A.7 Geostatistical Software

Pierre Goovaerts

A.7.1 Introduction

Geostatistical spatio-temporal models provide a probabilistic framework for data analysis and predictions that build on the joint spatial and temporal dependence between observations. Since its original development in the mining industry in the late 1950s and early 1960s, the geostatistical approach has been adopted in many disciplines, such as environmental sciences (remote sensing, characterization of contaminated sediments, estimation of fish abundance), meteorology (space-time distribution of temperature and rainfall), hydrology (modeling of subsurface hydraulic conductivity), ecology (characterization of population dynamics), agriculture (maps of soil properties and crop yields), and health (patterns of diseases and exposure to pollutants). Following the increasing popularity of geostatistics, the software market has expanded substantially since the late 1980s when it was restricted more or less to two public-domain applications running under DOS: Geo-EAS (Geostatistical Environmental Assessment Software, Englund and Sparks 1988) and the Geostatistical Toolbox (Froidevaux 1990). Nowadays geostatistical software encompasses a wide range of products in terms of price, operating systems, user-friendliness, functionalities, graphical and visualization capabilities. Several organizations, such as AI-GEOSTATS (www.aigeostats.org) or the Pedometrics commission of the International Union of Soil Sciences (www.pedometrics.org), provide a fairly complete list of geostatistical freeware and commercial packages on their website; the long list could intimidate any newcomer to the field and it is summarized in Table A.7.1. The following considerations should be taken into account when choosing a geostatistical package:

(i) Does the user need to have access to the source code (i.e. graduate student who plans to implement a new approach that is a variant of existing algorithms) or is (s)he content with a black-box product?
(ii) What are the characteristics of the data? Are the observations collected in 2D or 3D? Does the sampling domain span both space and time? Are the obser-

Handbook of Applied Spatial Analysis. Edited by M.M. Fischer and A. Getis ©2009 Springer

vations available at a limited number of discrete locations or over a large raster, such as DEM or satellite imagery?

(iii) What type of analysis is envisioned? A simple description of the major spatial pattern? Straightforward prediction (i.e. univariate kriging) at unsampled locations or more complex incorporation of secondary information? A modeling of local or spatial uncertainty?

(iv) What is the level of geostatistical expertise of the user? Does user-friendliness prevail over flexibility? Is the analysis restricted to geostatistics or does it involve several steps (for example sampling design) that require additional pieces of software? Would the user favor a completely automated approach where variogram modeling is done behind the scene?

Table A.7.1. List of main geostatistical software with the corresponding reference

Name	Code	Cost[a]	Reference
Agromet	C++	*F*	Bogaert et al. (1995)
AUTO-IK	Fortran	*F*	Goovaerts (2009)
BMELib	Matlab	*F*	Christakos et al. (2002)
COSIM	Fortran	*F*	ai-geostats website
EVS (C-Tech)		*H*	C Tech Development Corporation
GCOSIM3D/ISIM3D	C	*F*	Gomez-Hernandez and Srivastava (1990)
Genstat		*F,L*	Payne et al. (2008)
GEO-EAS	Fortran	*F*	Englund and Sparks (1988)
GeoR	R	*F*	Ribeiro and Diggle (2001)
Geostat Analyst		*H*	Extension for ArcGIS
Geostatistical Toolbox		*F*	Froidevaux (1990)
Geostokos Toolkit		*H*	ai-geostats website
GS+		*M*	Robertson (2008)
GSLIB	Fortran	*F*	Deutsch and Journel (1998)
Gstat	C,R	*F*	Pebesma and Wesseling (1998)
ISATIS (Geovariances)		*H*	www.geovariances.com
MGstat	Matlab	*F*	ai-geostats website
SADA (UT Knoxville)		*F*	Spatial analysis and decision assistance
SAGE2001		*M*	Isaaks (1999)
SAS/STAT		*H*	SAS Institute Inc. (1989)
S-GeMS	C++	*F*	Remy et al. (2008)
SPRING		*F*	Camara et al. (1996)
Space-time routines	Fortran	*F*	De Cesare et al. (2002)
STIS (TerraSeer)		*M*	AvRuskin et al. (2004)
Surfer		*M*	Golden Software, Inc.
Uncert	C	*F*	Wingle et al. (1999)
Variowin		*F*	Pannatier (1996)
VESPER		*F*	Minasny et al. (2005)
WinGslib	Fortran	*L*	www.statios.com

Notes: [a] Cost: *H* high, *M* moderate, *L* low, *F* free

Each issue is discussed briefly in this chapter and appropriate software, among the ones the author is familiar with, are suggested for the main types of situation.

A.7.2 Open source code versus black-box software

As the geostatistical community increases, more researchers, particularly in academia, start sharing source code that is either posted online or published in journals such as Computers and Geosciences. Table A.7.1 (column 2) lists the programming language, such as Fortran or C++, whenever the source code is provided. While some programs require only the availability of a compiler, other routines necessitate more expensive packages, such as Matlab. Some software (for example, STIS, S-GeMS), also supports a plug-in mechanism to augment their functionalities, allowing for the addition of new geostatistical algorithms or adding supports for new types of grids on which geostatistics could be performed (Remy et al. 2008).

The Stanford Center for Reservoir Forecasting (SCRF) has been instrumental in the last 20 years in making source code for common, as well as advanced, geostatistical algorithms available to the academic community. The first attempt was the publication in 1992 of the Geostatistical Software LIBrary (GSLIB), a collection of Fortran 77 codes and executable files that cover variogram analysis, spatial interpolation and stochastic simulation (Deutsch and Journel 1998). The programs are well documented and the user manual provides both theoretical background and useful application tips. User-friendliness was greatly improved in the subsequent C++ product S-GeMS (Stanford Geostatistical Modeling Software) which offers a graphical user interface that enables interactive variogram modeling and facilitates the visualization of data and results in up to three dimensions.

Users who are statistically and computer-literate can take advantage of the rich collection of classical and modem spatial techniques implemented in the open source statistical program R (Ripley 2001). In particular, Gstat offers a robust and flexible suite of univariate and multivariate geostatistical methods for estimation and simulation. Simulation comprises conditional or unconditional (multi-) Gaussian sequential simulation of point values or block averages, or (multi-) indicator sequential simulation. The GeoR package implements model-based geostatistical methods but is limited to small (500 to 1,000 observations) univariate 2D datasets (Ribeiro et al. 2003).

Although space-time geostatistical routines are rather limited, most of these programs are public-domain. The BMElib library is a Matlab numerical toolbox that implements space/time variography and estimation using the Bayesian Maximum Entropy (BME) theory. This library is fairly complete, but it requires a strong statistical background and the Matlab package. On the other hand, Cesare et al. (2002) modified some of the GSLIB Fortran 77 routines to estimate and model space-time variograms, as well as to accommodate the use of such models in traditional kriging interpolation. Two general families of models are incorporated in the programs: the product model and the product-sum model, both based on the decomposition of the space-time covariance in terms of a spatial covariance and a temporal covariance. The commercial software STIS (Space-Time Information System) is one of the rare examples of GIS software where a time stamp is assigned to each piece of information, allowing the incorporation of time in the spatial data analysis. The geostatistical treatment of space-time data in STIS is currently limited, however, to the repetition of a purely spatial analysis for each time step, prohibiting any prediction at unmonitored times.

A.7.3 Main functionalities

As a consequence of the wide variety of geostatistical applications and the continuous development of new algorithms, finding all the functionalities required by a specific application within a single product might become increasingly difficult. Most geostatistical studies, however, share a similar sequence of steps: exploratory data analysis to get familiar with the data, characterization and modeling of the pattern of spatial variation, interpolation to the nodes of a grid or over blocks (upscaling), and modeling of local and spatial uncertainty.

Exploratory spatial data analysis. Except for a few products focusing on specific tasks, such as variography (for example, Variowin, SAGE 2001), estimation (for example, AUTO-IK, Vesper) or stochastic simulation (for example, COSIM, GCOSIM3D), most software in Table A.7.1 provides basic data mapping and exploratory tools, such as the histogram and scatterplots. These programs, however, differ in their ability to handle 3-dimensional and space-time databases, as well as dynamic exploration and visualization of the data. The S-GeMS and Uncert software offer public-domain visualization tools for three-dimensional datasets, but they lack basic GIS capabilities, such as data queries or linked windows. Such features are incorporated in the C-tech product EVS which is designed to integrate seamlessly with ESRI's ArcView® GIS and ArcGIS® or to operate in a standalone mode. Licenses for this high-end software can be expensive, however. TerraSeer STIS is less expensive and has excellent browsing and linking capability for exploratory analysis of space-time datasets in two dimensions.

Variogram estimation and modeling. Quantifying and modeling the pattern of variation in the data is the cornerstone of any geostatistical analysis. A wide range of options is available at present: from fully automated computation and modeling of variograms to highly interactive programs that allow the detection and elimination of spatial outliers (for example, variogram cloud cleaning), the exploration of spatial anisotropy through variogram maps or surfaces, and the manual fits of variograms. One of the first interactive programs for variography was Variowin (Pannatier 1996) that is public-domain. It provides several variogram estimators, and computes both variogram map and variogram cloud in addition to the traditional variogram plot. This program is limited to small 2D datasets, however, and does not include any interpolation or simulation procedure.

GIS software, such as ArcView® Geostat Analyst or TerraSeer STIS, offer similar options with better visualization capabilities than Variowin and a series of kriging and simulation algorithms that can use the variogram model in subsequent analysis. In particular, the variogram cloud in STIS is linked with the location map, which facilitates greatly the detection of data pairs with undue influence on the computation of the variogram. Other unique features of this program are the flexibility in variogram modeling (for example model parameters can be estimated automatically under constraints on the nugget effect and type of basic models), the ability to compute variograms from areal data (for example counties) and to derive the point-support model accounting for the shape and size of these geographical units (deconvolution).

Other programs, such as ArcView® Geostat Analyst or Surfer, also offer an automatic variogram modeling procedure but they either lack transparency, lead to unsatisfactory fits or do not allow anisotropic modeling. The SADA variogram module allows automatic variogram modeling as well and its exploration of anisotropy through the rose diagram is very appealing. The general-purpose statistical package Genstat (Payne et al. 2008) offers a wide variety of variogram models and allows automatic modeling, but its command language and procedure library are challenging for all users who are not statistically and computer-literate.

The SAGE2001 software can be viewed as the 3D counterpart of the 2D stand-alone Variowin software. It is not free, but it has the capability of fitting 3D models automatically. Other commercial products, such as C-tech EVS and ISATIS, also provide an automatic 3D modeling procedure that is part of their kriging module. ISATIS is certainly the most flexible software since it allows identification of directions and scales of continuity through the unique 3D interactive variogram map. Public-domain software S-GEMS and UNCERT can compute variograms in three directions but only visual fitting is implemented. To the author's knowledge, there is currently no commercial software for the geostatistical treatment of space-time data, including the interpolation at unmonitored times and locations. Current public-domain software involves a lot of data manipulation and require expert knowledge in either the modeling of the variograms (De Cesare et al. 2002) or the use of the software itself (for example BMELib).

Spatial interpolation. Basic univariate kriging variants (simple, ordinary and universal kriging) are typically covered by geostatistical software. Products differ in their ability to handle irregular interpolation grids or uneven prediction supports (i.e. change of support through block kriging), their flexibility to set up a search strategy (for example stratified search windows), or the possibility of comparing various implementation schemes by cross-validation or jack-knifing. S-GeMS is an improvement over GSLIB and GEO-EAS because it allows the specification of user-defined interpolation grids instead of the traditional regular grids in mining applications. Point measurement supports and rectangular prediction supports only are implemented, which is not adequate for applications, such as those in epidemiology or the social sciences, where the units of measurement are irregular polygons. Such levels of complexity are handled in TerraSeer STIS where both measurement and prediction supports can be either points, polygons or raster cells. In addition, it is the only commercial software that implements Poisson kriging, an interpolation procedure that is tailored to the analysis of rate data, such as crime or mortality rates.

One of the key advantages of geostatistics over other spatial interpolation procedures is its ability to incorporate secondary information, which can be available at all locations where a prediction is sought (i.e. simple kriging with a local mean or external drift) or known at a limited number of locations (cokriging). All these algorithms are implemented in the public-domain GSLIB and in the commercial software ISATIS. Kriging with an external drift is lacking from S-GeMS, whereas cokriging is not implemented in STIS or C-tech EVS.

Probability mapping. An important contribution of geostatistics is the assessment of uncertainty about unsampled values, which usually takes the form of a map of the probability of exceeding critical values, such as regulatory thresholds. Such probabilities can be estimated using parametric (i.e. multi-Gaussian kriging) or non-parametric (i.e. indicator kriging) methods. Both sets of algorithms are available in S-GeMS as well as ISATIS. Indicator kriging is also implemented in SADA and the stand-alone AUTO-IK program (Goovaerts 2009).

Stochastic simulation. Stochastic simulation has certainly been one of the most active areas of research in geostatistics for the last decade. The basic idea is to generate a set of equiprobable representations (realizations) of the spatial distribution of attribute values and to use differences among simulated maps as a measure of uncertainty. Each simulated map looks more 'realistic' than the map of smooth kriging estimates because it reproduces the spatial variation modeled from the sample information. Simulation can be done using a growing variety of techniques that differ in the underlying random function model (multi-Gaussian or non-parametric), the amount and type of information that can be accounted for and the computer requirements.

S-GeMS implements the most common algorithms (i.e. sequential indicator and Gaussian simulations), as well as recent methods based on multiple point statistics. The most complete palette of simulation methods, covering both continuous and categorical variables, is in ISATIS. These two software packages also have modules to post-process the set of realizations, creating maps of averaged simulated values, the probability of exceeding critical thresholds or measures of differences among realizations. Table A.7.2 lists other products that include stochastic simulation, either as a stand-alone algorithm (COSIM, GCOSIM3D) or as part of the geostatistical module (STIS, Uncert).

Table A.7.2. List of functionalities for main geostatistical software (modified from the list on http://www.ai-geostats.org/)

Name	Data	V	K	CK	IK	MG	S	G
Agromet	2D	X	X	X				
AUTO-IK	2D	X			X			
BMELib	3D, ST	X	X	X			X	
COSIM	2D						X	
EVS (C-Tech)	3D	X	X		X			X
GCOSIM3D/ISIM3D	3D						X	
Genstat	3D	X	X	X				
GEO-EAS	2D	X	X					
GeoR	2D	X	X				X	
Geostat Analyst	2D	X	X	X	X	X		X
Geostatistical Toolbox	3D	X	X	X				
Geostokos Toolkit	3D	X	X	X	X		X	
GS+	2D	X	X	X			X	
GSLIB	3D	X	X	X	X	X	X	
Gstat	3D	X	X	X			X	
ISATIS	3D	X	X	X	X	X	X	X
MGstat	3D, ST	X	X					
SADA	3D	X	X		X			X
SAGE2001	3D	X						
SAS/STAT	2D	X	X					
S-GeMS	3D	X	X	X	X	X	X	
SPRING	2D	X	X		X		X	X
Space-time routines	2D, ST	X	X					
STIS (TerraSeer)	2D, ST	X	X			X	X	X
Surfer	2D	X	X					
Uncert	3D	X	X				X	
Variowin	2D	X						
VESPER	2D	X	X					
WinGslib	3D	X	X	X	X	X	X	

Notes: V variography, K kriging, CK cokriging, IK indicator kriging, MG multi-Gaussian kriging, S simulation, G GIS interface

A.7.4 Affordability and user-friendliness

A package can offer all geostatistical methods developed in the last 20 years, but it can scare away potential users by its price or design. In particular for academics, price and transparency typically drive the choice of geostatistical software. Consulting companies and federal agencies are likely to favor products that do not require advanced statistical background and provide all necessary functionalities within a single package. To appeal to users that are more task-oriented than method-oriented, several products such as SADA or STIS have task managers to guide the geostatistician through the sequence of steps required to accomplish the task at hand. For example, in SADA the task ‘Interpolate my data’ consists of eleven steps, starting with ‘See the data’ and ending at ‘Add to results gallery’. This public-domain software also offers integrated modules for using the results of the geostatistical analysis in human health risk assessment, ecological risk assessment, cost/benefit analysis, sampling design, and decision analysis. On the other hand, STIS includes a complete regression module that is useful for calibrating the trend model used in multivariate kriging procedures.

Another approach to improve user-friendliness is to automate some of the steps, in particular the variogram modeling procedure which is typically the stumbling block for the adoption of kriging over more traditional methods, such as inverse distance weighting. The key is to provide transparency and use reasonable default options; for example, the user should have access to the variogram model computed behind the scene and it is puzzling that the unrealistic linear model is still used as the default variogram in Surfer. For example, C-tech MVS/EVS uses expert systems to analyze the input data, construct a multidimensional variogram which is a best fit to the dataset being analyzed, and then perform kriging in the domain to be considered in the visualization. The user is provided with the option to specify values for parameters that control the variogram and kriging procedures, and the subsequent display and analysis of the data. A public-domain alternative for 2D interpolation is VESPER that allows the automatic computation and modeling of local variograms, followed by spatial interpolation. Such a procedure capitalizes on high sampling density to adapt the process spatially to distinct local differences in the level of variation in the field. For non-parametric geostatistics, AUTO-IK is a free computer code that performs the following tasks automatically: selection of thresholds for binary coding of continuous data, computation and modeling of indicator variograms, modeling of probability distributions at unmonitored locations (regular or irregular grids), and estimation of the mean and variance of these distributions.

A.7.5 Concluding remarks

Summarizing the pros and cons of the geostatistical software currently available on the market in a few pages is a daunting task given the large number and diversity of these products. This brief chapter by no means pretends to provide a complete overview of all software, but rather offers a few pointers to guide the choice of a suitable product based on the task at hand, the user’s expertise and financial resources. The main conclusion is that there is no such thing as a ‘best all-purpose software’. Creating a geostatistical model is rarely a goal per se, but rather a preliminary step towards decision-making, such as design of a sampling or remediation scheme. The current trend is to have software that is tailored to the characteristics of the data of interest (for example areal health data, 3D pollution data, space-time climatic data) as well as the type of decision-making envisioned (for example detection of cancer clusters, estimation of volume of contaminated sediments, location of new monitoring stations). This customization of the products should improve their user-friendliness and expand their use while reducing common mistakes in the application of the geostatistical methodology.

Acknowledgments. This research was funded by grant R44-CA132347-01 from the National Cancer Institute. The views stated in this publication are those of the author and do not necessarily represent the official views of the NCI.

References

AvRuskin GA, Jacquez GM, Meliker JR, Slotnick MJ, Kaufmann AM, Nriagu JO (2004) Visualization and exploratory analysis of epidemiologic data using a novel space time information system. Int. J. Health Geog 3(1):26

Bogaert P, Mahau P, Beckers F (1995) The spatial interpolation of agro-climatic data. Cokriging Software and Source Code. Agrometeorology Series Working Paper 12, FAO Rome, Italy

Camara G, Souza RCM, Freitas UM, Garrido J (1996) SPRING: integrating remote sensing and GIS by object-oriented data modeling. Comput Graph 20(3):395-403

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS: advanced functions for field-based applications. Springer, Berlin, Heidelberg and New York

De Cesare L, Myers DE, Posa D (2002) FORTRAN programs for space-time modeling. Comput Geosci 28(22):205-212

Deutsch CV, Journel AG (1998) GSLIB: geostatistical software library and user's guide (2nd edition). Oxford University Press, New York

Englund E, Sparks A (1988) Geo-EAS 1.2.1 user’s guide. EPA Report 60018-91/008. EPA-EMSL, Las Vegas [NV]

Froidevaux R (1990) Geostatistical toolbox primer, version 1.30. FSS International, Troinex, Switzerland

Gomez-Hernandez JJ, Srivastava RM (1990) ISIM3D: an ANSI-C three dimensional multiple indicator conditional simulation program. Comput Geosci 16(4):395-440

Goovaerts P (2009) AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. Comput Geosci 35(6):1255-1270

Isaaks E (1999) SAGE 2001: a spatial and geostatistical environment for variography. Isaaks, San Mateo [CA]

Minasny B, McBratney AB, Whelan BM (2005) VESPER version 1.62. Australian Centre for Precision Agriculture, The University of Sydney, NSW

Pannatier Y (1996) VARIOWIN: Software for spatial data analysis in 2D. Springer, Berlin, Heidelberg and New York

Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2008) GenStat for Windows (11th edition). VSN International, Hemel Hempstead

Pebesma EJ, Wesseling CG (1998) Gstat: a program for geostatistical modelling, prediction and simulation. Comput Geosci 24(1):17-31

Remy N, Boucher A, Wu J (2008) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, Cambridge

Ribeiro PJR, Diggle PJ (2001) GeoR: a package for geostatistical analysis. R NEWS 1(2):15-18

Ribeiro PJR, Christensen OF, Diggle PJ (2003) GeoR and geoRglm: software for model-based geostatistics. R-NEWS 1(2):15-18

Ripley BD (2001) Spatial statistics in R. R News 1:14–15

Robertson GP (2008) GS^+: Geostatistics for the environmental sciences. Gamma Design Software, Plainwell, Michigan

SAS Institute Inc. (1989) SAS/STAT user's guide 6(2) (4th edition). SAS Institute Inc., Cary [NC]

Wingle WL, Poeter EP, McKenna SA (1999) UNCERT: geostatistics, uncertainty analysis, and visualization software applied to groundwater flow and contaminant transport modeling. Comput Geosci 25(4):365-376