Introduction

Statistical packages such as MINITAB, SAS and SPSS have been available for many years and have made an enormous difference to the teaching of Statistics units. Teachers are now able to regularly include large data sets as part of the teaching and learning process. However, such packages are not equipped to deal with spatial data, where location as well as value is important. Such data arise naturally in the earth and environmental sciences. Examples are ore grade of a mineralisation and concentration of a pollutant. Spatial data usually exhibit some degree of spatial dependence and hence most of the standard statistical tests do not apply and the standard statistical software packages have only limited use in the analysis. Over the past thirty years many approaches have been made to the analysis of spatial data and this branch of mathematics is referred to as geostatistics. These methods include the modelling of the spatial dependence together with estimation and simulation algorithms. Modern technology has made possible the application of geostatistical techniques to real-world problems such as ore reserve estimation and contaminated sites rehabilitation. It has also stimulated the development of more sophisticated geostatistical methods.

In this paper we present a selection of these geostatistical methods, together with details of the software used in their implementation. We illustrate the outcomes by application to a permeability data set. We consider a sample data set, model the spatial dependence, and perform both estimation and simulation. We then compare our results with the full data set itself. At Edith Cowan University we have an active geostatistics research group and we teach geostatistics units at both the undergraduate and the postgraduate levels. In practice we use the technology detailed here in both our teaching and our research.

Data Sets

For illustrative purposes we will use the Berea data set which is a real two-dimensional data set arising from an oil and gas deposit in the United States (see cite: chujour and cite: bkc). These data comprise 1600 observations derived from air permeability measurements taken on a slab of Berea sandstone. We shall consider this as reality. As our sample data set we use the b2764 data set which is one of a number of random samples of 64 points taken by us from the full Berea set. Students in our introductory geostatistics unit receive the full Berea set, together with an individual sample set, for use in two of their assignments.

Summary statistics for Berea and b2764, obtained from MINITAB, are given in Figure 1 and Figure 2 respectively.
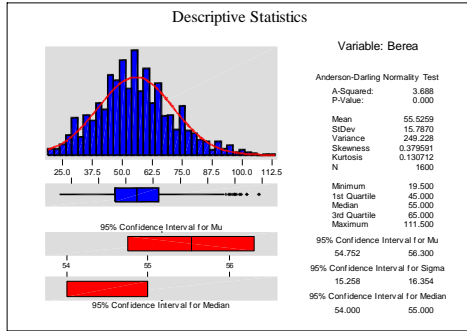
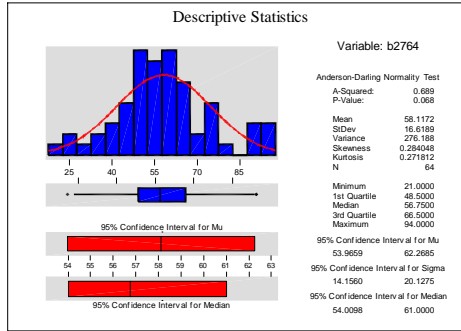Figure1: Berea Summary Statistics

Figure 2: b2764 Summary Statistics

While these summary statistics show that both data sets are essentially normally distributed and give an indication of the key values and values ranges, they do not capture the spatial dimension of the data. This aspect is illustrated in Figure 3 and Figure 4, which have been obtained from the 3PLOT program of GEOSTAT OFFICE cite: gstat .
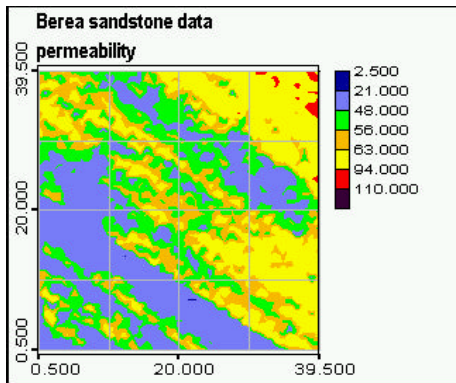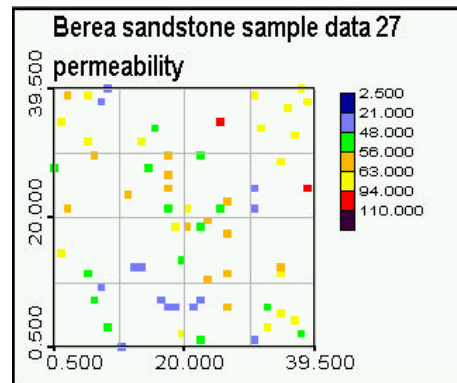

Figure 3: Berea Postplot

Figure 4: b2764 Postplot

Note that a table of summary statistics can be obtained from MINITAB and also from 3PLOT.

Variography

A measure of the spatial dependency is given by the experimental semivariogram $\gamma(\mathbf{h})$ defined by

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})]^2 \qquad \#$$

where $\mathbf{h}$ is the separation vector, $z(\mathbf{u})$ is the value at location $\mathbf{u}$ and $N(\mathbf{h})$ is the number of pairs at separation h. This is usually displayed as a plot of $\gamma(\mathbf{h})$ against . If $\gamma(\mathbf{h})$ varies with the direction of h then the data exhibit anisotropy. If the semivariogram depends only on the separation distance (referred to as the lag) then the data are isotropic. One way to check for anisotropy is to plot a variogram surface. This is a contour plot of semivariogram values by direction. The variogram

surface, for five lags with lag spacing 4, in Figure 5 has been generated by VARIOWIN cite: vario and the resulting values illustrated more effectively in Figure 6 via the contouring properties of 3PLOT.
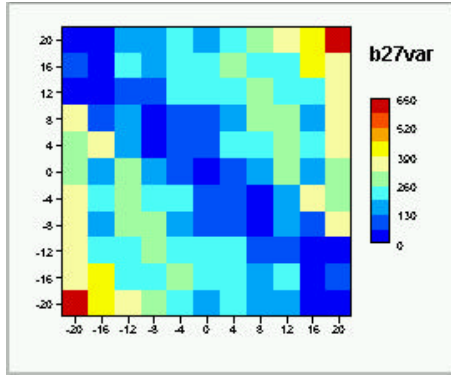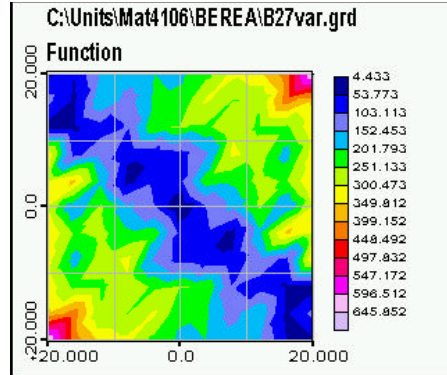

Figure 5: Variogram Surface (VARIOWIN)

Figure 6: Variogram Surface (3PLOT)

In this case the variogram surface clearly indicates anisotropy with the major axis of the anisotropy being in the 135°direction and the minor axis of the anisotropy being in the 45°direction. In VARIOWIN the angle is measured from East, as per the convention in mathematics. However, in most of the geostatistics software the angle is a bearing from North which reflects the practice in geology. The VARIOWIN package can then be used interactively to fit a variogram model based on an optimal fit in the two most important directions. The experimental semivariograms in these directions, together with our chosen variogram model is displayed in Figure 7. The experimental semivariograms were calculated for six lags with lag spacing 4. The model fitted is a linear combination of a nugget effect model with nugget parameter 2 and two exponential models with parameters detailed in Table 1.
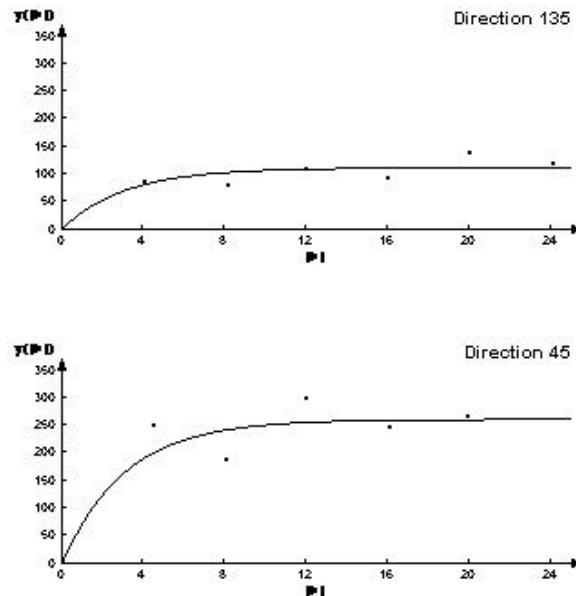
Figure 7: Experimental Semivariograms and Variogram Model

[B] caption

Direction	Partial Sill	Range	Anisotropy Factor
135°	110	9.25	1
45°	150	9.25	1000

[E]

It must be noted that VARIOWIN is a relatively simple package and can handle only two-dimensional data sets. For three-dimensional data sets we use VISOR cite: visor which is a much more powerful and sophisticated variography package. In addition, for some research projects with very large data sets, we use the GEMCOM mining software suite.

Estimation

The estimation method illustrated here is known as Ordinary Kriging (see cite: goov). This algorithm preserves the sample values at data locations and assigns a value at an unsampled location as a weighted average of sample data points. The sample values chosen, and the corresponding weights used in the estimation, vary from point to point and are determined by the spatial continuity as expressed in the variogram model used. To carry out Ordinary Kriging, usually referred to as OK, we use the GSLIB2 software library cite: gslib . This consists of Fortran subroutines and associated parameter files, which can be compiled and used as written for most teaching situations, but which can be customised to a particular research problem. In Figure 8 and Figure 9 we show the results from OK estimation using the routine KT3D, together with the estimation variance map. The smoothing properties of the interpolation can be seen in Figure 8 and that OK is an exact interpolator can be seen from Figure 9 by the zero variances at the sample data locations.
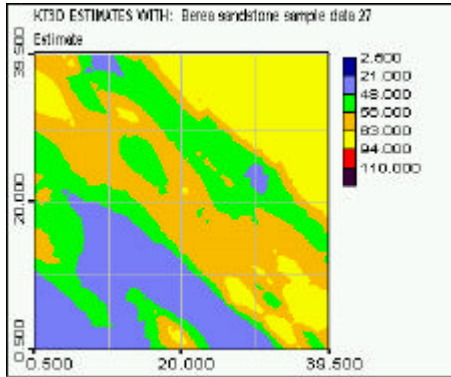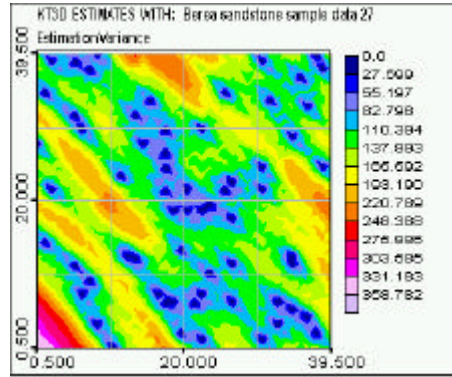
Figure 8: Estimates using OK

Figure 9: Variances from OK

Simulation

One problem with estimation techniques such as OK is that they exhibit conditional bias. This means that they tend to overestimate the lower values and underestimate the higher values. The result of OK is a smooth set, with a lower variance than the actual data set. There are situations where this conditional bias can lead to real difficulties. One example is the case of a contaminated site where the extremes of pollutant concentration can be vital.

Geostatistical simulation methods are used when one wishes to capture the variablity of the actual data set. There are many different algorithms in current use but we have chosen here the method of Sequential Gaussian Simulation (see cite: goov), referred to as SGSIM. Like Ordinary Kriging, the method SGSIM also preserves the values at the sample locations. Again we use GSLIB2 to carry out the simulations. Each implementation of the algorithm gives one realisation of reality. Four such realisations are shown in Figures 10 -13.
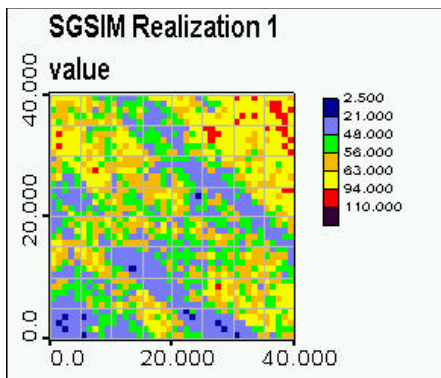

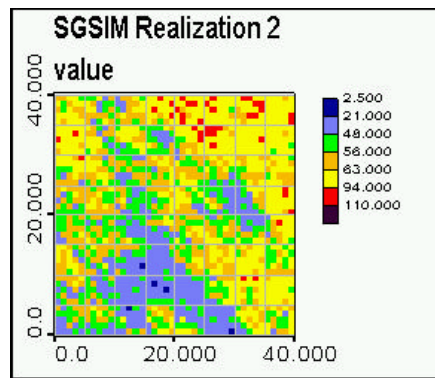Figure 10: Simulation using SGSIM

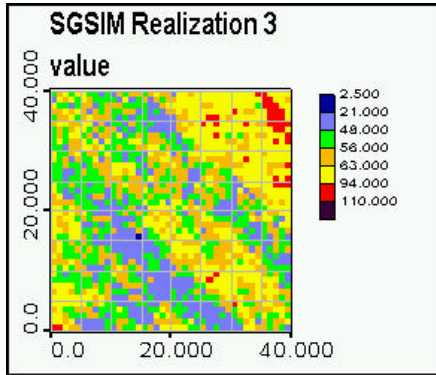

Figure 11: Simulation using SGSIM
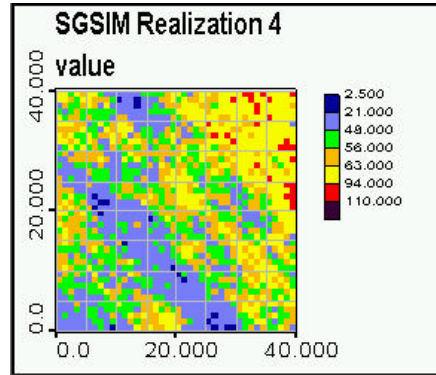
Figure 12: Simulation using SGSIM

Figure 13: Simulation using SGSIM

In practice a large number of such realisations are carried out and common features observed. Alternatively, an average can be found from these realisations. Figure 14 shows an averaging of 10 realisations for the Berea data set. In this case we obtain the same sort of smoothing that we observed in the estimation case. A better use of a number of realisations is to obtain a (spatial) map of the probability of exceeding a given cut-off value or the mean value above a given cut-off value. Figure 15 shows the probability of exceeding the cut-off value of 63.
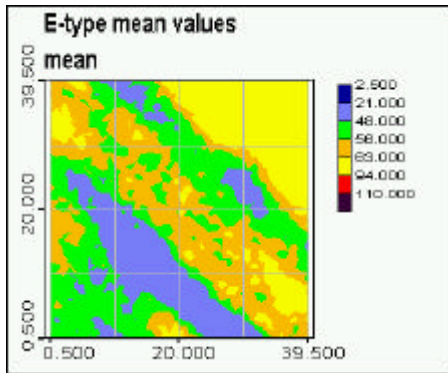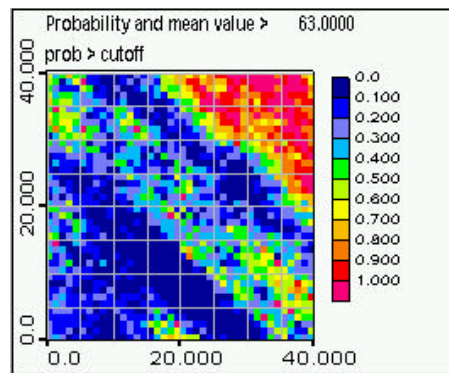

Figure 14: Average of 10 Simulations

Figure 15: Probability from10 Simulations

Conclusion

From their beginnings in the mining and petroleum industries, geostatistical techniques are now expanding rapidly and are being applied in many other areas, including soil science and environmental science. We have indicated here, in a small way, some of the features of typical geostatistical estimation and simulation techniques and the software used to implement these. These algorithms were simply not realistically applicable to large data sets without the advent of modern computing technology.

chujour   J. Chu and A.G. Journel. Conditional fBm simulation with dual kriging. in Dimitrakopoulos, R., (ed.) *Geostatistics For The Next Century*. 407 - 421. Kluwer Academic Publishers, 1994.

bkc   D. Kentwell, L., Bloom. and G. Comber. Geostatistical Conditional Fractional Simulation with Irregularly Spaced Data, in A McDonald and M. McAleer (eds). *MODSIM*97 *International Congress on Modelling and Simulation*, Vol. 3, 1190 - 1195, 1997

gstat   M. Kanevski et al, *GEOSTAT Office Sofware*. Nuclear Safety Institute, Moscow, Russia, 1998.

vario   Y. Pannatier,. *Variowin*: *Software for Spatial Data Analysis in* 2*D*. Springer Verlag, New York, 1996

visor   Snowden Mining Industry Consultants, *VISOR Interactive Variogram Analysis and Modelling Software*, Perth, 1998

goov   P. Goovaerts. *Geostatistics for natural resources estimation*. Oxford University Press, Oxford, 1997.

gslib   C. Deutsch and A. Journel. *GSLIB*: *Geostatistical Software Library and User's Guide*. 2nd ed., Oxford University Press, Oxford, 1998.