Using Public Domain Geostatistical and GIS Software for Spatial Interpolation

C. Varekamp, A.K. Skidmore, and P.A.B. Burrough

Abstract
A spatial interpolation system is presented that uses public domain geostatistical and GIS software from the U.S. and The Netherlands. This paper describes how component programs were linked together. The system was tested with a case study involving the mapping of forest soil variables. The system has the advantage that the program components are of low, or no cost, and are readily available on the internet.

Introduction
The objective of spatial interpolation is to generate a map of a variable that was measured at point locations. In order to produce an interpolated map, a model containing one or more parameters is used to define an interpolation function. The user will often chose the function based on the properties required from the desired output map, for example, smoothness of the interpolated surface.

Why would a user require an interpolated surface? There are two main reasons:

- Exploratory data analysis, undertaken to visualize the spatial pattern of variables, and to understand processes between multiple variables. Such an approach is necessarily exploratory in nature, but is particularly useful for comprehending relationships and designing models. A practical example is mapping the volume of contaminated soil in order to estimate clean-up costs.
- Generate input layers for GIS models of, for example, hydrological processes, forest growth, soil erosion, or climatic impacts.

Geostatistical interpolation has been used in combination with GIS to overcome problems of choropleth maps in GIS. Van Meirvenne and Hofman (1992) showed that many variables (such as pH) are continuous across map boundaries. Consequently, errors are introduced into GIS models when an average value for the polygon is read from a traditional choropleth map. Geostatistical techniques are used to interpolate soil properties when undertaking standard soil mapping, or when mapping heavy metal concentrations along rivers (e.g., Stein, 1991). Another example is the use of interpolated maps for modeling error propagation in a GIS (Heuvelink *et al.*, 1989).

At a practical level, much work has been done by Isaaks and Srivastava (1989) and Deutsch and Journel (1992) to develop public domain geostatistical software and guidelines for the analysis of data. However, geostatistical software tools in commercial GIS remain rudimentary; consequently, incomplete "tool-boxes" are offered in commercial GIS packages. A related problem is that few GIS analysts understand geostatistical techniques, even though the software appears easy to use in commercial packages. For example, we interviewed an analyst who had interpolated soil variables by pushing buttons in the order specified by a commercial GIS manual; however, a sub-optimal variogram model had been fitted to the data, and the analyst had little comprehension of the process or of the results.

There is clearly a need for lower cost and easy-to-use geostatistical software, as well as adequate documentation for analysts. Robust and accurate software for geostatistics is available. Examples include GSTAT, a multivariate geostatistical program by Pebesma (1993); GEO-EAS, a user-friendly package now distributed by the University of Lausanne, Switzerland but originally developed by Englund and Sparks (1988); and GSLIB, a complete geostatistical software library with 37 FORTRAN programs by Deutsch and Journel (1992). There are also general GIS packages available in the public domain, including the PC-Raster package by Van Deursen (1992) used here.

To summarize, the aims of this study were to

- identify suitable public domain software for geostatistical (and GIS) analysis,
- link these separate geostatistical and GIS programs into a system for spatial interpolation,
- test the system with a case study, and
- highlight how the system is useful for GIS analysts.

This paper aims to guide new, or even experienced, users in building their own spatial interpolation system using public domain software. In order to avoid yet another introduction to geostatistical techniques, we assume an elementary knowledge of this topic, such as presented in readily available books by Burrough (1986, Chapter 6) or Webster and Oliver (1990, Chapter 12).

System Description
The interpolation system proposed here links sub-program components developed by different geostatistical and GIS packages. The overall aim is to link these sub-programs in order to convert point data into maps. To achieve this, the geostatistical tools have been grouped into four broad functional classes:

- exploratory data analysis (EDA),
- parameterization,

C. Varekamp is with the Department of Water Resources, Wageningen Agricultural University, Nieuwe Kanaal 11, 6709 PA Wageningen, The Netherlands.

A.K. Skidmore (the corresponding author) is with the School of Geography, UNSW, Sydney 2052, Australia.

P.A.B. Burrough is with University of Utrecht, Netherlands Centre for Geoecological Research ICG, Faculty of Geographical Studies, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands.

Photogrammetric Engineering & Remote Sensing,
Vol. 62, No. 7, July 1996, pp. 845–854.

0099-1112/96/6207–845$3.00/0

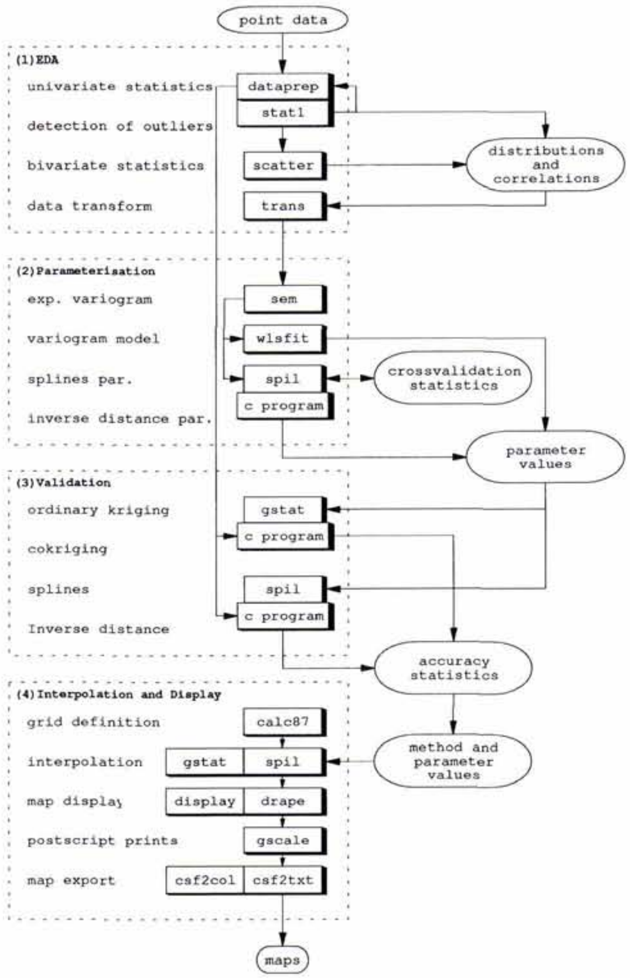

Figure 1. Flow chart showing the spatial interpolation process from point data to interpolated map. The dotted boxes represent the different tools as stages of the process, shaded boxes are computer programs, ovals are different types of data, and solid lines are flows of data.

- validation, and
- interpolation and display.

Figure 1 presents a flow chart of the spatial interpolation process from point data to interpolated map. The dotted boxes represent the broad functional classes, shaded boxes are computer programs, ovals are different types of data, and solid lines are flows of data. Functional stages 1 to 4 of the process are further subdivided; each subdivision requires a specific program component. For example, the first stage, exploratory data analysis (EDA), explores the statistical properties of the data (Tukey, 1977). The EDA function has been subdivided into tasks to generate univariate statistics, to detect outliers, to describe bivariate statistics, and to transform data. The four functional stages are described in detail below.

Table 1 lists the public domain computer packages and programs obtained from the Internet and libraries. For each program, a description of its function and the source are given. The programs are very different in nature. For example, there are programs which use "parameter" files that have no user interface whatsoever, and are distributed as FORTRAN-77 source code. But there are also interactive graphical model fitting programs with a menu driven user interface (e.g., program WLSFIT has a graphical user interface).

These disparate programs were successfully linked using MS-DOS batch files, with program execution following the direction of the flow chart in Figure 1. Simple user screens were created for the different stages in the spatial interpolation process. Figure 2 shows the user interface. There are five user screens, each with different menu options corresponding to the steps in the spatial interpolation process.

Each main functional class in Figure 1 is now described in detail.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was a precursory stage introduced to discover the statistical and mathematical properties of the data. Using EDA techniques, any necessary data transformations may be discovered, and appropriate models to apply to the data may be considered. EDA techniques became popular after a book by Tukey (1977) showed the applicability and usefulness of visualizing data and calculating simple statistics as a preliminary stage to confirmatory tests. In addition to the methods described here, standard statistical packages (such as SPSS, CSS, STATISTICA, MINITAB) offer a variety

TABLE 1. COMPUTER PACKAGES, COMPUTER PROGRAMS, FUNCTION OF PROGRAM, AND SOURCE.

Package	Programs	Function	Source
geo-eas	stat 1 dataprep trans scatter	univariate statistics data preparation arithmetic transformations scatterplots	Institute of Mineralogy, University of Lausanne, Switzerland. Anonymous ftp: eliot.unil.ch
pc-raster	sem wlsfit spil gstat calc87 display csf2col csf2txt	experimental variograms variogram fitting interpolation geostatistical interpolation map definition display of maps map file to column file map file to ASCII file	attn: Cees Wesseling, Dept. of Physical Geography, University of Utrecht, PO Box 80.115, 3508 TC Utrecht, The Netherlands. Anonymous ftp: pop.frw.ruu.nl http site address: http//www.frw.ruu.nl/pcraster.html
gslib	gscale	grey scale postscript maps	Geostatistical Software Library and User's Guide (Deutsch and Journel, 1992). ISBN 019 50 73924. Disks with 37 FORTRAN programs (source code).

of EDA techniques, while Haslett *et al.* (1990) describe software to dynamically link map data in one window to (soil) variable data in another window, thereby greatly assisting the visualization of outliers.

In this study, the point data were entered and pre-processed in a spreadsheet program (specifically Microsoft Excel, but any spreadsheet or data base program would suffice). The data were output from the spreadsheet in a simplified GEO-EAS ASCII format; a data format supported by many geostatistical programs.

Data Preparation

The *dataprep* program, provided by GEO-EAS, reads the simplified GEO-EAS ASCII file. The file can be sorted, if necessary, and a variable created to identify the observations. The *dataprep* program offers utilities for column extraction, row extraction, deletion of duplicate rows, and merging with other data files. After preparation, the data are input to the univariate statistics program.

Univariate Statistics

The GEO-EAS program *stat1* displays the number of observations and missing data, and calculates univariate statistics. Histograms may be viewed in order to investigate distributions of the statistics. Variables exhibiting a non-normal distribution may be transformed (see the *Data Transform* subsection below).

Detection of Outliers

An outlier is an observation that is well separated from the rest of the data (Bowerman and O'Connell, 1990). The program *stat1* may identify outliers using the coefficient of variation (CV). The CV is a dimensionless quantity that measures the amount of variation relative to the mean. A CV greater than 100 per cent may indicate the presence of samples with erratic values (Isaaks and Srivastava, 1989). The *stat1* program option "Examine" allows an analyst to easily identify record numbers that contain outliers. Rejection of outliers on a purely statistical basis is a dangerous procedure (D'Agostio and Stephens, 1986), and users should always double check how realistic the observations are. Finally, the *dataprep* program is used to remove outlying values from the data file.

Bivariate Statistics

A data set often contains not only the primary variable of interest, but also one or more secondary variables. The secondary variables often contain useful information about the primary variable. The secondary variable may be used by an interpolation method termed "cokriging" to improve the accuracy of estimation for the primary variable. To determine the possibility of cokriging, correlations between the primary and secondary variables are calculated using a GEO-EAS program called *scatter*. The *scatter* program prints scatter plots of variables and correlation statistics on the computer screen. After exploring linear relationships for all pairs of variables, the variables can be grouped into subsets containing only variables which are highly correlated (Pan *et al.*, 1993).

Data Transform

The *trans* program allows the user to perform arithmetic transformations when necessary. For example, left or right skewed distributions may be transformed to a normal distribution by calculating respectively the logarithm or exponent of the original data values (Tukey, 1977). Note that all these operations may be undertaken with general statistical packages such as SPSS, CSS, STATISTICA, MINITAB, etc.

Parameterization

Parameterization is the second stage in the interpolation system and is used to select and fit the parameters needed in the different interpolation procedures (Figure 1). These parameters are required to solve interpolation equations, as well as to calculate neighborhood parameters that determine which observation points are used for predicting the value of each cell in the GIS.

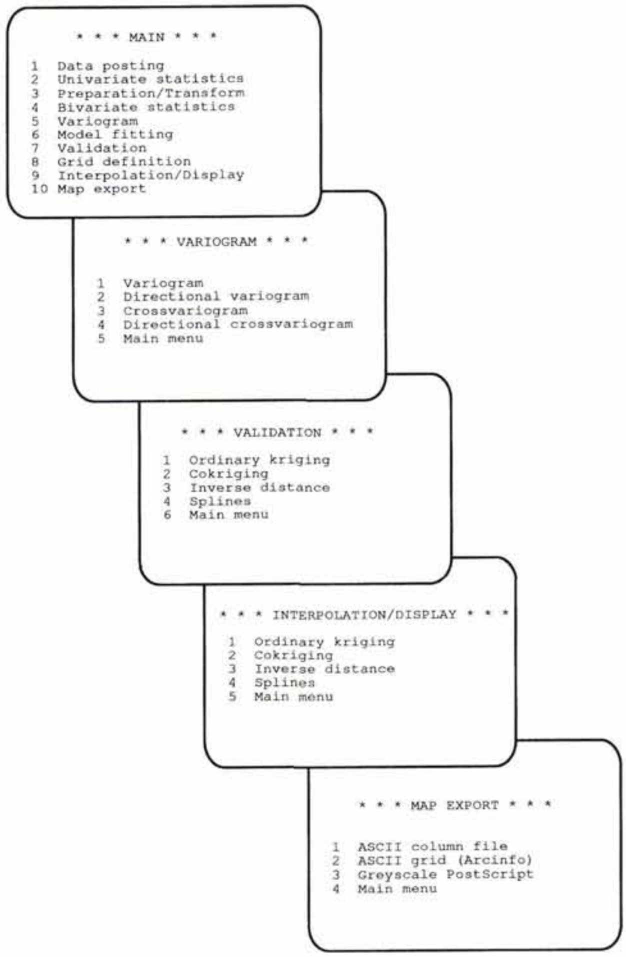

Figure 2. User interface: User screens with menu options of the spatial interpolation system.

It is useful to distinguish between two types of interpolation methods; one is based on a deterministic model and the other on a stochastic model. Data input to a deterministic model will yield just one outcome, and includes methods such as inverse distance, trend surface, moving average, splines, and Delauney triangulation. A stochastic model presupposes that the outcome is uncertain and is described by a probability function (Addiscott, 1993); methods include ordinary kriging, universal kriging, and cokriging.

For stochastic or "kriging" methods, a model is fitted to the experimental variogram; the parameters of the variogram model are then used in the interpolation procedure.

The terminology that is used to describe important features of the variogram is as follows (Isaaks and Srivastava, 1989):

- Range: The distance at which the variogram reaches a plateau. In other words, the point at which semivariance stops increasing with increasing lag.
- Sill: The plateau the variogram reaches at the range.
- Nugget Effect: The vertical jump from the value of 0 at the origin to the value of the variogram at extremely small separation distances.

Sometimes, we want to predict the average value of a variable within a local area. For example, soil data may be combined with a remote sensing image in a GIS to predict the mean soil variable for each remote sensing image pixel. In these cases, the stochastic method termed "block kriging" may be used.

Often, information will be available as secondary (and easy to measure) variables rather than the primary variable of interest. These secondary variables are usually spatially cross-correlated with the primary variable (Isaaks and Srivastava, 1989). The cokriging method also uses this information for the prediction of the variable of interest.

An advantage of stochastic methods is that interpolation errors can be calculated and mapped, in addition to the estimate of the variable.

Experimental Variogram

The *sem* program (from the PC-raster package) calculates experimental variograms of the variables. The user specifies in a "parameter" file the "lag" distance (which is often approximated by the sampling interval), and the maximum distance that will be considered between samples. Experimental variogram values are calculated and written into an ASCII file for further processing by the *wlsfit* and *gstat* software programs, both distributed in the PC-raster package.

Variogram Model

The user defines an input file containing the experimental variogram values. The operator may choose one of six variogram model types to be fitted to the experimental variogram (e.g., linear, spherical, Gaussian, exponential). The *gstat* program writes the experimental semivariances to a file. The *wlsfit* program is then used interactively to fit a model to the experimental variogram, using a weighted least-squares approximation (Cressie, 1985). The resulting model may be plotted on screen and printed.

The selection of a suitable variogram model is very much a matter of experience. It is often possible to select a few suitable model types. The *wlsfit* program is particularly useful as it calculates a goodness-of-fit statistic that can serve as a model selection criterion.

Splines Smoothing Parameter

When fitting splines to the data, the value of the so called "splines smoothing parameter" is needed. This parameter controls the roughness of the interpolated surface. A variable that is very continuous will not need much smoothing whereas, for a variable that behaves erratically, more smoothing is desirable. For this study, we adopted a technique called cross-validation to find an optimum value for the splines smoothing parameter. Cross-validation allows comparison of estimated and true values using only the information available in the sample data (Isaaks and Srivastava, 1989). An observed data value at a particular location is temporarily taken from the data set; the value at the same location is then estimated using the remaining samples. After doing this for all locations, and repeating the process using different "splines smoothing values," an optimum is found.

The cross-validation technique was implemented by linking the PC-raster *spil* program with a C program. The *spil* program cross-validates the set of parameter values, and writes estimated and true values into a simplified GEO-EAS format ASCII file. A C program was written at UNSW to calculate the Mean Absolute Error (MAE). By iterating between the *spil* and the C program and calculating MAE, the spline smoothing value that results in the smallest MAE is selected. The functions of *spil* have subsequently been implemented in the *gstat* program, and *spil* is no longer being supported.

Inverse Distance Parameter

This parameter is an important input to the inverse distance weighting method, because it determines the weight a nearby sample receives in proportion to the weight a faraway sample receives. As for the splines interpolation method, the inverse distance parameter is found by means of cross-validation.

Validation

Validation is the third stage in the interpolation process, and it allows the predicted values to be compared with the true values of a validation data set. Specifically, the objectives of this stage are to

- obtain information about the accuracy of the interpolation(s), and
- understand how the different methods and parameters affect the interpolation results.

In other words, it assists in empirically selecting an appropriate interpolation method, as well as appropriate parameter values for the chosen method.

The interpolation methods and parameters are validated using a separate, independent data set. This data set may be obtained after the interpolation and mapping work is complete, or by using data withheld from the original data set. In many cases, collecting an independent data set is difficult to arrange, so a withheld sample from the original data set is then a reasonable alternative.

Ordinary Kriging and Cokriging

The *gstat* program validates the parameter values for the stochastic methods. In a parameter file the user gives the variable, the parameter values, and the name of a validation file containing the validation data set. The *gstat* program writes interpolated values and the true values for the set of validation points into an ASCII file.

Splines and Inverse Distance

The PC-raster *spil* program validates the parameter values for the deterministic methods. The *spil* program writes interpolated values and true values for the set of validation points into an ASCII file.

A C program is executed for all methods to calculate validation statistics; the validation statistics measure the overall interpolation accuracy.

Interpolation and Display

The fourth and final stage in the interpolation process includes tools which interpolate the surface, as well as produce maps for display, analysis, or further modeling in a GIS. In a previous stage of the analysis, the interpolation accuracy of the different methods has been calculated. The aim in this stage is to produce interpolated maps using the best available method, as well as the optimal parameter values. The first step is to define a grid to interpolate.

Grid Definition

It is important to interpolate inside the observation area, as errors rapidly increase outside. The *calc* program, which is part of the PC-raster package, defines the grid to be interpolated by specifying the number of rows, number of columns, map projection, map origin, and cell size. Existing grids from other popular GIS systems such as ARC-INFO, ERDAS, GENASYS, IDRISI, and SURFER may be simply imported.

Interpolation

Interpolation requires a few extra lines added to the *gstat* and *spil* parameter files, which were also used in the parameterization and validation stages. The extra lines in the parameter file first specify the name of the file with grid definitions, and then specify the output file names to contain the interpolated maps. Interpolated maps are produced in the PC-raster file format. For stochastic methods, the interpolated map is accompanied by an error map showing the interpolation uncertainty.

Map Display

There are multiple ways to display the interpolated maps. The PC-raster *display* program produces color maps on the computer screen. A contouring option in *display* allows the

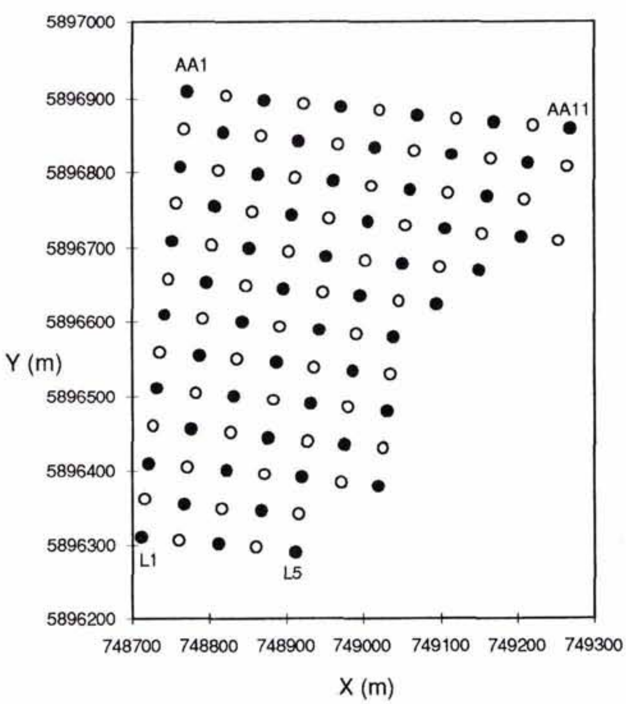

Figure 3. Sampling configuration. X(m) and Y(m) in UTM coordinates. The original data set consisted of an observation data set (dots) and a validation data set (circles).

user to add isolines to color map. The *drape* program allows one variable to be draped over another, and display of the result in a semi three-dimensional view. At present, the maps may be printed using the software drivers developed for Epson or HP DeskJet printers. We have also exported PC-raster data to commercial statistical and graphics programs to display data, including STATISTICA, MATLAB, SURFER, ARC-INFO, ERDAS, GENASYS, and (Aldus) FREEHAND.

Postscript Prints

For scientific publications, monochrome maps are preferable, with a proper legend and scale. The *gscale* program produces good quality grey-scale images in the postscript language (see examples below!).

Map Export

The *csf2txt* and *csf2col* programs, which are part of the PC-raster package, allow the user to export maps to other software packages for display and further GIS modeling.

Testing

The interpolation system described above was used to map forest soils at Nullica State Forest, located approximately 25 km northwest of Eden township, on the far south coast of New South Wales, Australia. The study area is relatively undisturbed, except for fire, some selective logging, and occasional grazing. Climate is mild, with an annual rainfall of about 1000 mm, and an average temperature of 15°C. The vegetation over the 500-m by 600-m study area comprises a typical transition of *Eucalyptus sieberi* (Silvertop Ash) on ridges, blending into stringybarks, and changing to gum forest in the valleys. The parent material was described by Beams (1980) as a rhyolite, part of the larger Upper Devonian Boyd Volcanic Complex. Within the study area, the parent material is essentially homogeneous. Forest soils in the vicinity of Eden have low nutritional status, are highly weathered, and are generally acidic (Lambert and Turner, 1983; Bridges and Dobbyns, 1991), and the rhyolite parent material on which the study area is located is one of the least nutritionally rich (Keith and Sanders, 1990).

Table 2. List of Soil Properties.

Acronym	Unit	Description
ph1 ph2 ph3	—	pH (1:5 soil-to-water extract, Rhoades, 1982)
ec1 ec2 ec3	dS m^{-1}	Electrical conductivity (1:5 soil-to-water extract, Rhoades, 1982)
p1 _ av p2 _ av p3 _ av	mg kg^{-1}	Available phosphorus (Ca-lactate extraction, Holford *et al.*, 1985)
k1 k2 k3	mmol kg^{-1}	Available potassium (1:5 soil-to-water extract, Rhoades, 1982)
ca1 ca2 ca3	mmol kg^{-1}	Available calcium (1:5 soil-to-water extract, Rhoades, 1982)
mg1 mg2 mg3	mmol kg^{-1}	Available magnesium (1:5 soil-to-water extract, Rhoades, 1982)
n1 n2 n3	mg kg^{-1}	Total nitrogen (Kjeldahl method, Heffernan, 1985)
p1 _ tot p2 _ tot p3 _ tot	mg kg^{-1}	Total phosphorus (Kjeldahl method, Heffernan, 1985)

A grid with cell size of 50 by 50 m was positioned over the 500- by 600-m study area such that the columns of the grid were oriented approximately north-south along a mountain ridge. For easy referencing, the grid rows are coded with characters, and the columns with numbers (Figure 3). For validation purposes, the original data set was split in half to form an observation data set (solid dots) and a validation data set (open circles). Soil samples were taken from three soil layers; three layers were common to all samples. Table 2 lists the acronyms of the soil properties analyzed. Note that the numerical suffix refers to the soil layer; for example, pH2 is the pH measured in soil layer 2. Phosphorus has an additional suffix: "_av" for available phosphorus and "_tot" for total phosphorus. For each variable, a short description of the method of analysis is given in Table 2.

Exploratory Data Analysis (EDA)

Using the *dataprep* program, the data file was split in half to create an observation data file and a validation data file. Table 3 presents univariate statistics calculated with the *stat1* program for some selected soil variables (i.e., pH1, ec1, ca1, p1_tot). Most variables exhibit a moderate positive skewness. The only variable showing a negative skewness was pH, which is not surprising because pH is derived from H+ concentration by taking the logarithm with base 10. Although the distributions are not perfectly normal, transformations of the data were not considered necessary.

We identified outliers using the *stat1* program. The *dataprep* program was used to remove the outliers from the observation data set. Table 4 lists the CV before the removal of outliers, as well as the CV after the removal of outliers. The CV was successfully reduced to less than 100 per cent for four out of seven variables in Table 4. The 50 percent drop in CV for available phosphorus in soil layer 3, after removing only one outlier from the database, suggests contamination for this particular soil sample.

The *scatter* program was used to generate a correlation

Table 3. Univariate Statistics for ph1, ec1, ca1, and p1 _ tot, Calculated with the *stat* 1 Program.

	ph1 —	ec1 dS m^{-1}	ca1 mmol kg^{-1}	p1 _ tot mg kg^{-1}
N	89	88	87	82
Mean	5.4	69.8	0.6	24.0
Variance	0.2	1211.8	0.3	203.1
Std. Dev.	0.5	34.8	0.6	14.2
CV	8.6	49.9	99.1	59.4
Skewness	−1.2	0.6	1.9	1.3
Kurtosis	4.9	2.5	7.2	4.6
Minimum	3.6	10.5	0.0	0.0
25th percentile	5.2	41.2	0.2	13.6
Median	5.5	63.9	0.4	20.5
75th percentile	5.7	90.6	0.7	29.3
Maximum	6.2	156.6	3.1	71.8

Table 4. Location Codes of Outliers, CV before the Removal of Outliers, CV after the Removal of Outliers, and Decrease in CV for the Soil Properties.

		CV (%)		
Variable	Location codes	before	after	decrease
p2 _ av	C1,C3,H1,H2,H4,I1	126.3	94.4	25.3
ca2	A3, C2	129.0	115.7	10.3
mg2	A2,A3,C2,K2,L5	110.7	71.2	35.7
p3 _ av	L2	244.3	120.9	50.5
k3	F5,G4	140.4	79.5	43.4
ca3	G4,J1,J2,J3,J7,K4,L2,L3	138.2	99.6	27.9
n3	B10,C10,G7	155.6	140.8	9.5

Table 5. Correlation Matrix for the Soil Properties in Layer 1.

	ph1	ec1	p1 _ av	na1	k1	ca1	mg1	n1
ec1	−0.40							
p1 _ av	−0.19	0.49						
na1	−0.28	0.53	0.26					
k1	−0.16	0.75	0.51	0.29				
ca1	−0.20	0.46	0.03	0.20	0.26			
mg1	−0.25	0.79	0.37	0.42	0.57	0.67		
n1	−0.14	0.61	0.44	0.33	0.59	0.14	0.46	
p1 _ tot	0.32	0.30	0.28	0.08	0.26	−0.03	0.24	0.55

matrix for the soil properties in layer 1 (Table 5). Underlined coefficients are greater than 0.5 and are statistically significant for $p < 0.001$; for example, note that cation concentrations are strongly correlated with electrical conductivity (EC). Such information is very useful when trying to estimate by cokriging an under-sampled variable (for instance, the variable may be expensive to measure) using a highly correlated intensively sampled variable. In order to provide an example for the following "parameterization" stage, magnesium is used to estimate calcium (for soil layer 1), and total phosphorus in layer 1 is used to estimate total phosphorus in layer 2.

Parameterization

Figure 4 shows experimental variograms and fitted models for the soil properties in soil layer 1. The spherical model was used for all soil variables; it gave the best model fit according to the *wlsfit* goodness-of-fit statistic. Figure 5 shows experimental cross variograms, and fitted spherical variogram models, for the variable combinations [ca1,mg1] and [p1_tot, p2_tot]. Parameter values for all variogram models are given in Table 6. A large difference in goodness-of-fit statistic (i.e., the SSD/SST ratio computed by the *wlsfit* program) among the variables was found. The model fit for p1_tot is considered to be good (SSD/SST=0.005), whereas the appropriateness of the fitted model for mg1 is suspect (SSD/SST=0.860). Most model fits result in a rather large value for the nugget parameter (C0). This could either mean that the variable varies within shorter distances than the sampling distance, or that the measurement error is large.

For p1_tot, the nugget parameter value (C0) is relatively small compared with the sill parameter value (C); for p1_tot, it is possible to model the spatial variation quite well. The range parameter a(m) is a measure for the distance up to which the spatial dependence extends. The spatial dependence extends to a few hundred metres for most variables, but is nearly 800 m for p1_tot.

Using the cross-validation technique described above, a value of 2 was found to be the optimal inverse distance parameter value for all soil properties in layer 1.

Validation

This stage (i.e., programs *gstat* and *spil* in Figure 1) used the "validation" part of the data set. Table 7 lists the resulting statistics; for each interpolation method, the mean absolute error (which estimates the model accuracy), mean error (which estimates bias), number of observation sites, and number of validation sites were calculated. For ca1, inverse distance is the most accurate method, although slightly more biased than ordinary kriging. For p1_tot, cokriging is most accurate, but much more biased than both ordinary kriging and inverse distance.

Interpolation and Display

Two grids were prepared with the *calc*; one grid with cells of 30 by 30 m, and another finer grid with cells of 10 by 10m. Block average predictions and variances were calculated over the coarser grid using ordinary kriging. To visualize the effects of different interpolation methods, ordinary kriging, cokriging, and inverse distance were used to interpolate on the fine grid.

Grey-scale maps of the block kriging interpolation for ph1 and p1_tot are shown in Figure 6; both the estimated concentrations of the soil variables as well as the variance are mapped using the *gscale* program. Such maps are very useful as input to other GIS models.

The *csf2txt* program exported the fine grid interpolations (i.e., ordinary kriging, cokriging, and inverse distance) as ASCII files, which were displayed using the MATLAB program; Figure 7 shows surface maps of ca1. Maps produced by the ordinary kriging and cokriging methods were "smooth," but characterized by low "peaks" repeated rapidly over short distances. This pattern of variation suggests the presence of

Table 6. Parameter Values and SSD/SST Statistic of (Cross-) Variogram Models.

Variable(s)	C0	C	a	SSD/SST
ph1	0.05	0.17	318	0.085
p1 _ av	3.15	26.10	174	0.105
na1	0.12	0.10	206	0.307
k1	0.12	0.04	173	0.804
ca1	0.11	0.08	273	0.194
mg1	0.15	0.03	222	0.860
n1	16500	5030	406	0.528
p1 _ tot	18.10	433	782	0.005
[ca1, mg1]	0.06	0.02	782	0.856
[p1 _ tot, p2 _ tot]	0.00	223	783	0.078

C0 is the nugget variance; C is the sill parameter, a is the range parameter measuring the distance to which the spatial dependence extends.

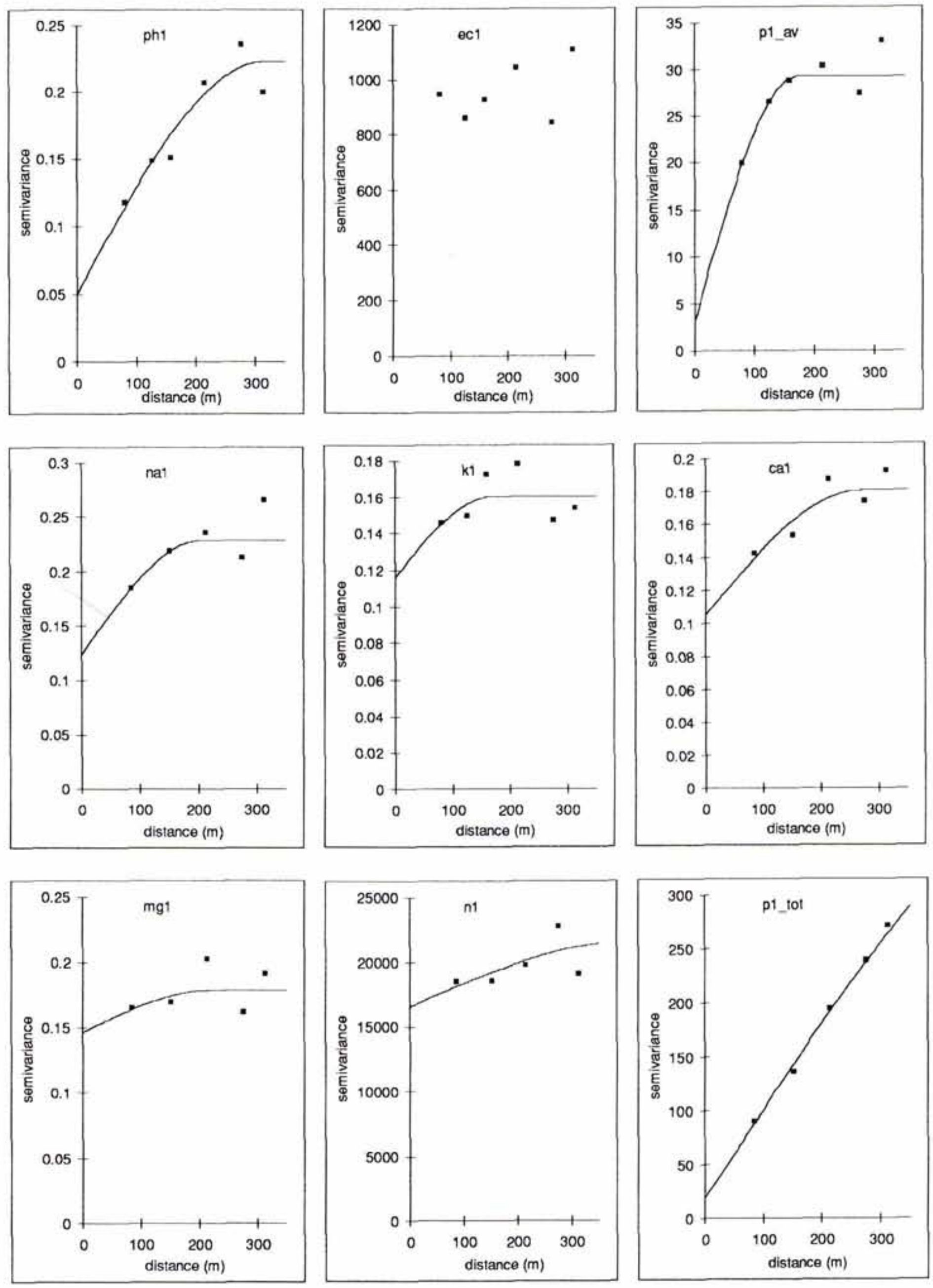

Figure 4. Experimental variograms and fitted spherical models for the soil properties in layer 1.

small environmental boundaries within the test area. Inverse distance maps have a less "smooth" surface, but with many steep peaks and pits (Figure 7).

Discussion

The main aim of this study was satisfied, that is, building a system for the spatial interpolation of point data using public domain software. The system was successfully tested with a case study that mapped the distribution of eucalypt forest soil variables.

For GIS analysts, another process starts after the data have been interpolated and converted to spatial maps. The interpolated maps may be visually analyzed, or input to models, and integrated with ancillary data. The public domain software used here allows flexible and easy coupling with commercial GIS systems; for example, maps have been easily exported to ARC-INFO and GENASYS. Figure 8 highlights an example, where the interpolated soil variables are draped over a digital elevation model (DEM); there is, for example, an obvious relationship between pH and terrain.

For further modeling in GIS, the estimated variance map representing the uncertainty of the prediction is important. For example, the cells in the northwest of the study area have a high variance, and are therefore uncertain (Figure 6). Such information may be used to estimate error in maps output by GIS models (e.g., Heuvelink *et al.*, 1989).

Turning to those readers who may wish to build a similar system, a diverse range of information is available on

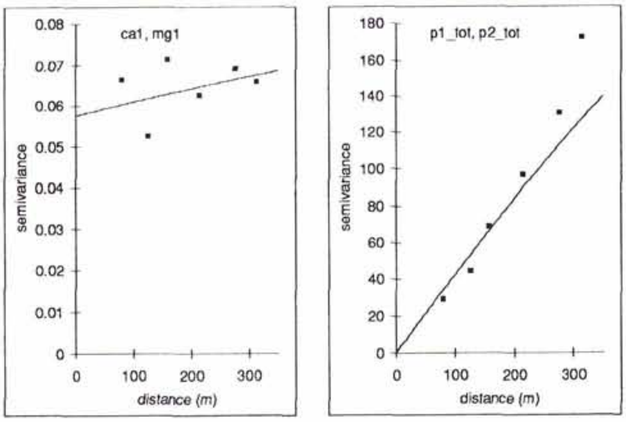

Figure 5. Experimental cross variograms and fitted spherical models for the variable combinations [ca1,mg1] and [p1 tot,p2_tot].

TABLE 7. MEAN ABSOLUTE ERROR (MAE), MEAN ERROR (ME), NUMBER OF OBSERVATION SITES (NOBS), AND NUMBER OF VALIDATION SITES (NVAL) FOR CA1 AND P1 _ TOT.

Variable	Interpolation Method	MAE	ME	Nobs	Nval
ca1	Ordinary kriging	0.316	0.0386	47	40
	Cokriging (covariable: mg1)	0.322	0.0529	47	40
	Inverse distance (α=2.0)	0.311	0.0400	47	40
p1 _ tot	Ordinary kriging	7.917	0.2436	44	38
	Cokriging (covariable: p2 _ tot)	7.741	0.7237	44	38
	Inverse distance (α=2.0)	8.066	0.1706	44	38

defining the interpolation process and selecting suitable program components. The most useful information sources accessed in this study are given in Table 8. In particular, the book by Isaaks and Srivastava (1989) contains practical advice and solutions when defining the separate stages of the interpolation process.

A difficult step is identifying suitable program components; the packages mentioned in this paper all have advantages and disadvantages. For example, GEO-EAS is useful for teaching and demonstration and has a simple data format which has become a standard for most other geostatistical packages. The disadvantage of GEO-EAS is that it is slow, uses EGA graphics, lacks some interpolation programs such as cokriging, and is not being updated. Source code was provided for the GSLIB package, and the package provides a lot of powerful routines. GSTAT allows a mask to be defined for the interpolation area, and can interpolate only those cells with non-missing values, thereby improving processing time. Another advantage of GSTAT is that it is linked to a powerful public domain GIS (PC-raster).

The software used in this project was obtained from the Internet and the book by Deutsch and Journel (1992). As a

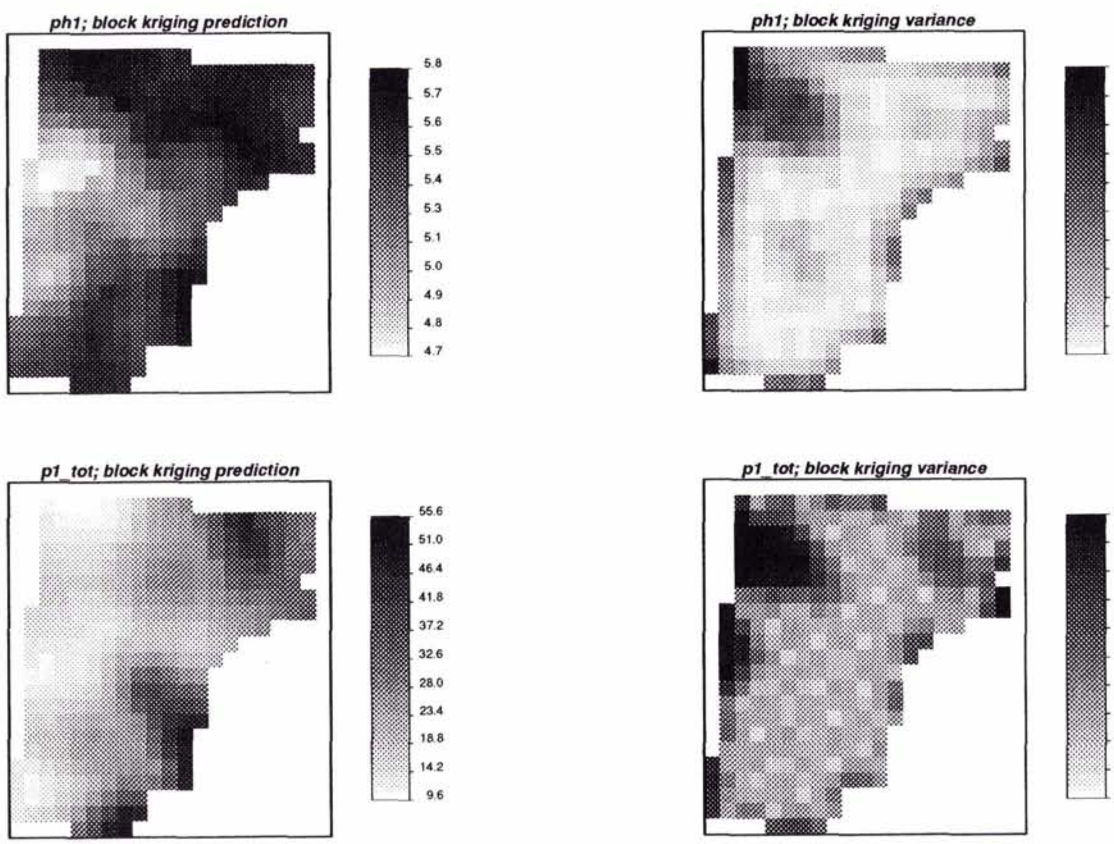

Figure 6. Grey-scale maps of the block kriging interpolation (left) and the block kriging variance (right) of ph1 and p1_tot. Interpolations are on a 30- by 30-m grid.

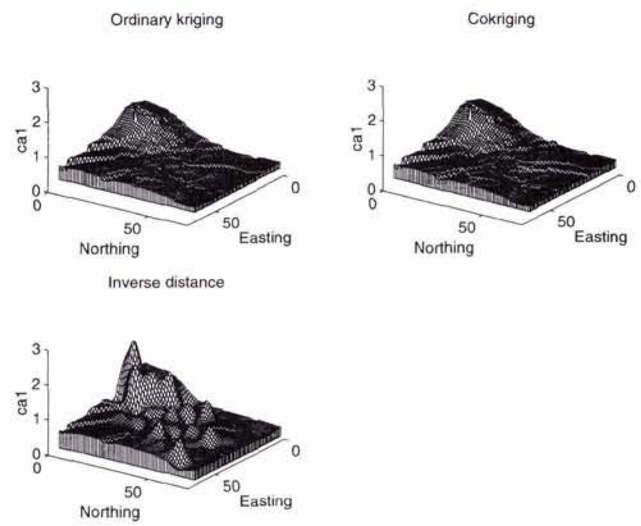

Figure 7. Surface maps of ca1 using three different interpolation methods. Interpolations are on a 10- by 10-m grid.

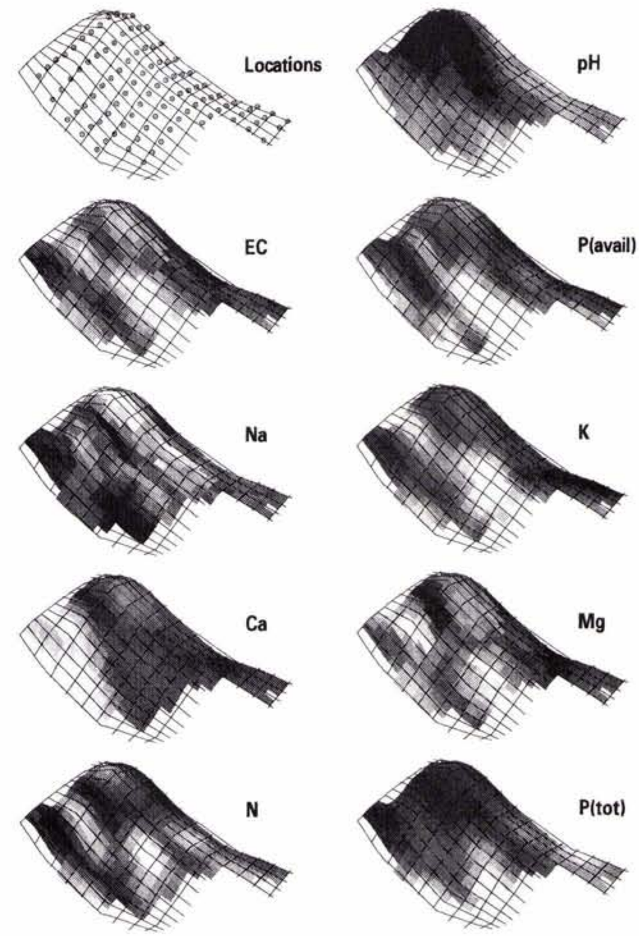

Figure 8. Interpolated variables from soil layer 1 draped over the digital elevation model (DEM)

result, the suite of programs used here is a reflection of what could be accessed through the Internet using ftp, or the library system, at the time the system was built. Some problems with file formats (i.e., input and output formats) were encountered when linking programs, due to the often very different nature of the programs, and the fact that the simple (and reasonably standard) GEO-EAS ASCII file format was not supported by all programs. Readers should consider using mostly GSLIB source code programs which support the GEO-EAS ASCII file format. A user friendly interface could then be built in a programming language such as C++ or FORTRAN, or by using a 4GL such as MATLAB or IDL.

The system developed here runs under MS-DOS (that is, it does not require Microsoft windows). Since completing this study, an exploratory variogram analysis tool called VARIOWIN has become available on the Internet. This public domain program has excellent help facilities and is very user friendly. The program can be obtained by anonymous ftp from the Institute of Mineralogy, University of Lausanne (see Table 1). Linking this program with a GSLIB derived system would make a powerful and user friendly interpolation system.

There have also been several developments in the PC-raster and GSTAT software since this paper was written. GSTAT now provides a single interface for cross-validation, inverse distance of kriging interpolation, and conditional simulation.

TABLE 8. INFORMATION SOURCES.

Source	Type of source	Comments
Burrough, 1986.	Book	Very useful for an elementary introduction to concepts of GIS, spatial interpolation and geostatistics.
Cressie, 1991.	Book	Not suitable for the beginner. A treatment of all methods and procedures.
Deutsch and Journel, 1992.	Manual: –Book with theory and detailed program description –Diskettes with 37 FORTRAN source code programs	Very practical and flexible. The FORTRAN source code allows the user to adapt programs and build user interfaces. Contains interpolation, simulation and utility programs.
Isaaks and Srivastava, 1989.	Book	A practical book on geostatistics with many case studies. The popular interpolation methods are treated in detail.
Lancaster and Salkauskas, 1986.	Book	Treats the basics of curve and surface fitting. Useful because it is more generic than geostatistics alone.
McBratney and Webster, 1986.	Article	Very helpful for deciding on appropriate variogram functions and fitting procedures.

In other words, GSTAT replaces blocks 2 and 3 in Figure 1 (the SPIL program is no longer used). The CALC program in PC-raster now allows for dynamic modeling, for example, with time series data (Wesseling *et al.*, 1996).

Conclusions

Readily available public domain geostatistical and GIS software have great potential for building a user friendly spatial interpolation system. Such a system was implemented using a diversity of public domain computer packages and programs. The set-up can easily be extended to include the interpolation of categorical variables and the conditional simulation of variables, using public domain software only. The use of IBM-PC batch file programs for the interpolation process works satisfactorily, though improved interfaces could be implemented using a 4GL (e.g., IDL or MATLAB) or programming languages such as C++ or FORTRAN.

References

Addiscott, T.M., 1993. Simulation modelling and soil behaviour, *Geoderma*, 60:15-40.

Bowerman, B.L., and R.T. O'Connell, 1990. *Linear Statistical Models: An Applied Approach*, PWS-KENT Publishing Company, Boston.

Bridges, R.G., and G.R. Dobbyns, 1991. The Dry Sclerophyll Silvertop Ash-Stringybark forests of south-eastern New South Wales, *Forest Management in Australia* (F.H. Mckinnel, E.R. Hopkins, and J.E.D. Fox, editors), Surrey Beatty & Sons Pty. Ltd., Western Australia.

Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resource Assessment*, Oxford University Press, Oxford.

Burrough, P.A., J. Bouma, and S.R. Yates, 1994. The state of the art in pedometrics, *Geoderma*, 62:311-326.

Cressie, N.A.C., 1985. Fitting variogram models by weighted least squares, *Mathematical Geology*, 17:563-586.

———, 1991. *Statistics for Spatial Data*, Wiley & Sons, New York.

D'Agostio, R.B., and M.A. Stephens, 1986. *Goodness-of-fit techniques*, Marcel Dekker, Inc., New York.

Deutsch, C.V., and A.G. Journel, 1992. *GSLIB: Geostatistical Software Library and User's Guide*, Oxford University Press, New York.

Dimitrakopoulos, R., 1993. Artificially intelligent geostatistics: A framework accommodating qualitative knowledge-information, *Mathematical Geology*, 25(3):261-279.

Englund, E., and A. Sparks, 1988. *Geo-EAS 1.2.1 Users Guide*, Report Number 60018-91/008, Environmental Protection Agency, EPA-EMSL, Las Vegas, Nevada.

Haslett, J., G. Wills, and A. Unwin, 1990. SPIDER — An interactive statistical tool for the analysis of spatially distributed data, *International Journal of Geographical Information Systems*, 4(3):28-29.

Heffernan, B., 1985. *A Handbook of Methods of Inorganic Chemical Analysis for Forest Soils, Foliage and Water;* A.R.A.C.I., C. CHEM. CSIRO Division of Forest Research, Canberra.

Heuvelink, G.B.M., P.A. Burrough, and A. Stein, 1989. Propagation of errors in spatial modelling with GIS, *International Journal of Geographical Information Systems*, 3(4):303-322.

Holford, I.C.R., J.M. Morgan, J. Bradley, and B.R. Cullis, 1985. Yield responsiveness and response curvature as essential criteria for the evaluation and calibration of soil phosphate tests for wheat, *Aust. J. Soil Res.*, 23:417-427.

Isaaks, E.H., and R.M. Srivastava, 1989. *Applied Geostatistics*, Oxford University Press, Oxford.

Kernighan, B., and D.M. Ritchie, 1988. *The C Programming Language*, Prentice-Hall Inc.

Lancaster, P., and K. Salkauskas, 1986. *Curve and Surface Fitting: An Introduction*, Academic Press, Inc., London.

Laslett, G.M., A.B. McBratney, P.J. Pahl, and M.F. Hutchinson, 1987. Comparison of several spatial interpolation methods for soil pH, *Journal of Soil Science*, 38:325-341.

McBratney, A.B., and R. Webster, 1986. Choosing functions for semivariograms of soil properties and fitting them to sampling estimates, *Journal of Soil Science*, 37:617-639.

Meirvenne, M. van, and G. Hofman, 1992. Combining GIS and geostatistics for quantitative soil texture mapping, *Proceedings Third European Conference on Geographical Information Systems* (J. Harts, H. Ottens, and H. Scholten, editors), Munich, 23-26 March, Vol II, pp. 1426-1433.

Oliver, M.A., and R. Webster, 1990. Kriging: A method of interpolation for geographical information systems, *International Journal of Geographical Information Systems*, 4(3):313-332.

Pan, G.C., D. Gaard, K. Moss, and T. Heiner, 1993. A comparison between cokriging and ordinary kriging: Case study with a polymetallic deposit, *Mathematical Geology*, 25:377-398.

Pebesma, E.J., 1993. *GSTAT 0.99h — Multivariate Geostatistical Toolbox*, National Institute for Public Health and Environmental Protection, Bilthoven, The Netherlands.

Rhoades, J.D., 1982. Soluble salts, *Methods of Soil Analysis: Part 2, Chemical and Microbiological Properties* (A.L. Page and A. Klute, editors), American Society of Agronomy, Inc., Madison, Wisconsin, pp. 167-178.

Rijksuniversiteit Utrecht, Department of Physical Geography, 1993. *MISC: Miscellaneous Software Modules*, Phys. Geography., Univ. of Utrecht, Utrecht, The Netherlands.

Stein, A., 1991. *Spatial Interpolation*, Doctoral thesis, Wageningen Agricultural University, Wageningen, The Netherlands.

Tukey, J, 1977. *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts.

Van Deursen, W.P.A., 1992. *The PC-Raster Package*, Dept. of Phys. Geogr., Univ. of Utrecht, Utrecht, The Netherlands.

Webster, R., and M.A. Oliver, 1990. *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, New York.

Wesseling, C.G., D.J. Karssenberg, P.A. Burrough, and W.P.A. van Deursen, 1996. Integrating dynamic environmental models in GIS: The development of a dynamic modelling language, *Transactions in GIS*, 1:40–48.

(Received 30 January 1995; accepted 17 July 1995; revised 3 October 1995)