gmGeostats: an R package for massive, parallelized geostatistical analysis

06/05/2014

Content

abstract A multivariate geostatistics R package is desired, that is able to deal with large data sets, massive interpolation/simulation grids and flexible variogram/spatial dependence structures, taking all the profit possible of parallelization capabilities. Apart of containing state of the art methods, it should be the framework to develop, test and disseminate our own geostatistical methods. It will be highly compatible/interfaced with our other packages (systemstats, compositions, gmDatabase and gmGeometallurgy), and specially adapted to the needs of modern and future geometallurgical analysis, regional geochemistry, and other main research topics at HIF. Eventually, should have graphical interfaces with QGIS, Rcmdr/RKWard, self-standing GUI (as an extra package?)

1 Data input, structure and spatial manipulation

Typical geostatistical software manage data sets with 2D and 3D spatial coordinates. In most of their applications, one data set is used for all steps of the geostatistical analysis. Most of the times, each variable is considered separated, even though some of them can be observed at the same locations. Two exceptions are notable. The first is MPS, where a training image must also be provided, i.e. as a sort of "training set". The second is collocated cokriging (a misnome), where a secondary variable densely sampled is used to interpolate a less measured one.

On the other hand, recent work of this group has shown how important is to consider that it is not always appropriate to consider the data as cartesian products of separate, univariate variables. Some sets of variables form natural groups, that should be analysed and interpreted together (parts in a composition, mass fractions in a particle size distribution, disjunctive indicators of facies, dip and strike of a direction, summaries of MLA images, etc). These sets of jointly defined variables are called here *layers*.

In this framework, we consider necessary to manage data of the following characteristics:

model	feature	structural tool	interpolation	simulation
Gaussian	continuous	covariance	Simple coKriging	Turning Bands (TB)
		covariogram	Ordinary coKriging	TB + ApproxChol
		residual covariance	Universal coKriging	TB + several
		MAF and rank-defficient covs	ScoK, OcoK, UcoK	TB
transGaussian	continuous	covariance	ScoK	Turning Bands (TB)
transition probs	discrete	transiograms		sequential indicator
indicator		indicator covariance	IScoK	cosimulation
				plurigaussian sim.
multipoint	discrete	training image	E-type from	MPS
	any	unsaturated GLM	simulations	MPS

- a project might contain several layers of information, and should be georeferenced (1D to 4D geographic space: borehole, plane, volume, volume× time);
- each **layer** can contain any number of collocated variables forming a system (a composition, a direction, a particle size distribution, a facies, etc): though missings should be allowed, in principle a layer should be (almost) fully observed;
- each layer might indicate the need of a **scaling function** (univariate or multivariate) and give its direct and inverse transformation function (analytical function, like an alr/ilr; Hermite polynomials for Gaussian anamorphosis; flow specification for multivariate Gaussian anamorphosis; isofactorial model?)
- it is necessary to distinguish between "**conditioning set**" (the data set that will provide the conditioning points in interpolation/simulation) and the "**training set**" (the data set that will provide the spatial structure information, e.g. the training image in MPS); note that, up to irruption of MPS, most geostatistical practice did not distinguish them;
- training sets might have relative geographic coordinates, but the scale and projection should be compatible with that of the data;
- for the moment, the range of foreseen applications suggest that it should be possible that each layer has its training set, or that each project has a unique multi-layer training set.

Further aspects of spatial data storage and manipulation that are often required, typical of GIS and similar software, are:

Domaining the geographic space can be partitioned in disjoint domains; contacts between domains might be "soft" (information can go throught them, data in both domains might contribute to interpolations-simulations) or "hard" (information should not go through the contact, interpolations-simulations should only use data on one side of the contact). Domains are usually large bodies, not necessarily continuous. These might be identified in several alternative ways: a wireframe surface (vector paradigm) or an facies-like indicator (raster paradigm).

Blocking the geographic space can be partitioned in disjoint blocks for which some averaged quantity is desired. These are mostly small, regular, continuous bodies, hence more easily identified as the interior of a grid. Blocks must often be discretized in a moderate number of internal points-segments-pixels-voxels.

Other spatial manipulations: rotation, erosion, dillation, global geometric anistotropy compensation.

2 Structural tool estimation

Up to now, most geostatistical sofwtare can manage only one type of structural tool: (cross-)variograms. MPS software is, of course, prepared to manage training images. However, the requirements on the kinds of data to manipulate and the new methods that we want to implement set these as too limiting. The following kinds of pairs [training data/structural function] should be available.

- the data used should be the training data;
- all layers enter the calculations after being applied the scaling function;
- it should be possible to use a domaining to restrict the pairs of locations (x_i, x_j) to enter the calculations only if they are in the same domain, or restrict the calculations to one single subdomain.

[real data/scalar-variogram] The basic tool of classical two-point geostats. The user should be able to select which variable treat Z, define a polar grid of lag distances h to search for pairs of data, and estimate the variogram

$$\hat{\gamma}_Z(h) = \frac{1}{2N(h)} \sum_{i,j}^{N(h)} (z(x_i) - z(x_j))^2.$$

[real data/scalar-covariance] Also a common tool of classical two-point geostats. The user should be able to select which variable treat Z, give its mean value μ, define a polar grid of lag distances h to search for pairs of data, and estimate the covariance function

$$\hat{C}_Z(h) = \frac{1}{N(h)} \sum_{i}^{N(h)} (z(x_i) - \mu)^2.$$

[real vector data/matrix-variogram] The basic tool of multivariate two-point geostats. The user should be able to select which layer(s) to treat, jointly in a column-vector $\mathbf{Z}$, define a polar grid of lag distances h to search for pairs of data, and estimate the variogram

$$\hat{\boldsymbol{\gamma}}_{\mathbf{Z}}(h) = \frac{1}{2N(h)} \sum_{i,j}^{N(h)} [\mathbf{z}(x_i) - \mathbf{z}(x_j)] \cdot [\mathbf{z}(x_i) - \mathbf{z}(x_j)]^t;$$

actual calculations might be better done by individual coordinates, to allow for some missing values in the vectors.

[compositional data/variation-variogram] The proposed tool of compositional two-point geostats. The user should be able to select a compositioanl layer to treat, jointly in a column-vector $\mathbf{Z}$, define a polar grid of lag distances h to search for pairs of data, and estimate the variogram

$$\hat{\boldsymbol{\tau}}_{\mathbf{Z}}(h) = [\hat{\tau}_{kl}(h)], \qquad \hat{\tau}_{kl}(h) = \frac{1}{2N_{kl}} \sum_{i,j}^{N_{kl}} \left(\ln \frac{z_k(x_i)}{z_l(x_i)} - \ln \frac{z_k(x_j)}{z_l(x_j)} \right)^2.$$

[categorical data/indicator-variogram] The classical tool of indicator two-point geostats. The user should be able to select which categorical variable treat Z, define a polar grid of lag distances h to search for pairs of data, and estimate the variogram

$$\hat{\gamma}_Z(h) = [\hat{\gamma}_{kl}(h)], \qquad \hat{\gamma}_{kl}(h) = \frac{1}{2N_{kl}(h)} \sum_{i,j}^{N_{kl}} (J_k(x_i) - J_l(x_j))^2;$$

where the disjunctive indicators show whether we are in one category K_k or not:

$$J_k(x_i) = \begin{cases} 1 & Z(x_i) = K_k \\ 0 & otherwise \end{cases}.$$

[categorical data/transiogram]

[categorical training image/MPS search list]

[multilayer training image/MPS search list] ???

3 Structural tool fitting/modelling

At the step of fitting/modelling, the following pairs of [structural function/model] should be offered:

[scalar-variogram or covariance/linear model of regionalization] The typical combination of valid simple models,

$$\gamma(h|\boldsymbol{\theta}) = \sum_{a=1}^{A} c_a \cdot \rho_a(h|\boldsymbol{\theta}_a)$$

with those unitary variograms $\rho_a(h|\boldsymbol{\theta}_a)$ included in table XXXX. The parameter vector $\boldsymbol{\theta}_a$ will include range and anisotropy, and can include other accessory scalars (periodicity and smoothness). Up to our present knowledge, periodicity and range must share the same anisotropy structure, and smoothness cannot be anisotropic.

[matrix-variogram or covariance/linear model of coregionalization] Also valid for compositions and indicators, as in the preceding case, it is a combination of valid simple models,

$$\gamma(h|\boldsymbol{\theta}) = \sum_{a=1}^{A} \mathbf{C}_a \cdot \rho_a(h|\boldsymbol{\theta}_a),$$

where each sill matrix $\mathbf{C}_a$ can be:

- a covariance matrix (full-rank or rank-deficient positive definite matrix) for real vector-valued layers

- a variation matrix for compositional layers (i.e. for $\boldsymbol{\tau}_{\mathbf{Z}}(h)$), possibly rank-deficient as well
- a covariance matrix with at least one singular eigenvalue, for indicators

Individual unitary variograms $\rho_a(h|\boldsymbol{\theta}_a)$ satisfy the same conditions as for the linear model of regionalization.

[**directional-covariance/complex regionalization**] Wackernagel book, paper from Sandra de Iaco

[**transiogram/Markov transition matrix**] Thesis Enayat

[**categorical MPS search list/saturated model**]

[**MPS search list/unsaturated JointSim model**] paper Strategic Mine Planning

4 Interpolation

Methods included the co/kriging following methods should be included:

- simple (co)kriging
- ordinary (co)kriging
- universal (co)kriging
- (co)kriging with a trend

Compositional particularities for a compositional layer, one should be able to manipulate missing values

Interpolation model specification

How to specify the interpolation grid

5 simulation