

1. Introduction

1.1. Expectations

Before you start reading this book, we would like to make it clear exactly what you can (and can't) expect from it and what we do (and don't) expect from the reader. This text is based on 28 years of courses taught to mining engineers, geologists, hydrologists, soil scientists, climatologists plus the occasional geographer, pattern recognition expert, meteorologist, statistician and computer scientist. Even, on one occasion, an accountant. Over those years, we have endeavoured to pare away all extraneous mathematics and concentrate on intuitive derivations where possible. Readers interested in rigorous mathematical proofs are urged to stop here and turn to the more theoretically based books (cf. Reference Texts in Bibliography). This book is *not* intended to turn out fully fledged geostatisticians. It is intended for people with problems to be solved which can be assisted by a geostatistical approach.

To read this book and benefit from it you need to be fairly comfortable with basic algebra. That is, with the notion of using symbols as shorthand for longer statements. We have worked hard to bring you a consistent notation throughout the book. Where notation is out of our control, we explain carefully what each symbol stands for and try not to use that symbol for anything else. This is not always possible. For example, *Student* (William Gosset) developed his distribution for the mean of a set of samples and called it the t distribution. Herbert Sichel developed an estimator for the mean of a lognormal distribution and called it (surprise) t .

Calculus — differentiation and integration — is discussed at various points in the text. The reader is not expected to do any calculus (as such) but is expected to know that the differential of x^2 is $2x$. The only other complication is the frequent use of simultaneous equations. We tend not to use matrix algebra in this book but will give the matrix form after explanations have been given in simple algebra. For example, linear regression is easier to understand if developed with algebra, but very simple to implement in spreadsheets or in packages such as MatLabTM if matrices are used.

If we haven't scared you off yet, be reassured by the fact that all the analyses are illustrated with real data sets in full worked answers. If you have the CD, the data sets are included along with software to reproduce the analyses (for the most part). If you are reading the hard copy, the data sets and software can be downloaded from the Web. There are exercises for you to try. Answers are available for you to check your results. Most of these exercises have been collected and used in classes or examinations at Final (Senior) Year and Master's levels.

It is our own fundamental regret that this book cannot contain the jokes, anecdotes and sheer *fun* that we have on the courses. We do advise you, however, to keep your sense of humour and common sense to the fore at all times while reading this book.

1.2. The problem to be solved

Geostatistics — as discussed in this book — was developed in geology and mining. However, the problem which it was developed to tackle is more general than geological applications. This text is intended as a basic introduction to statistical and geostatistical analysis of sample data which possesses a location as well as at least one observed value.

There is often confusion as to the intended objective of geostatistical techniques. We define them here as twofold:

1. to characterise and interpret the behaviour of the existing sample data;
2. to use that interpretation to predict likely values at locations which have not yet been sampled.

To set the scene for the rest of the book, let us imagine that there is a (more or less) continuous phenomenon which covers a study area (or volume).

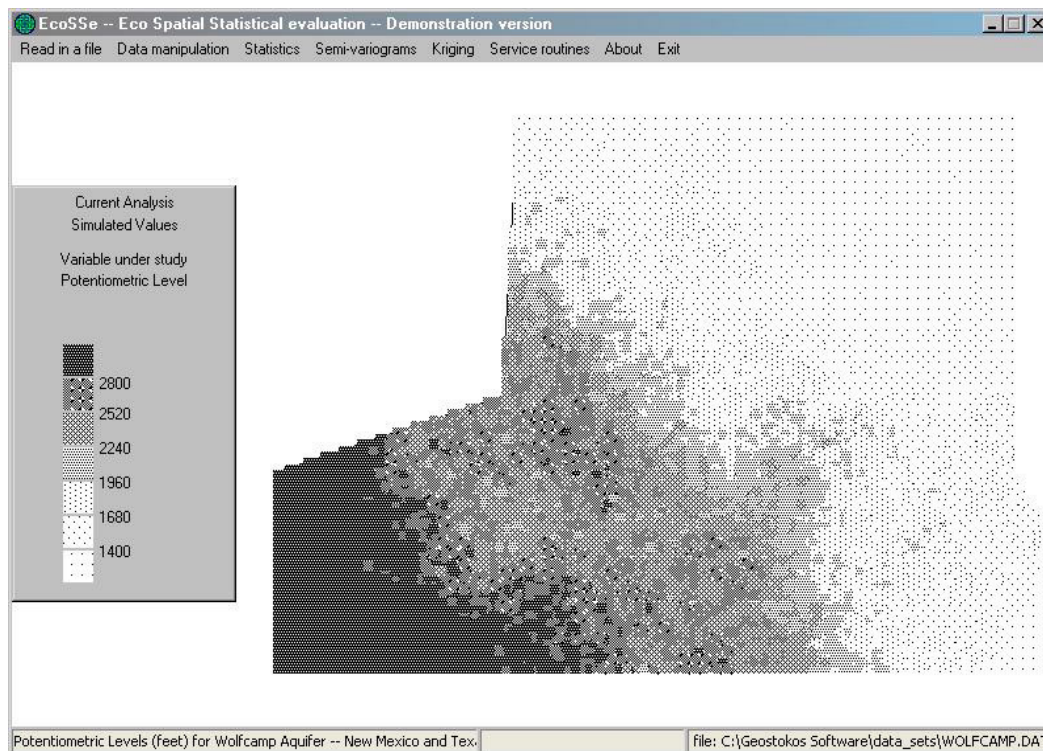


Figure 1.1: *Potentiometric levels over Wolfcamp study area*

Some samples have been taken over the study area and their locations noted. Measurements have been made on the samples taken. Our major task is to estimate the likely value at a location which has not been sampled.

There are many different ways to tackle this problem. This book covers just one approach which is based on a well defined set of assumptions. Other assumptions lead to other methods.

A lot of the criticism which is levelled at geostatistical estimation is founded on misconceptions about the capabilities and intentions of the method (cf. section Sceptics in Bibliography). We will tackle those as we come to them in the text. We will also discuss the shortcomings of the techniques which will be developed as and when appropriate. The

intention of this book is to give the reader an understanding of the statistical and geostatistical techniques which might be useful, not to lay down any laws and regulations on what should and should not be used.

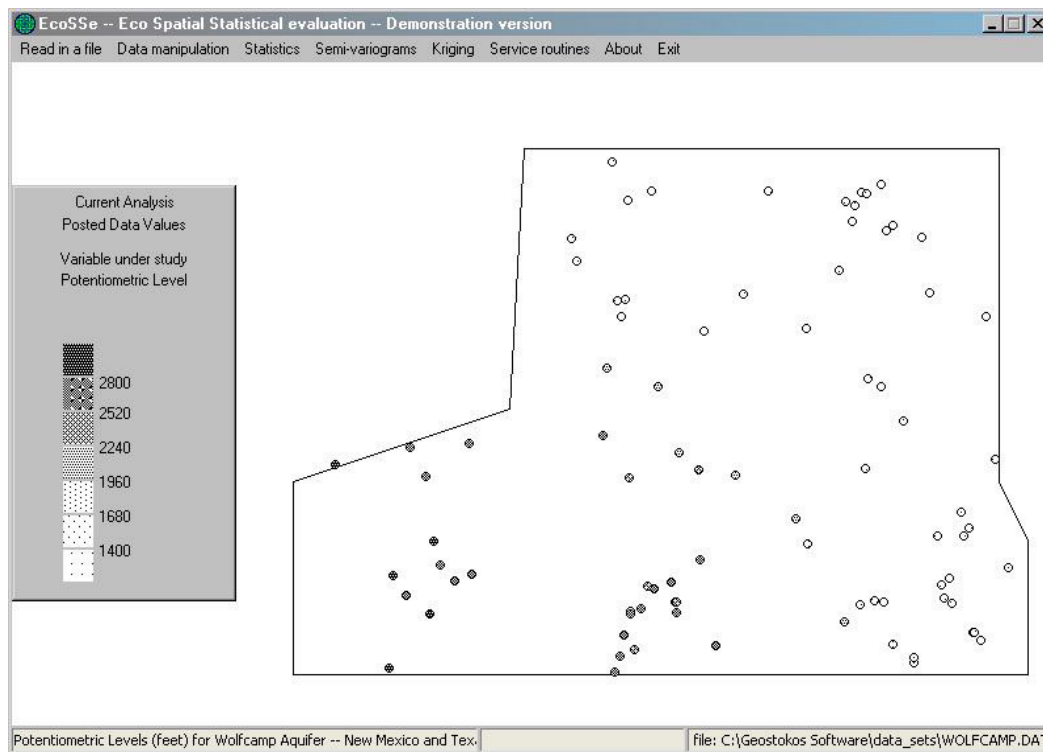


Figure 1.2: *Actual sample data available in Wolfcamp study*

The statistical portions of this book are intended to lay the groundwork for the geostatistical analysis. Much of this material can be found in foundation statistics books but not in the current context. The geostatistical portions of the book assume that you have mastered the statistical techniques described earlier. It is not advisable to ‘skip ahead’ on the assumption that what is being discussed has no relevance to your own interests. The development is extremely linear, in that one section leads into another. There are exceptions to this, of course. For example, if you will never have to deal with *skewed* data, you can skip the chapter on the lognormal distribution and its variants. If you will never deal with more than one measurement per sample, you can skip most of the Relationships chapter. If you never deal with data which has a trend in the values, you can skip all but the first few pages of that chapter.

1.3. Data sets

The sort of applications presented within the book are mainly geological with some hydrology and environmental case studies. The potential applications include any form of measurable spatial data and some which cannot be given a quantitative measure, such as rock type, land use etc. We have included applications of geostatistical techniques in the following fields (so far):

- Coal: a simulated set of data based on a real coal seam in Southern Africa. Boreholes drilled into the coal seam are measured for: thickness of coal (metres), energy content or ‘calorific value’ of coal (Megajoules per tonne); ash content (%) and sulphur content (%). Three co-ordinates in metres are available for the top of the coal seam where intersected by the drillhole.
- GASA: this data set is named for the Geostatistical Association of South Africa and was used in an illustration of geostatistical techniques at a meeting in April 1987 in Johannesburg. The sample data are taken from deep boreholes drilled into a typical Witwatersrand type gold reef. The measurements of interest are the grade of the gold in grams per tonne of rock (parts per million) and the thickness of the reef intersection in the borehole (centimetres). The 27 boreholes lie approximately 1 kilometre apart and constitute a typical data set for the planning and design of a new Wits gold mine. The values have been disguised by a factor but are otherwise unaltered. Co-ordinates are in metres.
- Samples: this data set is based on a Wits type gold mine some decades into production. The samples are chipped from the face of the reef in a working section of the mine (stope). As the face advances, new chip samples are taken. Values within a stope are traditionally estimated using the sample values from the face. This data is totally fictitious except for the locations of the samples, which are taken from a real Wits type gold mine.
- Copper: a simulation based on a stockpile of mined material in the former Soviet Union. Boreholes have been drilled into the dump. The drill core is cut every 5 metres and assayed for copper and cobalt content in percentage by weight. This is the only three dimensional set of tutorial data. Co-ordinates are in metres.
- Geevor: this is sample data from a hydrothermal tin deposit in Cornwall, England. The mineralisation appears as a continuous vein which is sub-vertical. Samples of around 1kg are chipped across the vein, which averages about 24 inches wide. Measurements are grade of tin in pounds of black tin (SnO_2) per ton of rock. The thickness of the vein or ‘lode’ is measured to the nearest inch. Co-ordinates are in feet along section and elevation above an arbitrary base level.

Clark, I., 1979, “Does geostatistics work?”, Proc. 16th APCOM, Thomas J O’Neil, Ed., Society of Mining Engineers of AIME Inc, New York, 213-225.

- Wolfcamp: measurements of water pressure (potentiometric level) in 85 water wells in the Texas panhandle. This data set was part of a study carried out by the Office for Nuclear Waste Isolation in the mid 1980s on a potential site for a high level nuclear waste repository. The Wolfcamp aquifer underlies the planned repository. One aspect of repository planning is to quantify the risks inherent in a breach of the storage facility. Should radionuclides leak into the local aquifers, the scope and speed of potential contamination has to be assessed. The pressure of fluid within the aquifer was one of several variables used to determine the travel path and speed of travel for escaped radionuclides.

Reference: Harper, W.V., and Furr, J.M., 1986. “Geostatistical analysis of potentiometric data in the Wolfcamp Aquifer of the Palo Duro Basin, Texas”, BMI/ONWI-

587, April, Office of Nuclear Waste Isolation, Battelle Memorial institute, Columbus, Ohio.

- Scallops: Scallop data were collected during a 1990 survey cruise off the east coast of North America. Scallop counts were obtained using a dredge. Any scallop smaller than 70 mm was termed a prerecruit. Total catch is the sum of prerecruits and recruits. Measurements included in the data file are:
 - National Marine Fisheries Service (NMFS) 4 digit strata designator in which the sample was taken;
 - sample number per year ranging from 1 to approximately 450;
 - location in terms of latitude and longitude of each sample in the Atlantic Ocean;
 - total number of scallops caught at the sample location;
 - number of scallops whose shell length is smaller than 70 millimeters;
 - number of scallops whose shell length is 70 millimeters or larger.

Reference: Ecker, M.D., and Heltshe, J.F. 1994. "Geostatistical estimates of Scallop Abundance", In, Case Studies in Biometry, Lange et al., editors. Wiley, New York

- Dioxin: A truck transporting dioxin contaminated residues dumped an unknown quantity of these wastes onto a farm road in Missouri. In November, 1983, the U.S. EPA collected samples of the site. In order to reduce the number of samples required, samples were composited along transects. The transects run parallel to the highway, and this direction is designated as the X-direction. The direction perpendicular to the highway is designated as the Y-direction. Data are TCDD concentration (tetrachlorodibenzo-p-dioxin) in micrograms per kilogram ($\mu\text{g}/\text{kg}$). Co-ordinates and transect length are given in feet.

Reference: Zirschy, J.H., and Harris, D.J. 1986. "Geostatistical analysis of hazardous waste site data". Journal of Environmental Engineering, 112:770-784.

- Organics: Data are Soil Organic Matter values (in grams per kilogram) derived from soil samples taken in a research field at the University of Nebraska West Central Research and Extension Center near North Platte, Nebraska, USA. Data were taken as part of experiments on variable-rate fertilizer technology. Co-ordinates are in metres.

Reference: Gotway, C.A. and Hergert, G.W. (1997). "Incorporating Spatial Trends and Anisotropy in Geostatistical Mapping of Soil Properties". Soil Science of America Journal, 61:298-309

- Velvetlf: Subsample of the number of velvetleaf weeds counted in 7 meter² area in a field in Nebraska. Data were collected by Gregg Johnson (see 2nd reference), as part of a research program in weed management at the University of Nebraska.

References: Data set taken from: Gotway, C.A., and Stroup, W.W. 1997. "A generalized linear model approach to spatial data analysis and prediction". Journal of Agricultural, Biological, and Environmental Statistics, 2:157-178.

Data collected by: *Johnsen, G.A., Mortensen, D.A., and Gotway, C.A. 1996. "Spatial and temporal analysis of weed seedling populations using geostatistics". Weed Science, 44:704-710.*

All of the above case studies appear somewhere within the text. The data files are available on the CD and can be downloaded from the Web. All, except samples and possibly copper, are small enough to tackle at desktop and hand calculator level. We strongly recommend that you carry out each analysis by hand at least once to reinforce the written text.

1.4. Software

If you have this book on CD, the disk also contains a 'demo' version of the Geostokos software created specially for teaching. This version has slightly more features than the EcoSse package and rather less than the full Geostokos Toolkit. It is a Windows based package which currently operates under Windows 95/98/2000/ME/XP and NT. Follow the installation instructions supplied with the package. All of the above data sets are supplied on the disk.

If you have this book in hard copy, you may download the software and data sets from the Web. Check your delivery package for current instructions. Full listings of the data sets (except for samples) are given in the Appendix.

The software is identical to the standard Geostokos EcoSse and Toolkit software packages except that it will only read the data files supplied with the book.

Cautionary note on Precision: *It is particularly important to remember that all of the worked answers given in this book have been computed using proprietary software, spreadsheets and/or hand calculators. All of these have different levels of precision in their makeup. Do not worry if your answer is out by anything up to a couple of percent compared to the one in this book. This is particularly so when you have to square, cube or raise numbers to a larger power.*

If you make a real mistake, your answer will be very different from ours. In most cases, mistakes during calculation lead to huge differences in the answers.