`GeoModels` Tutorial: simulation, estimation and prediction of positive spatial data using log-gaussian random fields

Moreno Bevilacqua
Christian Caamaño-Carrillo

Introduction

In this tutorial we show how to analyze geo-referenced spatial data with positive support using log-gaussian random fields (RFs) (Oliveira, 2006; Oliveira et al., 1997) with the `R` package `GeoModels` (Bevilacqua et al. (2018)). In particular, log-Gaussian processes have been widely used for the analysis of positive dependent data due to their well-known mathematical properties. The log-gaussian distribution is a flexible parametric model for positive data allowing right skewness.

We first load the R libraries needed for the analysis and set the name of the model in the `GeoModels` package:

```
rm(list=ls())
require(devtools)
install_github("vmoprojs/GeoModels")
require(GeoModels)
require(fields)
require(hypergeo)
model="LogGaussian"  # model name in the  GeoModels package
```

Simulation of log-Gaussian random fields

The definition of a log-Gaussian RF starts by considering a 'parent' Gaussian RF $Z = \{Z(\boldsymbol{s}), \boldsymbol{s} \in S\}$, where $\boldsymbol{s}$ represents a location in the domain S. In this tutorial, we assume $S = [0,1]^2 \subseteq \mathbb{R}^2$ and that Z is stationary with zero mean, unit variance and correlation function $\rho(\boldsymbol{h}) := \mathrm{cor}(Z(\boldsymbol{s}+\boldsymbol{h}), Z(\boldsymbol{s}))$.

Then a RF $V = \{V(\boldsymbol{s}), \boldsymbol{s} \in S\}$ with marginal distribution $LogGaussian(1, \sigma^2)$ can be derived by the transformation

$$V(\boldsymbol{s}) = exp\{\sigma Z(\boldsymbol{s}) - \sigma^2/2\} \tag{1}$$

where $\sigma > 0$ is a scale parameter. Under this specific parametrization, $\mathbb{E}(V(\boldsymbol{s})) = 1$ $\mathrm{var}(V(\boldsymbol{s})) = (\exp(\sigma^2) - 1)$ and the correlation function is given by:

$$\rho_V(\boldsymbol{h}) = \frac{\exp(\sigma^2\rho(\boldsymbol{h})) - 1}{\exp(\sigma^2) - 1}. \tag{2}$$

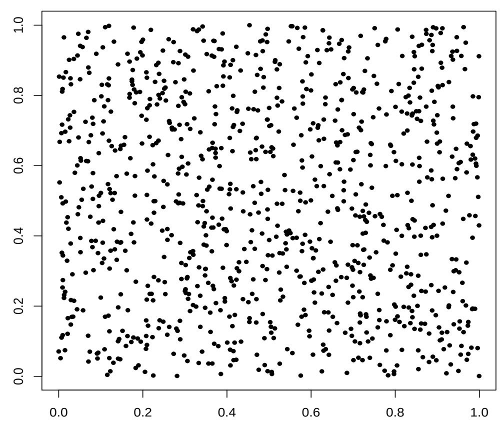
Then a non stationary version can be defined trough a multiplicative model as:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s})V(\boldsymbol{s}), \qquad \mu(\boldsymbol{s}) > 0 \tag{3}$$

with $\mathbb{E}(Y(\boldsymbol{s})) = \mu(\boldsymbol{s})$, $\text{var}(Y(\boldsymbol{s})) = \mu(\boldsymbol{s})^2(\exp(\sigma^2) - 1)$. A spatial regression model can be obtained by assuming that $\mu(\boldsymbol{s}) = e^{X(\boldsymbol{s})^T\boldsymbol{\beta}}$ where $X(\boldsymbol{s})$ is a k-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$ is a k-dimensional vector of (unknown) parameters.

Thus, in order to obtain a realization from a log-Gaussian RF we need to specify a regression mean parameters, a scale parameter and a parametric correlation model for $\rho(\boldsymbol{h})$. We first set the spatial coordinates

```
N=1000  # number of location sites
set.seed(24)
x = runif(N, 0, 1)
y = runif(N, 0, 1)
coords=cbind(x,y)  # spatial coordinates
plot(coords,pch=20,xlab="",ylab="")
```



Then we fix $k = 2$ and we build the matrix covariates and fix the regression mean parameters

```
X=cbind(rep(1,N),runif(N))     # matrix covariates
mean = -0.3; mean1=0.5         # regression parameters
nugget=0                       # nugget parameter
sill=0.5                       # scale parameter
```

where `mean` and `mean1` are respectively β_1 and β_2. Finally, we set the `sill` parameter of the log-Gaussian RF and the `nugget` parameter are respectively σ^2 and τ.

The names of the marginal parameters associated with the log-Gaussian model can be obtained with the function `NuisParam` (note that the option `num_betas` is the number of regression parameters involved):

```
NuisParam(model,num_betas=2)
[1] "mean"    "mean1"   "nugget" "sill"
```

For the correlation function we assume a special case of the isotropic Generalized Wendland class (Bevilacqua et al. (2019)) i.e the Askey model.

$$\rho(\boldsymbol{h};\alpha,\delta) := \begin{cases} (1-||\boldsymbol{h}||/\alpha)^{\delta} & ||\boldsymbol{h}||<\alpha \\ 0 & \text{otherwise} \end{cases}.$$

Using asymptotic arguments Bevilacqua et al. (2019) show that this correlation model has the same features of the exponential correlation model. Additionally it is compactly supported an interesting feature from computational point of view. We set the Askey model and the associated parameters. Note that the function `CorrParam` returns the names of the parameters associated for a given correlation model.

```
corrmodel = "Wend0"     ## correlation model
CorrParam(corrmodel)    ## names of correlation model parameter
[1] "power2" "scale"
scale = 0.2
power2 = 4
```

Here the `scale` parameter corresponds to α, the compact support of the correlation model.

We are now ready to simulate a log-Gaussian random field using the function `GeoSim`:

```
param=list(mean=mean,mean1=mean1,sill=sill, nugget=nugget,
           scale=scale,power2=power2)
set.seed(312)
data = GeoSim(coordx=coords, corrmodel=corrmodel, model=model,
          param=param, X=X,sparse=TRUE)$data
```

The simulation is performed using Cholesky decomposition. Note that the option `sparse=TRUE` allows to exploit specific algorithms for sparse matrices implemented in the `spam` package (Gerber et al. (2017)) when performing cholesky decomposition (Furrer and Sain (2010)).

Estimation of log-Gaussian random fields

The density of the bivariate random vector $(V(\boldsymbol{s}_i), V(\boldsymbol{s}_j))^T$ is given by

$$f_{\boldsymbol{V}}(v_i, v_j) = \frac{1}{2\pi\sigma^2 v_i v_j \sqrt{1-\rho^2(\boldsymbol{h})}} e^{-\frac{1}{2(1-\rho^2(\boldsymbol{h}))}\left[\left(\frac{log(v_i)}{\sigma}+\frac{\sigma}{2}\right)^2+\left(\frac{log(v_j)}{\sigma}+\frac{\sigma}{2}\right)^2-2\rho(\boldsymbol{h})\left(\frac{log(v_i)}{\sigma}+\frac{\sigma}{2}\right)\left(\frac{log(v_j)}{\sigma}+\frac{\sigma}{2}\right)\right]}, \quad (4)$$

and the bivariate densities of Y can be derived from (4) as

$$f_Y(y_i, y_j) \quad = \quad (\mu_i \mu_j)^{-1} f_V(y_i/\mu_i, y_j/\mu_j). \tag{5}$$

Given a realization $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$ of Y, then, the pairwise likelihood function is defined as:

$$pl(\boldsymbol{\theta}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} log(f_Y(y(\mathbf{s}_i), y(\mathbf{s}_j)))w_{ij}$$

where w_{ij} are non-negative weights, not depending on $\boldsymbol{\theta}$. An efficient way to specify the weights from computational and efficient viewpoint is based on neighborhoods:

$$w_{ij}(k) = \begin{cases} 1 & \mathbf{s}_i \in N_k(\mathbf{s}_j) \cup \mathbf{s}_j \in N_k(\mathbf{s}_i) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Here $N_k(\mathbf{s}_l)$ is the set of the neighbors of order $k = 1, 2, \ldots$ of the point $\mathbf{s}_l$ and in this case $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma, \alpha, \delta)^T$. The pairwise likelihood estimator $\hat{\boldsymbol{\theta}}_{pl}$ is obtained maximizing (5) with respect to $\boldsymbol{\theta}$. In the `GeoModels` package we can choose the fixed parameters and the parameters that must be estimated. Pairwise likelihood estimation is performed with the function `GeoFit`:

```
start=list(mean=mean,mean1=mean1,scale=scale,sill=sill)
fixed=list(nugget=nugget,power2=power2)
# Maximum composite-likelihood fitting of the Weibull random field:
fit = GeoFit(data=data,coordx=coords, corrmodel=corrmodel,model=model,
    X=X, optimizer="BFGS",start=start,fixed=fixed,neighb=3)
```

The object `fit` include informations about the pairwise likelihood estimation

```
fit
#################################################################
Maximum  Composite-Likelihood Fitting of Log Gaussian Random Fields
Setting: Marginal Composite-Likelihood
Model: LogGaussian
```

```
Type of the likelihood objects: Pairwise
Covariance model: Wend0
Optimizer: BFGS
Number of spatial coordinates: 1000
Number of dependent temporal realisations: 1
Type of the random field: univariate
Number of estimated parameters: 4
Type of convergence: Successful
Maximum log-Composite-Likelihood value: -2422.28
Estimated parameters:
   mean     mean1     scale      sill
-0.3431    0.5871    0.2020    0.5103
##################################################################
```

Checking model assumptions

Given the estimation of the mean $\widehat{\mu(\boldsymbol{s})} = e^{X_1(\boldsymbol{s})\hat{\beta}_1 + X_2(\boldsymbol{s})\hat{\beta}_2}$, the estimated residuals

$$\widehat{v(\boldsymbol{s}_i)} = y(\boldsymbol{s}_i)/\widehat{\mu(\boldsymbol{s}_i)} \qquad i = 1, \ldots, N \tag{7}$$

can be viewed as a realization of V a stationary RF with marginal distribution $Loggaussian(1, \sigma^2)$ with unit mean and correlation function given by (2).

The estimated residuals can be computed using the `GeoResiduals` function:

```
res=GeoResiduals(fit)  # computing residuals
```

Then the agreement of the marginal distribution assumption on the residuals with the theoretical model can be graphically checked with the `GeoQQ` function (Figure 1 left part):

```
GeoQQ(res)
```

The covariance model assumption can be checked comparing the empirical and the estimated semivariogram using the `GeoVariogram` and `GeoCovariogram` functions (Figure 1 right part). In particular the function `GeoVariogram` compute the empirical semivariogram:

```
### checking model  residuals assumptions: covariance model
vario = GeoVariogram(data=res$data,
coordx=coords,maxdist=0.3) # empirical variogram
GeoCovariogram(res,show.vario=TRUE, vario=vario,pch=20)
```

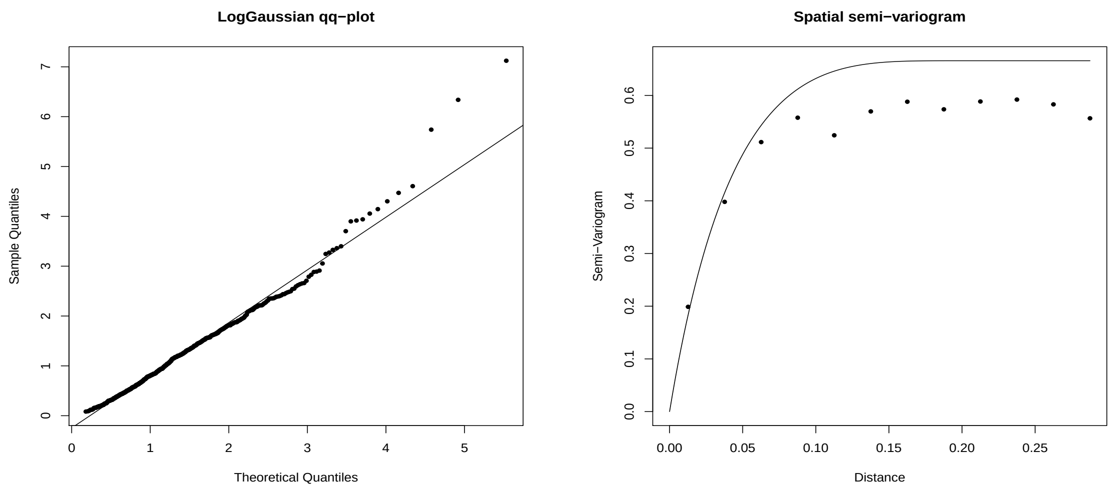


Figure 1: Left: QQ-plot for the log-Gaussian model residuals. Right: empirical vs estimated semi-variogram function for the residuals

Prediction of log-Gaussian random fields

The optimal linear prediction of log-Gaussian RF at a location $\boldsymbol{s}_0$ is given by (Bevilacqua et al. (2020)):

$$\widehat{Y(\boldsymbol{s}_0)} = \widehat{\mu(\boldsymbol{s}_0)} \left(1 + \sum_{i=1}^{N} \lambda_i [\widehat{V(\boldsymbol{s}_i)} - 1] \right) \tag{8}$$

where the vector of weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ is given by $\boldsymbol{\lambda} = R^{-1}\boldsymbol{c}$.

Here $\boldsymbol{c} = (cor(V(\boldsymbol{s}_0), V(\boldsymbol{s}_1)), \ldots, cor(V(\boldsymbol{s}_0), V(\boldsymbol{s}_n)))'$ and $R = [\text{cor}(V(\boldsymbol{s}_i), V(\boldsymbol{s}_j)]_{i,j=1}^{N}$ is the (estimated) correlation matrix associated to (2).

We first set the spatial locations to predict and the associated covariates. In this example, we choose a regular fine grid in order to construct a prediction map.

```
# locations to predict  and associated covariates
xx=seq(0,1,0.013)
loc_to_pred=as.matrix(expand.grid(xx,xx))
Nloc=nrow(loc_to_pred)
Xloc=cbind(rep(1,Nloc),runif(Nloc))
```

Then the optimal linear prediction (8), using the estimated parameters, can be performed using the `GeoKrig` function (computation can be time consuming):

```
param_est=as.list(c(fit$param,fixed))
pr=GeoKrig(data=data, coordx=coords,loc=loc_to_pred, X=X,Xloc=Xloc,
```

```
        corrmodel=corrmodel,model=model,mse=TRUE,
       sparse=TRUE,param=param_est)
```

and we can compare the map of simulated data with the kriging prediction (and associated mean square error) with the following code (see Figure 2):

```
par(mfrow=c(1,3))
quilt.plot(x, y, data,main="Data")
map=matrix(pr$pred,ncol=length(xx))
image.plot(xx, xx, map,xlab="",ylab="",main="Kriging")
map_mse=matrix(pr$mse,ncol=length(xx))
image.plot(xx, xx, map_mse,xlab="",ylab="",main="MSE")
```

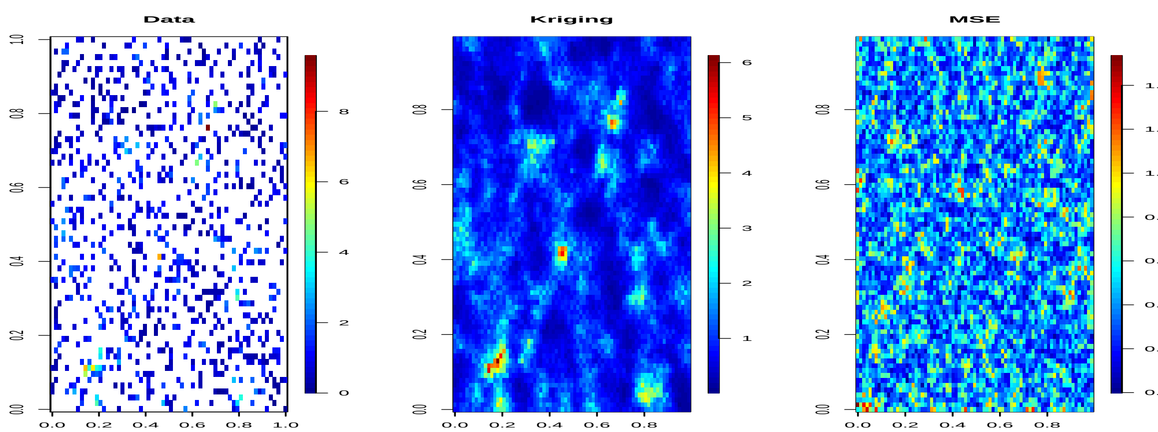


Figure 2: From left to right: colored map of observed data, kriging prediction and associated mean squared error

References

Bevilacqua, M., C. Caamaño-Carrillo, and C. Gaetan (2020). On modeling positive continuous data with spatiotemporal dependence. *Environmetrics 31*(7), e2632.

Bevilacqua, M., T. Faouzi, R. Furrer, and E. Porcu (2019). Estimation and prediction using generalized Wendland functions under fixed domain asymptotics. *The Annals of Statistics 47*, 828–856.

Bevilacqua, M., V. Morales-Oñate, and C. Caamaño-Carrillo (2018). *GeoModels: A Package for Geostatistical Gaussian and non Gaussian Data Analysis.* R package version 1.0.3-4.

Furrer, R. and S. R. Sain (2010). spam: a sparse matrix R package with emphasis on mcmc methods for Gaussian Markov random rields. *Journal of Statistical Software 36*, 1–25.

Gerber, F., K. Moesinger, and R. Furrer (2017). Extending R packages to support 64-bit compiled code: An illustration with spam64 and GIMMS NDVI3g data. *Computer & Geoscience 104*, 109–119.

Oliveira, V. D. (2006). On optimal point and block prediction in log-gaussian random fields. *Scandinavian Journal of Statistics 33*, 523–540.

Oliveira, V. D., B. Kedem, and D. A. Short (1997). Bayesian prediction of transformed gaussian random fields. *Journal of the American Statistical Association 92*, 1422–1433.