it perhaps too noisy to infer the true underlying spatial structure. All calculation parameters should be considered carefully.

2.1.3 Variogram Calculation Challenges

Tabular deposits are extensive in two dimensions and limited in a third dimension. Some considerations include (1) unfolding or stratigraphic coordinates are important to allow the variogram to follow undulations in the tabular structure, (2) careful setting of the variogram tolerance for variograms within the plane of the structure is important - using a limited bandwidth perpendicular to the plane of continuity.

Variogram volumes can be useful. The concept is to consider polar coordinates - a lag of zero is at the center of the plot. Then different lags in diferent

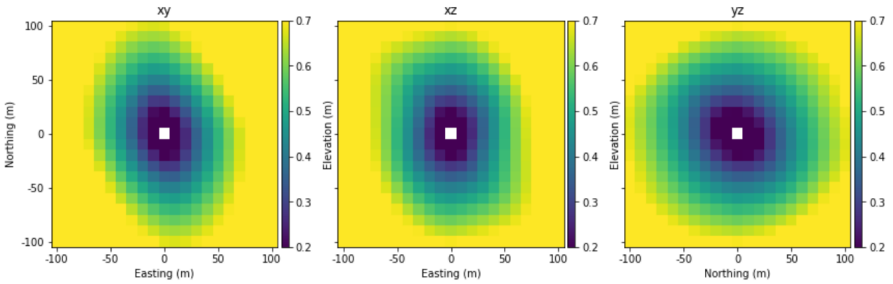


Figure 2.7: Example of three sections through a variogram volume. The three maps are XY, XZ and YZ slices (variogram maps) slices through the origin.

directions can be considered in Cartesian or radial coordinates. Contours of the variogram values in different planes may be informative on the directions and magnitude of anisotropy. Only sections through the origin are meaningful - any other is anchored to a lag not showing on the plot. Figure 2.7 shows an example. This variogram volume was constructed from a *neutral* model kriged with an isotropic variogram. Any anisotropy shown is due to the data. Directional variograms would be calculated in the principal directions identified from the variogram volume.

Standardization of the variogram should be discussed with some preliminary remarks made about the sill. Standardizing the variogram should almost always be performed - making the expected sill equal to one makes interpretaion and other analysys straightforward. Two situations where we would *not* standardize the sill are (1) when checking variogram reproduction in simulation - fluctuations in the variance are expected and part of variogram reproduction, and (2) when performing exploratory data analysis with different data types of the same variable in the same domain. Understanding the difference in variance is important.

This would also be the setup and explanation of the data considered for the exercise. Perhaps a demo of variogram calculation.

2.1.4 Exercise W2-1

Variograms are very important in a geostatistical study. They specify the spatial correlation that the grades are expected to have. Directional variograms must be calculated with parameters that lead to stable and reliable experimental variograms. The objective of this exercise is to calculate and interpret variograms. Any software can be used. Consider the bitumen grade in the oilsands.dat file. Please consider the bitumen grades directly rather

than the normal scores transform.

1. Review the spatial arrangement of data with relevant plots (such as a plan view and cross sections). Comment on the anticipated directions for variogram calculation. Discuss the selection of variogram parameters such as the angle tolerances, lag spacing, and lag tolerance. Pay particular attention to the parameters that will be required for horizontal variogram calculation.
2. Calculate and interpret the vertical variogram. What is the vertical range? What vertical bandwidth do you consider relevant for horizontal variogram calculation?
3. Calculate and interpret an omnidirectional horizontal variogram. Perform a sensitivity to the vertical angle and bandwidth parameters, which are very important in such relatively flat stratigraphic formations.
4. Discuss the selection of directions for horizontal directional variograms. Calculate and plot a horizontal variogram map. Choose the bin size reasonably big and choose relatively few lags: many variogram maps are too noisy to be useful. Take care and try to get a stable variogram.
5. Calculate and interpret horizontal directional variograms. Choose major and minor orthogonal horizontal directions for the variograms.

2.2 Variograms II

2.2.1 Robust Variogram Estimators

There are two primary reasons why the *variogram is not a good estimator of the variogram*:

1. Clustered data and the proportional effect: high valued areas may be drilled preferentially so there are more short scale pairs in those areas. The proportional effect is when the local variance is dependent on the local mean. High values areas are more variable when the distribution of grades is positively skewed. Thus, the early lags in the variogram calculation are higher than expected making the variogram appear to show less structure. Figure 2.8 illustrates this.
2. The unequal use of extreme values in different lags: when an extreme value enters a variogram lag, the squared difference increases significantly. This is not wrong - in fact, fairly representing extreme values is essential in modern geostatistics. The challenge occurs when the extreme values enter different lags differently; the variogram becomes noisy. This happens because of the specific data configuration and calculation parameters.

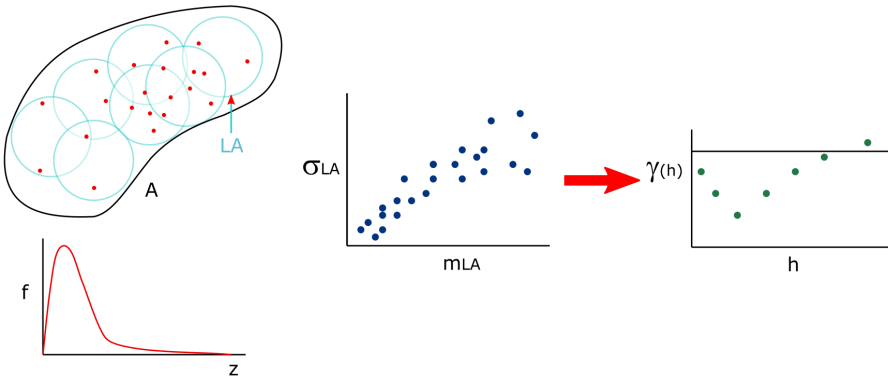


Figure 2.8: Illustration of how the proportional effect and clustered data impacts the variogram. The upper left sketch shows a stationary area A with data denoted by red dots and local areas denoted by blue circles. The distribution on the lower left shows a positively skewed distribution. The central scatterplot shows one example of the proportional effect, that is, the dependence of the standard deviation on the mean. To the right is a variogram showing increased variability for the closely spaced lag due to clustered data in high valued areas and the proportional effect.

A robust variogram estimator with the longest history is the pairwise relative variogram [11, 3]:

$$2\gamma_{PR}(\mathbf{h}) = E\left\{\frac{\left(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right)^2}{\left(\left(Z(\mathbf{u}) + Z(\mathbf{u}+\mathbf{h})\right)/2\right)^2}\right\}$$

this has proven effective in many geological environments with highly skewed distributions and clustered data. A historical challenge was that the sill is unknown; however, we can easily determine the sill by sampling a large number of random paris and calculating the pairwise relative variogram. This is done automatically in most modern software. A criticism of the pairwise relative variogram is that it is not theoretically justified. There is some validity to this criticism.

A more defendable alterternative is to consider the correlogram [34, 56]:

$$2\gamma_{CO}(\mathbf{h}) = 1 - \frac{E\left\{[Z(\mathbf{u}) - m(\mathbf{u})][Z(\mathbf{u}+\mathbf{h}) - m(\mathbf{u}+\mathbf{h})]\right\}}{\sqrt{E\left\{[Z(\mathbf{u}) - m(\mathbf{u})]^2\right\} E\left\{[Z(\mathbf{u}+\mathbf{h}) - m(\mathbf{u}+\mathbf{h})]^2\right\}}}$$

the use of lag dependent means and variances can make the correlogram more stable than the variogram. The correlogram does not always work and tends to unreasonably dampen the influence of trends and zonal anisotropy. The correlogram appears more theoretically defendable , but future volume

variance relations and kriging do not consider lag dependent parameters. Also, the correlogram does not capture zonal anisotropy very well.

A third alternative is to consider the normal score variogram (then back transform with MCS for each lag assuming equal representivity). The back transform of the normal scores variogram will be explained. After normal score transform there is no proportional effect and no outliers. The experimental normal score variogram may be very stable. If we were to back transform the exact pairs going into each lag of the normal score variogram, then the result would not change, but we assume that the correlation is representative and back transform a series of bivariate distributions with the same marginal distribution. This leads to a stable variogram back in original units. This is theoretically correct and fairly common at the present moment [22].

In practice, the pairwise relative variogram is hard to beat. The back transformed normal score variogram is a theoretically defendable alternative, but in presence of challenging data the pairwise relative variogram may provide the best input to kriging.

2.2.2 Variogram Interpretation

Variogram interpretation consists of explaining the variability over different distance scales. The variogram is a chart of variance versus distance or geological variability versus direction and Euclidean distance. Based on seeing other variograms understanding site-specific considerations, we must establish the reasonableness of the variogram

Some variogram terminology is important. The sill is the scalar variance of the data used in variogram calculation, see [29]. The sill is one if the variogram has been standardized or normal score data are used. In presence of stationary data, the variogram would flatten off at the variance. It is not practical to think of the sill as where the variogram flattens off because the experimental variograms may go above the sill variance or stay below the sill variance in different directions. The precise choice of the sill does not influence ordinary kriging, but has a significant influence on other types of kriging and on simulation. Consider the variance of the data or one if the variogram has been standardized.

The range is the distance at which the variogram reaches the sill (the first time) *if* it reaches the sill. The range is a very blunt instrument since the variogram can approach the sill at different slopes. Figure 2.9 illustrates a case where a variogram with a long range represents a variable with less continuity. The range is a well known parameter, but must be considered with the entire variogram shape.

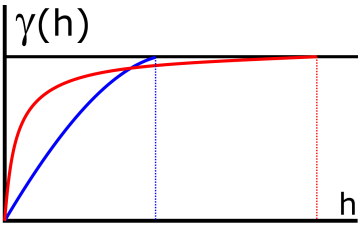


Figure 2.9: Illustration of two variograms with different ranges. The red variogram has a twice the range, but reflects a variable with less continuity because of its short scale behavior.

The nugget effect is a variogram fitting parameter that is a discontinuity at the origin of the variogram. See the Lesson [7]. Some data have nuggety characteristics where some samples have a nugget of mineralization and adjacent samples do not; then, the variogram will show an abrupt jump at short distance. The nugget effect is not the variability wihin a sample; it is the variability between adjacent samples that are non-overlapping. Compositing the data to larger scale leads the nugget effect to be smaller since there is more averaging within the sample lengths. The nugget effect is a fitting parameter of variogram model when the variogram is extrapolated back to a lag of zero. It is a scalar parameter, that is, independent of direction.

The four elementary signatures of variogram interpretation are shown on Figure 2.11:

1. Trends. When the variogram goes about the expected sill by a significant amount (20%), then the variogram is indicating a negative correlation at large distance which is called a trend.
2. Cyclicity. When the variogram oscillates above the sill and below the sill at a fixed length scale of periodicty, the we call this cyclicity (or periodicity or the hole effect). It would be observed drilling through a stacked succession of different units that have alternating high and low values spaced about the same amount. The distance where we see the first trough should be at twice the first peak and the second peak should be at three times the the distance of the first peak. Figure 2.10 shows two examples of cyclical variograms.
3. Geometric anisotropy. When the shape of the variogram is the same in different directions, but the length scale is different. An affine transformation of coordinates - stretching and squeezing after rotation - would make the directional variograms look approximately the same. Geometric anisotropy is a natural characteristic of many phenomenon. For example, in a stratigraphic environment, we would naturally expect a horizontal to vertical anisotropy of 100:1 (sometimes less and sometimes more, but that is a good rough estimate). Anisotropy is common in vir-

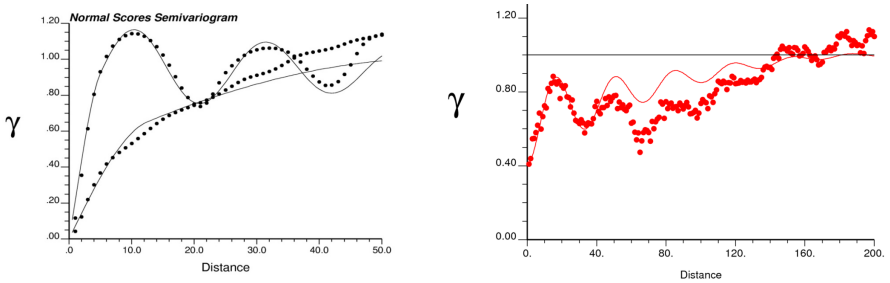


Figure 2.10: Two examples of cyclical variograms from a scanned rock (left) and outcrop measurements (right).

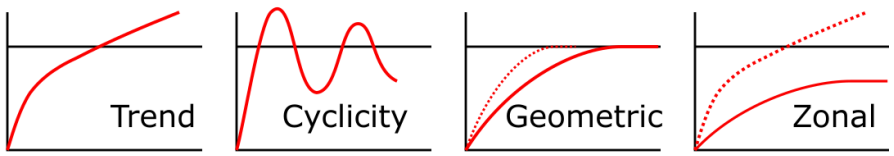


Figure 2.11: The four elemetary signatures of variogram interpretation.

tually all depositional and diagenetic influenced spatial distributions.

4. Zonal anisotropy. The apparent sill in one or more directions falls significantly below the expected sill. The classic example is a tabular deposit where there is some internal "zonation" within the formation, that is, higher values at the top (or bottom or middle) that persist across the entire study area. Then, the horizontal variogram will see persistent positive correlation over large distances - the horizontal sill will not reach the expected sill. The vertical variogram will show something else - likely a trend or large scale cyclicity.
   The zonal anisotropy can also manifest as the vertical variogram not reaching the expected sill. This would occur in presence of horizontal trends.

A geometric anisotropy at one scale may appear as a zonal anisotropy at another scale. The variogram and its interpretation depends strongly on the size of the domain.

A normal variogram that leaves a low and normal nugget effect, rises toward the sill, then plateau's at the sill is very rare. Real variograms show combinations and superpositions of all of the variogram signatures mentioned.

A strong trend or zonal anisotropy may indicate that trend modeling, removal of the trend, then modeling the detrended variable may be in order. Increasingly, with the improved trend modeling and modeling with a trend

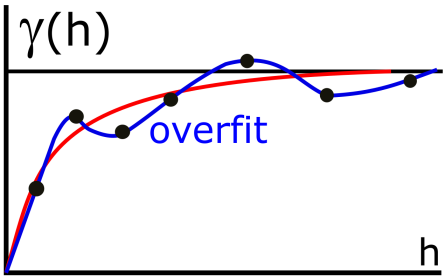


Figure 2.12: Sketch of experimental variogram points (black dots) that are overfit (blue line) and fit reasonably (red line).

algorithms available, this is a recommended approach.

2.2.3 Variogram Modeling

Variogram modeling amounts to fit a smooth function to the variogram. The variogram model is closer to the true variogram than the points themselves! Variogram modeling is not like curve fitting. The true variogram (if it were available) would be a smooth function with distance. The variogram is an expected behaviour, but the variability that would be seen in the underlying random function. Another important aspect of variogram modeling is shown on Figure 2.12, that is, the goal is not a strictly close fit to the data - some smoothing through variations in the experimental variogram is desired. The true variogram is going to be smooth and the variogram model is aimed at the true underlying variogram and not over-fitting the experimental points.

There are three main reasons to model a variogram:

1. Need to know the variogram for all distances and directions ($\gamma(\mathbf{h})\forall\mathbf{h}$) - not just the ones calculated. The variogram function must be extrapolated in distance and direction. We focus on the principal directions and the lags we can calculate, but we need the variogram for all possible lags in kriging and simulation.
2. Incorporate additional geological knowledge (analogue information or information on directions of continuity and so on).
3. The variogram model must be positive definite (a legitimate measure of distance) for subsequent estimation and simulation. This is easy to understand mathematically, but for practical purposes it means that the variogram as a distance measure is a valid one. Example of non-physical distances will be presented.

The need for a positive definite function may be the least of the three reasons

presented above, but it drives the choice of a mathematical model. There are an infinite number of positive definite variogram models, but we choose to use a few functions that are commonly encountered and mathematically simple, see Figure 2.13. We can combine them in multiple structures to give us all the flexibility we could ever imagine. The four most common variogram shapes:

1. Nugget effect: a value of zero at $\mathbf{h} = 0$, then a constant. This is for that aspect of the regionalized variable that is completely random.
2. Spherical variogram: this is the most common of the family of geometric variogram models, that is, one minus the volume of intersection of two geometric objects:

$$Sph(h) = 3/2h - 1/2h^3 \text{ for } h \leq 1; \quad 1, \text{ otherwise}$$

   any geometric shape could be used, but the resultant variogram shape from the spherical variogram is sufficient.
3. Exponential variogram: one minus a standardized exponential decay function is also commonly seen for quite eratic variables such as permeability or hydraulic conductivity.

$$Exp(h) = 1 - e^{(-3h)}$$

4. Gaussian variogram: the exponential function could be raised to any power between 0 and 2. The limit variogram function at a power of 2 would lead to the Gaussian variogram (named because of the mathematical form and not because of any intrinsic connection to the Gaussian distribution):

$$Gauss(h) = 1 - e^{(-3h^2)}$$

   This variogram is parabolic near the origin and suitable for continuous variables such as thickness and surface elevations.

Modeling with a linear model of regionalization(LMR) for different structures and positive definiteness:

$$Z(\mathbf{u}) = \sum_{i=0}^{nst} b_i Y_i(\mathbf{u}) \qquad \gamma(\mathbf{h}) = \sum_{i=0}^{nst} c_i \gamma_i(\mathbf{h})$$

where $i = 0$ is nugget by convention. Each structure is defined by a shape and six anisotropy parameters (three angles and three ranges - see Section 1.1). The common shapes are described above. Considering that anisotropy is outside of variogram modeling, that is, it is specified per structure and it modifies how the "standardized isotropic distance" is calculated is relevant.

Variogram modeling amounts to picking a single (lowest) isotropic nugget effect, choosing the same number of variogram structures for all directions

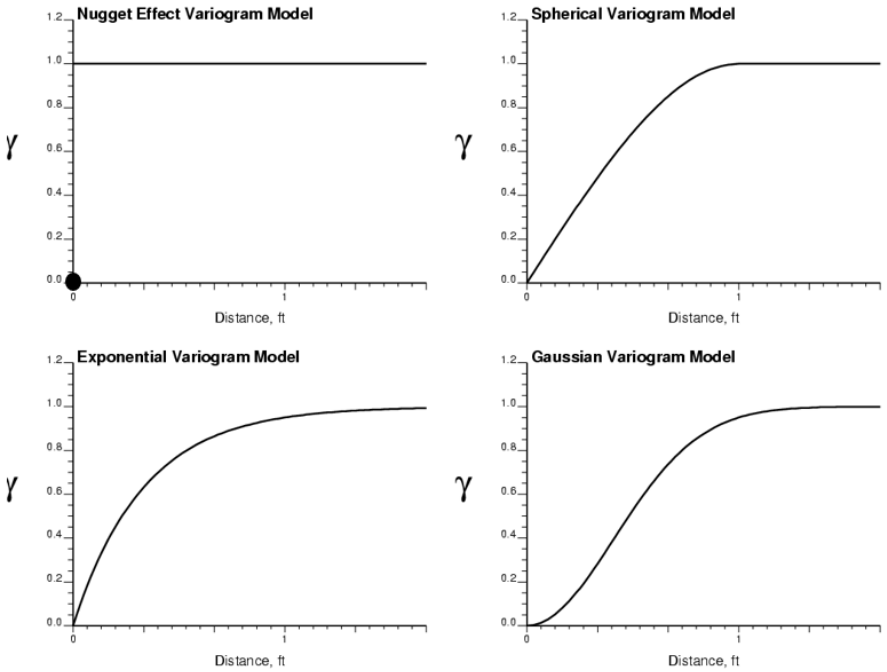


Figure 2.13: The four commonly used variogram model shapes.

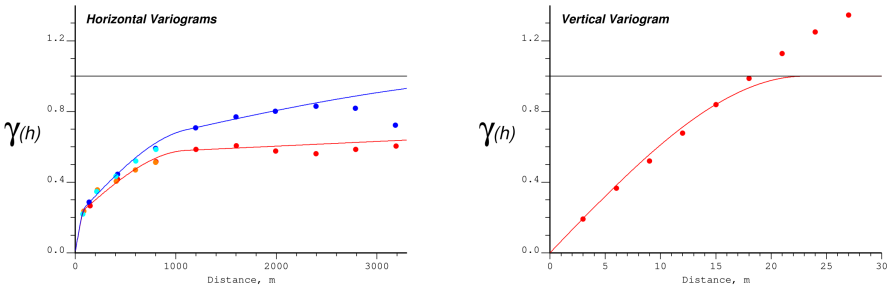


Figure 2.14: Standardized horizontal and vertical variograms for the bitumen grade in an oilsands deposit.

based on most complex directions, ensuring that the same sill parameter is used for all variogram structures in all directions, and allowing a different range parameter in each direction. A zonal anisotropy by setting a very large range parameter in one or more of the principal directions.

Figure 2.14 shows experimental and model variograms for bitumen grade in an oilsands deposit. This fit was chosen manually with some iteration. The parameters of the fit are specified in the table below. Note that there are three spherical structures (type 1). The total number of parameters for a variogram model includes the number of structures, the nugget effect, the sill and eight parameters per structure (three angles, three ranges, the type/shape and the contribution). The sill is not always specified independently although that would be considered best practice.

```
3    0.0   1.0                  -number of structures, nugget, sill
1    0.2  45.0    0.0    0.0    -type, contribution, ang1, ang2, ang3
        100.0  100.0   23.0    -  major, minor, tertiary ranges
1    0.35 45.0    0.0    0.0    -type, contribution, ang1, ang2, ang3
       1100.0 1100.0   23.0    -  major, minor, tertiary ranges
1    0.45 45.0    0.0    0.0    -type, contribution, ang1, ang2, ang3
      25000.0 5000.0   23.0    -  major, minor, tertiary ranges
```

An important topic in data poor areas of geostatistics is how to infer a variogram in presence of sparese data. This is particularly relevant for variogram models in stratigraphic settings (petroleum reservoirs) with few wells. An estiamte of the zonal anisotropy and the geometric anisotropy of each structure is required.

There are software that mimics the iterative procedure performed by the varigoram modeler, varying the different parameters that define a variogram model. The nugget effect, sill contributions, structure types and ranges, are fit to experimental variogram points in up to three directions simultaneously.

The user can fix any subset of parameters. This is useful when the experimental variograms are well defined, but expert judgement is always to be considered.

A variogram model is specified by the number of structures, the nugget effect and (for each structure, the contribution, the shape and six anisotropy parameters (three angles and three ranges). This is a very efficient compression of the variogram information.

When there is doubt as to the most correct variogram model, then we consider (1) fidelity to the geology and the regionalized variable, (2) simplicity - simpler is better,

Variogram modeling is one of the most important steps in a geostatistical study - all spatial inference depends on model of spatial variability / continuity.

2.2.4 Exercise W2-2

The objective of this exercise is to gain experience and confidence in the experimental calculation, interpretation and modeling of variograms for a number of data sets.

1. Consider the directional horizontal and vertical variograms experimentally calculated in exercise W2-1 for bitumen with the oilsands.dat data set. Fit the variograms. Tabulate your variogram model parameters and comment on the uncertainty in the variogram model parameters.
2. Calculate, interpret and model the variogram of Au in skarn2d.dat (2-D).
3. Calculate, interpret and model the variogram of Au in Misima.dat (3-D).

2.3 Change of Support

2.3.1 Scales of Relevance

Change of support or volume variance relations relates to understanding what happens at a larger scale given data and information at a smaller scale. Theoretically we may think of going down in scale, but the reality is that we drill and observe variability at a small scale and we wish to predict what is happening at a large scale of relevance. The data within reasonably station-

ary domains provides a representative distribution through declustering and a representative variogram model. These statistical parameters relate to the scale of the data. Engineering considerations lead us to consider a selective mining unit (SMU) that effectively represents a scale relevant for resource and reserve assessment. The SMU scale (often denoted v) may be 4 to 9 orders of magnitude larger than the scale of the data. We aim to predict what would be the representative distribution at that larger scale. This has significant consequences on resource and reserve estimation.

This is particularly important when there are no block data [46]. A theoretical prediction of the change of support may all we have. The main reasons why we consider this topic include (1) the resources and reserves we compute and report are to reflect future mining and production, (2) there are no block data available early in the pre- and feasibility studies of a deposit, yet we are to represent that scale, (3) resources depend very strongly on the production volume and the inevitable smoothing that occurs as we consider statistics representing larger volumes, (4) we would like to know how much smoothing is acceptable in the application of conventional geostatistical (or any) estimation algorithm, that is, we would often want to calibrate our resource estimation to anticipate future selectivity and provide unbiased block estimates, and (5) we would like a check on advanced simulation, UC, MIK and other prediction algorithms.

The choice of an SMU (v) size is most relevant to open pit mining. It could be used for underground, but the less constrained selection of open pit is more amenable to the underlying assumptions. The considerations include (1) the mining method - more flexible means smaller SMU size, (2) the mining equipment - smaller and more flexible means smaller SMU size, (3) blasting practice - more agressive blasting with larger movement means larger SMU size, (4) production sampling - more and better production sampling would imply a smaller SMU size, (5) visual control on the ore and the availability to support mine operations in the visual control - better control would imply a smaller SMU size, and (6) various other mining operations including the attention of management, mining at night near contacts, the bonus and incentivization schemes - as expected. The Lesson on change of support is a useful reference [32].

2.3.2 Volume-Variance Relations

The first key principle is additivity of variance expressed with dispersion variances:

$$D^2(v, A) = D^2(v, V) + D^2(V, A) \text{ where } |v| < |V| < |A|$$

A demonstration will be given that proves the additivity of variance. The key principle is that the variance of composites within the stationary domain is equal to the variance of composites within the SMU volume and the variance of SMU volumes within the domain. The variance of SMU volumes within the domain is a goal of volume variance relations.

The variance is often written as $D^2(\cdot, A) = \sigma^2$, that is, the conventional variance of a variable is interpreted as the dispersion variance of point samples within the domain.

Second, the volume averaged variogram is defined for two different volumes, perhaps at different locations:

$$\overline{\gamma}(V, v) = \frac{1}{|V|\,|v|} \int_V \int_v \gamma(x - y) dx dy$$

This could be used for general calculations including estimation variance and for populating block kriging equations with data of different support. For change of support calculations, the volume average variogram of a volume with itself, that is, of v if $v = V$ is particularly important. It is the expected variance of points within the SMU volume:

$$\overline{\gamma}(v, v) = D^2(\cdot, v)$$

Attempts have been made to solve the integrals requried for the volume averaged variogram, but the modern approach is to discretize the volume and solve for the integral numerically. A discretization of 5x5x5 is adequate (Lesson on discretization could be referred to [57]). If the number of discretization points is made too big, then numerical precision errors can become significant. The discretization in the direction of drilling should consider the composite length relative to the SMU size. For example, if the composites are 5m long and the SMU is 15m large, then the discretization should be set to three - an even multiple (or close to one) of the number of composites for the volume.

The reduction in variance could be summarized as a ratio:

$$f = \frac{D^2(v, A)}{D^2(\cdot, A)} = \frac{D^2(\cdot, A) - \overline{\gamma}(v, v)}{D^2(\cdot, A)} = \frac{1 - \overline{\gamma}(v, v)}{\sigma^2}$$

So, for change of support the mean stays the same, the variance reduces by a predictable amount and the shape of the distribution becomes slightly more Gaussian.

Before reviewing change of shape options, the fact that compositing dramatically reduces the nugget effect is practically important. The variance of the average of n random numbers is σ^2/n. This classic result will be demonstrated. FIn the context of compositing, the nugget effect will go down by a multiple of the length being composited, for example, if 1m samples are

composited to 4m, then the nugget effect of the 4m samples is 1/4 that of the 1m samples. The structured part of the variance remains mostly unchanged because of the small scale of averaging. The nugget effect could be estimated by compositing to different lengths and observing the change in variance. The nugget effect for composites at a certain length would be $C_{0,l} = 2(\sigma_l^2 - \sigma_{2l}^2)$ considering the variance at the length and at doube the length.

2.3.3 Change of Distribution Shape

The classical techniques should not be used. Assuming the shape of the distribution does not change (the affine correction) is entirely unrealistic. Assuming the values are lognormal and there is some kind of preservation of lognormality (the indirect lognormal correction - ILC) is also entirely unrealistic. The discrete Gaussian model (DGM) is more complicated, but provides very stable estimates of distribution shapes for larger volumes.

As a recall. The affine correction $z_v = \sqrt{f}(z - m) + m$ assumes no shape change and is very unrealistic due to an artificial minimum and no convergence to a Gaussian distribution as the scale goes up. The Indirect Lognormal Correction (ILC): $z_v^{(1)} = a\ z^b$

$$b = \sqrt{\frac{ln(f \cdot CV^2 + 1)}{ln(f \cdot CV + 1}} \quad a = \frac{m}{\sqrt{f \cdot CV^2 + 1}} \left[\frac{\sqrt{CV^2 + 1}}{m} \right]^b$$

$$z_v^{(2)} = z_v^{(1)} \cdot \frac{m}{m^1}$$

provides for unrealistic shape change in upscaling, but is surprisingly useful for downscaling. For upscaling, the DGM more general and flexible.

The discrete Gaussian model may not be mathematically simple, but it is powerful and very practical. Discrete Gaussian model (DGM) is based on Hermite polynomials:

$$z_v = \sum_{p=0}^{n_p} \phi_p H_p(y)$$

$$H_p(y) = \frac{1}{\sqrt{p!}g(y)} \frac{d^p g(y)}{dy^p}$$

$$H_0(y) = 1 \;\; H_1(y) = -y \;\; H_2(y) = \frac{1}{\sqrt{2}}(y^2 - 1) \;\; H_3(y) = \frac{-1}{\sqrt{6}}(y^3 - 3y)$$

recursive equation for more:

$$H_{p+1}(y) - \frac{-1}{\sqrt{p+1}} y H_p(y) - \sqrt{\frac{p}{p+1}} H_{p-1}(y)$$

since the Hs are independent and standard the variance :

$$\sigma^2 = \sum_{p=1}^{n_p} \phi_p^2$$

The diffusion and mosaic models for change of support (r values are determined by trial and error):

$$\phi_{v,p} = r_D^p \phi_p \quad \phi_{v,p} = r_M \phi_p$$

The diffusion and mosaic models are characterized by:

$$\frac{\gamma_1(\mathbf{h})}{\sqrt{\gamma(\mathbf{h})}} = \frac{C_1}{\sqrt{C}} \qquad \frac{\gamma_1(\mathbf{h})}{\gamma(\mathbf{h})} = \frac{C_1}{C}$$

the barycentric model is in between the two. A measure of dissemination (1=D, 0=M) could be calculated:

$$\beta(\mathbf{h}) = \left(\frac{\gamma_1(\mathbf{h})}{C_1} - \frac{\gamma(\mathbf{h})}{C}\right) / \left(\sqrt{\frac{\gamma(\mathbf{h})}{C}} - \frac{\gamma(\mathbf{h})}{C}\right)$$

$$\phi_{v,p} = \left(\beta_e r_B^p + (1-\beta_e) r_B\right) \phi_p$$

Most practitioners use the diffusion model and stop there. The use of the mosaic model or a blend between the two (the barycentric model) may be useful in certain cases. In practice, much upscaling is done by simulating at a high resolution, then averaging to larger volumes.

2.3.4 Exercise W2-3

The objective of this exercise is to learn how to scale variances with average variogram values and to review the common change of shape models. Average variogram or *gammabar* values tell us the variance at any scale. The discretization required for stable numerical integration is a consideration. Average variogram values can be calculated between two disjoint volumes V and v; however, classic histogram and variogram scaling requires the average variogram to be calculated for V = v, that is, for the same volume and itself.

1. Write a short review of the key theoretical results needed for variogram scaling: (1) the definition of the average variogram or average covariance, (2) the definition of the dispersion variance and the link to the average variogram, and (3) the additivity of variance.
2. Consider the Au variable from Misima.dat. Consider the variogram model from a previous exercise and a 10m by 10m by 10m block size. Create a plot with the average variogram versus discretization level starting with 1x1x1 and going to 6x6x6 in increments of 1.

3. Calculate the average variogram for block sizes of 5x5x5, 10x10x10, 15x15x15, 25x25x15 and 50x50x15. Plot and tabulate (1) the average variogram versus block size, and (2) the block variance versus block size.
4. Consider a block size of a 10m by 10m by 10m. Calculate the scaled distributions using the (1) affine, (2) indirect lognormal, and (3) discrete Gaussian models. Plot the original Au histogram and all of the scaled histograms. Comment on the results.

2.4 Kriging I

2.4.1 Linear Estimation

The context is to estimate a continuous rock property at an unsampled location using the available data in the domain. The data are denoted as a set:

$$n: \quad z(\mathbf{u}_\alpha) = z_\alpha, \alpha = 1, \ldots, n$$

statistical parameters have been inferred from the data including:

$$F(z), \;\; m, \;\; \sigma^2, \;\; \gamma(\mathbf{h}), \quad \text{and } C(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h})$$

An unsampled location is denoted $\mathbf{u}_\square$. A linear estimator at the unsampled location is written:

$$z^*_\square - m = \sum_{\alpha=1}^{n} \lambda_\alpha [z_\alpha - m]$$

This assumes that the mean is stationary. The weights also depend on the unsampled location and should more properly be written $\lambda_\alpha(\mathbf{u}_\square)$. We are *not* fitting a linear function to the data; we are computing the best linear estimate at each unsampled location. Taken together, the best estimates form a highly non-linear surface. We could consider the residual from the mean $Y = Z - m$ to simplify notation:

$$y^*_\square = \sum_{\alpha=1}^{n} \lambda_\alpha y_\alpha$$

A good estimate could be calculated considering an inverse distance scheme for the weights:

$$\lambda_\alpha = \frac{\left(\frac{1}{h_{\alpha,\square}+0.5}\right)^{1.5}}{\sum_{\beta=1}^{n}\left(\frac{1}{h_{\beta,\square}+0.5}\right)^{1.5}}$$

where $h_{\alpha,\square}$ is the standardized Euclidean distance (see Chapter 1) between data α and the unsampled location $\square$. The small constant of 5% to the

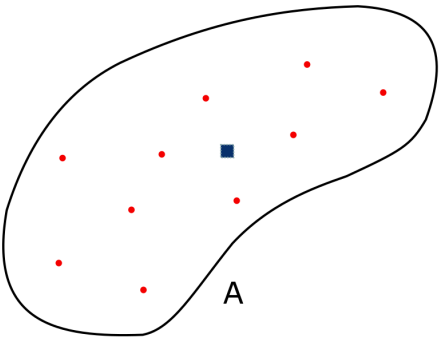


Figure 2.15: Classic sketch of kriging setup.

range is for stability and to mimic a small nugget effect. The power of 1.5 has proven itself in cross validation studies although a higher power would lead to less smooth estimates. Considering global inverse distance would lead to a very reasonable (*good*) artifact free model. The goal of geostatistics to calculate the *best* estimate.

There are many different criteria for the *best* estimate, but the quadratic loss $L(e) = e^2$ is practical and unbiased (a recall of the theorem that "among all estimators, the one that minimizes the squared error criterion is unbiased" will be reviewed in class. Minimizing mean squared error (MSE) is equivalent to maximizing the coefficient of determination R^2. So, our definition of a *best* estimate is one that minimizes error variance (MSE):

$$\sigma_E^2 = E\left\{[Y^* - Y]^2\right\}$$

Considering absolute error leads to a bias. Introducing asymmetric penalties for positive and negative errors would also lead to a bias. Risk aversion or opportunity seeking would be considered after predicting a local distribution of uncertainty (see Chapter 4).

2.4.2 Estimation Variance and Simple Kriging

The error variance can be expanded into three terms:

$$\begin{aligned}
\sigma_E^2 &= E\{Y_\square^{*2} - 2Y_\square^* Y_\square + Y_\square^2\} \\
&= E\left\{\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta Y_\alpha Y_\beta - 2\sum_{\alpha=1}^{n}\lambda_\alpha Y_\alpha Y_\square + Y_\square^2\right\} \\
&= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\lambda_\alpha\lambda_\beta C_{\alpha\beta} - 2\sum_{\alpha=1}^{n}\lambda_\alpha C_{\alpha\square} + \sigma^2
\end{aligned}$$

the first term assesses the redundancy between the data - greater redundancy leads to higher error. The second term assesses the closeness between the data and the unsampled location - the closer the data to the location being estimated leads to lower error. The third term is the stationary variance if there are no nearby relevant data - a lower the stationary variance (homogeneous domain) is better - everything else being equal.

The kriging equations are derived by taking the derivative of the error variance with respect to each of the weights and setting them equal to zero:

$$\frac{\delta\sigma_E^2}{\delta\lambda_\alpha} = 2\sum_{\beta=1}^{n}\lambda_\beta C_{\alpha\beta} - 2C_{\alpha\square} = 0,\ \alpha = 1,\ldots,n$$

$$\sum_{\beta=1}^{n}\lambda_\beta C_{\alpha\beta} = C_{\alpha\square} \quad \alpha = 1,\ldots,n$$

Solving these equations leads to a set of weights $\lambda_\alpha, \alpha = 1,\ldots,n$ that lead to the minimum error variance estimate. This formalism was called Simple Kriging (SK) by Georges Matheron because of the fundamental essence/simplicity/lack of constraints. This form of estimator is also called the Kalman Filter, Optimal Interpolation and the Normal Equations in different contexts.

A demonstration in class at this point shows both cross validation and estimation on a grid. The estimates are smooth and converge to the mean far from data.

2.4.3 Properties of Kriging and Ordinary Kriging

There are many remarkable properties of kriging that include:

- Kriging minimizes the error variance by construction: check the second derivative of the error variance with respect to the weights - the result is $2\sigma^2$ for all weights which is always positive - therefore we are minimizing.
- The variogram/covariance model for the minimization of error variance is data driven (the experimental variogram points) and model driven (interpolation and extrapolation of the variogram to all distances). This makes it very hard to find an estimator with lower error variance.
- Existence and uniqueness: provided there are no data at exactly the same location and provided that we are using a positive definite covariance function $C(\mathbf{h})$, the existence and uniqueness of the solution can be established.
- Exactitude: the estimate $z^*(\mathbf{u}_\square) = z(\mathbf{u}_\beta)$ if $\mathbf{u}_\square = \mathbf{u}_\beta$. This can be shown easily from the existence and uniqueness of the kriging estimate. In these days of machine learning (ML) it is worth emphasizing this property. ML estimates do not have the property of exactitude.

- Block kriging: any shaped blocks can be kriged by discretizing them with points, replacing the variance by the block variance $\sigma^2 = D^2(v, A)$, and replacing the right hand side covariances by averages to the discretization point locations: $C_{\alpha\square} = \overline{C}(\alpha, v_\square)$.
- Linearity of kriging: kriging can be applied with point and block data. The block data could be any dimension or geometry - contiguous or not. Also, kriging a block directly with one system of equations or kriging each discretization point, then averaging the point estimates leads to exactly the same result.
- Kriging with kriged estimates gives the same estimate. Clearly the minimized estimation variance or the kriging variance is lower, but the estimate is exactly the same. Consider kriging at any location $\mathbf{u}$ using n data. Then, krige at location $\mathbf{u}'$ with the same n data. Finally, krige at location $\mathbf{u}$ using the n data plus the estimate at $\mathbf{u}'$. The final estimate and first estimate are exactly the same.
- Simplified $\sigma^2_{SK} = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha\square}$. The result of simple kriging and the equation for the minimization of the error variance can be combined and the minimized error variance shown with this simpler equation.
- Kriged estimates are smoother than the random function, that is, $Var\{Y^*_\square\} = \sigma^2 - \sigma^2_{SK}$. The variance should be σ^2, but it is reduced by the kriging variance. The higher the kriging variance, the smoother the estimate. This makes sense since the kriging variance is higher when there are fewer nearby conditioning data.
- Covariance reproduction, that is, $E\{Y_\alpha Y^*_\square\} = C_{\alpha\square}$. Although the variance of the kriged estimates is too small, the covariance between all kriged estimates and all of the data is correct. This is sometimes used as a basis for sequential simulation; however, a better basis would be the recursive decomposition of the multivariate.

Ordinary kriging, discussed below, starts to impose constraints on the kriging estimator. Many of the properties listed above will only apply approximately, but the estimator may be improved.

Ordinary Kriging

The standard linear estimator:

$$z^*_\square - m = \sum_{\alpha=1}^{n} \lambda_\alpha [z_\alpha - m]$$

is reorganized to collect the terms applied to the mean:

$$z^*_\square = \sum_{\alpha=1}^{n} \lambda_\alpha z_\alpha + \left[1 - \sum_{\alpha=1}^{n} \lambda_\alpha \right] \cdot m$$

The global mean m may be well known, but it may not be appropriate locally. The weight to the mean is constrained to equal zero while minimizing the error variance.

$$z^*_\square = \sum_{\alpha=1}^{n} \lambda_\alpha z_\alpha$$

$$\text{minimize: } \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C_{\alpha\beta} - 2\sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha\square} + \sigma^2$$

$$\text{subject to } 1 - \sum_{\alpha=1}^{n} \lambda_\alpha = 0$$

The Lagrange formalism is introduced, that is, to minimize $f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$, we minimize $h(\mathbf{x}) + \mu g(\mathbf{x})$ in an unconstrained fashion. In our case:

$$h(\lambda, \mu) = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C_{\alpha\beta} - 2\sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha\square} + \sigma^2 + \mu \left[1 - \sum_{\alpha=1}^{n} \lambda_\alpha\right]$$

Taking the $n+1$ derivatives and setting them to zero leads to:

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} - \mu/2 = C_{\alpha\square}, & \alpha = 1, \ldots, n \\ \sum_{\beta=1}^{n} \lambda_\beta = 1 \end{cases}$$

Georges Matheron called this ordinary kriging (OK). In presence of non-stationarity OK can outperform SK by not assigning any weight to the global mean. OK also provides an opportunity for us to intervene in the smoothing of kriging by limiting the search.

2.4.4 Exercise W2-4

The least squares formalism known as kriging in geostatistics is very important for geostatistical modeling.

1. Derive the estimation variance in terms of the covariance. Explain where the assumption of stationarity comes into the derivation.
2. Derive the simple kriging equations by taking the derivative of the estimation variance with respect to the weights. Explain why the kriging weights do not depend on the data values and why this is important.
3. Consider the configuration below. The global mean is 1.3 thickness *units* and the variance is 0.2 thickness $units^2$. Calculate the simple kriging estimate and the simple kriging variance at the unsampled location given that the isotropic standardized covariance function is $C(h) = exp(-3h/250)$. Show all steps clearly.

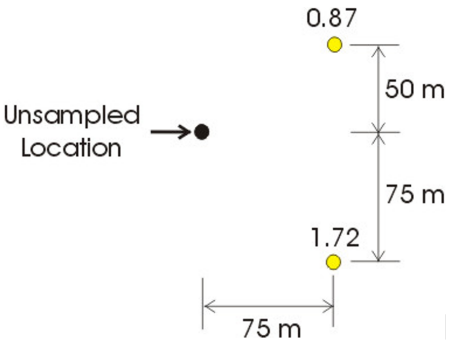


Figure 2.16: Data configuration for kriging exercise Question 3.

Chapter 3

Kriging and Simulation

The link between kriging and simulation is through the multivariate Gaussian distribution, but the history of geostatistics has made the gradual transition between these topics. A lecture plan for the third week:

Four Days	**Five Days**
• Constrained kriging • Primal and dual kriging • Measures of performance	• Constrained kriging • Primal and dual kriging • Measures of performance
• Kriging paradigms • Local uncertainty • MultiGaussian kriging	• Kriging paradigms • Local uncertainty • MultiGaussian kriging
• Monte Carlo simulation • Simulating correlated variables • Sequential Gaussian simulation	• Monte Carlo simulation • Simulating correlated variables • Sequential Gaussian simulation
• Implementation of SGS • Checking local uncertainty • Checking realizations	• Implementation of SGS • Review and Demonstration
	• Checking local uncertainty • Checking realizations • Review and Demonstration

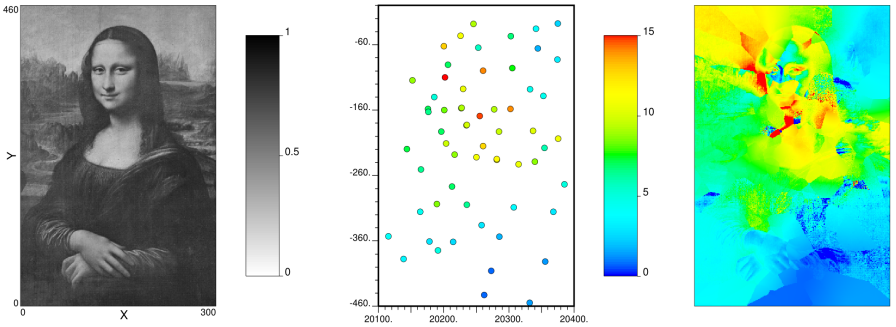


Figure 3.1: Strange example of using a gray scale image of the Mona Lisa as an external drift variable in kriging.

3.1 Kriging II

3.1.1 Constrained Kriging

Ordinary and universal kriging where the mean is non-stationary [37, 43]:

$$m(\mathbf{u}) = \sum_{l=0}^{L} a_l f_l(\mathbf{u})$$

where the a coefficients are unknown and the f_l terms are specified functionals of location. By convention $f_0(\mathbf{u}) = 1$.

The path forward is to remove the mean from the estimator and enfoce unbiasedness:

$$E\left\{\sum_{\alpha=1}^{n} \lambda_\alpha z(\mathbf{u}_\alpha)\right\} = E\left\{z(\mathbf{u}_\square)\right\}$$

The OK framework from the previous chapter is extended:

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} + \sum_{l=0}^{L} \mu_l = C_{\alpha\square}, & \alpha = 1, \ldots, n \\ \sum_{\beta=1}^{n} \lambda_\beta f_l(\mathbf{u}_\beta) = f_l(\mathbf{u}_\square), & l = 0, \ldots, L \end{cases}$$

Kriging with external drift (KED) is where one or more of the functionals come from an external data source such as geophysics. KED is quite widely used and recommended. Most implementations of UK where the functionals are monomials of the coordinates are not recommended. The results are very sensitive to the search and can become unstable particularly in extrapolation with a small search.

CCG paper 2006-119 documents and summarizes an interesting perspective on kriging with constraints (OK, UK/KT or KED). The perspective is based

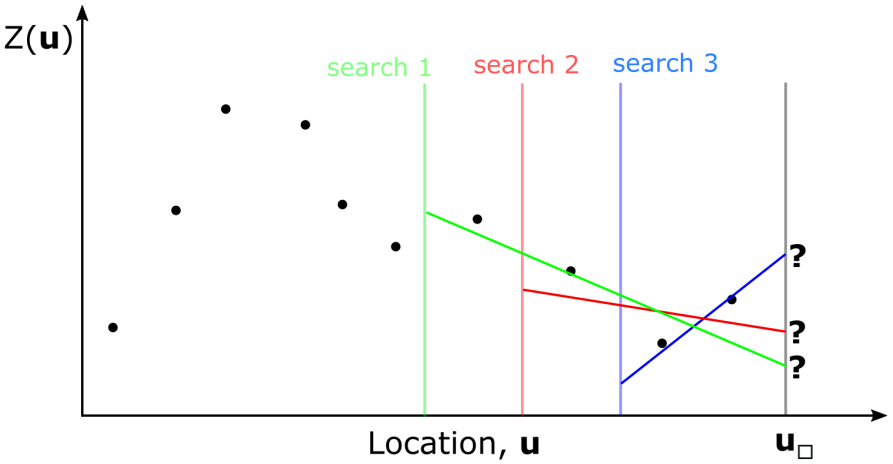


Figure 3.2: Illustration of the instability in UK estimates depending on the search radius chosen.

on an understanding from the early days of geostatistics, but not well documented. The perspective is that these constrained forms of kriging (OK, UK/KT or KED) are exactly equivalent to (1) estimation of the trend modeling parameters (the a_l terms above) by linear regression, then (2) simple kriging considering the mean estimated in the first step. This is an interesting perspective since it places all types of kriging as variants of simple kriging.

3.1.2 Primal and Dual Kriging

Kriging solves an optimization problem. The best weights are computed that minimize the expected error variance. Most optimization problems can be expressed in a primal form and a dual form. The development above is all in the *primal* form. Expressing an optimization problem in a dual formalism is revealing and potentially computationally advantageous.

Consider vectors of data, weights, right hand sides and the matrix for the left hand side:

$$z^T = [z_1, z_2, \ldots, z_n, 0, \ldots, 0]$$

$$\lambda^T = [\lambda_1, \lambda_2, \ldots, \lambda_n, \mu_0, \ldots, \mu_L]$$

$$k^T = [C_{1\square}, C_{2\square}, \ldots, C_{n\square}, f_{0\square}, \ldots, f_{L\square}]$$

$$K = \left[\frac{C_{\alpha\beta}}{f_{\alpha l}} \middle| \frac{f_{\beta l}}{0} \right]$$

Kriging can be expressed given this fundamental matrix notation. The estimator in the primal form:

$$\begin{aligned} z^* &= \lambda^T z \\ K\lambda &= k \\ \lambda &= K^{-1}k \\ \lambda^T &= k^T K^{-1} \end{aligned}$$

The estimator in the dual form:

$$\begin{aligned} z^* &= k^T K^{-1} z \\ &= z^T K^{-1} k \\ &= d^T k \end{aligned}$$

$$\begin{aligned} d^T &= z^T K^{-1} \\ d &= K^{-1} z \\ Kd &= z \end{aligned}$$

The dual form is computationally efficient since the weights do not depend on the unsampled location. They are a function of the data values. Estimats can be computed by recalculating the right hand side vector and a simple vector multiplication. Moreover, the Lagrange multipliers mean something - they are the drift terms (the a_l values). Unfortunately, in the dual form the kriging weights have no easily interpretable meaning and the kriging variance cannot be calculated.

3.1.3 Kriging Measures of Performance

The estimate from kriging always gets better with more data used, see [23] and many other examples. Nevertheless, there are many kriging metrics of kriging performance that are considered including:

1. Kriging variance (KV): $\sigma_K^2(\mathbf{u})$ is the ultimate measure of performance - it is the measure minimized by the kriging equations. No other estimator will have lower KV. Figure 3.3 shows the typical case. SK is better than OK for a relatively small number of data, then OK outperforms because the non-stationary mean is more reliable. Interestingly, the error never increases.
2. Kriging efficiency (KE):

$$KE = \frac{BV - KV}{BV} = \frac{D^2(v, A) - \sigma_K^2}{D^2(v, A)} = 1 - \frac{\sigma_K^2}{D^2(v, A)}$$

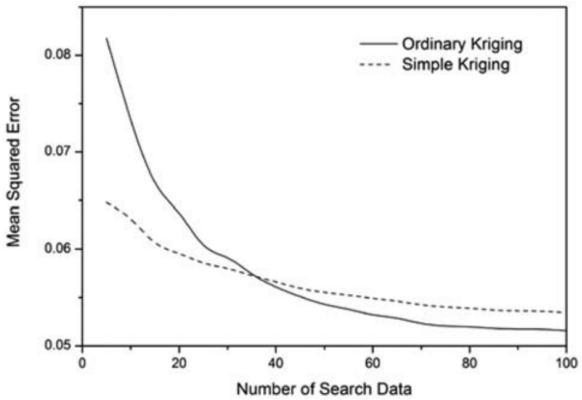


Figure 3.3: Typical example of the estimation variance versus the number of data used. Error goes down with the number of data used.

This is like a local R^2 measure. It is also perfectly inversely related to KV and does not bring any new information - although it is dimensionless.

3. Statistical efficiency (SE): $\frac{GSKV(\mathbf{u})}{\sigma_K^2(\mathbf{u})}$. Efficiency in statistical regression is related to how low the error variance is relative to the lowest it could be. The lowest possible is the global simple kriging variance.
4. Slope of regression (SR) will be developed more below, but it is less than one in OK with a limited search. Ideally, the SR is one, but we accept it being less to avoid excessive smoothing.
5. Negative weight measure (NW)

$$nwm(\mathbf{u}) = \sum_{\text{all neg wts}} \frac{|\lambda_\alpha|}{n} \cdot 100$$

such that one weight of -0.1 in 10 would give a *nwm* of one. Negative weights are not necessarily bad - they lead to local extrapolation in the behaviour of the regionalized variable. Of course, large negative weights and outliers certainly cause problems. In general, set negative estimates to zero and move forward. If the number and magnitude of the negative estimates are significant, then it is likely that the variogram is inconsistent with the spatial distribution of the variable.

6. Weight to the mean (WM) - this is really only for SK. The more weight to the mean, the less the data are important.

Slope of regression:

$$z = a + bz^* \qquad b = \rho\frac{\sigma_Z}{\sigma_Z^*}$$

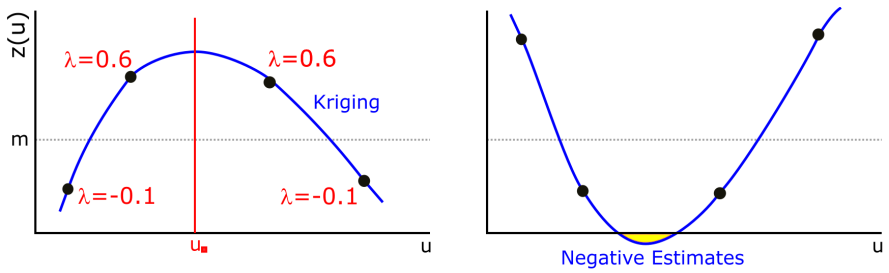


Figure 3.4: Two schematic examples of kriging with negative weights. The example on the left shows an estimate at an unsampled location higher than the surrounding data due to large positive weights to the close data and negative weights to data that are screened. The example on the right shows how the same weights could lead to negative estimates.

$$
\begin{aligned}
b &= \frac{Cov\{z, z^*\}}{\sigma_z \sigma_{z^*}} \frac{\sigma_Z}{\sigma_{Z^*}} = \frac{Cov\{Z, Z^*\}}{\sigma^2_{z^*}} \\
&= E\{ZZ^*\} - m^2 / E\{Z^*Z^*\} - m^2 \\
&= \frac{E\{\sum_{\alpha=1}^n \lambda_\alpha z_\alpha z\} - m^2}{E\{\sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_\alpha \lambda_\beta z_\alpha z_\beta\} - m^2} \\
&= \frac{\sum_{\alpha=1}^n C_{\alpha\square}}{\sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_\alpha \lambda_\beta C_{\alpha\beta}}
\end{aligned}
$$

for all variants of kriging. For SK :

$$
\begin{aligned}
\sigma^2_{SK} &= \sigma^2 - \sum_{\alpha=1}^n \lambda_\alpha C_{\alpha\square} \qquad (3.1) \\
&= \sigma^2 - 2\sum_{\alpha=1}^n \lambda_\alpha C_{\alpha\square} + \sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_\alpha \lambda_\beta C_{\alpha\beta}
\end{aligned}
$$

$$
b = \frac{\sigma^2 - \sigma^2_{SK}}{\sigma^2_{SK} - \sigma^2 + 2(\sigma^2 - \sigma^2_{SK})} = \frac{\sigma^2 - \sigma^2_{SK}}{\sigma^2 - \sigma^2_{SK}} = 1
$$

For OK with one data:

$$
b = \frac{C_{1\square}}{\sigma^2} \leq 1
$$

this shows that the slope is less than one and significantly less than one when the weight to the implicitly calculated local mean increases.

Metrics of performance should always be reviewed, but the results are unsatisfying. Kriging larger blocks is always better according to the measures, but are no different than smaller blocks in practice. The block size should be chosen for practical engineering reasons, adapting to geological boundaries

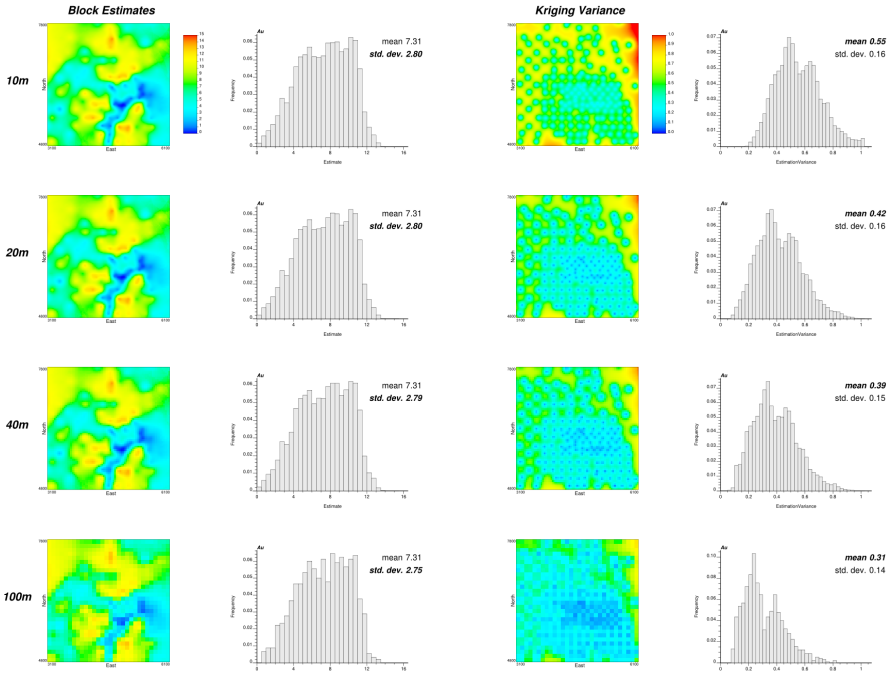


Figure 3.5: Illustration of how block estimates do not change significantly with block size, but the estimation variance reduces significantly.

and for computational considerations. Using more data is always better, but this depends on the purpose of the estimate.

3.1.4 Exercise W3-1

The least-squares formalism known as kriging in geostatistics is very important for geostatistical modeling. In this exercise you will apply different kriging variants under cross validation and for kriging a grid. Data set to be used for mining is the gold grade in Misima.dat.

1. Run cross validation for simple kriging using the declustered mean with 40 closest data using your variogram model. Plot a scatterplot between the estimated and true values for each number of search data. Comment on the scatterplot, the mean error, mean squared error, variance of the estimates and correlation between the true values and estimates. Rerun the cross validation with 4 data and compare.
2. Now run cross validation for ordinary kriging with 40 closest data using your variogram model. Plot scatterplots and comment on your results

as before. Comment on the differences in results between simple and ordinary kriging. Rerun the cross validation with 4 data and compare. Pay particular attention to the conditional bias with ordinary kriging.

3. Estimate on a grid with simple point kriging and ordinary point kriging using 40 samples. Plot slices through the models, histograms of the estimates and comment on the results.
4. Estimate on a grid with simple block kriging and ordinary block kriging using 40 samples. Plot slices through the models, histograms of the estimates and comment on the results.

3.2 Kriging and Local Uncertainty

3.2.1 Kriging Paradigms

The four paradigms of kriging are differentiated by the goal of estimation:

- Visualization or implicit modeling of geology using distance functions or indicators: the goal for this type of estimate is an artifact free estimate that is smooth and natural. A global search should be used if possible, ordinary kriging is preferred and the dual formalism is also preferred for computational efficiency.
- Final estimate when no (or very little) additional data is coming available to help the design or decision: the goals for this type of estimate is to have the best resolution possible (down to 1/4 of the data spacing), avoid any conditional bias and get the best estimate possible - minimum MSE and minimum Type I/II errors. The smoothing of the estimate is of no consequence. A reasonably large search, ordinary kriging, points or small block estimation with 25 (2-D) to 50 (3-D) data.
- Interim estimate - long range resource estimates awaiting more data in the future: the goal is to match what will be mined in the future and to avoid excessive smoothing of the estimates. Conditional bias is of no consequence to this type of estimate. Ordinary kriging is used with a limited search. Volume variance relations are often used to predict the variance or grade tonnage curve that would be encountered in the future and to help choose a limited number of data for the kriging. The search is set fairly large, but the number of data are calibrated for the SMU v variance $D^2(v, A)$, that is, the variance of the estimates is matched to the anticipated variance (or the grade tonnage curve at the chosen cutoff. Figure 3.6 illustrates this. The estimate is not a good final estimate, but the purpose is to match future production - not provide the best local estimate.
- Probabilistic prediction for uncertainty calculations or simulation: the

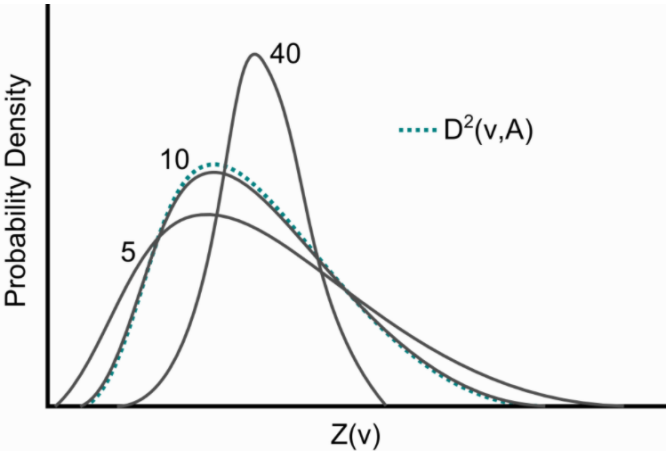


Figure 3.6: Illustration of histogram matching taken from Geostatistics Lessons.

> goal is accuracy and precision of the predicted probabilities, artifact free estimation and consistency with the underlying random function. For Gaussian variables, simple kriging (SK) should be used with a large search and a large number of data: 25 (2-D) to 50 (3-D) data. For indicators, OK is used with a consistent large search across all thresholds.

The key is the difference between a final estimate and an interim estimate anticipating the information effect. Some more details on setting the kriging plan can be found in the Lesson [20].

The block size is an important consideration in kriging. The natural size is about 1/4 data spacing. Estimating blocks smaller than that does not improve the delimitation of boundaries between high and low values. Larger blocks make boundaries fuzzy and imprecise. Small blocks or even points should be used for final estimates, then boundaries can be traced between zones of high and low values that reflect the selectivity of the operation. Final selectivity is better represented by boundaries than by large blocks. The concept of an SMU is for long range interim resource estimation and not final estimation.

The number of data to use in kriging is another consideration. Interestingly, according to all metrics of performance (MSE, R^2, Type I/II errors,...) more data is always better. The advice given above is for 25 (2-D) to 50 (3-D) since the improvement past that number is negligible, yet the computer cost keeps increasing. The close data receive almost all of the weight in kriging and increasing the number beyond these values do not change the estimates much. In fact, global kriging with all of the data, particularly in the dual formalism, is a very useful approach to estimation.

The search radius is often set to the variogram range. Unfortunately, the variogram range is an ambiguous parameter since the variogram often approaches the sill in an asymptotic manner. The search range should be set fairly large, then the number of data actually used in the kriging restricted by the maximum per drill hole (2 to 3 - to avoid the string effect and sub-optimal estimates), maximum per octant (4 to 8 - to avoid artifacts caused by clustered data) and the absolute maximums of 25 to 50. The final estimates can be clipped by a maximum data spacing, that is, a high resolution data spacing model should be constructed and blocks that are in areas with a too-large data spacing should be set to missing.

There is no theory for multiple pass searches. Some practitioners like the use of multiple pass searches to avoid excessive smoothing, build in classification criteria and avoid estimating too far. The preference of this author is to make the best estimate possible for the paradigm under consideration, then clip and classify according to a high resolution data spacing model.

3.2.2 Local Uncertainty

The principle behind kriging is to compute the *best* estimate, which seems reasonable, but all estimates are associated with some error and the uncertainty is not quantified. The modern approach is to quantify the uncertainty at all unsampled locations and use that uncertainty for decision making and simulation. Consider a stationary domain A and a set of data (n):

$$(n): \quad \{z(\mathbf{u}_\alpha) = z_\alpha, \alpha = 1, \ldots, n\}$$

Kriging aims to calculate the best estimate at all unsampled locations:

$$\{z^*(\mathbf{u}_\square), \forall \mathbf{u}_\square \in A\}$$

The minimized estimation variance or the kriging variance provides some information on local uncertainty; a location with high kriging variance is likely more uncertain than one with low kriging variance. The kriging variance does not depend on the data values, that is, it does not consider the proportional effect and does not provide details of the shape of the distribution of uncertainty.

The modern approach is to compute the local uncertainty at each unsampled location:

$$\left\{f_{Z(\mathbf{u}_\square)|(n)}(z), \forall \mathbf{u}_\square \in A\right\}$$

Recall the principles from Section 1.3.1 on Bayes' Law. The definition of a conditional probability can be applied to compute the conditional distribution at an unsampled location:

$$f_{z(\mathbf{u}_\square)|(n)}(z) = \frac{f_{Z_\square, Z_1, \ldots, Z_n}(z_\square, z_1, \ldots, z_n)}{f_{Z_1, \ldots, Z_n}(z_1, \ldots, z_n)}$$

That is, a conditional distribution at an unsampled location is the $n+1$-variate multivariate distribution of the unsampled value with all of the data divided by the n-variate multivariate distribution of the conditioning data.

The simplest possible way of visualizing this is to consider one data and one unsampled location some distance apart. Scanning over data in a stationary domain leads to pairs that inform the bivariate distribution and the marginal distribution required to infer the conditional distribution.

The challenge is to infer the $n+1$ and n variate distributions as n approaches a reasonable number like 25 or 50. The curse of dimensionality (a phrase coined by Richard Bellman) summarizes how such high dimensional distributions cannot be inferred from a finite number of data. The number of replicates required would increase exponentially to infer a high dimensional distribution. There is a need for a parametric distribution.

There are an infinite number of parametric distributions for univariate random variables. Some of these are based on a probabilistic generating mechanism, but most are based on satisfying the requirements of a CDF (non-decreasing between 0 and 1) or a PDF (non-negative and integrates to 1). There are also an infinite number of parametric distributions for independent variables. The multivariate distribution of n independent variables is simply the product of the univariate distributions. In geostatistics, the $n+1$ locations under consideration are related together spatially through the underlying geological processes that led to the present day distribution. The options close in and there is only one parametric multivariate distribution that is mathematically tractible with arbitrarily correlated variables. The multivariate Gaussian (MG) distribution is unique in mathematics and probability.

3.2.3 MultiGaussian Kriging

The multivariate Gaussian distribution is truly unique and widely exploited in statistics and geostatistics. The setup of MG including back transform will be presented. This was developed by Verly in the 80s [60]. The multivariate Gaussian distribution has remarkable properties:

$$f\left(y;\mu,\Sigma\right)=\frac{1}{\left(\sqrt{2\pi}\right)^{d}\left|\Sigma\right|^{1/2}}e^{\frac{-(y-\mu)^{T}\Sigma^{-1}(y-\mu}{2}}$$

The value y is a vector position in a d dimensional space. There are two parameters: (1) μ - a d dimensioned vector of mean values in each coordinate/dimension, and (2) Σ - a dxd dimensioned matrix of variance-covariance values between each coordinate/dimension. As explained in Chapter 1, the generating mechanism of this distribution is summarized by the central limit theorem - the sum of many independent identically distributed variables tends

to a Gaussian distribution. Regardless of whether our variables satisfy this generating mechanism, we adopt this distribution because of its remarkable properties.

The first of many remarkable properties is that all lower $c < d$ distributions are also Gaussian if the d-variate distribution is Gaussian. To the extreme, the univariate distribution of each of the d variables must be Gaussian. All bivariate distributions are bivariate Gaussian and so on. The shape of any marginal or conditional distribution is Gaussian - all we need is the mean vector and variance-covariance matrix to define it completely.

The second remarkable property is that all conditional mean values are linear functions of the conditioning data. In many cases, a linear estimate is adopted for simplicity, but in the case of the multivariate Gaussian distribution, a linear combination of conditioning data *is* the conditional mean.

A third remarkable property of the multivariate Gaussian distribution is that the conditional variance (of any conditional distribution) does not depend on the conditioning data - the conditional variance only depends on the variance-covariance structure between the conditioning data and the variable under consideration.

These properties all lead to the final and most important remarkable property of the multivariate Gaussian distribution. The conditional mean and variance (covariance) values for a multivariate Gaussian distribution are given by the normal equations that are known as simple kriging (SK) in geostatistics.

$$m_c = \sum_{\alpha=1}^{n} \lambda_\alpha y_\alpha$$

$$\sigma_c^2 = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha,\square}$$

$$\sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} = C_{\alpha\square} \quad \alpha = 1, \ldots, n$$

This is an endless source of confusion for classically trained geostatisticians. The belief is that simulation and uncertainty evolved from kriging. That is not true. The normal equations (referred to as SK above) are the analytical solution arising from the MG distribution. *If* the random variable is multivariate Gaussian, then any conditional distribution is (1) Gaussian in shape, (2) a mean value given by a linear combination of the conditioning data, and (3) a conditional variance that depends on the variace-covariance structure of the data. Conditional distributions are easily calculated if the multivariate distribution is multivariate Gaussian.

The premise of modern geostatistics is that two assumptions are required: (1) a careful choice of stationarity - a geological population of data and unsampled locations that belong together, and (2) a normal score transform of all the variables in the stationary domain. Then, all locations within the stationary domain are assumed to be multivariate Gaussian. All local conditional distributions are calculated by the normal equations. The local Gaussian distributions must be back transformed.

The back transformation of a local distribution must be accomplished through a quantile-quantile transformation via the global distribution. Most of us wish for a transformation that would consider only the mean and variance, but the entire shape of the local distribution must be informed. This involves defining a set of probability values randomly or by a regular discretization (preferred at this point). The L probability values are defined by:

$$p_l = \frac{l}{L+1}, \quad l = 1, \ldots, L \quad \text{using } L = 200$$

Then, considering the local Gaussian distribution, the original units distribution is defined by the back transformed quantiles:

$$z_l = F^{-1}\left(G\left(G^{-1}(p_l) \cdot \sigma_c(\mathbf{u}) + m_c(\mathbf{u})\right)\right) \quad l = 1, \ldots L \;\; \forall \mathbf{u}$$

The back transformation is sketched in Figure 3.7 where two local conditional distributions are back transformed. In Gaussian units, the red and the green distributions have nearly the same variance, but the distributions in original units are quite different due to the skewness of the original units distribution. Given $z_l, l = 1, \ldots L$, any summary measure of uncertainty can be computed. The mean, variance, probability to be within an interval of the mean, any other probability values can be computed from these values. The local values could be corrected by volume variance relations; however, the local change of support is not the same everwhere as the global change of support.

There are three key ingredients for this approach to work. First, the stationary domain and data must be geologically coherent and belong together. Second, the distribution $F(z)$ that represents the stationary domain must be truly representative of the stationary domain. Finally, the variogram model of the normal scores that inform the covarnace model must also be representative of the stationary domain.

3.2.4 Exercise W3-2

The objective of this exercise is to become familiar with how kriging can be used to quantify uncertainty without resorting to simulation. Concepts here will be useful for understanding Gaussian simulation algorithms.

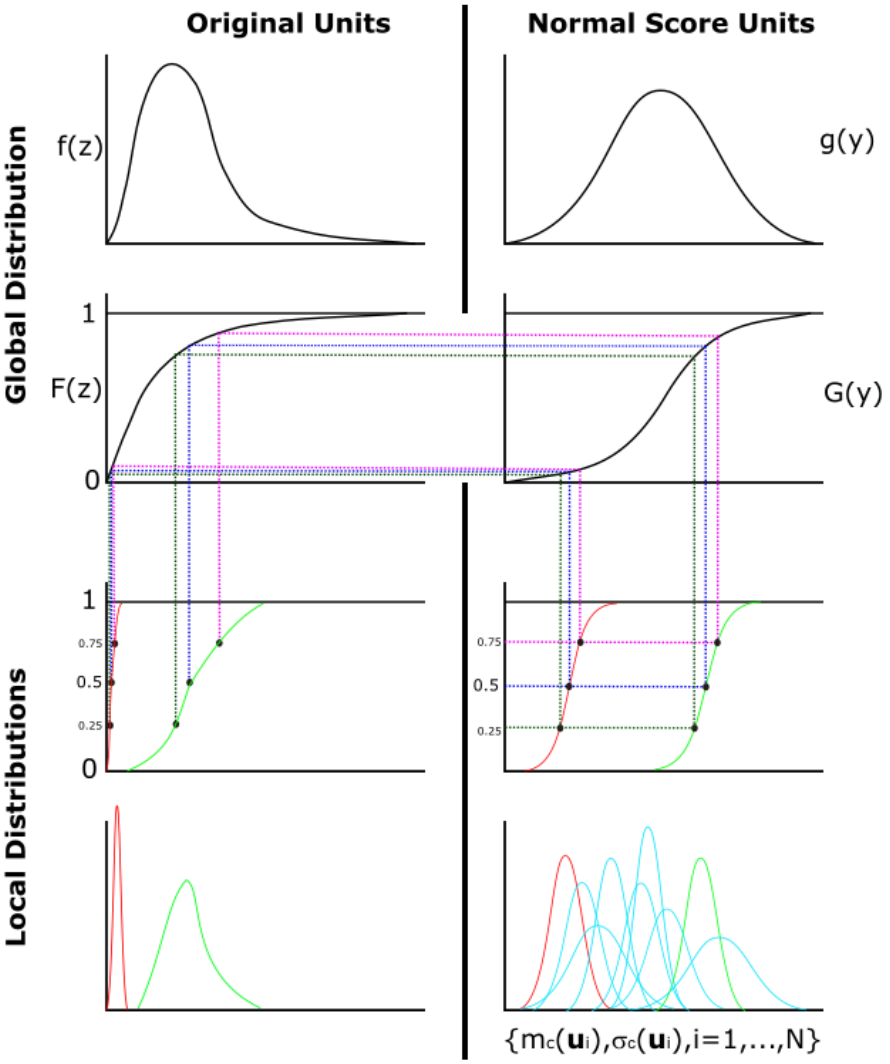


Figure 3.7: Illustration of the back transformation of local conditional distributions in Gaussian units to original units.

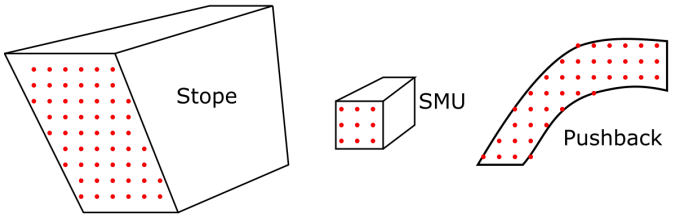


Figure 3.8: Schematic illustration of a stope, SMU and pushback with some discretization points.

Recall the data configuration used previously for kriging by hand (Figure **??**). For this exercises, consider that the thicknesses were found to fit an exponential distribution with a mean of 1 thickness unit. The cumulative distribution function of the exponential distribution is given as:

$$F(x;\lambda) = 1 - exp(-\lambda x), \quad \text{when } x > 0 \quad and \;\; 0 \quad \text{otherwise}$$

where λ is the inverse of the mean. We are interesting in the uncertainty in thickness at the unsampled location shown.

1. Transform the thickness values to Gaussian units.
2. Calculate the conditional mean and conditional variance at the unknown location in Gaussian units.
3. Back transform 99 evenly spaced percentiles $(0.01, 0.02, \ldots, 0.99)$ to establish the local conditional distribution at the unsampled location in units of thickness. Plot the distribution and comment on the shape.
4. Using these back-transformed values, calculate the mean thickness and a 90% probability interval for the thickness at the unsampled location.

3.3 Simulation I

3.3.1 Monte Carlo Simulation

Local uncertainty is useful for drill hole planning and placement, but we are often interested in the uncertainty that involves multiple values/locations at the same time. We are often interested in uncertainty in a value that depends on many locations at the same time: a stope, an SMU, a push back, the production for a month (quarter or year), the entire life of mine (LOM) resources, see Figure 3.8 for a sketch. A challenge is to convert local uncertainty to uncertainty in many locations at the same time. *Many* locations implies thousands to millions.

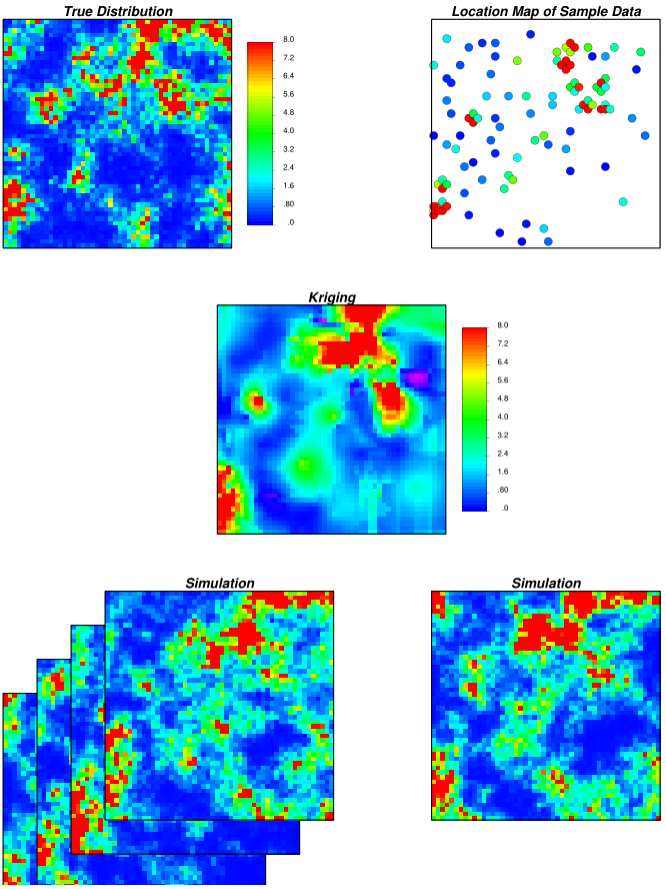


Figure 3.9: Illustration of the principle of simulation. The truth is variable at all scales and sampled at relatively few locations. Kriging leads to a too-smooth distribution. Simulation leads to multiple realizations that capture the heterogeneity and the uncertainty.

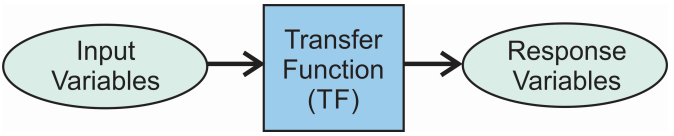


Figure 3.10: Illustration of the principle of MCS. Known uncertainty in input variables is transfered to response or output variables.

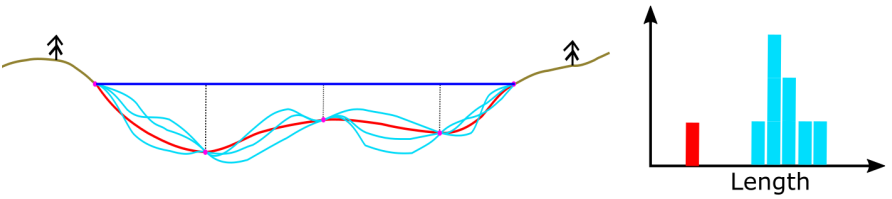


Figure 3.11: Schematic illustration of cable length required by kriging and simulation. The message is that variability may lead to results skewed in one direction. A smooth representation may be biased in another direction.

Monte Carlo simulation is the novel and unassailable technique to solve this problem. A classic reference with historical perspective is [31]. On reflection, there is no other approach to transfer uncertainty from a point scale to larger scale. Think about the curse of dimensionality.

Consider n input variables that are arbitrarily correlated and a response that is some function of all n at the same time. There is uncertainty in each of the n variables that must be transferred through to the response. see Figure 3.10. An evident approach is to discretize all n input variables by some number (10 to 100) and compute the response for the combinatorial of input variables. The problem is that that the number is nearly infinite. A full combinatorial of all options is not possible.

Sampling in a systematic manner would not be reasonable. Sampling all low values for every variable or all high values for every variable would lead to highly improbable results. Sampling the median of all variables is unlikely to be the median of the response. The variability matters and the variability may lead to a systematic change in the results. Examples based on cable length (see Figure 3.11) and flow simulation will be discussed.

The concept of Monte Carlo Simulation (MCS) is to sample a reasonable number of realizations in a fair manner. The key idea emerged from the researchers on the Manhattan Project. There was an effort to directly draw from arbitrary distributions, but that quickly gave way to a two step procedure: (1) draw a random number uniform between 0 and 1, p, then (2) choose that quantile from the distribution being sampled: $z_p = F^{-1}(p)$. Over many

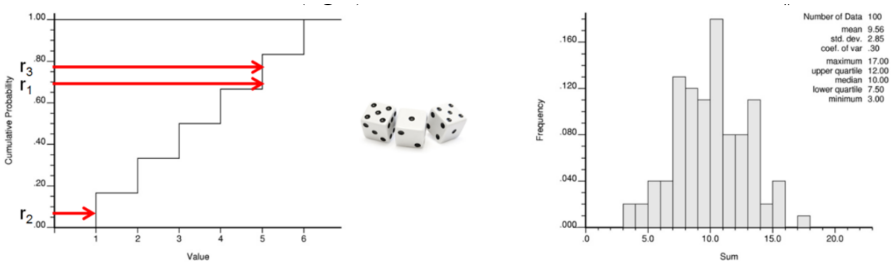


Figure 3.12: Consider three dice as the input variables and the sum shown on the top faces as the response variable. MCS could be used for this. One outcome with three random numbers (r_1, r_2, r_3) is shown on the left and 100 response values is shown on the left.

realizations the simulated values would exactly reproduce the distribution being sampled.

The three dice example is one I have used many times. Consider three dice thrown at random (the input variables) and the sum of the three dice (the response variable). In this simple example there are only 216 possible outcomes (6x6x6), but we could still apply MCS to understand the outcome, see Figure 3.12.

Transfering uncertainty in variables that are independent is very straightforward. Sample each input variable, calculate the response, and repeat many times. The distribution of the response can be constructed to an arbitrary precision. The challenge is when the variables are correlated.

3.3.2 Simulating Correlated Variables

The concept of MCS is well suited to independent variables. Simulating many dependent variables is essential in geostatistics. There may be a variety of mathematically clever solutions, but for understanding purposes, there is nothing that comes close to sequential simulation.

Consider MCS from a bivariate distribution of two correlated variables. The Z_1 value and the Z_2 value cannot be drawn independently - any direct or inverse relationship would not be respected. If the same random number is used for both the Z_1 and Z_2 value, then the relationship would be a perfect direct relationship and the scatter would not be respected. As above, recall the principles from Section 1.3.1 on Bayes' Law. The definition of a conditional probability can be applied to compute the conditional distribution at

an unsampled location:

$$f_{Z_2|Z_1}(z) = \frac{f_{Z_1,Z_2}(z_1, z_2)}{f_{Z_1}(z_1)}$$

leading to:

$$f_{Z_1,Z_2}(z_1, z_2) = f_{Z_2|Z_1}(z) \cdot f_{Z_1}(z_1)$$

The path forward is clear. Regardless of the structure between Z_1 and Z_2, the bivariate distribution can be decomposed into a sequence of two univariate distributions (recall Figure 1.19 for an example): (1) draw z_1 from the marginal distribution of Z_1, then (2) draw z_2 from the conditional distribution of $Z_2|Z_1 = z_1$. Any bivariate distribution would be preserved by this procedure. The order of the variables does not matter - z_2 could be drawn first, then z_1. The key point is that after drawing one variable it must condition the next.

This concept can be generalized to any number of variables. Consider N variables:

$$\begin{aligned} P(A_1, \ldots, A_N) &= P(A_N|A_1, \ldots, A_{N-1}) \\ &\quad P(A_1, \ldots, A_{N-1}) \\ &= P(A_N|A_1, \ldots, A_{N-1}) \\ &\quad P(A_{N-1}|A_1, \ldots, A_{N-2}) \\ &\quad P(A_1, \ldots, A_{N-2}) \\ &\vdots \end{aligned}$$

In practice, the n data are considered as samples already drawn from the multivariate distribution. Simulation starts at $n+1$ and proceeds to N. Also, in practice, as the number of conditioning data increases the number that are actually considered must be limited. A Markov screening like assumption is considered where the nearby values that would influence the conditional distribution are used (the 25 to 50 data mentioned previously). In practice, the derivation of a conditional distribution using 25 to 50 data remains a challenge.

3.3.3 Sequential Gaussian Simulation (SGS)

SGS may not stand the test of time as the most efficient implementation, but for understanding and for flexibility - it is unparalleled. A review of SGS is presented.

Consider the sequential simulation paradigm presented in the previous section. Consider also that the regionalized variable under consideration has

been transformed to normal scores, assumed multivariate Gaussian, and fully parameterized by a stationary assumption for the mean and a normal scores variogram model that informs the covariance structure. Then, the same methodology as used for MG kriging could be used to derive the conditional distribution at each step in the sequential simulation approach. Consider the following steps:

1. Establish a representative distribution $F_Z(z)$ using the best declustering and debiasing availble. The data are normal score transformed using this distribution. The mean is assumed stationary here - a trend model could intervene in the transformation, but that will be seen in the next chapter.
2. Calculate, interpret and model a variogram of the normal score transformed data. This provides the variance/covariance values needed for the multivariate Gaussian distribution.
3. Define a sequence or path through all of the locations to be simulated. This could be any path, but a random path is often considered to avoid any artifacts in the simulated realization.
   (a) Calculate the conditional distribution at the current location considering all original data and previously simulated values - limit our consideration to the nearest 25-50 values. This amounts to applying SK to the nearby data.
   (b) Draw a random value from the conditional distribution. Draw a random Gaussian deviate and consider the conditional mean and variance from the previous step.
   (c) Add the simulated value to the set of data and continue to the next location.
4. Repeat the step above as many times as required for a reasonable number of realizations (200?).
5. Back transform all normal score values to original units.

This procedure is well established and known as Sequential Gaussian Simulation (SGS). There are alternatives based on unconditional simulation and conditioning by kriging, but they amount to the exact same result - realizations from a multivariate Gaussian distribution. There are computational considerations, but theoretically the results are the same. SGS is easy to understand.

3.3.4 Exercise W3-3

Simulation is the only method for the joint quantification of uncertainty between multiple variables at multiple locations. This exercise will use the same data sets used for the kriging exercise. Consider Au in the Misima data.

1. Using the declustered distribution, normal score transform variable to Gaussian units. Calculate and model directional variograms for the normal score transformed variable. Use the directions and models from the kriging assignment as a starting point and modify as appropriate.
2. Run point simple kriging of the normal score variable with a mean of zero, a search up to the variogram range and up to 40 data. Plot some slices of the simple kriging estimates and the simple kriging variance.
3. Use sequential Gaussian simulation to simulate 200 realizations of the normal score variable. Do not have the simulation program transform the data; use your normal score transformed values. Plot slices through the first two realizations and comment on the results. Average all 200 realizations on a block-by-block basis to calculate the e-type estimate and variance. Plot maps of the e-type mean and variance and compare with the point simple kriging results from Question 2.
4. Use sequential Gaussian simulation to simulate 200 realizations letting the program transform and back transform. Average all realizations on a block-by-block basis to calculate the e-type mean and variance. Plot slices through the first two realizations, e-type mean and variance in original units.

3.4 Simulation II

3.4.1 Implementation of SGS

All Gaussian simulation algorithms have the same prerequisites. Any large scale trend must be modeled and the variable detrended before proceeding with simulation. The data must be correctly positioned and outliers must be managed. Simulation is more robust with respect to outliers than kriging, but the tails of the distribution are reproduced by simulation and outliers must be managed carefully.

Another consideration is that simulation assigns values at the scale of the data. The values are averaged to blocks after simulation. There should be 9 values or so (see Lesson on discretization) per block for reliable block values. The number of discretization points may need to be greater if the property being simulated is non-linear and a physics based upscaling is being applied.

All Gaussian simulation algorithms require the data to be transformed to a standard normal distribution. The reference distribution for transformation must be truly representative of the domain under consideration. Declustering and debiasing are applied, see Chapter 1.

The number of data used to condition the local distributions as simulation

proceeds must be sufficient. The 25 in 2-D and 50 in 3-D rules expressed above should be considered. The default in legacy SGS codes (like the GSLIB sgsim program) may be set low for testing/speed. The number of data should not be compromised.

The sequence of visiting the grid nodes for simulation is random to avoid any artifacts. Not completely random - a multiple grid search is a good idea. Simulating a coarse grid, then an intermediate grid, then the final grid helps to enforce long range variogram structure - especially zonal anisotropy. The order within each grid is random.

Sequential Gaussian Simulation (SGS) works well if sufficient data are used in the path. Some published examples showing SGS does not work consider implementation choices that are poor: no surprise that a technique does not work well if incorrectly applied. Although SGS can work well, an alternative to Gaussian simulation is to (1) generate unconditional realizations, then (2) condition by kriging. Unconditional simulation can be very fast with turning bands, spectral techniques and perhaps moving average (convolution) techniques depending on the variogram. Then, conditioning by kriging can be fast if global dual kriging can be applied. The results of this two-step procedure may be more computationally efficient than SGS, but the multivariate distribution being sampled is the same if parameterized the same. Figure 3.13 shows a small 1-D example of this workflow (see the Lesson [30] for more details).

3.4.2 Checking Local Uncertainty

Local estimation is straightforward - we look to minimize the mean squared error (MSE) and maximize the coefficient of determination (R^2) with an unbiased estimator. In the context of simulation or the prediction of uncertainty, the local expected value should be checked against the true values in a classical estimation cross validation mode. There may be systematic biases that are masked by the variability/uncertainty that are caught by checking the expected value.

In the case of uncertainty, accuracy and precision are important. [15] and have to be checked in addition to the expected value. Accuracy is the correctness of the probabilities, that is, do the proportions meet the predicted meaning. Consider some probability intervals, e.g., $p_i = 0.1, 0.2, \ldots, 0.9$, then:

$$p_{low_i} = \frac{1 - p_i}{2} \quad p_{high_i} = \frac{1 + p_i}{2}$$

$$\hat{p}_i = \frac{\text{Number in } (p_{low_i}, p_{high_i}]}{\text{Total Number}}$$

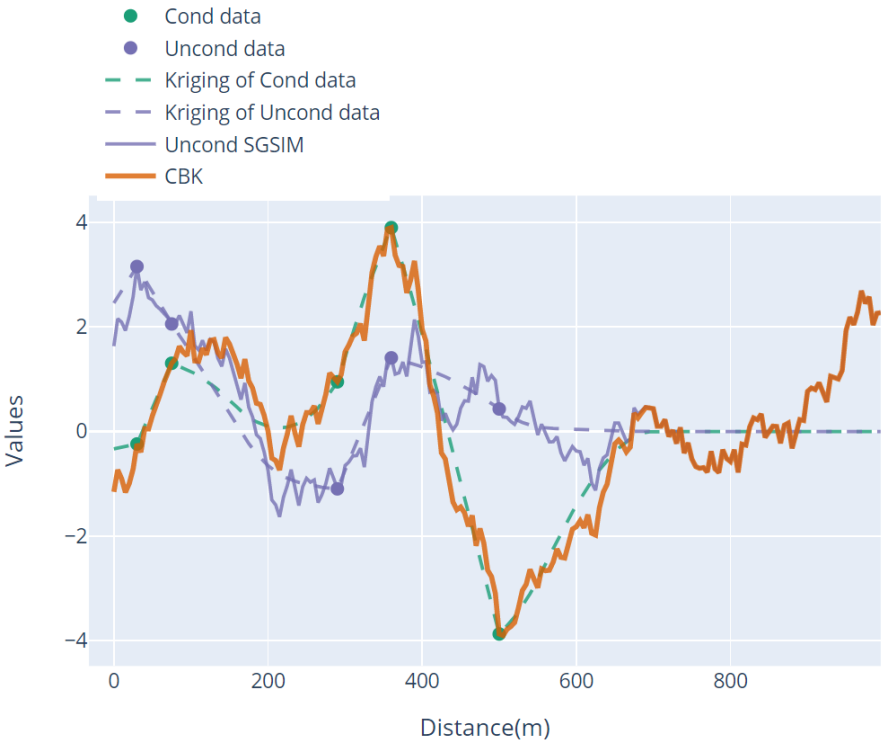


Figure 3.13: Illustration coming from Geostatistics Lessons [30] for conditioning by kriging.

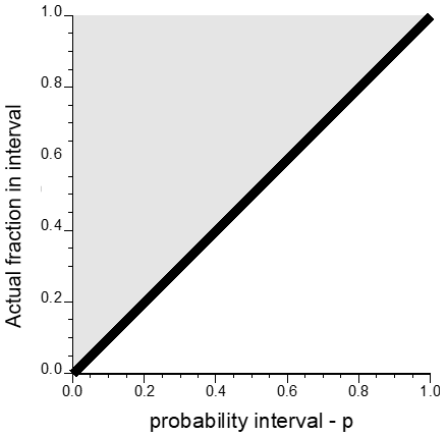


Figure 3.14: Illustration of an accuracy plot. The experimental points should fall on the 1:1 45 degree line. Points above the line are accurate, but needlessly high variance. Points below the line are inaccurate.

The probabilities are accurate if $\hat{p}_i \geq p_i \forall i$. Categorical variables are checked by the probability $\pm$ a tolerance.

This general approach to check accuracy applies to continuous or categorical variables and parametric or non-parametric distributions. In the context of Gaussian distributions the checking is more straightforward since the local distributions are fully defined by a conditional mean and conditional variance. There is no need to simulate to assess the local uncertainty. Recall that simulation is to transfer local uncertainty to multilocation or multivariate uncertainty - the local uncertainty is defined analytically in a multivariate Gaussian context. The *accplt_ns* program is designed especially for normal scores (Gaussian) values. A schematic of an accuracy plot is shown on Figure 3.14. Note that experimental points should fall on (or above) the 1:1 line for accuracy.

Accuracy is the number one consideration in checking uncertainty. If distributions of uncertainty are not accurate, then they are not useful. Once accuracy is achieved we turn our attention to precision. Precision is the narrowness of the uncertainty. Everything being equal we would prefer more precise distributions. Considering the global distribution at each location would be accurate, but not precise. Considering a single estimate at each location would be precise (no variability at all), but not accurate. As mentioned, we must ensure accuracy and then seek the most precise distributions possible.

Measures of precision for continuous and categorical variables:

$$P_{cont} = \frac{\frac{1}{n}\sum_{i=1}^{n}\sigma_{c,i}^2}{\sigma^2}$$

$$P_{cat} = \frac{\frac{1}{n}\sum_{i=1}^{n}H_{c,i}}{H_{max}}$$

The checking software or a custom piece of code should be used to calculate the precision. If the entire workflow is designed correctly, then the practitioner will be naturally led to accurate distributions that are as precise as possible. It should not be possible to randomly perturb the parameters of uncertainty calculation and get to better results.

As more data is acquired, predictions of uncertainty should stay accurate and become more precise. At times, uncertainty increases with more data due to a change in the conceptual geological model. This should not happen often. Parameter uncertainty early on should be large enough to capture this uncertainty.

3.4.3 Checking Simulated Realizations

Checking and validating the results of numerical modeling should be considered to the greatest extent possible. There are minimum acceptance criteria that have been documented. Consider the Lesson [19] and other tradecraft details.

The results of simulation must always be checked carefully - regardless of how well the practitioner understands the theory and software for simulation. Real geology is complex and unlikely to exactly follow our implicit assumptions. A checklist:

- Data reproduction - all data should be reproduced at their locations. Some differences are expected when the simulated locations do not coincide with the data locations or there are dense conditioning data relative to the spacing of the simulated nodes.
- No visual artifacts - although visual inspection is not definitive, careful visual inspection of the realizations can reveal numerical artifacts, edge effects, high grades in known low grade areas (and vice versa), unrealistic continuity or randomness and other issues. The pattern of variability away from the drill holes should look like at the drill holes.
- Cross validation in estimation and uncertainty mode - as described in the previous section should be considered. The expected values should be unbiased and the uncertainty should be accurate and precise.
- Statistics reproduction $(F, \gamma, \rho, \ldots)$ - the histogram, variogram and other statistical parameters such as the correlation coefficient to secondary

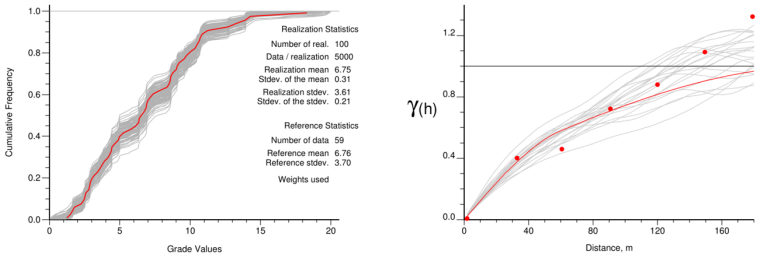


Figure 3.15: Illustration of statistics reproduction for the histogram and variogram. These results show excellent reproduction.

data other simulated variables should be reasonably reproduced. The reference declustered distribution is important. The variograms of the simulated realizations should be closer to the experimental points than the fitted model. An example of statistical reproduction is shown in Figure 3.15. In presence of parameter uncertainty, check against base case.

- Swath plots - considering the average properties in principal directions are used to check reproduction of trends and the data. The "swaths" are perpendicular to the direction vector with tolerance. The average of the data in each swath is compared to the realizations. Figure 3.16 shows an example.
- Average of many realizations equal to kriging - simulation after all does not bring new information to the spatial distribution. Realizations reflect reasonable variability and quantify uncertainty, but kriging provides a solid estimate of the expected value. The correlation between the average of many realizations and kriging should exceed 0.97.
- Reconciliation with production data - this is always the gold standard. Researchers at a university are developing algorithms and are not working at a mine. Unsurprisingly, they do not "talk to their metallurgists" since there are none. At a real operation, however, it is essential that the resource model be reconciled with short term models and with production from the mill.

Correct to kriging if required:

$$\hat{z}(\mathbf{u};l) = z(\mathbf{u};l) \cdot \frac{z_{OK}(\mathbf{u})}{\frac{1}{L}\sum z(\mathbf{u};l)}$$

This correction is particularly useful if a trend model is not being considered and a good ordinary kriged model captures the trend.

Correct the histogram (perhaps iteratively):

$$z_1(\mathbf{u};l) = F_{rep}^{-1}\left(F_L(z(\mathbf{u};l))\right)$$

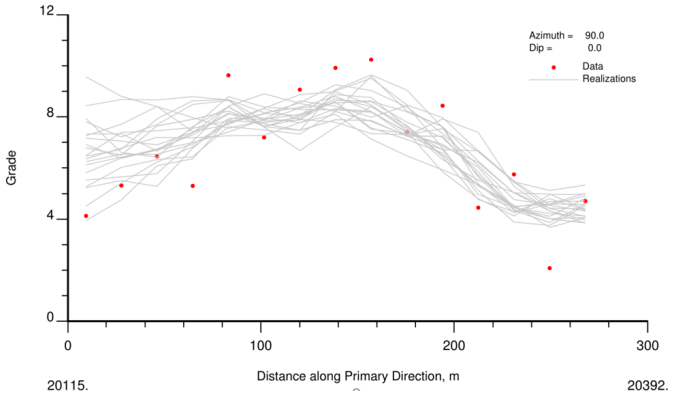


Figure 3.16: Illustration of a swath plot. These results are somewhat concerning. the realizations are too high to the left and too low in the middle. Trend modeling may be in order.

$$z_2(\mathbf{u};l) = z(\mathbf{u};l) + (z_1(\mathbf{u};l) - z(\mathbf{u};l))\,\frac{\sigma_K^2(\mathbf{u})}{\sigma_K^2 max}$$

This was considered more in the past. Correcting the values to the kriged model is preferred in modern applications.

3.4.4 Exercise W3-4

This exercise will be used to investigate some model checks for kriging and simulation and post-processing to enforce statistic reproduction in simulated models. Consider the gold grade in the Misima data.

1. Perform simple kriging cross validation of the normal score transformed variable in Gaussian units. Create an accuracy plot and comment.
2. Using the back-transformed realizations of the variable of interest simulated in the previous exercise, check histogram reproduction of 10 realizations. The mean, variance and shape of the simulated distributions should be compared with the declustered distribution.
3. Now using at least 10 of the normal score (not back-transformed) realizations, check the variogram reproduction along your principal directions and compare with the input normal score variogram.
4. On account of departures from stationarity, it is often convenient to correct the realizations to reproduce a good OK model (using 40 data or so). Perform this correction and comment on the results.

Chapter 4

Multivariate and Categorical

The Chapter covers *Multivariate, Categorical and Post Processing.* A lecture plan for the fourth week:

Four Days	**Five Days**
• Linear model of coregionalization • Cokriging • Collocated cokriging	• Linear model of coregionalization • Cokriging • Collocated cokriging
• Overview of multivariate • Decorrelation • Trend modeling	• Overview of multivariate • Decorrelation • Imputation
• Overview of categorical • Indicators • MPS and HTPG	• Review of multivariate • Trends • Review and geometallurgy
• MIK and UC • Post processing • Classification and localization	• Overview of categorical • Indicators • MPS and HTPG
	• MIK and UC • Post processing • Classification and localization

4.1 Cokriging

4.1.1 Linear Model of Coregionalized Variables

This is particularly relevant for completely unequally samples variables such as data from different drilling or sampling that are never measured at the same location, see Figure 4.1 for a sketch. This would be the case with exploration and production samples or different drilling techniques. Cross variograms and covariances are explained, interpretation summarized and modeling with the LMC explained.

Consider K variables $\{Z_k(\mathbf{u}), \mathbf{u} \in A\}$ within a reasonable stationary domain A. There are K direct variograms and $K^2 - K$ cross variograms that measure the spatial relationship between all pairs of variables, see sketch on Figure 4.2. The generalized expression for direct and cross variograms is given by:

$$2\gamma_{k,k'}(\mathbf{h}) = E\left\{\left[Z_k(\mathbf{u}) - Z_k(\mathbf{u}+\mathbf{h})\right]\left[Z_{k'}(\mathbf{u}) - Z_{k'}(\mathbf{u}+\mathbf{h})\right]\right\}$$

The generalized expression for direct and cross covariances is given by:

$$C_{k,k'}(\mathbf{h}) = E\left\{Z_k(\mathbf{u})Z_{k'}(\mathbf{u}+\mathbf{h})\right\} - m_k m_{k'}$$

In presence of unequally sampled data, the cross covariance must be calculated instead of the cross variogram (recall Figure 4.1). Note that the cross variogram requires both variables at both locations. For this reason, the cross covariance is almost always required when multiple data types are used in cokriging. The cross covariance could be fitted directly or it could be *flipped over* with an interpolated cross covariance of collocated values $C_{k,k'}(0)$.

The calculation principles for cross variograms and cross covariances are very similar to direct variograms as explained in Chapter 2. The choice of lag spacing must consider the spacing of both data types and the possibility of pairing different measurements. The cross covariance in the forward and backward direction are not the same and may need to be calculated and combined outside the software. For example, $C_{k,k'}(\mathbf{h} = 20m45^o)$ is not the same as $C_{k,k'}(\mathbf{h} = 20m225^o)$ although it is the same as $C_{k',k}(\mathbf{h} = 20m225^o)$. The difference is sometimes called the lag effect. Most software assumes $C_{k,k'}(\mathbf{h}) = C_{k',k}(\mathbf{h})$ so the experimental covariances may have to be combined before interpretation and fitting.

Regarding interpretation, the sill of the cross variogram is the cross covariance of collocated values $C_{k,k'}(0)$:

$$\gamma_{k,k'}(\mathbf{h}) = C_{k,k'}(0) - C_{k,k'}(\mathbf{h})$$

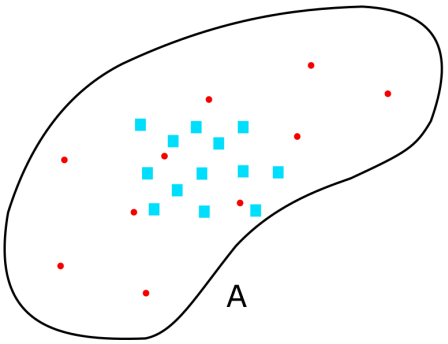


Figure 4.1: Schematic illustration of two data types that are not sampled at the same location, e.g., production and exploration sampling.

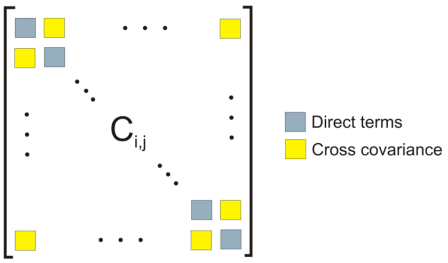


Figure 4.2: Schematic illustration of the direct and cross relationships between all pairs of variables.

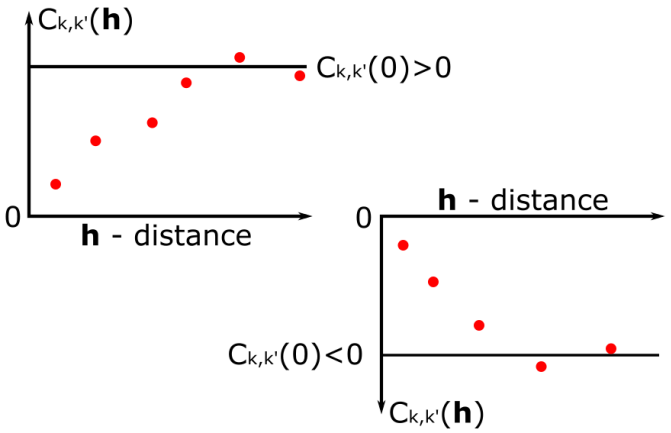


Figure 4.3: Schematic illustration of cross variograms between positively correlated variables (left) and negatively correlated variables (right).

The sill of a cross variogram could be positive or negative - depending on whether the two variables have a direct or inverse relationship, see Figure 4.3 for an illustration of both cases. A cross variogram going below zero may be unexpected.

All direct and cross variograms must be modeled simultaneously to ensure that they are jointly positive definite. The linear model of coregionalization (LMC) is widely used for this purpose:

$$Z_k(\mathbf{u}) = m_k + \sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u}) \quad k = 1, \ldots, K$$

Y_is are independent standard factors with $i = 0$ corresponding to no spatial structure.

$$Var\{Z_k(\mathbf{u})\} = \sum_{i=0}^{nst} a_{k,i}^2$$

$$C_{Z_k, Z_{k'}}(0) = \sum_{i=0}^{nst} a_{k,i} a_{k',i}$$

Given this linear model, the resulting form of the direct and cross variograms is given by:

$$2\gamma_{k,k'}(\mathbf{h}) = \sum_{i=0}^{nst} a_{k,i} a_{k',i} \gamma_i(\mathbf{h})$$

The a values are often combined into c values in the writing of the LMC:

$$2\gamma_{k,k'}(\mathbf{h}) = \sum_{i=0}^{nst} c_{k,k',i}\gamma_i(\mathbf{h})$$

The challenge is to choose a flexible and reasonable pool of nested structures, then set the parameters to fit the experimental points (and all of the known sill values). There are a variety of iterative algorithms that automate this procedure. Relatively recent modeling efforts consider the *very large* LMC with 10 or more structures. Specialized software would be required.

4.1.2 Cokriging

As mentioned, the LMC and cokriging is relevant for completely unequally samples variables such as different drilling or sampling (recall Figure 4.1). The data should be considered as *data events* that have a location, a measured value, a variable or data type identifier, and (optionally) a different volume support and error content. Consider n data indexed as $\alpha = 1, \ldots, n$ that all could be of different type and, perhaps, all of different type than the variable/data being estimated. The cokriging estimator at an unsampled location is written:

$$z^*_\square - m_\square = \sum_{\alpha=1}^{n} \lambda_\alpha \left[z_\alpha - m_\alpha\right]$$

Much information is implicit in this notation. The unsampled location $\square$ has a particular volume support and a particular variable or, perhaps, the best data type for a variable is being estimated. Each data has the attributes mentioned in the paragraph above.

Minimizing the error variance, just like kriging, leads to equations just like kriging. With no constraints, we would have simple cokriging:

$$\sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} = C_{\alpha\square} \quad \alpha = 1, \ldots, n$$

This looks exactly like the simple kriging in Chapter 2. The difference is that the covariance values between different data events and the covariance values between each data event and the location/variable being estimated come from the appropriate direct or cross variogram / covariance model from the LMC. Simple cokriging is a technique that works well. Many practitioners prefer ordinary kriging.

A *traditional* form of ordinary cokriging involved setting multiple constraints: (1) the sum of weights to the primary variable of the same type as being

estimated is constrained to sum to one, and (2) the sum of weights to other variables or data types is constrained to sum to zero. This is a terrible idea. The secondary data cannot be given any influence - any positive weight would have to be compensated for by a negative weight. This led some early practitioners in geostatistics to conclude that cokriging is not useful: an incorrect conclusion.

An alternative is to consider standardized ordinary cokriging where the estimator is written with standardized variables:

$$\frac{z^*_\square - m_\square}{\sigma_\square} = \sum_{\alpha=1}^{n} \lambda_\alpha \frac{[z_\alpha - m_\alpha]}{\sigma_\alpha}$$

Then, the sum of all the weights is constrained to be one. This is a hybrid between simple and ordinary kriging. It seems to work well.

A common question is *what correlation between variables is required to make cokriging a useful and worthwhile endeavour?* A correlation less than 0.25 is almost always insignificant. A correlcation above 0.9 may mean the variables are almost the same. The greater the number of secondary variables / data, the lower our standards for the correlation. Given an exhaustive grid of secondary data coming from geophysics, a corelation of 0.4 would be very significant and important. The special case of an exhaustive grid of secondary data for predicting a primary is often handled with collocated cokriging.

Another concern to address is the volume support difference between different data. Some claim that all data must be at the same support. This is not true. If data are at different support, then the experimental direct and cross variograms are multi-support, that is between different supports. That is fine since they are the correct multi-support measures needed in cokriging.

4.1.3 Collocated Cokriging

There are two important situations when there are exhaustive grids of secondary data for the prediction of a primary variable: (1) exhaustive secondary variable from data collection such as geophysics/seismic, and (2) when variables are being modeled in a sequential fashion, that is, model Z_1, then $Z_2|Z_1$, then $Z_3|Z_2, Z_1$ and so on. In presence of exhaustive secondary data, the collocated data are surely the most important. There are many historical references on this including [1, 39, 54, 2]. The more recent Lesson is helpful summary of the variants [53].

In the original version of collocated cokriging two main assumptions are considered. First, only the collocated secondary variable is used - although there is an exhaustive grid of secondary data, the collocated is surely the most

important. Second, the required cross variogram model is assumed to by the same shape as the primary variable variogram, that is, we assume an intrinsic model between the primary Z variable and the secondary Y variable. If both variables are standard this is written:

$$\gamma_{YZ}(\mathbf{h}) = \rho_{YZ}(0) \cdot \gamma_Z(\mathbf{h})$$

the direct primary variogram is scaled so that the cross variogram has correct sill. This is remarkably simple in implementation - no LMC is required. The primary variable variogram and the correlation to the secondary data are required.

This original form of collocated cokriging (sometimes referred to as Markov Model I) has some problems. The calculated variance is often too high. In a sequential Gaussian framework this increased variance compounds and the final histogram is not reproduced. Another problem is that the shape of the cross variogram may be more similar to the secondary variable variogram. The variogram of the secondary variable is often well informed because it is on an exhaustive grid.

The so-called Markov Model II was devised to allow the cross variogram to be more similar in shape to the secondary variable. Recall the Markov Model I:

$$\rho_{PS}(\mathbf{h}) = \rho_{PS}(0)\rho_P(\mathbf{h})$$

$$\rho_S(\mathbf{h}) = \rho_P(\mathbf{h})$$

The Markov Model II takes the following form where there is a residual variogram/correlogram $\rho_R(\mathbf{h})$ to allow fitting the primary variable variogram that is often less well informed than the secondary variable variogram.

$$\rho_{PS}(\mathbf{h}) = \rho_{PS}(0)\rho_S(\mathbf{h})$$

$$\rho_P(\mathbf{h}) = \rho_{PS}^2(0)\rho_S(\mathbf{h}) + \left(1 - \rho_{PS}^2(0)\right)\rho_R(\mathbf{h})$$

The Markov Model II can work well when the secondary variable dominates. This model does not fix the variance inflaction.

A modification to fix the variance inflation is to extend the number of secondary data used to all of the primary locations as well, see Figure 4.4. This provides a consistent mathematical model and a variance that is not inflated. This leads to improved results in sequential simulation. The kriged values are not much different.

Another important practical aspect of collocated cokriging is to consider the case of one primary variable and many secondary variables. The secondary variables could come from direct measurements or from previously simulated variables. Most software is setup to krige or simulate one primary variable

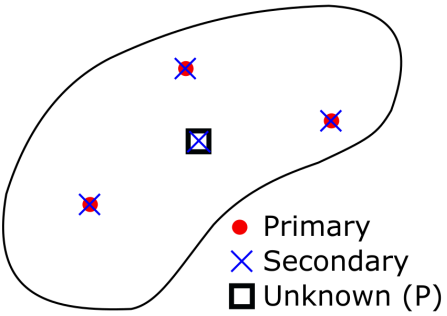


Figure 4.4: Schematic illustration of the setup for collocated cokriging and intrinsic collocated cokriging.

with one secondary variable. The use of a super secondary variable is very useful and practical to combine all secondary variables into one and to facilitate the use of commercial software. See the Lesson on the subject [63].

Consider a primary variable denoted Y and multiple secondary variables denoted $X_j, j = 1, \ldots, n_s$. The idea is to combine all n_s secondary variables at an unsampled location into a single *super* secondary variable denotes X_{SS}. This is done by a type of cokriging:

$$X_{SS} = \frac{\sum_{j=1}^{n_s} \lambda_j X_j}{\sigma_{SS}}$$

The equations to calculate the weights are exactly cokriging equations. The correlations between all secondary variables and between the primary and secondary variables are required (the primary variable, of course, is not required - just the correlation structure):

$$\sum_{j=1}^{n_s} \lambda_j \rho_{i,j} = \rho_{i,Y}, \quad i = 1, \ldots, n_s$$

$$\sigma^2_{cond} = 1 - \sum_{j=1}^{n_s} \lambda_j \rho_{j,Y}$$

If the secondary data have been normal score transformed, then the shape of the super secondary variable distribution is likely Gaussian. The variance of the super secondary variable, however, is not one. It can be calculated as:

$$\sigma^2_{SS} = 1 - \sigma^2_{cond} = \sum_{j=1}^{n_s} \lambda_j \rho_{j,Y}$$

This is used to standardize the super secondary variable. Also, the correlation between the primary variable and the super secondary variable is always

positive and greater than the correlation to any single secondary variable. It is calculated as:

$$\rho_{SS,Y} = E\{X_{SS}Y\} = \frac{\sum_{j=1}^{n_s} \lambda_j \rho_{j,Y}}{\sigma_{SS}} = \frac{\sigma_{SS}^2}{\sigma_{SS}} = \sigma_{SS}$$

This greatly facilitates the use of ICCK in modeling secondary variables. Other important related techniques are Bayesian Updating and Error Ellipses that will be covered in lectures if time permits. The Lessons on Bayesian Updating [64] and Error Ellipses [27] could be reviewed for combining distributions of uncertainty.

Permanence of ratios is an interesting solution to the "ABC" problem discussed in Chapter 1, see [40] and [49]. The idea is to consider a probability distance:

$$r = \frac{P(\overline{A})}{P(A)} = \frac{1 - P(A)}{P(A)} \quad \in [0, \infty)$$

$$r_{12} = \frac{1 - P(A|D_1, D_2)}{P(A|D_1, D_2)}$$

$$r_1 = \frac{1 - P(A|D_1)}{P(A|D_1)} \qquad r_2 = \frac{1 - P(A|D_2)}{P(A|D_2)}$$

Assume that the contribution of D_2 to A with knowledge of D_1 is the same as not knowing D_1. D_1 and D_2 could be interchanged:

$$\frac{r_{12}}{r_1} = \frac{r_2}{r} \qquad \frac{r_{12}}{r_2} = \frac{r_1}{r}$$

$$r_{12} = \frac{1 - P(A|D_1, D_2)}{P(A|D_1, D_2)} = \frac{r_1 r_2}{r}$$

$$P(A|D_1, D_2)\left(\frac{r_1 r_2}{r} + 1\right) = 1$$

$$\begin{aligned} P(A|D_1, D_2) &= \frac{1}{1 + \frac{r_1 r_2}{r}} = \frac{r}{r + r_1 r_2} \\ &= \frac{\frac{1-P(A)}{P(A)}}{\frac{1-P(A)}{P(A)} + \frac{1-P(A|D_1)}{P(A|D_1)}\frac{1-P(A|D_2)}{P(A|D_2)}} \end{aligned}$$

$$P(A|D_1, \ldots, D_n) = \frac{\left(\frac{1-P(A)}{P(A)}\right)^{n-1}}{\frac{1-P(A)}{P(A)} + \prod_{i=1}^{n} \frac{1-P(A|D_i)}{P(A|D_i)}}$$

This approach is useful when a full multivariate distribution is not available or easily estimated.

4.1.4 Exercise W4-1

This exercise will introduce some more advanced multivariate simulation techniques. Consider skarn2d.dat with gold and copper as a secondary variable.

1. Normal score transform gold and copper, calculate and model a reasonable normal score variogram.
2. Using appropriate declustering weights, normal score transform both copper and gold. Make a cross plot of the normal score transformed variables and comment on the correlation and relationship between these variables.
3. Use sequential Gaussian simulation to simulate 100 realizations of gold. Have the simulation program perform the transformation using the declustering weights. Carefully choose the distribution tails for the back transformation based on the variable histograms. Average all 100 realizations on a block-by-block basis to calculate the e-type mean and variance. Plot maps of the first two realizations, e-type mean and variance in original units.
4. Simulate 100 realizations of copper conditional to the gold realizations. Average all 100 realizations on a block-by-block basis to calculate the e-type mean and variance. Plot maps of the first two realizations (plot them next to the first two gold realizations), e-type mean and variance in original units.

4.2 Multivariate

4.2.1 Overview of Multivariate Techniques

The theory and practice of multivariate geostatistics is dispersed in papers, presentations, books and courses including [62]. Barnett's guidebook is a good source. The book by Rossi and Deutsch is descriptive and contains some interesting case studies. The multivariate statistics book of Johnson and Wichern is a good theoretical reference. One important way for practitioners to understand multivariate geostatistics is through the workflows to accomplish certain tasks. The workflows on Figure 4.5 are considered important in modern mining multivariate geostatistics.

The first six workflows are related to data preparation, exploration and modeling. The final six workflows are more related to spatial modeling.

Workflow 1.a: Transform Compositional Data. Many of the data considered in multivariate geostatistics are compositional, that is, they are non-

I. Data Preparation
   a. Transform Compositional Data
   b. Transform to Normal Scores
   c. Summarize Multivariate Relationships
   d. Aggregate Data Variables
   e. Clusters Data Observations
   f. Fit Response Surface

II. Spatial Prediction
   a. Estimate with Unequally Sampled Data
   b. Simulate with Colocated Cokriging
   c. Impute Missing Data
   d. Simulate with Linear Decorrelation
   e. Simulate with Nonlinear Decorrelation
   f. Simulate with Stepwise Conditional Transform

Figure 4.5: Hierarchy of workflows for multivariate geostatistics. Although incomplete, this provides some structure to become acquainted with the subject.

negative and sum to unity. A filler variable may be required if all constituents are not measured. Considering the variables in standard workflows without special treatment will not enforce the summation constraint. The use of a ratio or logratio transformation enforces the constraint.

Workflow 1.b: Transform to Normal Scores. Virtually all multivariate techniques require the data to be standardized and, ideally, univariate standard normal. The normal score transform is commonly applied to each variable one at a time. The original data should be despiked, that is, ties or constant values caused by detection limit or significant digits should be broken. A representative distribution is required. Declustering and/or debiasing is applied as required. Unequal sampling may require different declustering, but a unique set of declustering weights is appropriate for equally sampled data. An upper and lower tail specification is important for the back transformation. Fitting the distribution with polynomials or with kernel smoothing may be useful. Sometimes spikes should be preserved and not despiked - this makes distance-based multivariate techniques more accurate.

Workflow 1.c: Summarize Multivariate Relationships. The main goal is to understand multiple variables for decisions of stationarity, ensuring quality control, workflow selection and so on. Understanding multivariate data is challenging because we cannot visualize the data in high dimensions. Cross plots of the data are interesting, but only show two axes at the same time. Alternative attributes such as color, size and symbol shape could be used, but

those representations are not mutually perpendicular to the other dimensions. Cross plots of normal scores could be created by the scatnscores program that includes a test of bivariate Gaussianity.

A correlation matrix of all secondary and primary variables is a nice display. Some optimal ordering of the variables could be considered (see Maryam Hadavand's work). Understanding a large correlation matrix is not easy. A useful supplement is to plot the variables with their MDS coordinates to more easily see the ones that are close, far and clustered. Multidimensional Scaling (MDS) embeds entities in a series of lower dimensional spaces to preserve the pairwise distances as accurately as possible. The first requirement is pairwise distances:

$$d_{i,j} = \sqrt{\sum_{k=1}^{K} (x_{k,i} - x_{k,j})^2} \qquad \text{or } d_{i,j} = 1 - \rho_{i,j}$$

first dimension is best 1-D, second is best 2-D and so on. This is another good way to summarize multivariate relationships. See the Lesson on MDS [44].

Workflow I.d: Aggregate Data Variables. There are two main reasons to aggregate or combine variables. One reason is to facilitate techniques that consider a single or a few secondary data such as intrinsic colocated cokriging. Multiple variables are combined into a *super* secondary variable and the information from all is passed to subsequent calculations (see above). A second reason to aggregate variables is to reduce the number used in fitting a response surface. Some response surface fitting techniques are prone to overfitting, yet we do not want to remove variables from the analysis. Aggregating closely related variables aims to preserve the information from them while reducing some noise due to averaging. As described above, a linear aggregation scheme with cokriging is commonly applied. This would be done with the normal scores. A hierarchical modeling workflow would require a different data aggregation operation at each step in the workflow.

Workflow I.e: Cluster Data Observations. Clustering will assemble the data observations into different groups where, ideally, the observations in a group are close together and the groups are far apart. The main goal is to understand the data better and, perhaps, to assemble the data into a revised set of stationary populations. The data should be normal score transformed before clustering. A hierarchical clustering is often applied first to look for natural groups. A dendogram display with the data ribbon plot (the cluster program of Ryan Barnett) provides a useful display. This is done with the results of hierarchical clustering, then alternative clustering techniques such as k-means and Gaussian Mixture Model (GMM) clustering is considered.

If the number of clusters is specified from prior knowledge (probably the den-

dogram of natural clustering), then a k-means or Gaussian Mixture Models (GMM) clustering algorithm would often create better clusters. The advantage of GMMs is the use of anisotropy in the cluster space. No non-linearity is considered. The spatial consistency of the resulting clusters must be considered. Although the spatial coordinates could easily be used in clustering, the results are not usually meaningful. The clusters can be easily visualized.

Workflow I.f: Fit Response Surface. A response surface provides a prediction of a dependent variable given a set of predictor or independent variables. An assessment of the importance or sensitivity of the response to the predictor variables is also provided. There are multiple reasons to fit a response surface. A common application is in post processing multiple realizations with the results of a transfer function (the response variable) and posterior parameters summarizing the realizations (the predictor variables). Another application is in fitting geometallurgical response variables from other assay measurements.

Non-informative predictor variables should be screened at the start to avoid over-fitting. A linear or quadratic response surface is easily fit, but most modern approaches would consider them over fit particularly with relatively few observations and a large number of predictor variables. It is reasonable to start with a linear and quadratic response surface fit with sabor (Zagayevskiy) or similar program. Increasingly, non-linear fits with techniques like ACE, random forests, gradient boosting or neural networks are becoming popular. Gradient boosting can improve the fit in many cases. Sensitivity analysis with tornado charts is important to show an easy-to-understand. Show prediction on cross plot and note residual R squared (compare to ρ^2).

Workflow II.a: Estimate with Unequally Sampled Data. Conventional cokriging is rarely applied. One case where cokriging still proves remarkably useful is to create estimated models in presence of multiple data types that have not been sampled at the same locations, e.g., production sampling and exploration data or legacy and modern data. This requires a linear model of coregionalization between primary and secondary data. Note that the cross covariance between the primary and secondary data is required since a cross variogram cannot be computed directly.

Standardized ordinary cokriging or simple cokriging are applied and traditional ordinary cokriging constraints is avoided since secondary data are forced to have no importance. Should cross validate with primary only and with primary+secondary to show that there is no bias and to show that the estimates are improved. Should create kriged model with primary only to compare with cokriged model to further confirm no bias and a reduction in smoothing with more informaiton.

Workflow II.b: Simulate with Colocated Data. A hierarchical approach

to simulate many variables remains viable. The first primary variable is simulated with all secondary data. The second is simulated with an aggregated variable consisting of the first and all secondary. The third considers the first two primary and the secondary, and so on. The primary variables are ordered in decreasing order of available data and importance. For all primary variables in order: (1) combine all secondary data and previous primary variables, (2) note updated correlation coefficient, and (3) apply Intrinsic Colocated Cokriging (ICCK). There are multiple realizations of the super secondary variables when simulating the second and subsequent primary variables. The use of ICCK in preference to conventional colocated cokriging avoids variance inflation and a lack of histogram reproduction. This workflow makes a strong multivariate Gaussian assumption between all primary and secondary data, but virtually any software can be used.

The following workflows are discussed in upcoming lectures, but a summary is given here for completeness.

Workflow II.c: Impute Missing Data. The decorrelation workflows II.d and II.e require full valued data sets with no missing values. Data imputation techniques have evolved to account for multivariate relationships and spatial correlation. Multiple imputation creates multiple datasets where each dataset is used to create one geostatistical realization. There are parametric and non-parametric methods. Details are described below.

Workflow II.d: Simulate with Linear Decorrelation. The only technique to decorrelate data is principal component analysis (PCA). This venerable technique has been applied to decorrelate data, reduce dimensionality and lead to a greater understanding of the truly independent factors in a system of variables. The principal components are ordered in decreasing order or variance. This permits some to be eliminated from geostatistical calculations; PCA is chosen if dimension reduction is important. The principal components can be restandardized by dividing by their standard deviations. This "sphering" step can be followed by a rotation back to the original basis to avoid mixing of the variables.

Any rotation could be considered once the variables are sphered. The MAF rotation is one that aims to decorrelate at an arbitrary **h** vector and to order the resultant factors according to autocorrelation (hence the name min/max autorcorrelation factors). This is useful if there is an important lag effect and cross correlation in the data. This MAF rotation could also be applied after the II.e non-linear workflows.

Workflow II.e: Simulate with Nonlinear Decorrelation. Complex non-linear relationships are not removed by the linear decorrelation workflows. The PPMT approach considers further transformations to correct the multivariate distribution to a truly multivariate normal distribution. Other

transforms are possible, but the PPMT has well established codes and has been tested.

A MAF spatial decorrelation step could be added after the PPMT transform if considered necessary. The resulting factors are simulated independently. Ratios and data imputation may be required before this transformation.

Workflow II.f: Simulate with Stepwise Conditional Transform. This workflow predates the PPMT in II.e. The transform proceeds in a stepwise fashion from the most sampled and most important variables to the less sampled/less important variables. The transform requires a fitting of the multivariate distribution using Gaussian mixture models or similar if there are more than a few variables. PPMT is more flexible except in cases of trends and secondary data; in these cases the stepwise transform is simpler to apply.

4.2.2 Decorrelation

PCA was introduced in the first chapter, but there are important extensions that greatly facilitate modern multivariate modeling. The series of Lessoms will be useful [4]. Principal Component Analysis (PCA) rotates the coordinate space until all values appear uncorrelated. This helps with:

1. Understand contributions to pure components
2. Reduce dimensionality
3. Decorrelate variables for subsequent calculations

consider the correlation matrix of standardized data:

$$\Sigma : C_{i,j} = \frac{1}{n} \sum_{\alpha=1}^{n} y_{\alpha,i} y_{\alpha,j}$$

Spectral decomposition $\Sigma = VDV^T$ to compute PCs (P):

$$P = YV \qquad Y = PV^T$$

Dimension reduction sphereing (standardize):

$$W = YVD^{\frac{-1}{2}}$$

Spectral decomposition sphereing (back to original frame):

$$X = YVD^{-\frac{1}{2}}V^T$$

Minimum Maximum Autocorrelation Factors (MAF) rotate again to impart properties for a particular lag ($\mathbf{h}$). MAF is useful for decorrelating at a lag different from 0 [58, 6].

- Spectral decomposition: $\Gamma_X = V_x D_x V_x^T$
- $M = XV_X$ Could also use W and Γ_W

Projection Pursuit Multivariate Transformation (PPMT) considers sphered components (MAF could apply afterwards - just like PCA or sphereing). PPMT is covered in Lesson [5]. The process involves:

1. Search for projection vector where data are most non-Gaussian
2. Normal score transform along that vector
3. Iterate until convergence

In presence of secondary data the PPMT factors become related through the secondary data and must be simulated in sequence considering previously simulated. This is very counterintuitive.

Imputation has to fit somehere are the is lilely the place. Multiple Imputation (MI) of missing data:

1. Order locations and variables in decreasing order of info.
2. Establish conditional distribution 1 based on collocated data
3. Establish conditional distribution 2 based on spatial data
4. Merge conditional distributions
5. Sample merged distribution and continue

The many implementation details of this would be found in other sources.

4.2.3 Trend Modeling and Modeling with a Trend

A continuous regionalized variable $Z(\mathbf{u}), \mathbf{u} \in A$ is not intrinsically composed of a deterministic trend and stochastic residual; nevertheless, there are large scale volumes of higher and lower values within a nominally stationary domain. Dividing the domain into smaller domains that appear more stationary may be an option, but the number of data reduces and artificial discontinuities are introduced at domain boundaries. Maintaining reasonably large domains for modeling and explicitly considering a trend has proven useful in many case studies.

A numerical model of the deterministic trend is required. Kriging for the trend has not been successful because kriging aims for data reproduction and tends to the global mean at the margins of a domain. Geological variables are rarely amenable to a simple polynomial or functional trend shape. A weighted moving window average has proven effective. Some important implementation details to consider: (1) a length scale for the moving window specified for the primary direction of greatest continuity, (2) a Gaussian shape to the weighting function, (3) anisotropy in the kernel length scale somewhat less

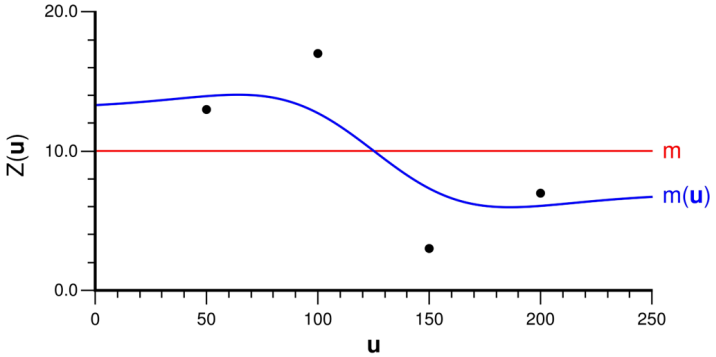


Figure 4.6: Example of a constant mean and a plausible trend model.

than that of the regionalized variable - often a square root of the anisotropy ratio to the maximum direction of continuity, (4) the weight to each data is the kernel weight multiplied by the declustering weight, and (5) a small background weight to all data of, say, one percent. The only free parameter is the length scale in the primary direction. Despite some worthwhile attempts to automate the calculation of this parameter it is set by experience and the visual appearance of the final model. A value one third of the domain size may be reasonable. Once the trend is modeled, we must simulate with the trend. The blue trend $m(\mathbf{u})$ on Figure 4.6 comes from a weighted moving window average.

Modeling a trend in an optimal fashion is a worthwhile objective. The goal is to avoid over and under fitting. The assessment of a trend model is often based on visual criteria. Considering the correlation between the variable $Z(\mathbf{u})$ and $Z(\mathbf{u}) - m(\mathbf{u})$ can help to determine an appropriate trend model. This correlation should be a minimum. We consider plotting this correlation versus a smoothing parameter to determine the parameter that minimizes this correlation. Some research (J. Qu's PhD thesis) has indicated that the smoothing parameter should be larger than the value that minimizes the correlation - perhaps as much as twice.

Creating a residual as $R(\mathbf{u}) = Z(\mathbf{u}) - m(\mathbf{u}), \mathbf{u} \in A$ is not good practice. Z and m are related in complex ways causing R and m to be dependent. If R is modeled independently then artifacts will be introduced in the $R + m$ back transform. A stepwise conditional transform (SCT) [42] of the original variable conditional to the trend has proven effective. As shown in Figure 4.7, this transform completely removes the dependence on the trend.

Independent modeling proceeds and the reverse transform introduces the dependency between the original variable and the trend. The SCT considers a fitted Gaussian Mixture Model (GMM) between the normal scores of the

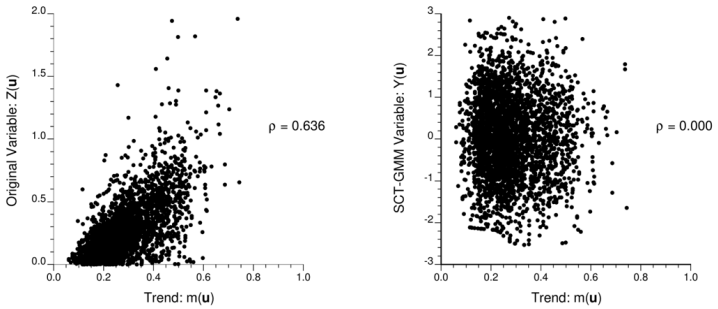


Figure 4.7: Example of how a variable is transformed conditional to a trend to be independent of the trend.

original variable and the normal scores of the trend. These normal score transforms are an intermediate step and are easily reversed. The workflow of trend modeling, data transformation, simulation and back transformation can be largely automated. There are, of course, validation steps that require the attention of a professional.

Modeling the relationship between a trend and the variable (in Gaussian units) with a Gaussian Mixture Model, then modeling, then back transformation seems to work very well. That is, the large and short scale structure, the histogram and the variogram appear reasonably reproduced.

4.2.4 Exercise W4-2

The objectives of this exercise are to learn how to normal score transform data, calculate a correlation matrix and apply principal component analysis to determine orthogonal combinations of variables which account for a large amount of variation. Use the skarn data with W, CaF2, Au, Cu, and Bi.

1. Normal score transform the data. Note that you do not need to use declustering weights for this exercise, but you may optionally also complete this exercise with declustering weights to see if there are any differences in the correlations.
2. Calculate a correlation matrix for the normal score transformed data. Also plot the normal score bivariate distributions and comment on the results.
3. Run principal component analysis on the normal score transformed variables to construct orthogonal linear combinations of the normal score data. Sphere the results (the standardize option). Plot the correlation matrix and bivariate distributions. Comment on the results.

4. See how the linear combinations are constructed and visualize the loading of each variable on to the principal components to see which variables explain the greatest amount of variance in the data set.
5. Run PPMT with the original data. Plot the correlation matrix and bivariate distributions of the PPMT factors. Comment on the results.

4.3 Categorical Variables

4.3.1 Overview of Boundaries and Surfaces

An overview of surfaces, boundaries and categories in a hierarchical modeling framework will be presented. This is a taxonomy lecture like that at the start of multivariate.

Surfaces should be modeled across the entire domain and unlimted with respect to holes and edges.

Boundaries are 2-D and 3-D limits that intersect previously modeled regionalized variables. Lesson on multicategory SDF is interesting [55].

The signed distance function (SDF) is the sign dependent shortest distance to something different. SDF values could be interpolated by global inverse distance, global kriging or radial basis functions (RBF). Then, the resulting model is thresholded at zero. There are significant problems with boundary and edge effects. The length of the drillholes has an influence (a parameter than modifies the slope may be considered). A gap parameter (the c parameter) could be introduced to capture uncertainty. Recent work has shown that it is better to resort to indicators and to calibrate the thresholding of the indicator according to a nearest neighbor model.

Multiple point statistics (MPS) are to go beyond the variogram (2 point, but not considere multiple). The idea is to extract patterns from a training image and to impose them on a simulated realization. The definition of the conditional distribution is used to derive conditional probabilities. Values are simulated from the conditional probabilities.

4.3.2 Categorical Indicators

Categorical variables are discrete lablels that have no inherent numerical meaning. They should, of course, have a geological meaning relevant to the modeling context. The spatial continuity of the discrete variables must be

quite large; otherwise, the variable is really a continuous variable of the proportion of the categories. The categories are also mutually exclusive, that is, each location belongs to one category or another. The list of categories must also be exhaustive even if one category is "everything else".

Although the categories do not have intrinsic numerical meaning they are often ordered. Perhaps by a degree of alteration, a concentration of an important species or some other geological variable or position. The $k = 1, \ldots, K$ categories should be ordered according to their meaning if available to facilitate some of the modeling algorithms.

Categorical variables are based on some geological characteristic of the rock that is large scale. We normally do not consider more than 5 to 7 at one time - perhaps we consider categories in a hierarchical manner where there less than 5-7 at any one scale. The continuous variables within the categories should have different statistics (univariate or spatial). The idea is to have reasonable consistency within each category and have the variation between the categories. The categories must also be spatially consistent/coherent with some degree of predictability. Finally, there must be enough data within each category for inference of the required statistical parameters.

Anything we know should not be left to chance. The referred order of techniques starts with deterministic mapping (digitized surfaces or solid models) where our geological understanding is explicitly reproduced in the model. Given uncertainty and the demands of explicit digitization, it is common to consider volume or distance functions for boundaries.

Process based models mimicking geological processes are rarely applied in mining, but have a place in sedimentary settings with wide data spacing. Object based models simulating geological features are also rarely applied in mining, but have a place when the geological units form clearly defined shapes. Cell based geostatistical models such as indicator simulation, truncated Gaussian or multiple point statistics are often used when the scale of the geological features is smaller than the data spacing.

Indicators were made for categories - lending them numerical meaning. In the context of $k = 1, \ldots, K$ categories the indicator transform is defined as:

$$i(\mathbf{u}_\alpha; k) = \begin{cases} 1, & \text{if } z(\mathbf{u}_\alpha) = k \\ 0, & \text{otherwise} \end{cases} \quad , \quad k = 1, \ldots, K$$

Indicators have a numerical meaning and we can calculate statistics. The expected value is the mean or proportion:

$$E\{I(\mathbf{u}; k)\} = proportion of category k = p_k$$

We would use declustering for to calculate the proportions in practice. The

variance of an indicator is a simple function of the proportion.

$$E\{[I(\mathbf{u};k) - p_k]^2\} = p_k - p_k^2 = p_k(1 - p_k)$$

The variance of an indicator is useful, but entropy is a natural measure of uncertainty. The entropy is a maximum when $p_k = 1/K$:

$$H = -\sum_{k=1}^{K} p_k \ \ lnp_k \qquad H_{max} = -ln(p_k) = ln(K)$$

Each indicator variable could be considered in variography. The variogram of each indicator provides a measure of continuity for that category. The definition of indicator variograms:

$$2\gamma(\mathbf{h}) = E\{[I(\mathbf{u};k) - I(\mathbf{u}+\mathbf{h};k)]^2\} \quad k = 1, \ldots, K$$

Indicator variograms are related and cannot be modeled independently. Conider two categories - the indicator variograms must be the same!

Another feature of indicator variogams is that they must increase linearly. The number of transitions for the second lag is double the number for the first - provided the intervals are more than two units thick, which should be the case.

Indicator kriging [36] could be considered at an unsampled location. Often, ordinary kriging is considered with a reasonably large search - to provide stable estimates.

$$i^*(\mathbf{u}_\square;k) - p_k(\mathbf{u}_\square;k) = \sum_{\alpha=1}^{n} [i(\mathbf{u}_\alpha;k) - p_k(\mathbf{u}_\alpha;k)] \quad k = 1, \ldots, K$$

Cokriging does not work well because fitting an LMC is impossible. A locally varying mean or trend model is useful. A large and consistent search is a good idea. There is no smoothing in the estimation of an indicator; the probabilities capture the uncertainty.

The indicator kriging estimates may not satisy order relations, that is, they may be less than zero and above one (due to negative kriging weights) snd they may not sum to one due to different variograms. Standard practice is to reset negative estimates to zero, estimates above one to one and to restandarize the resultant probabilities. This works and no alternative has proved better.

We should resist the idea of taking the most probable category since that would lead to the majority category getting even more. A deterministic model could start by populating the highest probability of the smallest proportion category first (until the correct proportion), then proceed to the majority category. Better to use simulation.

Indicator kriging can be embedded in a sequential simulation framework leading to sequential indicator simulation (Alabert and [16] (SIS). SIS has been extensively used for categorical variables. The following lecture on HTPG will supercede SIS in many cases, but the historical and practical importance is significant.

The results of SIS tend to be noisy with short scale variability that is unrealistic. Image cleaning is often applied (Schnetzler and [17]). The maximum a posterior selection (MAPS) algorithm visits each grid node and replaces the current assignment with (perhaps) an improved one based on the surrounding values.

4.3.3 Hierarchical Truncated PluriGaussian

Categorical variables are usually modeled first and utilized to define stationary domains for the modeling of the continuous variables. Truncated pluri-Gaussian simulation (TPGS) is flexible. The utilization of underlying Gaussian latent variables for the simulation of categories allows for the use of the well established Gaussian simulation techniques. The truncation rules utilized to map the continuous variables to the categorical variable allow the introduction of geological constraints. The practical application of TPGS is often limited to the utilization of no more than three Gaussian latent variables. This is mostly attributed to the current practice on the definition of truncation rules using truncation masks. This limitation is addressed by the hierarchical truncated pluri-Gaussian (HTPG) technique. HTPG utilizes a tree structure for the truncation of the Gaussian latent variables facilitating its definition based on geological expertise. The developed methodology allows for an arbitrary number of latent variables to model an arbitrary number of categories. As a result, the developed method better explores the potential of the truncated Gaussian method.

Details of HTPG are best understood by reviewing CCG Guidebook 23 that accompanies the lecture notes for the class. This 100 page Guidebook reviews the background, steps through the theory and provides a full worked example. A high level overview is presented here.

There are five essential steps for the application of truncated Gaussian methods: (1) definition of a truncation rule; (2) mapping of spatial continuity from categorical to continuous space; (3) imputation of continuous data subject to categorical observations; (4) simulation of the continuous variables at modeling nodes; and (5) truncation of the simulated models to generate categorical realizations. The first step is perhaps the most important.

Figure 4.8 illustrates the key concept behind HTPG. A hierarchical truncation

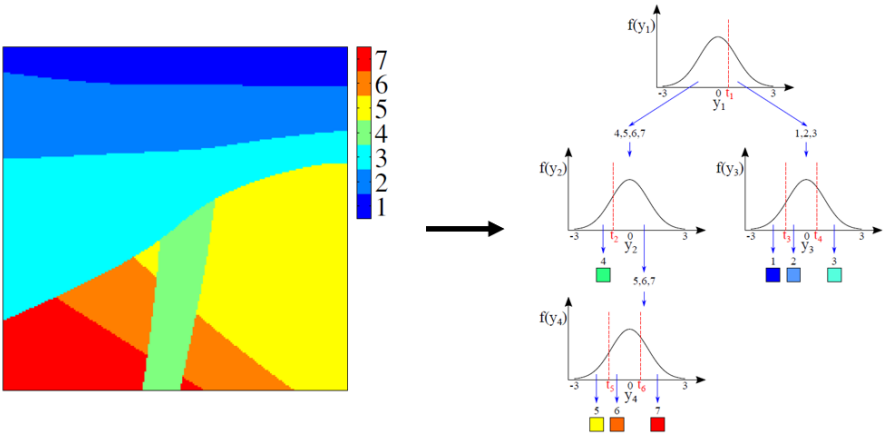


Figure 4.8: A schematic cross section showing seven categories and a corresponding HTPG truncation tree for categorical variable simulation.

tree is setup to reflect the deemed relevant geological relationships. This may be done by consideration of the geological setting. Considering transition probabilities, multidimensional scaling and other statistical tools could help. If there are no compelling geological evidence, then a tree could be setup where one category is separated from all others - one at a time. A total of $K-1$ Gaussian variables would be required. In practice, without detailed geology, small proportion and unusual categories are placed at the top of the tree.

Another consideration with all categorical variable modeling techniques is the use of a trend model in the proportions. Some details of trend modeling were discussed *yesterday* in Section 4.2.3. The trend model for the proportion of each category should be constructed in a consistent manner.

Variograms of the latent Gaussian variables are required for HTPG. These are inferred given the truncation rule, trend model and the variograms of the indicator residuals (residuals from the trend model). This is an inverse problem solved by Monte Carlo Simulation (see the Guidebook). The variograms of the latent Gaussian variables almost always appear more spatially correlated than the indicator variograms; recall that indicator variograms must increase linearly whereas the variograms of the Gaussian variables do not have this constraint.

Simulation of the latent Gaussian variables required conditioning data values. We have the result of the truncation process - the categories, but we do not have the underlying latent Gaussian variables. This is a missing data problem solved by data imputation. The missing Gaussian deviates are imputed to

reproduce the categorical data, the histograms and the variograms. This is a tricky problem solved by an iterative Gibbs Sampler. See the explanation and software covered in the Guidebook.

The remaining practical steps of HTPG are quite straightforward. The simulation of the latent variables is accomplished with SGS or with an unconditional algorithm followed by conditioning. The simulated Gaussian values are truncated and categorical variable realizations are constructed. They must be checked and validated to the greatest extent possible. A k-fold validation scheme should be implemented and the accuracy and precision of the predictions should be checked.

4.3.4 Exercise W4-3

This exercise will introduce some categorical variable modeling techniques using a large indicator data set. Indicator kriging and sequential indicator simulation will be used. Consider the indicator variable in sic.dat.

1. Plot a location map of the data and run cell declustering with a reasonable cell size: 15000 m is likely reasonable. What are the declustered proportions? What is the declustered variance?
2. Calculate and model directional indicator variograms. As there are only 2 categories (0 and 1), the variograms for each category will be identical. Pick a reasonable azimuth for the major direction based on the location map.
3. Use indicator kriging with the declustered proportions and modeled variograms to estimate proportions over the domain and plot the results. Do they match the data?
4. Create 100 realizations of the domain using sequential indicator simulation. Post-process the realizations to calculate the e-type mean and entropy. Plot the first 2 realizations, the e-type mean and entropy. Comment on your results and compare the e-type mean with the indicator kriging maps.

4.4 Setup and Post Processing

4.4.1 Indicator Kriging and Uniform Conditioning

Multiple indicator kriging (MIK) has the promise of estimating local conditional distributions of uncertainty without a strong dependence on the multivariate Gaussian distribution [36]. The Lesson on MIK is a nice summary

[8].

1. Choosing thresholds is the first step of MIK. The number of thresholds is normally chosen between 7 to 15; considering too many may induce more order relation problems and too few thresholds would result in low resolution of the predicted distributions. Some criteria for choosing the thresholds include: (1) commonly start with nine thresholds defining deciles of the global distribution, (2) Move the thresholds to correspond to interesting inflection points on the cumulative distribution function, (3) Remove some low thresholds depending on the cutoff grade and add some high thresholds to define approximately equal quantity of metal in the upper classes, (4) Move the thresholds so one matches the specified cutoff grade, and (5) Intervals should have enough data for a robust estimation; perhaps a minimum of 5% of the data.
2. Indicator variograms are required for every threshold. These variograms are often well behaved as the data are only 0s and 1s. Nevertheless, carefully choosing orientations and variogram parameters is still necessary to acquire stable variograms. The indicator variograms should be standardized to make it easier to analyze, compare and model the indicator variograms. Any licit variogram structure that rises linearly at the origin such as exponential and spherical can be used for modeling the variograms. It is convenient to use the same type for all indicator variograms to ensure consistent changes. The indicator variograms, after all, relate to the same underlying continuous variable. Transitions between the indicator variograms are expected to be smooth in terms of variance, anisotropy, and ranges. To check these transitions from one indicator to the next, the ranges, nugget effects and anisotropy can be plotted versus the threshold number. There should be no abrupt discontinuity between variograms of consecutive thresholds.
3. Indicator kriging (perhaps Markov-Bayes), order relations correction
4. Post processing

Uniform conditioning is another technique to predict the uncertainty in SMU grades given local panel estimate conditions.

1. Consider four scales (see Figure 4.9)
2. Basic data scale statistics
3. Ordinary kriging at the panel scale
4. Calculate the distribution of SMU grades within each panel

Localization will be mentioned here.

1. Distribution of SMU grades within each panel - could come from UC, MIK, MG or simulation
2. Choose SMU grades - as many as there are SMUs within the panel
3. Assign values based on a localizing variable

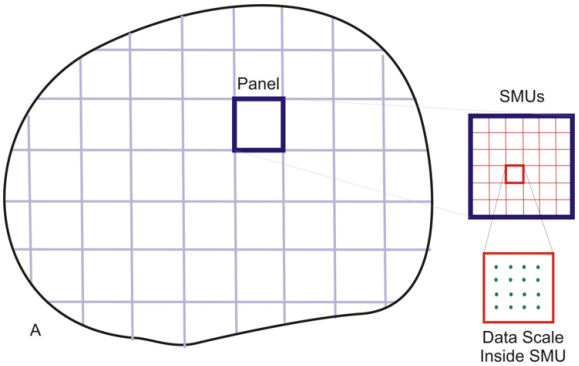


Figure 4.9: Illustration of the four scales involved in uniform conditioning: the data, the SMU, panel and domain.

The advantage of localization is a unique single model that respects the SMU scale distribution. Local accuracy is compromised, but is reasonable at the larger panel scale.

4.4.2 Post Processing

Setup of model - parameter uncertainty, data uncertainty, simulate realizations, process through a transfer function, report uncertainty and sensitivity.

Local uncertainty, block averaging (upscaling), data spacing (Lesson [50]) and localization

Take expected value as late as possible. Calculate on all realizations, then take the average as late as possible.

Grade control - calculate profit for all possibilities and all realizations, then take the decision that maximizes the expected profit. The resolution of the prediction should be at a resolution of one quarter of the data spacing to provide the best possible boundary.

4.4.3 Model Setup

Post processing is straightforward with estimation or deterministic models. Categories representing the geology is modeled first, then the grades are modeled within categories. The estimates are clipped by maximum drill hole spacing and the results are validated and checked to the greatest extent possible.

Post processing simulated realizations is more challenging given the multiple models.

A high resolution data spacing model is important to clip estimates/simulated values and to understand how uncertainty depends on data spacing. In 2-D a constant n approach is used - the area within:

$$dhs = \sqrt{\frac{\pi r^2}{n}}$$

where r is the radius half way between the distance that includes n and $n+1$. Often, a value of n between 10 to 20 is considered. In cases of irregular 3-D spacing of data, a constant volume approach is often considered. A large volume oriented with the overall anisotropy is considered, the number of data within that volume are counted, the equivalent square spacing relative to the plane of continuity is computed by calibration.

Simulated realizations are often scaled up to nominal production volumes of months, quarters or years. Then, the uncertainty is calibrated to the drill hole spacing. Uncertainty is often summarized as the probability to be within 15% of predicted ($P \in \pm 15$). Uncertainty is not simply related to the data spacing. Other modifying factors include (1) the geological complexity - perhaps summarized by H, (2) the proportional effect, (3) the distance to the closest drill hole, and other factors. A response surface model (RSM) could be built to quantify how each factor is contributing.

Uncertainty of large volumes can be artificially small if parameter uncertainty is not considered. Stochastic variations between high and low values average out and large scale values converge to a similar value. Parameter uncertainty with the multivariate spatial bootstrap [61] is essential to accurately reflect large scale uncertainty. Each realization being simulated considers different input distributions.

The drill hole spacing and uncertainty response curve is considered carefully to understand when there is little response and when uncertainty is responding quickly to more data. Figure 4.10 shows the learning curve concept of how uncertainty responds to more data.

Classification decisions must always depend on the qualified or competent person. Uncertainty such a "quarterly production volumes must be within 15% of predicted 90% of the time or more." Sometimes drill holes have to be removed to understand uncertainty at a larger spacing. Sometimes we must drill a simulated realization to understand how uncertainty responds to drill hole spacings that have not been seen.

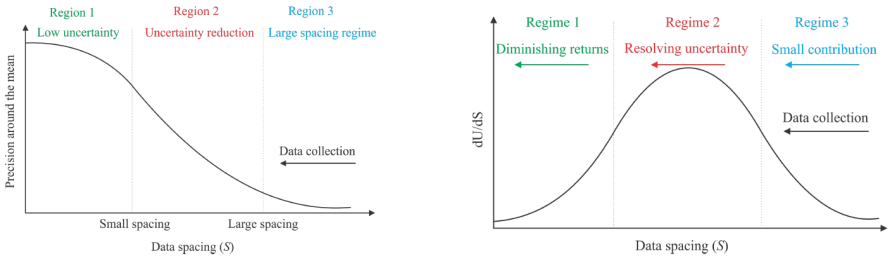


Figure 4.10: Illustration of how uncertainty often responds to more data. Consider moving from large data spacing to close data spacing (right to left in the figure above) - the uncertainty responds slowly, then more quickly as spatial features are resolved, then more slowly again as higher resolution details are slowly learned.

4.4.4 Exercise W4-4

The objectives of this exercise are to learn how to post process a set of multiple realizations. Data spacing, localization and other topics will be summarized.

1. Consider the multiple realizations of Au for the Misima data (earlier exercise). Block average to a 20x20x10m scale. Visualize the results and calculate the expected grade tonnage curve.
2. Calculate the local variance and local coefficient of variation. Visualize and compare to the local data spacing.
3. Consider the multiple realizations of Au for the Misima data (earlier exercise). Block average to a 40x40x10m scale. Visualize the results and calculate the expected grade tonnage curve. Compare the results to Question 1.
4. Localize the block averaged results to a 20x20x10m scale with a panel scale of 100x100x20m scale. Show the results and comment.

Appendix A

Additional Information

Software Setup

Any software could be used. There are commercial software that performs the requested calculations. The GSLIB software, updated versions and additional software from the Centre for Computational Geostatistics (CCG) are provided for all participants in the Citation program. The preferred approach to run all computer exercises (all but three) is to run the entire exercise in a Jupyter notebook, then export the notebook as an html file and submit that as the solution. There are many online resources for this and the wealth of information online cannot be repeated here. Some tips are provided.

The first step to getting setup for running the Citation exercises is to ensure that you have a clean installation of Anaconda. Depending on your comfort with this platform, you may have to remove old versions or setup a new environment. You will also have to have a version of GIT installed on your computer (https://git-scm.com/) since this is used by pygeostat to access the executables. If necessary, you could copy the executables to the correct location and avoid the use of git. You should bookmark (http://www.ccgalberta.com/pygeostat/welcome.html) which is the pygeostat documentation since this will be referred to extensively.

Many notebooks that illustrate different calculations are provided with the course material. These should be reviewed. There are many ways to launch a notebook. The simplest way is to open the Anaconda Prompt and install pygeostat: pip install pygeostat. The small batch file `CitNotebook.bat` could be copied to your installation directory (C:/Users/Clayton/anaconda3 on my computer) and edit appropriately. This is to open a jupyter notebook in the

right place.

An important step in the Citation is to review the introduction and learning notebooks that have been provided: `Py1Basics.ipynb`, `Py2GSLIB.ipynb` and `Py3Pygeostat.ipynb`. For those that still use legacy GSLIB code for certain plotting functions - review the notebook and example in the `Postscript` directory.

Support

There is none. Questions can be sent to the instructor of the class and any designated teaching assistants; however, everyone will have to struggle through learning how to bend the computer to your will.

Plots and Colors

The color scales and choices available at (https://colorbrewer2.org) should be considered. Colors should also be standardized across a project and, perhaps, an entire organization looking the same commodity.

Bibliography

[1] A. S. Almeida and A. G. Journel. Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology*, 26:565–588, 1994.

[2] O. Babak and C. V. Deutsch. An intrinsic model of coregionalization that solves variance inflation in collocated cokriging. *Computers & Geosciences*, 2007.

[3] J. Bai and C. V. Deutsch. The pairwise relative variogram. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/pairwiserelative.

[4] R. M. Barnett. Principal component analysis. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/principalcomponentanalysis.html.

[5] R. M. Barnett. Projection pursuit multivariate transform. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/lineardecorrelation.html.

[6] R. M. Barnett. Sphereing and min/max autocorrelation factors. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/errorellipses.

[7] F. A. Camana and C. V. Deutsch. The nugget effect. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2019. Retrieved from http://www.geostatisticslessons.com/lessons/nuggeteffect.

[8] D. Carvalho and C. V. Deutsch. An overview of multiple indicator kriging. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/mik.

[9] J. P. Chilès and P. Delfiner. *Geostatistics: modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics, 2nd edition, April 2012.

[10] A. P. Chiquini and C. V. Deutsch. A simulation approach to calibrate outlier capping. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/simulationcapping.

[11] M. David. *Geostatistical Ore Reserve Estimation*. Elsevier, Amsterdam, 1977.

[12] C. Deutsch. Calculating effective absolute permeability in sandstone/shale sequences. *SPE Formation Evaluation*, pages 343–348, September 1989.

[13] C. Deutsch and A. Journel. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, New York, 2nd edition, 1995. In Press.

[14] C. V. Deutsch. Kriging in a finite domain. *Mathematical Geology*, 25(1):41–52, 1993.

[15] C. V. Deutsch. Direct assessment of local accuracy and precision. In E. Y. Baafi and N. A. Schofield, editors, *Fifth International Geostatistics Congress*, pages 115–125, Wollongong, Australia, Sept. 1996.

[16] C. V. Deutsch. Cleaning categorical variable (lithofacies) realizations with maximum a-posteriori selection. *Computers & Geosciences*, 24(6):551–562, 1998.

[17] C. V. Deutsch. A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. *Computers & Geosciences*, 32(10):1669–1681, 2006.

[18] C. V. Deutsch. Cell declustering parameter selection. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/celldeclustering.

[19] C. V. Deutsch. Checking simulated realizations - mining. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/checkingmin.

[20] C. V. Deutsch and J. L. Deutsch. Introduction to choosing a kriging plan. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/introkrigingplan.

[21] C. V. Deutsch and J. L. Deutsch. An application of Bayes theorem to geostatistical mapping. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2018. Retrieved from http://www.geostatisticslessons.com/lessons/bayesmapping.

[22] C. V. Deutsch and P. Kumara. Transforming a variogram of normal scores to original units. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/convertnsvariograms.

[23] J. Deutsch, J. Szymanski, and C. Deutsch. Checks and measures of performance for kriging estimates. *The Journal of the Southern African Institute of Mining and Metallurgy (SAIMM)*, 114:223–230, Mar. 2014.

[24] J. L. Deutsch. Experimental variogram tolerance parameters. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/variogramparameters.

[25] J. L. Deutsch. Variogram calculation for tabular deposits. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/tabularvariograms.

[26] M. V. Deutsch. The angle specification for gslib software. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/errorellipses.

[27] O. Erten and C. V. Deutsch. Combination of multivariate gaussian distributions through error ellipses. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/errorellipses.

[28] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997. In Press.

[29] E. Gringarten and C. V. Deutsch. Methodology for variogram interpretation and modeling for improved petroleum reservoir characterization. In *1999 SPE Annual Technical Conference and Exhibition Formation Evaluation and Reservoir Geology*, Houston, TX, Oct. 1999. Society of Petroleum Engineers. SPE Paper # 56654.

[30] Z. Hadavand and C. V. Deutsch. Conditioning by kriging. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/conditioningbykriging.

[31] J. Hammersley and D. Handscomb. *Monte Carlo Methods*. John Wiley & Sons, New York, 1964.

[32] B. E. Harding and C. V. Deutsch. Change of support and the volume variance relation. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2019. Retrieved from http://www.geostatisticslessons.com/lessons/changeofsupport.

[33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[34] E. Isaaks and R. Srivastava. Spatial continuity measures for probabilistic and deterministic geostatistics. *Math Geology*, 20(4):313–41, 1988.

[35] E. Isaaks and R. Srivastava. *An Introduction to Applied Geostatistics.* Oxford University Press, New York, 1989.

[36] A. Journel. Non-parametric estimation of spatial distributions. *Math Geology*, 15(3):445–68, 1983.

[37] A. Journel. *Fundamentals of Geostatistics in Five Lessons.* Volume 8 Short Course in Geology. American Geophysical Union, Washington, D.C., 1989.

[38] A. Journel and C. J. Huijbregts. *Mining Geostatistics.* Academic Press, New York, 1978.

[39] A. G. Journel. Markov models for cross-covariances. *Mathematical Geology*, 31(8):955–964, 1999.

[40] A. G. Journel. Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. *Mathematical Geology*, 34(5), 2002.

[41] D. G. Krige. A statistical approach to some mine valuations and allied problems at the witwatersrand. Master's thesis, University of Witwatersrand, South Africa, 1951.

[42] O. Leuangthong and C. V. Deutsch. Stepwise conditional transformation for simulation of multiple variables. *Mathematical geology*, 35(2):155–173, 2003.

[43] D. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, New York, 1969.

[44] S. A. Mancell and C. V. Deutsch. Multidimensional scaling. In J. L. Deutsch, editor, *Geostatistics Lessons.* 2019. Retrieved from http://www.geostatisticslessons.com/lessons/mds.

[45] G. Matheron. Traité de géostatistique appliquée. Vol. 1 (1962), Vol. 2 (1963), ed. Technip, Paris, 1962.

[46] H. Parker. The volume-variance relationship: a useful tool for mine planning. In P. Mousset-Jones, editor, *Geostatistics*, pages 61–91, New York, 1980. McGraw Hill.

[47] H. Parker. Statistical treatment of outlier data in epithermal gold deposit reserve estimation. *Math Geology*, 23(2):175–199, 1991.

[48] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[49] F. L. Pereira and C. V. Deutsch. Permanence of ratios. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/permanenceofratios.

[50] F. A. C. Pinto and C. V. Deutsch. Calculation of high resolution data spacing models. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2017. Retrieved from http://www.geostatisticslessons.com/lessons/dataspacing.

[51] M. Pyrcz and C. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, New York, 2nd edition, 2014.

[52] M. J. Pyrcz and C. V. Deutsch. Transforming data to a gaussian distribution. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2018. Retrieved from http://www.geostatisticslessons.com/lessons/normalscore.

[53] M. Samson and C. V. Deutsch. Collocated cokriging. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/collocatedcokriging.

[54] L. E. Shmaryan and A. G. Journel. Two Markov models and their application. *Mathematical Geology*, 31(8):965–988, 1999.

[55] D. A. Silva. Signed distance function modeling with multiple categories. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2015. Retrieved from http://www.geostatisticslessons.com/lessons/signeddistancefunctions.

[56] R. Srivastava and H. Parker. Robust measures of spatial continuity. In M. Armstrong, editor, *Geostatistics*, pages 295–308. Reidel, Dordrecht, Holland, 1988.

[57] C. A. Sunday and C. V. Deutsch. Choosing the discretization level for block property estimation. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/discretization.

[58] P. Switzer. Inference for spatial autocorrelation functions. In G. Verly et al., editors, *Geostatistics for Natural Resources Characterization*, volume 1, pages 127–140. Reidel, Dordrecht, Holland, 1984.

[59] J. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

[60] G. Verly. The multiGaussian approach and its applications to the estimation of local reserves. *Math Geology*, 15(2):259–86, 1983.

[61] J. D. Vincent and C. V. Deutsch. The multivariate spatial bootstrap. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2019. Retrieved from http://www.geostatisticslessons.com/lessons/spatialbootstrap.

[62] H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer-Verlag, Berlin, 1995.

[63] D. Yang and C. V. Deutsch. Aggregating multiple secondary variables into a super secondary variable. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2019. Retrieved from http://www.geostatisticslessons.com/lessons/supersecondary.

[64] H. Zhang, O. Erten, and C. V. Deutsch. Bayesian updating for combining conditional distributions. In J. L. Deutsch, editor, *Geostatistics Lessons*. 2020. Retrieved from http://www.geostatisticslessons.com/lessons/bayesianupdating.