ELSEVIER

Computers & Geosciences 30 (2004) 683–691

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Multivariable geostatistics in S: the gstat package☆

Edzer J. Pebesma*

Department of Physical Geography, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, Netherlands

Received 2 July 2003; received in revised form 17 March 2004; accepted 18 March 2004

Abstract

This paper discusses advantages and shortcomings of the S environment for multivariable geostatistics, in particular when extended with the gstat package, an extension package for the S environments (R, S-Plus). The gstat S package provides multivariable geostatistical modelling, prediction and simulation, as well as several visualisation functions. In particular, it makes the calculation, simultaneous fitting, and visualisation of a large number of direct and cross (residual) variograms very easy. Gstat was started 10 years ago and was released under the GPL in 1996; `gstat.org` was started in 1998. Gstat was not initially written for teaching purposes, but for research purposes, emphasising flexibility, scalability and portability. It can deal with a large number of practical issues in geostatistics, including change of support (block kriging), simple/ordinary/universal (co)kriging, fast local neighbourhood selection, flexible trend modelling, variables with different sampling configurations, and efficient simulation of large spatially correlated random fields, indicator kriging and simulation, and (directional) variogram and cross variogram modelling. The formula/models interface of the S language is used to define multivariable geostatistical models. This paper introduces the gstat S package, and discusses a number of design and implementation issues. It also draws attention to a number of papers on integration of spatial statistics software, GIS and the S environment that were presented on the spatial statistics workshop and sessions during the conference *Distributed Statistical Computing 2003*.

Keywords: Kriging; Cokriging; Linear model of coregionalisation; Open source software; S language; Stochastic simulation

1. Introduction

S is a high-level language for data analysis and graphics. Currently, it has one commercial implementation, S-Plus (S-Plus home page: http://www.insightful.com/), Becker et al., 1988; Chambers, 1998 and an open-source implementation, called "R" (Ihaka and Gentleman, 1996; Bivand, 2000; R home page: http://www.r-project.org/Comprehensive R archive network: http://cran.r-project.org/ and mirrors). Geostatistics (Isaaks and Srivastava, 1989) is not a new subject to the S community, and several S packages or libraries are available. Some of these were developed for teaching purposes, and some have very advanced functionality. Still, all of the currently available S packages lack features that are commonly used in applied geostatistics, notably block kriging, kriging in a local neighbourhood, multivariable variogram modelling, cokriging and cosimulation. This paper introduces that gstat S package, which fills this gap.

Gstat (Pebesma and Wesseling, 1998; gstat home page: http://www.gstat.org/) used to be a stand-alone computer program that provides all these features, but with no graphics capabilities of its own: it has an interactive user interface for variogram modelling, but

☆Code available from server at http://www.iamg.org/CGEditor/index.htm.

*Tel.: +31-30-2533051; fax: +31-30-2531145.

E-mail address: e.pebesma@geog.uu.nl (E.J. Pebesma).

0098-3004/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2004.03.012

uses the gnuplot graphics program for visualising variograms. The gstat stand-alone program works well with several GIS systems, as it can read and write point and/or grid map data to and from more than 20 GIS formats. Graphical user interfaces that use gstat as a back-end have been developed within PCRaster, Idrisi32 and ArcGIS environments.

The S (R/S-Plus) environment has much to offer for multivariable geostatistical analysis. The Trellis/Lattice graphics functions allow visualising high-dimensional data by creating structured, composite graphs. The gstat S package now offers the major geostatistical functionality of the gstat stand-alone program to S users, provides new functions for fast modelling of arbitrarily many cross and direct variograms, and provides a number of useful functions for plotting spatial point data, multiple grid maps, and multivariable or directional variograms. In the following, "gstat" will refer to "the gstat S package".

2. DSC2003 and spatial statistics in S

During the conference *distributed statistical computing 2003* (DSC2003) held in Vienna on March 19–22, 2003, a 1-day workshop and three paper sessions were devoted to spatial statistics, and the handling of spatial data in S environments, R in particular. The overview given by Bivand (2003) shows that at least six other R packages deal with geostatistics; three packages deal with point pattern analysis; one package deals with lattice (polygon) data and 10 packages with interfacing R to GIS formats (e.g. Bivand, 2000), of which one uses the generic spatial data abstraction layer GDAL (Gdal home page: http://www.remotesensing.org/gdal/). All of these packages share a need for S data structures that are aware of their spatial topology. An initiative for a public mailing list and a CVS repository aimed at dealing with spatial data and spatial statistics in S was started as a result of this workshop.

3. Multivariable geostatistics

Multivariable geostatistics involves the simultaneous prediction (or simulation) of multiple variables based on single or multiple predictors, as well as the modelling of all necessary direct and cross variograms. This section is meant to introduce notation for the multivariable geostatistical model as implemented in `gstat`, as briefly as possible, but necessary for the explanation of the functionality of the `gstat` package for S. Further theory is also found in various papers and text books, e.g. Cressie (1993), Ver Hoef and Cressie (1993) and Wackernagel (1998).

3.1. Univariable prediction

Let $Z(s)$ be a vector of length n with observations $Z(s_1),\ldots,Z(s_n)$ observed at spatial locations s_i arbitrarily spread in R^1, R^2 or R^3. The variability in observations $Z(s)$ is usually thought of as consisting of a trend and a residual, and the trend is modelled as a linear function

$$Z(s) = \sum_{j=0}^{p} X_j(s)\beta_j + e(s) = X\beta + e(s) \tag{1}$$

with $X_j(s), j>0$, the p explanatory or predictor variables, with β_0 usually being an intercept and $X_0(s)\equiv 1$, with β the vector with unknown regression coefficients, and with $e(s)$ the residual vector. For spatial data, residuals are usually spatially correlated, and given the covariance matrix V of $e(s)$, best linear unbiased prediction (kriging) of $Z(s_0)$ at an unobserved location s_0 is obtained by

$$\hat{Z}(s_0) = x(s_0)\hat{\beta} + v'V^{-1}(Z(s) - X\hat{\beta}) \tag{2}$$

with $x(s_0)$ the row of X that would have corresponded to $Z(s_0)$, with $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Z(s)$ the generalised least-squares estimate of the trend coefficients where X' denotes the transpose of X, and with $v = (\mathrm{Cov}(Z(s_0), Z(s_1)), \ldots, \mathrm{Cov}(Z(s_0), Z(s_n)))'$ where $\mathrm{Cov}(\cdot,\cdot)$ denotes covariance.

The corresponding prediction error variance is

$$\sigma^2(s_0) = \sigma_0^2 - v'V^{-1}v + (x(s_0) - v'V^{-1}X)(X'V^{-1}X)^{-1} \times (x(s_0) - v'V^{-1}X)', \tag{3}$$

where σ_0^2 is $\mathrm{Var}(Z(s_0))$.

3.2. Multivariable prediction

Multivariable prediction involves the joint prediction of multiple, both spatially and cross-variable correlated variables. Consider m distinct variables, and let $\{Z_i(s), X_i\ \beta^{\mathrm{i}}, e_i(s), x_i(s_0), v_i, V_i\}$ correspond to $\{Z(s), X, \beta, e(s), x(s_0), v, V\}$ of the ith variable. Next, let $\mathbf{Z}(s) = (Z_1(s)',\ldots,Z_m(\mathrm{s})')'$, $\mathbf{B} = (\beta^{1\prime},\ldots,\beta^{m\prime})'$, $\mathbf{e}(s) = (e_1(s)',\ldots,e_m(s)')'$,

$$\mathbf{X} = \begin{bmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & X_m \end{bmatrix},$$

$$\mathbf{x}(s_0) = \begin{bmatrix} x_1(s_0) & 0 & \ldots & 0 \\ 0 & x_2(s_0) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & x_m(s_0) \end{bmatrix}$$

with 0 conforming zero matrices, and

$$\mathbf{v} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,m} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,m} \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,m} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ V_{m,1} & V_{m,2} & \cdots & V_{m,m} \end{bmatrix},$$

where element i of $v_{k,l}$ is Cov($Z_k(s_i)$, $Z_l(s_0)$), and where element (i,j) of $V_{k,l}$ is Cov($Z_k(s_i)$,Z_l (s_j)).

The multivariable prediction equations equal Eqs. (2) and (3) when all matrices are substituted by their multivariable forms (see also Ver Hoef and Cressie, 1993), and when in (3) σ_0^2 is substituted by Σ with Cov($Z_i(s_0)$, $Z_j(s_0)$) in its (i,j)th element. Note that (3) is now a prediction error covariance matrix.

The implementation of this model in gstat does not pose restrictions to the number of variables m, and each variable can have its own set of predictor variables, number of observations, and unique observation locations. Covariances are specified by ways of variogram functions and cross variogram functions.

3.3. Extensions

Gstat provides a number of highly useful extensions to the straightforward application of Eqs. (2) and (3):

Kriging in a local neighbourhood: Instead of using all data, only data in a local neighbourhood around s_0 are used for predicting $\mathbf{Z}(s_0)$, where neighbourhood can be defined for each variable in terms of distance to s_0 or in terms of the number of nearest observations. There are at least two good reasons for restricting kriging to a local neighbourhood. First, the system $V^{-1}X$ becomes prohibitively large when data are abundant ($n \gg 10^3$) or when sequential simulation is used to simulate large fields. Second, the assumption of spatially constant trend coefficients in Eq. (1) may need to be relaxed to apply only to local neighbourhoods. Gstat takes care of cases where one or more of the variables are missing in a local neighbourhood, defined by a distance criterium. An efficient, scalable quadtree-based neighbourhood algorithm (Hjaltason and Samet, 1995; Quadtree demos: http://www.cs.umd.edu/~brabec/quadtree/index.html) is used to select data in a local neighbourhood.

Block kriging or simulation: Instead of predicting $Z(s_0)$ (point kriging), block kriging (Journel and Huijbregts, 1978) aims at predicting the average of $Z(\cdot)$ over a larger support (area or volume) $B_0 : Z(B_0) = |B_0|^{-1} \int_{B_0} Z(s)\,\mathrm{d}s$, with $|B_0|$ the area (or volume) of B_0. Blocks B_0 may be rectangular or irregular (specified by a number of points discretising B_0). Although the interest was originally limited to mining applications, block kriging is now widely used in environmental applications when spatially aggregated predictions for larger areas are required, or when point support predictions are too inaccurate.

Simple and ordinary kriging: In certain cases, the trend coefficients can be assumed known, e.g. when an other mechanism, such as an external deterministic model takes care of estimating them. In this case, called simple kriging, β is substituted for $\hat{\beta}$ in Eq. (2), and the third term on the right-hand side of Eq. (3) disappears. Another simplified version of universal kriging is ordinary kriging, which contains only an intercept ($p = 0$).

Shared trend coefficients and colocated cokriging: When two variables measure the same phenomenon with different devices, they will show different variability, but share a common mean value. In this case, they should be treated as two variables, having a common mean (or trend) coefficient(s). Gstat allows the sharing of any two (or more) coefficients across pairs of variables. The simplest case of this corresponds to standardised ordinary cokriging with one single unbiasedness constraint, or colocated ordinary cokriging (Goovaerts, 1997; note that this is different from Wackernagel's (1998) interpretation of colocated ordinary cokriging). Simple colocated cokriging is a special case of simple cokriging with a neighbourhood size of one for secondary variables.

Generalised linear models: Regression models for count data or for presence/absense (1/0) data are usually dealt with by generalised linear models. Gotway and Stroup (1997) extended these models to the case where residuals are spatially correlated in which case residuals have mean-related non-stationary covariances. Prediction of residuals for several variance functions are implemented in gstat.

Debugging results: Near-singularities may occur for a number of reasons, such as near-zero distances between data points, or linear dependencies among columns of a (locally formed) matrix X. Gstat has many debug modes for obtaining information on all aspects of the systems, and can verify that estimated condition numbers of V and $X'V^{-1}X$ stay below a threshold.

3.4. Sequential simulation

Sequential simulation (Johnson, 1987; Gómez-Hernández and Journel, 1993) involves the generation of many independent realisations of a Gaussian (or in case of indicator simulation, binary) random field, conditional to observed data, that honour the variogram (covariance) of the random field. Gstat uses the sequential simulation algorithm because it is versatile, efficient, and suitable for large to very large fields (number of nodes $\gg 10^6$).

Traditionally, simulation algorithms only involved the simulation of the residual part of Eq. (1), although some attempts to stretch this have been reported (Goovaerts, 1997). This can be seen as the simulation equivalent of simple kriging. Gstat implements a wider class that allows to account for statistical uncertainty on trend coefficients, using the algorithm reported (although somewhat hidden) by Abrahamsen and Espen Benth (2001). For each realisation, it involves the simulation of trend coefficients, followed by simulating residuals with respect to the trend coefficients drawn. It is the simulation equivalent of universal kriging. For the simulation of trend coefficients, the multivariate normal distribution with mean $\hat{\beta}$ and covariance $(X'V^{-1}X)^{-1}$ is used.

3.5. Variogram modelling

All methods mentioned above assume that the residual covariance is known. A common convention is to enter the covariance by ways of the variogram. Gstat calculates direct sample variograms, cross variograms (“classical” cross variograms for variables that have identical locations, pseudo-cross variograms (Ver Hoef and Cressie, 1993) when locations do not coincide), and can fit nested variogram models to sample variograms. In fitting direct and cross variogram models, it can also guarantee that the fitted model obeys the linear model of coregionalisation (Goovaerts, 1997), ensuring that cross covariance matrices are always positive definite. Furthermore, gstat can calculate and visualise directional variograms, variogram clouds, and provides identification through interactive examination (for example of extreme points) in the variogram cloud.

Variogram models may consist of simple models such as the Nugget, Exponential, Spherical, Gaussian, Linear, Power model, or the nested sum of one or more basic models. Each simple model can have its own 2D or 3D geometric or zonal anisotropy parameters defined. The gstat R package also includes the Matérn class (strongly recommended by Stein, 1999), but does not automatically fit its smoothness parameter.

4. Implementation

4.1. The S environment

S is a functional language: functions are called with data and specifications as the function arguments. The gstat S package provides a set of functions, most of which are listed in Table 1. These functions consist of about 1000 lines of S code, and for a part they hide calls to the underlying 40,000 lines of C code in the gstat

Table 1
User functions in package gstat

`Gstat`	Add variable definition to gstat object
Variogram modelling	
`variogram`	Calculate sample variogram, directional sample variograms, or direct and cross variograms
`fit.variogram`	Fit variogram model coefficients to sample variogram
`fit.lmc`	Fit a linear model of coregionalisation to direct and cross variograms
`variogram.line`	Calculates variogram values from a variogram model
Prediction/simulation	
`predict.gstat`	Spatial prediction or simulation, see also Fig. 3
`krige`	Univariable wrapper around `gstat` and `predict.gstat`
`krige.cv`	Leave-one-out or *n*-fold cross-validation wrapper for `krige`
`zerodist`	Detect observation pairs with identical locations
Plotting	
`bubble`	Bubble scatter plot for data or residuals (using colour for sign, size for value)
`plot.variogram`	Plot sample variogram (optional with number of point pairs) and fitted model; uses conditioning plots for directional or multivariable variograms (Fig. 2)
`plot.variogram.cloud`	Plot variogram cloud, with options for interactive point pairs identification
`plot.point.pairs`	Plot point pairs, identified by `plot.variogram.cloud`, in a map
`image.data.frame`	Draw image for (x,y,z) values, stored in columns of a data frame
`map.to.lev`	Stack data in the form $(x,y,z_1,z_2,\ldots,z_n)$ to a form, suitable for plotting with `levelplot`
`mapasp`	Calculate aspect ratio for geographically correct levelplot

package. Data in S are typically stored in *data frames*, tables which contain in each column one variable, of categorical (factor) or numerical mode.

4.2. Examples

In the following, examples are given that use the data set of heavy metal pollutions in the topsoil of a floodplain along the Meuse river near Stein, Netherlands (Burrough and McDonnell, 1998). This data set is supplied with gstat.

4.3. Formula interface

The gstat S package uses the S formula interface, (Chambers and Hastie, 1992), which is also found in the regression and ANOVA functions (`lm`, `aov`), generalised linear models (`glm`), and many other regression modelling or prediction methods. The first function argument is a formula, like `y~x1+x2`, to express that variable `y` depends on `x1` and `x2`, and possibly in a later argument the data frame that contains `y`, `x1` and `x2` as columns. Formulas may contain mathematical functions of variables (e.g., `sqrt(y)` instead of `y`), complex relationships (like interactions, `x1:x2`, or nested effects), and dependent variables may be factor (nominal) variables in which case they are automatically converted into the necessary set of dummy (0–1) regressor variables.

Gstat uses one formula to define how the response depends on the predictor variables, and a second formula to define the spatial coordinates. Suppose we model zinc concentrations $z(s)$ as a linear regression function of distance to the river D, $\log(z(s)) = \beta_0 + \beta_1 D(s) + e(s)$, we can calculate the residual variogram of `log(zinc)` as a function of `dist` with spatial coordinates in `x` and `y`, found in data frame `meuse` by (“>” is the S prompt):

```
> zn.vgm = variogram(log(zinc) ~ dist,
  ~ x + y, meuse)
```

which saves the results in `zn.vgm`, to be shown, plotted or fitted:

```
> zn.mod = fit.variogram(zn.vgm,
  model = vgm(1, ‘Exp‘, 300, 1))

> plot(zn.vgm, model = zn.mod)
```

fits an exponential variogram model and plots sample variogram and fitted model (Fig. 1a). By default, ordinary least-squares residuals are used, but generalised least-squares residuals given a variogram models are optional. Note that `plot` is a generic function: as its first argument is of class `variogram`, in reality the function `plot.variogram` of the gstat package is called; this function adds a number of options useful to plotting variograms.

Univariate universal kriging on locations defined in `meuse.grid`, using a fitted (residual) variogram model `zn.mod` is obtained by

```
> zn.krg = krige(log(zinc) ~ dist, ~ x + y,
  meuse, meuse.grid, zn.mod)

> levelplot(var1.pred ~ x + y, data = zn.krg,
  asp = mapasp(zn.krg))
```

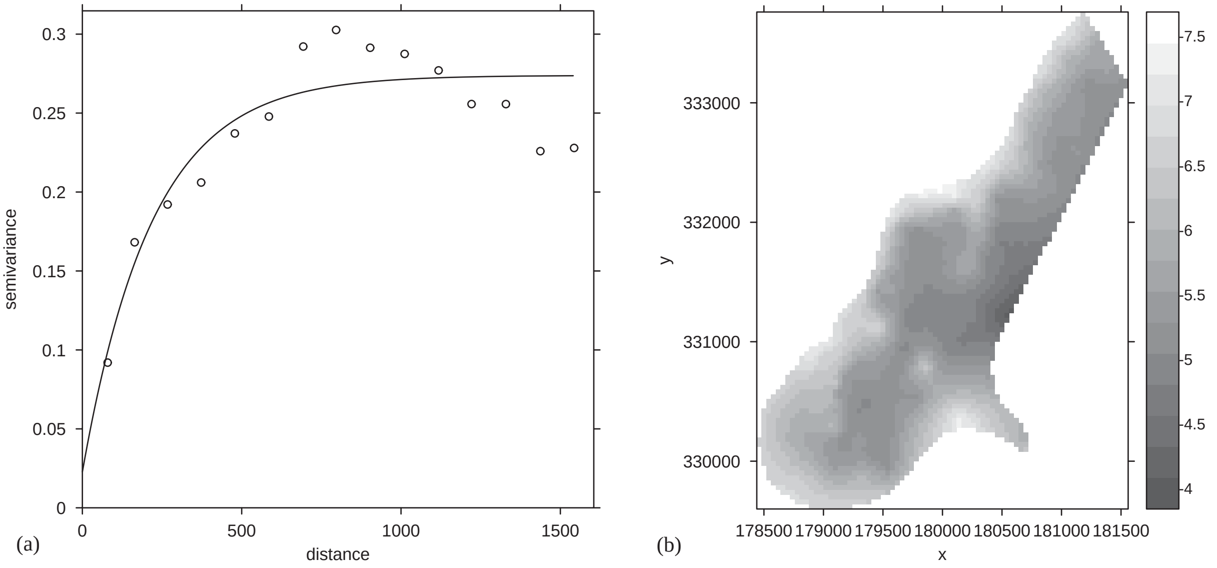

Fig. 1. Sample variogram and fitted model for log(zinc) residuals (a); universal kriging predictions for log(zinc) (b).

for which the plot is shown in Fig. 1b. Alternatively, 50 conditional simulations are obtained by

```
> krige(log(zinc) ∼ dist, ∼ x + y, meuse,
  meuse.grid, zn. mod, nmax = 20, nsim = 50)
```

where nmax refers to the neighbourhood size, limited for fast sequential simulation.

For multivariable prediction or simulation, we need to specify for each variable at least two formula's and a data frame. All this information is stored in an object of class gstat, which is built one variable at a time, by a function (surprisingly) called gstat:

```
> meuse.g = gstat(id = ‘log-zn‘,
  formula = log(zinc) ∼ 1,
  locations = ∼ x + y, data = meuse)

> meuse.g = gstat(object = meuse.g,
  id = ‘log-cu‘, formula = log(copper) ∼ 1,

  locations = ∼ x + y, data = meuse)

...
```

that can accumulate an arbitrary number of variables. Suppose meuse.g is filled with the four heavy metal variables measured in the meuse data set, then the five commands

```
> meuse.g = gstat(meuse.g, model = vgm(1, ‘Sph‘,
  900, nugget = 1), fill.all = T)

> x = variogram(meuse.g, cutoff = 1000)

> meuse.fit = fit.lmc(x, meuse.g)

> plot(x, model = meuse.fit)

> meuse.cok = predict(meuse.fit,
  newdata = meuse.grid)
```

fill all variogram models with the same initial (Nugget + Spherical) variogram model, (ii) calculate sample variograms and cross variograms, (iii) fit a linear model of coregionalisation to direct and cross variograms, (iv) plot the variograms and fitted models (Fig. 2), and (v) store four-variate cokriging predictions and prediction error (co)variances in meuse.cok.

The prediction function, predict.gstat, is the prediction and simulation engine of gstat. Depending on the data it is fed with, it decides what to do; Fig. 3 shows the decision tree for this. The list of user functions in package gstat is shown in Table 1.

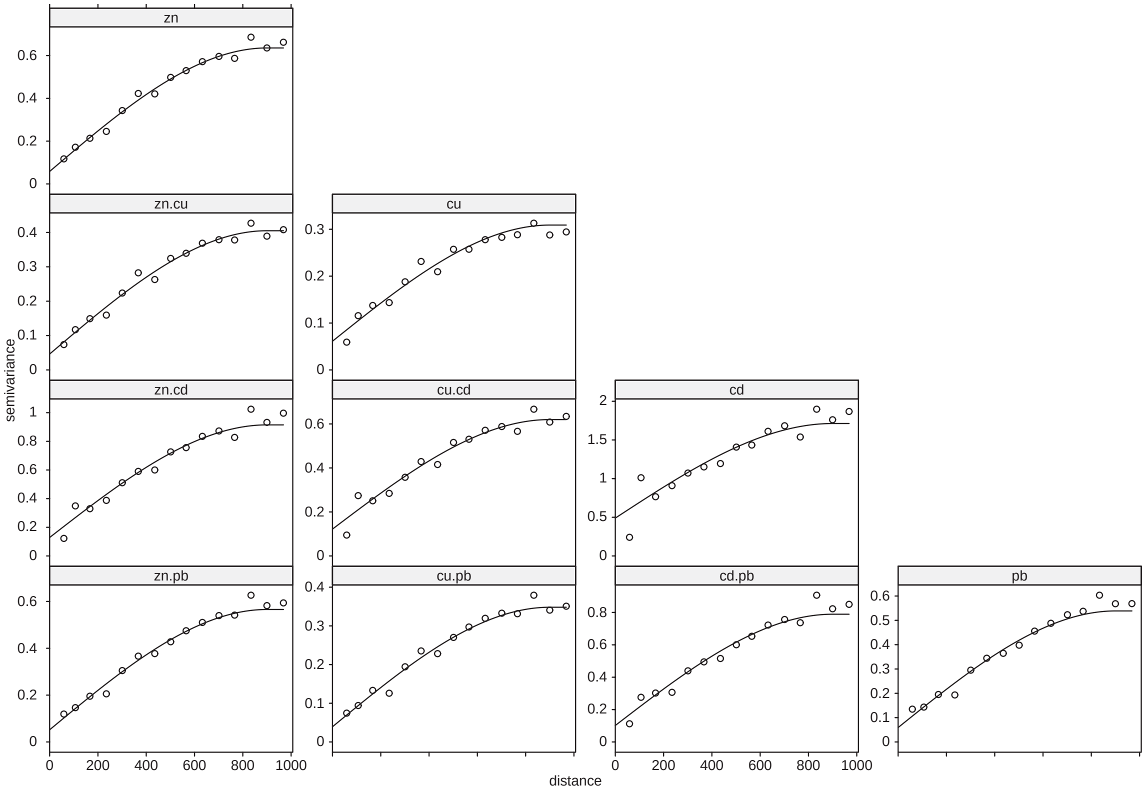

Fig. 2. Direct sample variograms (diagonal), cross variograms (off-diagonal) and fitted linear model of coregionalisation for four heavy metal variables in meuse data set.

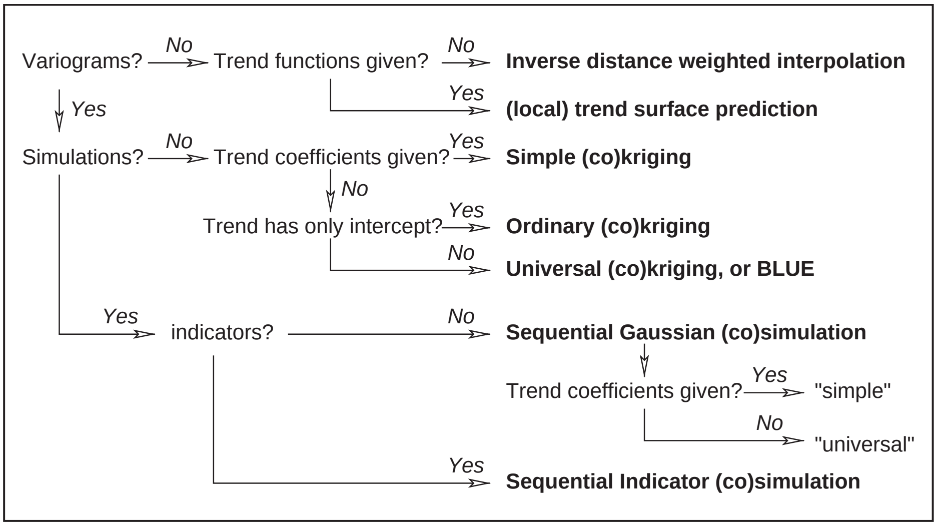
Variograms?
No
Trend functions given?
No
Inverse distance weighted interpolation
Yes
(local) trend surface prediction
Yes
Simulations?
No
Trend coefficients given?
Yes
Simple (co)kriging
No
Trend has only intercept?
Yes
Ordinary (co)kriging
No
Universal (co)kriging, or BLUE
Yes
indicators?
No
Sequential Gaussian (co)simulation
Trend coefficients given?
Yes
"simple"
No
"universal"
Yes
Sequential Indicator (co)simulation

Fig. 3. Decision tree for `predict.gstat` (or `krige`); each of prediction/simulation methods may apply to points, rectangular blocks, or irregular blocks, and may use all data or a selection of local data in a local neighbourhood around each prediction location.

The `location` argument is necessary because S data frames do not register which columns contain spatial coordinates. As this is not likely to ever change, a more elegant solution would be to use a data class that is aware of its own spatial topology, in which case the `location` formula could be left out altogether.

4.4. C *code*

The gstat C code used for the gstat package consists of approximately 25,000 lines of "native" gstat code, and 14,000 lines of C code in the Meschach matrix, library (Stewart and Leyk, 1994; Meschach home page: http://www.math.uiowa.edu/~dstewart/meschach/) used by gstat. Because originally gstat was written as a stand-alone program (Pebesma and Wesseling, 1998), a large part of the effort of writing the gstat S package was dedicated towards making the code suitable as a callable library. This involved removing many static variables, re-initialising the full state of the library after every call from S, and writing wrapper functions around all log, warning and error messages.

Two important optimised algorithms are implemented in the gstat C code. The first is a fast neighbourhood search algorithm, based on the PR-bucket quadtree search index structure (Hjaltason and Samet, 1995). The second is the realisation of many simulated random fields in a single call following a single random path through the simulation locations, re-using the expensive results, i.e. the neighbourhood selection and $V^{-1}X$.

All variogram models are defined in the gstat packages are in the gstat C code, and provides not an easy way to use variogram functions defined in S. Adding a new variogram function to the gstat C code is straightforward, though.

5. Relation to other geostatistics packages

Ripley (2001) gives a short overview of available R packages for spatial statistics. Geostatistics packages on CRAN (R home page: http://www.r-project.org/ Comprehensive R archive network: http://cran.r-project.org/ and mirrors) include `spatial`, `sgeostat`, `geoR/geoRglm` (geoRhome page: http://www.est.ufpr.br/geoR/), `fields` and `RandomFields`. Most of these packages provide variogram modelling, trend surface analysis and/or universal kriging. None of them provides kriging in a local neighbourhood, block kriging, cokriging, or three-dimensional kriging. S-Plus has a commercial module, S + SpatialStats, that provides block kriging. Large parts of the `geoR/geoRglm` (geoR home page: http://www.est.ufpr.br/geoR/) code address the uncertainty of estimated covariance parameters in a Bayesian framework (also called *model-based* kriging; Diggle et al., 1998), an issue that seems to be relevant especially for smaller data sets (Moyeed and Papritz, 2002).

6. Code availability

For R, the gstat package can be installed from CRAN (R home page: http://www.r-project.org/Comprehensive R archive network: http://cran.r-project.org/ and mirrors), which means that a single mouse click on Windows version or a single command for Unix versions is sufficient to install the package on computers with an internet connection. For S-Plus, the gstat library is available in binary form for Windows versions of S-Plus, and in source code form for Unix/Linux versions of S-Plus from the gstat home page (http://www.gstat.org/). Installation instructions are also found there.

7. Conclusions

The gstat package provides a robust and flexible suite of univariable or multivariable geostatistical methods. From the following five items:

- One-, two- or three-dimensional,
- point, regular block, or irregular block,
- univariable, multiple (uncorrelated), or multivariable (correlated) cokriging,
- (co)kriging, unconditional or conditional (co)simulation,
- using a global or a local neighbourhood,

any combination (e.g. three-dimensional universal irregular block cosimulation) can be obtained by the gstat package. Also, routines are available for very fast fitting of large numbers of direct and cross variograms. The objection to cokriging or cosimulation that the modelling of a large number of (cross) variograms is prohibitively tedious can now only be put in the past tense. The open-source gstat extension package makes the S environment (the R or S-Plus programs) a very powerful environment for (multivariable) geostatistics.

The package offers several methods for handling one or more exhaustive grids of secondary information for prediction or simulation of a primary variable:

- secondary variables can be treated as explanatory or predictor variables, leading to linear regression or universal kriging prediction (sometimes referred to as external drift kriging);
- secondary variables can be treated as (realisations of) random fields, leading to a cokriging formulation;
- colocated ordinary or simple cokriging can be used, limiting the availability of the secondary variable to that of the prediction location.

A discussion on structural differences between these approaches is found in Rivoirard (2002).

8. Discussion

8.1. S *visualisation*

One major reason why S is a suitable environment for doing multivariable geostatistics with gstat is its graphics capabilities. The gstat package gratefully uses the Trellis/Lattice functions to visualise its results, notably

- `xyplot` for visualising directional variograms and multivariable (direct and cross) variograms (e.g. Fig. 2), and to visualise spatial data and cross-validation residuals;
- `levelplot` for visualising (multiple) grid maps, using the aspect argument to make them geographically correct (1 km north equals 1 km east, a convention that even `S+SpatialStats` ignores);
- `image` for fast display of many grid maps; and
- `plot` and `identify` to identify extreme point pairs in a variogram cloud.

The graphics functions in Table 1 are no more than simple wrapper functions around the S graphics functions, but may be among the most critical ones to make a multivariable analysis successful.

8.2. *Gstat stand-alone features missing in the* S *package*

The major functionality of gstat is made available in the package, but a number of advanced features are missing. Most of them can be added easily once a common set of S data structures for spatial data (grids, lines) is defined. Gstat stand-alone features missing in the S package are: *Stratified mode*: the gstat program has an efficient way of dealing with a stratification, where each stratum has its own data, variogram and prediction locations. *Variogram maps*: two-dimensional variogram maps, calculated on a regular grid are not yet implemented in the S package. *Efficient variogram calculation for gridded data*: knowing the gridded topology of data, sample variograms can be calculated in $O(N)$, instead of $O(N^2)$. *Multi-step simulation* (Gómez-Hernández and Journel, 1993): the gstat code can use a recursively refining random visiting sequence (Pebesma and Wesseling, 1998) for sequential simulation, but needs to know the grid topology of prediction locations; currently a simple random path is chosen. *Edges*: open or closed polygons can be defined to further constrain the search neighbourhood. *Quadrant/octant search neighbourhoods*, *variogram distance*: these are other methods to refine search neighbourhoods based on direction or correlation. *Latin hypercube sampling of Gaussian random fields* (Pebesma and Heuvelink, 1999) is an issue that should be easy to re-implement in S.

8.3. *Handling spatial data in* S

Prediction locations are often gridded, and observations sometimes are. As noted above, a number of efficiency gains can be obtained when the grid topology of data, if present, is available to gstat. Storing prediction results as grids (2D matrices) can be wasteful, because large part of the area may be filled with NAs.

Currently, gstat resolves coordinates and explanatory variables at prediction locations using `model.matrix`, which requires both observation data and prediction locations to be in a data frame. Storing output of `predict.gstat` as grids might be beneficial when they are plotted with `image`, but not when plotted with `levelplot`. The conversion of table data to gridded data is close to $O(N)$ (see function `xyz2img` in package gstat).

Currently, an open-source effort (r-spatial project page: http://www.sourceforge.net/projects/r-spatial/) is being taken to provide spatial classes for R (and potentially S-Plus), for point, grid, and polygon data, and gstat supports this. It requires prior specification which variables in a data frame refer to spatial coordinates, and removes the need to specify coordinates in subsequent gstat library function calls.

Acknowledgements

The development of the gstat S package was supported financially by the Dutch National Institute for Coastal and Marine Management (RIKZ); Richard Duin (RIKZ) played a very stimulating role in the work presented here. Roger Bivand not only cotriggered the writing of the package presented here, but also organ ised the workshop and three sessions on spatial statistics during the DSC2003 conference (DSC2003: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/). One anonymous reviewer provided valuable comments.

References

Abrahamsen, P., Espen Benth, F., 2001. Kriging with inequality constraints. Mathematical Geology 33 (6), 719–744.

Becker, R.A., Chambers, J.M., Wilks, A.R., 1988. The New S Language. Chapman & Hall, London 702pp.

Bivand, R.S., 2000. Using the R statistical data analysis language on GRASS 5.0 GIS data base files. Computers & Geosciences 26, 1043–1052.

Bivand, R.S., 2003. Approaches to classes for spatial data in R. In: Hornik, K., Leisch, F. (Eds.), Proceedings of the Third International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X; available from DSC2003: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, New York, NY 431pp.

Chambers, J.M., 1998. Programming with Data. Springer, New York 469pp.

Chambers, J.M., Hastie, T.J., 1992. Statistical Models in S. Chapman & Hall, London 428pp.

Cressie, N.A.C., 1993. Statistics for Spatial Data, Revised Edn. Wiley, New York. 900 pp.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. Applied Statistics 47 (3), 299–350.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, NY 483pp.

Gómez-Hernández, J.J., Journel, A.G., 1993. Joint sequential simulation of multiGaussian fields. In: Soares, A. (Ed.), Geostatistics Tróia, Vol. 92. Kluwer, Dordrecht, pp. 85–94.

Gotway, C.A., Stroup, W.W., 1997. A generalized linear model approach to spatial data analysis and prediction. Journal of Agricultural, Biological and Environmental Statistics 2 (2), 157–178.

Hjaltason, G., Samet, H., 1995. Ranking in spatial databases. In: Egenhofer, M.J., Herring, J.R. (Eds.), Advances in Spatial Databases—Fourth Symposium, SSD'95, Lecture Notes in Computer Science, Vol. 951. Springer, Berlin, pp. 83–95. See also Quadtree demos: http://www.cs.umd.edu/~brabec/quadtree/index.html.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

Isaaks, E., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY 561pp.

Johnson, M.E., 1987. Multivariate Statistical Simulation. Wiley, New York 230pp.

Journel, A.G., Huijbregts, Ch.J., 1978. Mining Geostatistics. Academic Press, London 600pp.

Moyeed, R.A., Papritz, A., 2002. An empirical comparison of kriging methods for nonlinear spatial point prediction. Mathematical Geology 34 (4), 365–386.

Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics 41 (4), 303–312.

Pebesma, E.J., Wesseling, C.G., 1998. Gstat, a program for geostatistical modelling, prediction and simulation. Computers & Geosciences 24 (1), 17–31.

Ripley, B.D., 2001. Spatial statistics in R. R News 1 (2), 14–15.

Rivoirard, J., 2002. On the structural link between variables in kriging with external drift. Mathematical Geology 34 (7), 797–808.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York, NY 247pp.

Stewart, D.E., Leyk, Z., 1994. Meschach: matrix computations in C. Proceedings of the Centre for Mathematics and its Applications, Vol. 32. Australian National University 240pp. See also Meschach home page: http://www.math.uiowa.edu/~dstewart/meschach/.

Ver Hoef, J.M., Cressie, N.A.C., 1993. Multivariable spatial prediction. Mathematical Geology 25 (2), 219–240.

Wackernagel, H., 1998. Multivariate Geostatistics; An Introduction with Applications, 2nd Edition. Springer, Berlin 291pp.