Choice and implementation of software for the online teaching of Geostatistics

L. M. Bloom and U. A. Mueller
Edith Cowan University, Perth, Australia
l.bloom@ecu.edu.au, u.mueller@ecu.edu.au

Abstract At Edith Cowan University (Perth, Western Australia) we offer the course *Postgraduate Certificate in Geostatistics*. This consists of three one-semester long units, one of which has also an undergraduate version, which together provide an introduction to spatial descriptive statistics, variography for modelling spatial continuity and a number of geostatistical estimation and simulation techniques. All three units make extensive use of technology. The majority of the students enrolled in the postgraduate course live outside Perth, many working for mining or petroleum companies. Such students are unable to come to on-campus sessions and are also unable to use the laboratory-based computer software packages for which the University has an on-site licence. Even some of the locally based undergraduate students are unable to attend the on-campus sessions due to part-time work commitments. We therefore need to use software that we can reasonably expect the students to have or which is available at low cost or, preferably, as public domain software. In addition, we want geostatistical estimation programs that we can customise, and we also have the desire for the students to use software in a way that is transparent and not simply as a 'black-box'. Further, the packages must be relatively simple to use and, since we make the course materials available online, they must be able to be readily communicated to the students. In this paper we discuss, in context, our specific software requirements, the software choices we have made, the advantages and disadvantages of the implementation of these decisions, and their implications for the future offerings of these units.

1. Introduction

Whenever one offers on-line (or even in distance mode) a Unit that requires the student to carry out calculation and modelling there is the question of which software package to use. The choice will be based on a range of factors, including cost and ease of use, not simply on which package is the 'best' for the task in hand. Often quite sophisticated (and expensive) software packages are available but these will usually require the student to commit to considerable training time and may not be worthwhile in the overall context. For instance, any student employed by a major company may well be expected to work subsequently on a quite different package.

At Edith Cowan University (Perth, Western Australia) we offer the (full fee-paying) course *Postgraduate Certificate in Geostatistics*. This consists of the three one-semester long units Introduction to Geostatistics (MAT5106), Geostatistical Methods (MAT5114) and Modelling and Simulation (MAT5115). The first unit, for which we also offer an undergraduate version (MAT3106), provides an introduction to geostatistical inference and estimation, while the second and third units are in-depth explorations of geostatistical estimation and simulation techniques respectively. All three units make extensive use of technology (see [1] for an overview of some of the outcomes). The majority of the students enrolled in the postgraduate course live outside Perth, many working for mining or petroleum companies. Such students are unable to come to on-campus sessions and are also unable to use the laboratory-based computer software packages for which the University has on-site licences. These students are provided with unit materials on-line and we therefore need to use software that we can reasonably expect the students to either have or to be able to easily obtain at low or no additional cost. In some cases we provide the software on the unit website for downloading by the students and such packages must be able to be readily handled by

the students. We do however make the assumption that students have access to a Windows PC environment.

In this paper we discuss in detail our specific software requirements, the software choices we have made and, in context, the advantages and disadvantages of the implementation of these decisions. Finally, we discuss possible options for the future offering of these units.

In geostatistics we consider the analysis of spatially dependent data, where location as well as value is important. Such data arise naturally in the earth, petroleum and environmental sciences. The spatial dependence means that values at nearby locations are more likely to be similar than values located far apart. Any analysis of such data needs to model this spatial dependence, also known as spatial continuity. The aim of the analysis is the inference of the distribution at locations where the true value is unknown and to provide a measure of the associated uncertainty.

2. Software Needed

In this section we give a description of the tasks for which we require appropriate computer software, together with an indication of the packages chosen and why we have made our particular choices.

2.1 Descriptive analysis of the sample data

In order to adequately describe the sample data, it is necessary to generate sample statistics, and some graphical plots to get an idea of the distribution of the data values. For this purpose we have written the unit material for the first unit (MAT5106) and its undergraduate counterpart (MAT3106) in terms of the MINITAB® Statistical Software (Release 14) package [6]. Output from MINITAB takes the form given in Fig. 2.1 for summary statistics together with a histogram, a boxplot and depicted confidence intervals for the mean and median, and in Fig. 2.2 for a lognormal probability plot.

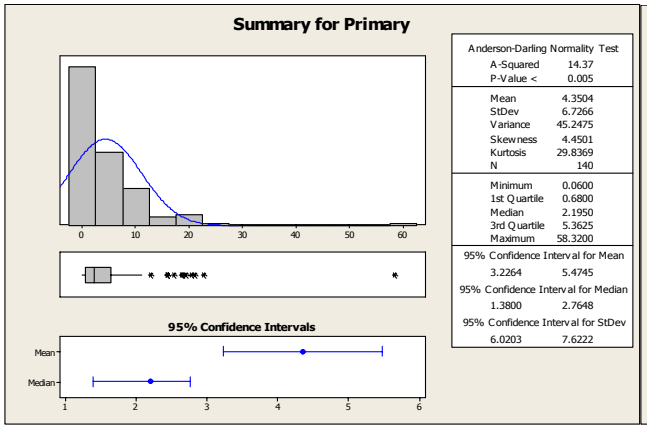

Fig. 2.1 Graphical Summary

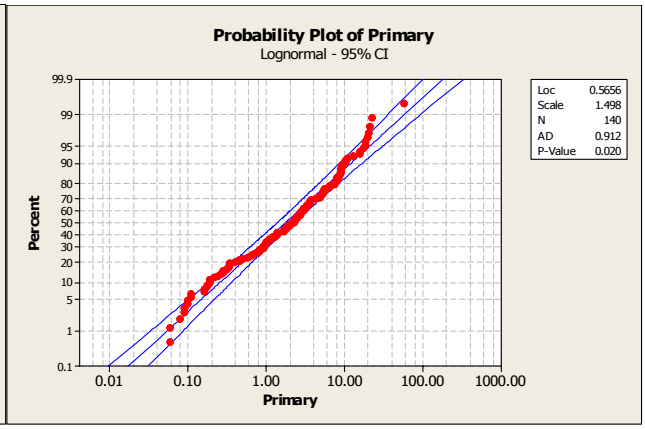

Fig. 2.2 Lognormal Probability Plot

MINITAB is a sophisticated package which is at the same time extremely user-friendly. It is a common choice in teaching statistics and referred to in many standard statistical textbooks. At Edith Cowan University we use it also for our standard second and third year undergraduate statistics units. However, although the University has an on-campus licence for MINITAB, this

does not cover its provision to off-campus students. There is a 30 days time-limited version available by downloading from the website (http://www.minitab.com), or on CD from the supplier, but this is not time enough for the entire course. For this reason we make the assumption that the students have access to the Microsoft Office suite of programs and encourage them to use EXCEL for the basic descriptive statistical analysis.

Since we are dealing with spatial data, we need to obtain also a spatial description of the sample data. We need to ascertain where the high and low values are located, examine for any connectivity in the data and look for evidence of a spatial trend in the data. This is done by means of a Postplot, an example of which is given in Fig. 2.3, where we have used the public domain program 3Plot (3Plot98), which is part of the commercial suite GEOSTAT Office [3]. Such a plot gives both sample locations and a summary of their values. In the example here the legend reflects the deciles of the chosen data set. In addition, 3Plot can be used to obtain (very limited) summary statistics, as indicated in Fig. 2.4.

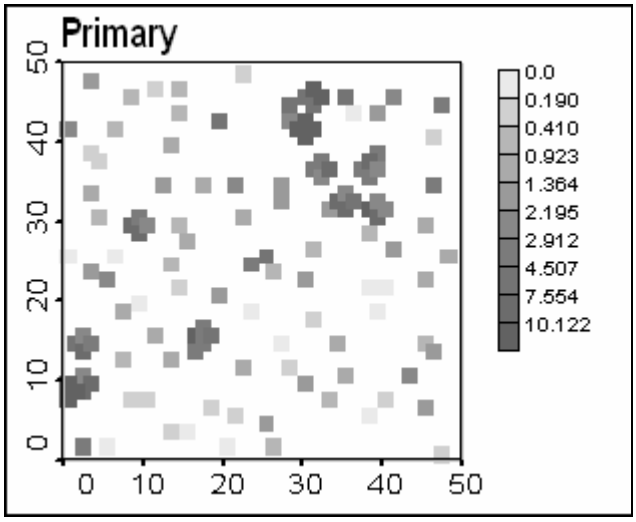

Fig. 2.3 Data Postplot

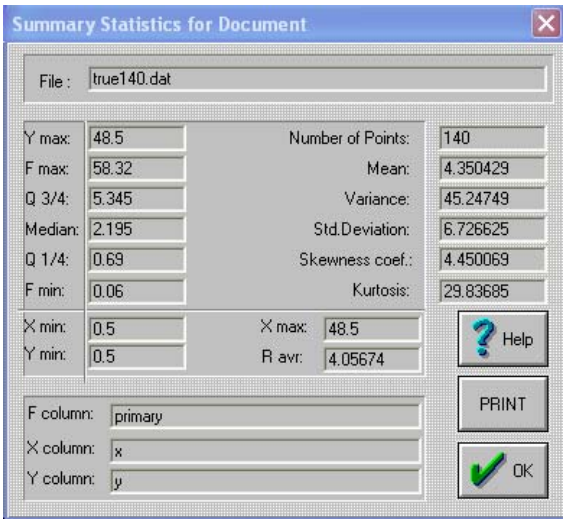

Fig. 2.4 Summary Statistics

Variography

Since we are dealing with spatial data it is essential to model the spatial continuity. This is done by means of an appropriate semivariogram model for a univariate data set (and by means of a co-semivariogram model for a multivariate data set). The experimental semivariogram for attribute (variable of interest) *z* and lag class **h** is defined (see [4] for more detail) by

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

where $\mathbf{h}$ is the separation vector between locations $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha + \mathbf{h}$, $z(\mathbf{u}_\alpha)$ and $z(\mathbf{u}_\alpha + \mathbf{h})$ are the values of the attribute at locations $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha + \mathbf{h}$ respectively and $N(\mathbf{h})$ is the number of pairs at separation $\mathbf{h}$. The semivariogram is then displayed as a plot of $\gamma(\mathbf{h})$ against **h.** Since $\gamma(\mathbf{h})$ is large when the attribute values at locations $\mathbf{u}_\alpha$ and $\mathbf{u}_\alpha + \mathbf{h}$ are different, the semivariogram is actually a

measure of dissimilarity. If $\gamma(\mathbf{h})$ depends only on $|\mathbf{h}|$ then the spatial continuity is the same in all directions and we say the attribute is isotropic. In practice, the semivariogram is modelled by means of a linear combination of standard functions. The most common of these (in isotropic form) are the nugget effect model, the spherical model and the exponential model. These are defined (in isotropic form) as follows:

Nugget Effect: $g(|\mathbf{h}|) = 0$ if $|\mathbf{h}| = 0$ and $g(|\mathbf{h}|) = 1$ otherwise

Spherical: $g(|\mathbf{h}|) = 1.5|\mathbf{h}|/a - 0.5(|\mathbf{h}|/a)^3$ if $|\mathbf{h}| \le a$ and $g(|\mathbf{h}|) = 1$ otherwise

Exponential: $g(|\mathbf{h}|) = 1 - exp(-3|\mathbf{h}|/a)$

where the range parameter *a* defines the distance within which spatial correlation is evident. In the non-isotropic case the experimental semivariograms in the directions of maximum and minimum spatial correlation need to be calculated and jointly modelled.

The package we use to model the experimental semivariogram is VARIOWIN (VARIOWIN2.2) [7]. This software was originally a commercial package but is currently available as (non-maintained) Public Domain software. In fact, the package consists of the three separate modules. These are **Prevar2D**, which calculates the distances between all possible pairs of sample locations and generates a pair comparison file (pcf) file, **Vario2D with PCF**, which uses the pcf file in the calculation of the desired semivariograms and generates a variogram (var) file, and **Model,** which makes use of the var file and is used to model the relevant semivariograms. VARIOWIN has the definite advantage of enabling the student to do the modelling interactively but the disadvantages of being geared only to 2D modelling and not being able to handle large data sets. An example from VARIOWIN for a 2D (univariate) data set is given in Fig. 2.5. This shows an experimental semivariogram, together with an isotropic model in the case where it has been determined that the spatial continuity does not vary with direction. The VARIOWIN model parameters window is also shown in Fig. 2.5, indicating that a nested structure consisting of a linear combination of a nugget effect model and two spherical models has been used. Note here that the model fit at the early lags is most important.

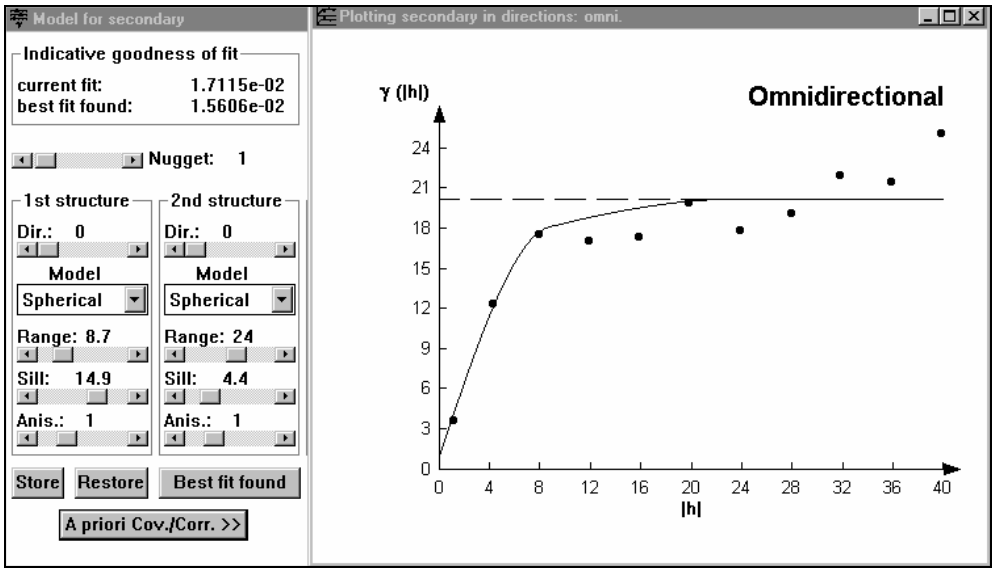

Fig. 2.5 Experimental semivariogram and an isotropic fitted model

We cannot use VARIOWIN for 3D modelling. For this purpose we choose to use the program **gamv.exe** from the package GSLIB (GSLIB2), see [2], which is a Library of Fortran programs. Each program in this library is executed in the MS-DOS Window, or for Windows XP machines in the Command Prompt window, and operates by calling on a user-defined parameter file, which is simply a text file. In fact, **gamv.exe** can be used for generating semivariograms from 1D and 2D data also. However it provides only the experimental semivariogram values for the chosen lags. In order to carry out any semivariogram modelling it is necessary to do so using either trial and error together with the non-interactive GSLIB option **model.exe** or to find some other way to carry out interactive modelling. One possibility is to program EXCEL for this but then some level of programming skills are necessary and not all of our students have this ability or background. Another possibility is to create an input file compatible with **Model** from VARIOWIN. This is the easier approach but requires an understanding of the actual structure of the input file. In fact, this is the advised approach as this skill is utilised elsewhere in the unit when standarised semivariograms are required. We want to standardise the semivariograms by dividing the semivariogram value at each lag by the total variance of the sample data set. This is not the same as the standardised semivariogram provided by VARIOWN where the standardisation is achieved by dividing the semivariogram value at each lag by the variance of the actual data used at that lag. To carry out the procedure we import the var file into EXCEL or MINITAB and replace the relevant column in the var file by our semivariogram values obtained from **gamv.exe**. However, whatever method is chosen, it is rather cumbersome and lends itself to errors.

An example **gamv.exe** parameter file for a 1D (univariate) data set is given in Fig. 2.6. The program can also be used to generate other measures of spatial continuity, but we will not consider these in this paper.

```
Parameters for GAMV
*******************
START OF PARAMETERS:
pH5.dat                          -file with data
1   0   0                       -columns for X, Y, Z coordinates
1   2                           -number of variables,col numbers
0     1.0e21                    -trimming limits
pH5v.out                        -file for variogram output
30                              -number of lags
1.0                             -lag separation distance
0.5                              -lag tolerance
1                                -number of directions
90.0  40.0 10.0   0.0  90.0  50.0    -azm,atol,bandh,dip,dtol,bandv
0                                 -standardize sills? (0=no, 1=yes)
1                                -number of variograms
1   1   1                       -tail var., head var., variogram type

type 1 = traditional semivariogram
2 = traditional cross semivariogram
```

Fig. 2.6 Gamv.exe Parameter File

The corresponding experimental semivariogram and suggested model (dashed curve) are shown in Fig. 2.7. The model shown there is a linear combination of a nugget effect model and an exponential model. In this case the output from **gamv.exe** has simply been imported into MINITAB and trial and error has been used to obtain a suitable model.

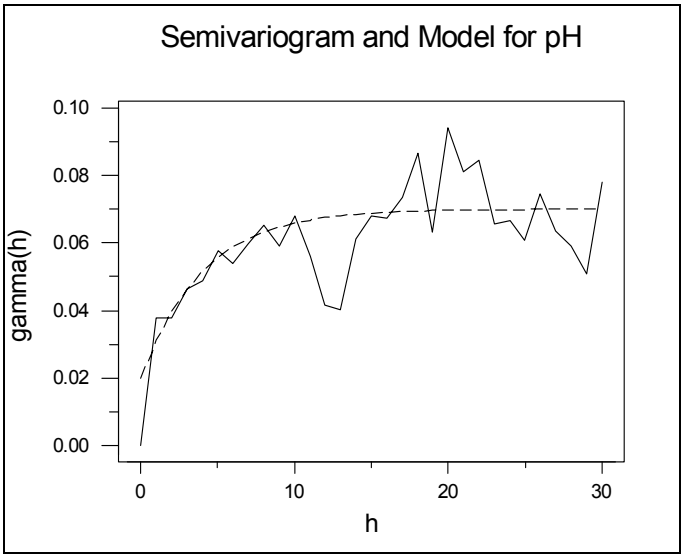

Fig. 2.7 1D experimental semivariogram and model

2.2 Estimation

The estimation method covered in MAT5106 and MAT5114 is Kriging (see [4] for details) which is essentially a linear regression approach. The Kriging Estimator is given by

$$Z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) \left[Z(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha) \right]^2$$

where

$Z^*(\mathbf{u})$ is the estimate of the true (unknown) random variable $Z(\mathbf{u})$

$n(\mathbf{u})$ is the number of sample values to be used in the estimation at location $\mathbf{u}$

$\lambda_\alpha(\mathbf{u})$ is the weight to be assigned to the sample value $z(\mathbf{u}_\alpha)$, considered as a realisation (value) of $Z(\mathbf{u}_\alpha)$.

$m(\mathbf{u}) = E\{Z(\mathbf{u})\}$ and $m(\mathbf{u}_\alpha) = E\{Z(\mathbf{u}_\alpha)\}$ is the mean (expected value) of $Z(\mathbf{u})$ and $Z(\mathbf{u}_\alpha)$ respectively.

The objective is to find an unbiased estimate that minimises the error variance $Var\{Z^*(\mathbf{u}) - Z(\mathbf{u})\}$. This minimum value at location $\mathbf{u}$ is referred to as the Kriging Variance at that location. Kriging is actually a suite of estimation methods and we access different types of kriging depending on the assumptions made about how the mean $m(\mathbf{u})$ varies over the study region. For example, if we assume that $m(\mathbf{u})$ is a known constant over the whole study region we have

Simple Kriging and if we assume that $m(\mathbf{u})$ is locally constant, but unknown, throughout the study region then we have Ordinary Kriging. If we apply a user-defined function for $m(\mathbf{u})$ then we have Kriging with a Trend, previously referred to as Universal Kriging. To carry out kriging (for univariate data) and co-kriging (for multivariate data) we again use GSLIB, together with an appropriate user-generated parameter file. In the units MAT5106 and MAT3106 we use a short 2D version **kb2d.exe,** which has a simplified parameter file and carries out the actual estimation and minimum variance calculation for the Simple Kriging (SK) and Ordinary Kriging (OK) choices. In each case the output consists of estimates and estimation variances on a user-determined estimation grid and the program 3Plot is used to generate a postplot or, in the case of gridded data, a mosaic plot, and MINITAB or EXCEL can be used to generate the summary statistics. An example of a mosaic plot for each of the kriging estimates and the corresponding kriging variances is given in Fig. 2.8 and Fig. 2.9 respectively.

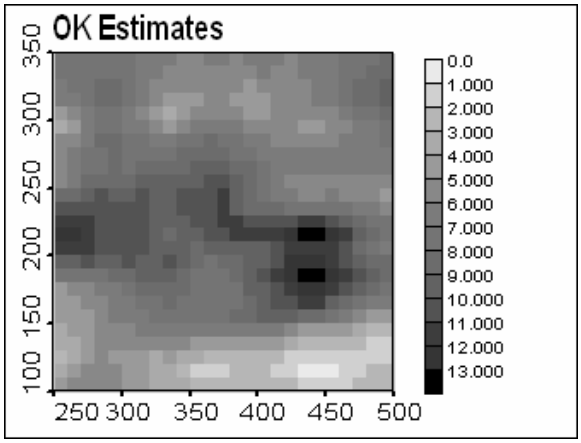

Fig. 2.8 Kriging Estimates

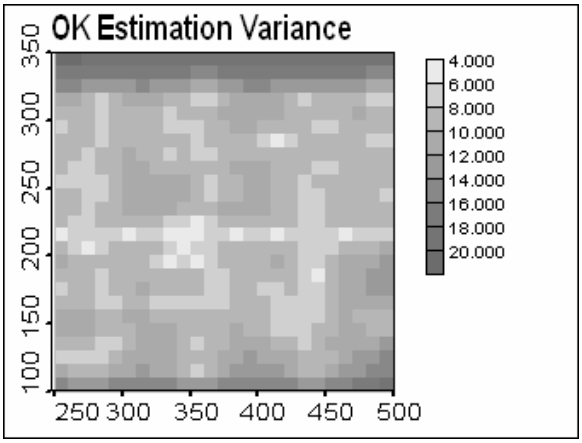

Fig. 2.9 Kriging Variances

In MAT5114 we use the more flexible (and more complicated) GSLIB program **kt3d.exe**, which can be used for 1D, 2D or 3D data. This program can also be used to carry out block (rather than point) kriging as well as Kriging with a Trend and its processing options include both cross-validation and jackknifingto help assess the chosen semivariogram model.

2.3 Simulation

The simulation algorithms are covered in MAT5115, and these all rely on Monte Carlo simulation. The main aim is an exploration of the spatial variability of data distributions compatible with a particular data set and its key statistics; that is, we consider so-called conditional simulation where at sample locations the estimated values of the attribute of interest are equal to the sample values. Several different simulation algorithms are explored. Some simulate the whole set at the same time while others use sequential simulation (as a consequence of Bayes' rule) to simulate one location after another along a randomly determined simulation path. All algorithms are implemented in FORTRAN and come as part of GSLIB. Theoretical discussion is supported with examples in EXCEL workbooks.

The sequential algorithms were introduced as a response to the need to simulate on a very large scale. For attribute of interest z the algorithms in this family all have the same structure, where *(n)* indicates that the simulation is conditioned by *n* sample values. For each realisation

1. Define a random path visiting each location to be simulated exactly once.
2. At each location $\mathbf{u}'$ determine the parameters of the conditional cumulative distribution function $F(\mathbf{u}', z|(n))$. These parameters are derived using Kriging.
3. Draw a random deviate from $F(\mathbf{u}', z|(n))$ and append it to the data set.
4. Move to the next location along the random path and repeat steps 3 and 4.

Loop until all locations are simulated.

Input parameters for this class of algorithms are as follows: data set, variogram models for the variogram of interest, search parameters, search type, kriging type. Irrespective of the simulation algorithm used, the output consists of equiprobable realisations that then require further processing. Such processing includes the computation of summary statistics for each realisation, an approximation of the expected value and variance by location and the evaluation of the algorithm using both local as well as global performance measures. These calculations are carried out using both EXCEL and GSLIB routines.

The various processing options are discussed in the session notes for the sample set and associated simulation output used in the unit. Problems that arise are the adaptation to the specifics of the data set used by the students in the assignments. These include ensuring that descriptive statistics are computed on the whole output and not just part of it. A specific example is block misclassification analysis. Here realisations selected on some performance criterion, for example having the semivariogram closest to the model semivariogram specified in the parameter file in the least squares sense, are compared to the exhaustive data, point by point or block by block. The instances of finding a value above or below the actual value are recorded and subsequently totalled. EXCEL spreadheets are used for this purpose and it is a common mistake for students to fail to adjust formulae so that they match the specific circumstances, by either using the total number of blocks from the template in the calculation of the proportions of misclassified blocks, or failing to adjust the comparison sheet to count all blocks.

3. Student Implementation

We first examine the student interaction with our chosen software packages and some of the problems that arose. These range from the lack of extended access to the package to difficulties with the use of the package itself.

The difficulty with MINITAB is the fact that the publicly available version is time-limited (30 days) and machine specific. Some students have access to a number of computers and have downloaded the available MINTAB to each in sequence. Although we assume that students have access to the Microsoft Office suite and have made it clear that students may have a choice of package (for example EXCEL) to compute the standard statistics, there is no guarantee that these students are able to use the statistics routines in EXCEL. We can of course provide notes on the use of EXCEL for this purpose (and do so in some cases). One decision we need to make is whether to continue to use MINITAB at all for the postgraduate units.

We provide the 3Plot program in the form of zipped files, with downloading instructions. This aspect works well and, for its limited purpose, so does 3Plot. However, the program can be

tedious to use as the dialog boxes retain no memory from one use to the next. This means that when displaying a number of plots one has to constantly re-enter the relevant parameters. This is not in itself arduous but is a source of irritation to students who are familiar with the slickest of computer graphics.

The VARIOWIN package is also provided in a zipped file, with instructions for its installation. This package is very successful and the students really enjoy the interactive semivariogram modelling facility it provides. The restriction to 2D is not a particularly onerous one since this is the main illustration for this level of course. However, VARIOWIN was originally written as a lead-in to an earlier version of GSLIB and has a feature in **Model** that provides the formulation of the chosen model ready for input into a GSLIB1 kriging parameter file, and this is not relevant (and can even be misleading) for students using the GSLIB2 version. In addition, although VARIOWIN provides the facility to calculate and model a number of alternative measures of spatial continuity, it does not present these in standard form but rather adjusts them so that they have a direct comparability with the semivariogram. This makes it difficult to model these alternative measures directly and causes difficulty for our students.

EXCEL is not provided by us, as the assumption is that students have access to a computer with Microsoft Office. Some problems that arise with the use of EXCEL is the failure of students to customise the spreadsheets to the data set that is being investigated. Here we have problems with students not being familiar with programming a spreadsheet or writing macros. This may be more relevant for the undergraduate students in MAT3106 than for the postgraduate students in the other units.

However, it is the use of GSLIB that causes the most difficulty. The students doing these units are not computing majors and are more used to using software that is menu driven. Such students are unable to deal with the format of the parameter files and are often not very clear about the exact nature of the options that are provided. An example of this is the specification of the semivariogram parameters, where amongst other things the major direction of continuity needs to be specified. The information from VARIOWIN is given in terms of mathematical angles (measured anticlockwise from East) while in the GSLIB parameter files the geological conventions for specifying angles are used (azimuth angles, measured clockwise from North). The parameter file for any particular routine contains all the user specified information required to execute the program. An example of this is the input of the semivariogram model. In addition, even though we endeavour to explain the various parameter options available in a package and detail about the various packages is provided in the user's guide, there are problems with translating this abstract information to the concrete case of a data set. This is exacerbated by the nature of the feedback provided from the programs. This consists in the first instance of echoing the input information in the DOS or Command Prompt window, and secondly through information written to the debugging file. The estimation and simulation routines in GSLIB provide information on the calculations that are being carried out at each estimation node, and the amount of detail provided depends on the debugging level chosen by the user. For a user with a reasonable mathematical background it is this file which helps make sense of the output. For a user with a weak mathematical background it will not be of much use. In particular, it does not safeguard the user against mis-specification of data columns or making the wrong choice in specifying the simulation or estimation grid. There is no interactivity and so it is hard to validate choices. In addition, the student usually generates a GSLIB parameter file by editing a given parameter file. The implementation is unforgiving and the inadvertent inserting of an unseen character during the editing can lead to an error measage indicating, for example, that the specified sample data file does not exist. For the student who can see the file in the relevant directory this is source of intense frustration.

An additional implementation problem lies in the fact that some of our students already have rudimentary knowledge of some of the material they are required to learn and whatever incorrect method they have learnt prior to starting the course gets perpetuated. The classical example here is the separate modelling of the semivariogram in the major and minor directions followed by a later collation into one model, which often turns out to be incorrect.

4. Conclusion

In this section we review our software choices and give an indication of our future direction. MINITAB of course works well and, being a commercial package, is regularly updated. Although VARIOWIN and 3Plot work well at present, they each have a limited life span and it will soon become necessary to find appropriate replacements. One possibility for replacing 3Plot would be to provide a facility in either MINITAB or EXCEL to do data posting and mosaic maps. It would also be good to replace GSLIB by a more interactive package, but those available do not meet the criteria we specified earlier.

An alternative to the use of free-ware would be to use an integrated, interactive package. There are several packages of this type available, one of which is ISATIS (see [5]), which we use in our geostatistical research work. However, this approach has two major drawbacks. The first difficulty is that it is a substantial task for the students to familiarise themselves with the running of such a package. This is particularly relevant for the undergraduate students in MAT3106 who study only the one geostatistics unit. The second, greater, difficulty is that such packages are geared to industry and are extremely expensive to purchase. Even when there is a free educational version, as for the case of ISATIS, there is a high maintenance charge and the educational licence does not permit us to make the software available to students who are unable to come to the campus and do campus based laboratory work. As the vast majority of our postgraduate students work in remote locations in Australia and some even are based overseas, this is not for us a practical solution.

The most likely solution is for us to utilise a package such as EXCEL and to write as many as possible of the calculation routines as macros, in the way it has been done in parts of MAT5114 and MAT5114. However, we will also explore the possibility of obtaining an educational licence for ISATIS and arranging for at least our postgraduate students to login remotely.

References

[1] Bloom, L.M., Mueller, U. and Shi, B. (1999) Application of technology in the analysis of spatial data, in Yang, W-C et al (eds), *Proceedings of the Fourth Asian Technology Conference in Mathematics*, ATCM Inc, USA, p 338 – 346.

[2] Deutsch, C. and Journel, A. (1992) *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press.

[3] GEOSTAT OFFICE '98 (1998), http://users.podolsk.ru/scher/eng/gsoffice/gscontent.html

[4] Goovaerts, P. (1997) *Geostatistics for natural resources estimation*, Oxford University Press.

[5] ISATIS (2001), GEOVARIANCES, http://www.geovariances.fr

[6] MINITAB® Statistical Software for Windows (2004), Release 14, http://www.minitab.com

[7] Panatier, Y. (1996) *Variowin Software for Spatial data Analysis in 2D*, Springer.