It is worth mentioning the commercial module **S+SpatialStats** for S-PLUS, which covers all the areas mentioned here, and is **not** available for R. The prospect of such a module, and later of further development of it, has dampened enthusiasm for user-contributed spatial statistics code over much of the last decade. It is worth bearing in mind that a great deal of the current wealth of packages for R emanates from the work of users filling gaps they saw in S and its commercial offspring.

Bibliography

[1] B. D. Ripley. *Spatial Statistics*. Wiley, 1981.

Brian D. Ripley
University of Oxford, UK
ripley@stats.ox.ac.uk

geoR: A Package for Geostatistical Analysis

by Paulo J. Ribeiro Jr and Peter J. Diggle

geoR is a package to perform geostatistical data analysis and spatial prediction, expanding the set of currently available methods and tools for analysis of spatial data in R. It has been developed at the Department of Mathematics and Statistics, Lancaster University, UK. A web site with further information can be found at: http://www.maths.lancs.ac.uk/~ribeiro/geoR.html.

Preliminary versions have been available on the web for the last two years. Based on users' feedback and on our own experiences, we judge that the package has been used mainly to support teaching material and to carry out data analysis and simulation studies for scientific publications.

Package **geoR** differs from the other R tools for geostatistical data analysis in following the model-based inference methods described in (3).

Spatial statistics and geostatistics

Spatial statistics is the collection of statistical methods in which spatial locations play an explicit role in the analysis of data. The main aim of *geostatistics* is to model continuous spatial variation assuming a basic structure of the type $Y(x) : x \in \mathbb{R}^d, d = 1, 2$ or 3 for a random variable Y of interest over a region. Characteristic features of geostatistical problems are:

- data consist of *responses* Y_i associated with *locations* x_i which may be non-stochastic, specified by the sampling design (e.g. a lattice covering the observation region A), or stochastic but selected independently of the process $Y(x)$.

- in principle, Y could be determined from any location x within a continuous spatial region A.

- $\{Y(x) : x \in A\}$ is related to an unobserved stochastic process $\{S(x) : x \in A\}$, which we call the *signal*.

- scientific objectives include prediction of one or more functionals of the stochastic process $\{S(x) : x \in A\}$.

Geostatistics has its origins in problems connected with estimation of reserves in mineral exploration/mining (5). Its subsequent development, initially by Matheron and colleagues at École des Mines, Fontainebleau (8) was largely independent of "mainstream" spatial statistics. The term "kriging" was introduced to describe the resulting methods for spatial prediction. Earlier developments include work by Matérn (6, 7) and by Whittle (10). Ripley (9) re-casts kriging in the terminology of stochastic process prediction, and this was followed by significant cross-fertilisation during 1980's and 1990's (eg the *variogram* is now a standard statistical tool for analysing correlated data in space and/or time). However, there is still vigorous debate on practical issues such as how to perform inference and prediction, and the role of explicit probability models.

The Gaussian model

The currently available functions on **geoR** assume a basic model specified by:

1. a signal $S(\cdot)$ which is a stationary Gaussian process with

    (a) $\mathrm{E}[S(x)] = 0$,

    (b) $\mathrm{Var}\{S(x)\} = \sigma^2$,

    (c) $\mathrm{Corr}\{S(x), S(x-u)\} = \rho(u)$;

2. the conditional distribution of Y_i given $S(\cdot)$ is Gaussian with mean $\mu + S(x_i)$ and variance τ^2;

3. $Y_i : i = 1, \ldots, n$ are mutually independent, conditional on $S(\cdot)$.

Covariate information can be incorporated by assuming a non-constant mean $\mu = F\beta$ where F is a matrix with elements of the type $f_j(x_i)$, a measurement of the jth covariate at the ith location. The model can be made more flexible by incorporating the family of Box-Cox transformations (1), in which case the Gaussian model is assumed to hold for a transformation of the variable Y.

The basic model parameters are:

- β, the mean parameters,

- σ^2, the variance of the signal,

- τ^2, the variance of the noise,

- ϕ, the scale parameter of the correlation function.

Extra parameters provide greater flexibility:

- κ is an additional parameter, required by some models for correlation functions, which controls the smoothness of the field,

- (ψ_A, ψ_R) allows for geometric anisotropy,

- λ is the the Box-Cox transformation parameter.

Package features

The main features of the package are illustrated in the PDF document installed at 'geoR/docs/geoRintro.pdf'.
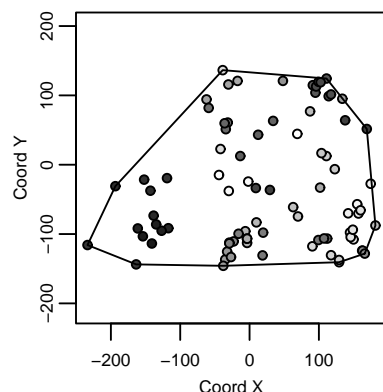

Figure 1: Output from the function `points.geodata` for the `wolfcamp` data.

There are functions available which can be used at different stages of geostatistical data analysis. Here we use the dataset `wolfcamp`, included in the package distribution, for a brief illustration of the package resources.

For **exploratory data analysis**, **geoR** uses R's graphical capabilities to produce plots of the data including their spatial locations. The most relevant functions are `plot.geodata` and `points.geodata`. Figure 1 shows an output from the latter.

Empirical variograms are used to explore the spatial structure of the data. Residual variograms can be obtained from an ordinary least squares "detrend", internal to the functions. Figure 2 shows directional and omni-directional variograms for the `wolfcamp` data assuming a first order polynomial trend on the coordinates. The main function for empirical variogram calculation is `variog`.
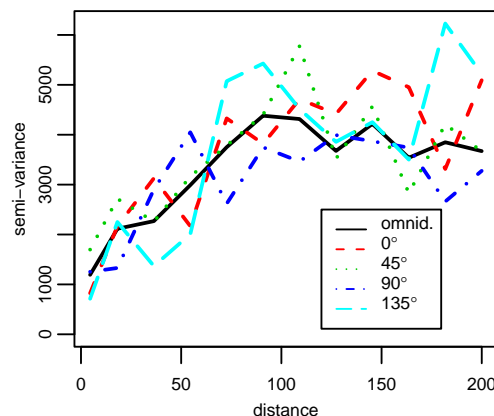

Figure 2: Directional and omni-directional variograms for the `wolfcamp` data.

Parameter estimation can be performed using different paradigms. Likelihood-based estimation (maximum and restricted maximum likelihood) is implemented by the function `likfit`. Alternatively, variogram-based estimation can be used. This consists of fitting a chosen parametric model to the sample variogram. The fitting can be done by "eye", ordinary least squares or weighted least squares. Figure 3 shows the empirical variogram and correlation function models fitted by different methods.
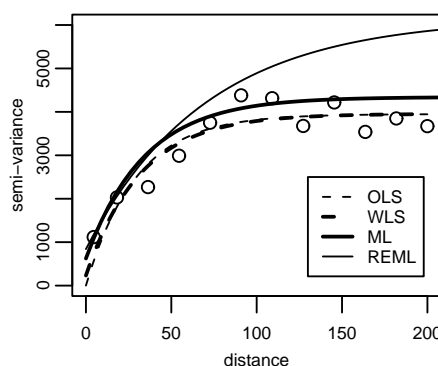

Figure 3: Directional and omni-directional variograms of the `wolfcamp` data.

The function `proflik` computes 1-D and 2-D profile likelihoods for the model parameters as illustrated by Figure 4. The profiles can be used to assess the variability of the estimates.
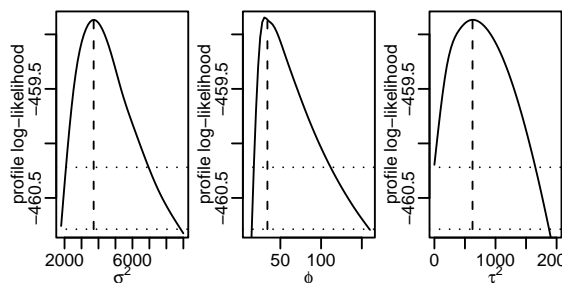
Figure 4: Profile likelihoods for the model parameters.

Spatial prediction by "plugging-in" the model parameters is performed by the function `krige.conv`. Depending on the input options the results correspond to methods known as *simple , ordinary, universal* and *external trend kriging*.

Model validation tools are implemented by the cross-validation function `xvalid`. Envelopes for the empirical variograms are computed by the functions `variog.model.env` and `variog.mc.env`.

Bayesian inference takes the parameter uncertainty into account when performing spatial prediction. Simulations of the posterior distribution $[S(x)|Y]$ allow inferences on linear and non-linear functionals of the signal $S(x)$ (Figure 5).
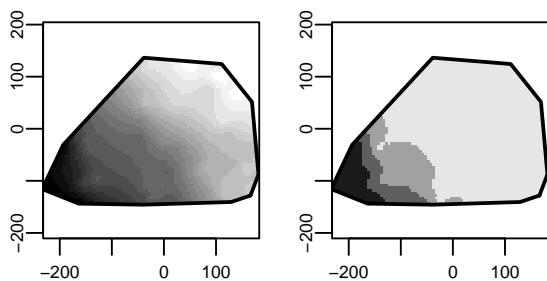

Figure 5: Predicted values over the area (left panel) and estimated probabilities of having values greater than 850 (right panel). The grey levels in right panel correspond to breaks $(0, 0.1, 0.5, 0.9, 1)$.

Simulation of (transformed) Gaussian random fields can be obtained by using the function `grf`. This function is intended for quick simulations with a small number of data locations. For simulations using large number of locations we recommend the package **RandomFields**.

Typically, the *methods* functionality is used to plot and/or print results returned by the main functions in the package. Additional details on the package resources can be found at the package's web-page and its documentation.

Future developments

The file 'docs/ToDo.geoR' is included in the package distribution and contains a list of features for which implementation is planned.

There is joint work with Ole Christensen (Aalborg University, Denmark) in progress to implement the non-Gaussian spatial models proposed by Diggle et al. (4). These models generalise the Gaussian model previously mentioned in a similar way that generalised linear models extend the classical linear regression model.

Acknowledgements

Bibliography

[1] G. E. P. Box and D. R. Cox. An analysis of transformations (with discussion). *Journal of Royal Statistical Society series B*, 26:211–252, 1964. 16

[2] N. Cressie. *Statistics for Spatial Data – revised edition*. Wiley, New York, 1993.

[3] P. J. Diggle and P. J. Ribeiro Jr. Model-based geostatistics. XIV Sinape (Simpósio Nacional de Probabiulidade e Estatística, Caxambu, MG, Brasil, July 2000. 15

[4] P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model based geostatistics (with discussion). *Applied Statistics*, 47:299–350, 1998. 17

[5] D. G. Krige. A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master's thesis, University of Witwatersrand, 1951. 15

[6] B. Matérn. *Spatial Variation*. Technical report, Statens Skogsforsningsinstitut, Stockholm, 1960. 15

[7] B. Matérn. *Spatial Variation*. Springer, Berlin, 2nd edition, 1986. 15

[8] G. Matheron. Principles of geostatistics. *Economic geology*, 58:1246–1266, 1963. 15

[9] B. D. Ripley. *Spatial Statistics*. Wiley, 1981. 15

[10] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954. 15

Paulo J. Ribeiro Jr
Universidade Federal do Paraná, Brasil

and *Lancaster University, UK*
Paulo.Ribeiro@est.ufpr.br

Peter J. Diggle
Lancaster University, UK
p.diggle@lancaster.ac.uk

Simulation and Analysis of Random Fields

by Martin Schlather

Random fields are the d-dimensional analogues of the one-dimensional stochastic processes; they are used to model spatial data as observed in environmental, atmospheric, and geological sciences. They are traditionally needed in mining and exploration to model ore deposits, oil reservoirs, and related structures.

The contributed package **RandomFields** allows for the simulation of Gaussian random fields defined on Euclidean spaces up to dimension 3. It includes some geostatistical tools and algorithms for the simulation of extreme-value random fields.

In the following two sections we give an example of an application, and a summary of the features of **RandomFields**.

A brief geostatistical analysis

To demonstrate key features we consider soil moisture data collected by the Soil Physics Group at the University of Bayreuth (see Figure 1), and perform a simple geostatistical analysis.
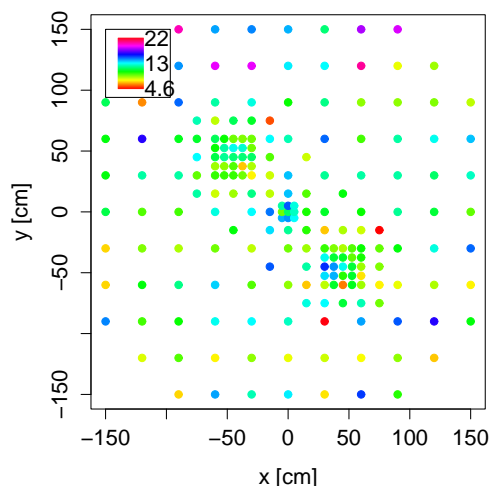

Figure 1: Measured soil moisture content (%)

The coordinates of the sample locations are assumed to be stored in `pts` and the moisture measurements in `d`. (See the example to the data set `soil` for the complete code.)

In geostatistics the variogram γ is frequently used to describe the spatial correlation structure. It can be expressed in terms of the (auto-)covariance function C if the random field is stationary: $\gamma(h) = C(0) - C(h)$ for $h \in \mathbb{R}^d$. We assume isotropy, i.e. γ depends only on the distance $|h|$. Then, we can find variogram values by

```
ev <- EmpiricalVariogram(pts, data=d,
        grid=FALSE,
        bin=c(-1, seq(0, 175, l=20)))
```

and select an appropriate model by

```
ShowModels(0:175, x=x, y=y, emp=ev)
```

see Figure 4 where x and y equal `seq(-150, 150, l=121)`. The parameters of the variogram model (here the Whittle-Matérn model) might be fitted by eye, but we prefer maximum likelihood,

```
p <- mleRF(pts, d, "whittle", param=rep(NA,5),
        lower.k=0.01, upper.k=30).
```

Now,

```
Kriging("O", x=x, y=y, grid=TRUE,
        model="whittle", par=p, given=pts,
        data=d)
```

yields the expected moisture content on the grid given by x and y (Figure 2).