`GeoModels` Tutorial: simulation, estimation and prediction of spatial binomial data using binomial random fields.

Moreno Bevilacqua
Christian Caamaño-Carrillo

August 21, 2024

Introduction

In this tutorial we show how to analyze spatial discrete data using random fields with binomial marginal distribution using the R package `GeoModels` (Bevilacqua et al. (2018)).

We first load the R libraries needed in this tutorial.

```
rm(list=ls());
require(GeoModels);
require(fields);
```

Binomial random fields

We first define a Bernoulli random field. A simple approach for defining a Bernoulli random field $B = \{B(\mathbf{s}), \mathbf{s} \in A\}$ is by thresholding a zero mean and unit variance Gaussian random field $G = \{G(\mathbf{s}), \mathbf{s} \in A\}$ and with correlation function $\mathrm{cor}(G(\mathbf{s}), G(\mathbf{s}+\mathbf{h})) = \rho(\mathbf{h})$, namely

$$B(\mathbf{s}) = \begin{cases} 1 & G(\mathbf{s}) < \mu(\mathbf{s}) \\ 0 & \text{otherwise} \end{cases}.$$

where $\mu(\mathbf{s})$ is a non random function.

With this definition we have that the marginal probability of success is given by

$$p(\mathbf{s}) := \Pr(B(\mathbf{s}) = 1) = \Phi\left(\mu(\mathbf{s})\right), \tag{1}$$

where Φ is the univariate standard Gaussian cumulative distribution function.

The main idea to obtain a random field with marginal binomial distribution is to sum for each $\mathbf{s}$, $n(\mathbf{s})$ independent copies of B that is we define a random field $Y = \{Y_{n(\mathbf{s})}(\mathbf{s}), \mathbf{s} \in A\}$ as:

$$Y_{n(\mathbf{s})}(\mathbf{s}) := \sum_{i=1}^{n(\mathbf{s})} B_i(\mathbf{s}) \tag{2}$$

where $n(\mathbf{s})$ is positive natural number. Then for each site $\mathbf{s}$, the binomial random field Y represents the number of successes in $n(\mathbf{s})$ independent trials .

It turns out that the marginal distribution is binomial *i.e*:

$$\Pr(Y_{n(\mathbf{s})}(\mathbf{s}) = u) = \binom{n(\mathbf{s})}{u} p(\mathbf{s})^u (1-p(\mathbf{s}))^{n(\mathbf{s})-u}, \quad u = 0, 1, 2, \ldots, n(\mathbf{s})$$

where mean and variance are given by $\mathbb{E}(Y_{n(\mathbf{s})}(\mathbf{s})) = n(\mathbf{s})p(\mathbf{s})$, $\text{var}(Y_{n(\mathbf{s})}(\mathbf{s})) = \mathbb{E}(Y_{n(\mathbf{s})}(\mathbf{s}))(1-p(\mathbf{s}))$ respectively. In addition, it can be shown that the correlation function is given by

$$Cov(Y_{n(\mathbf{s}_i)}(\mathbf{s}_i), Y_{n(\mathbf{s}_j)}(\mathbf{s}_j)) = min(n(\mathbf{s}_i), n(\mathbf{s}_j))(p_{11}(\mathbf{h}) - p(\mathbf{s}_i)p(\mathbf{s}_j)) \tag{3}$$

where $p_{11}(\mathbf{h})$ is the probability of joint success at locations $\mathbf{s}_i$ and $\mathbf{s}_j$ in the Bernoulli RF that is

$$p_{11}(\mathbf{h}) = Pr(B(\mathbf{s}_i) = 1, B(\mathbf{s}_j) = 1) = \Phi_2(\mu(\mathbf{s}_i), \mu(\mathbf{s}_j), \rho(\boldsymbol{h}))$$

and $\Phi_2(\cdot, \cdot, \rho(\boldsymbol{h}))$ is the bivariate standard Gaussian CDF with correlation $\rho(\boldsymbol{h})$. Note that when $p(\mathbf{s})$ and $n(\mathbf{s})$ do not depend on $\mathbf{s}$ then the random field is stationary

Stationary Binomial random fields

We first set the spatial coordinates:

```
set.seed(269);
N=1000;
coords=cbind(runif(N),runif(N));
```

Let us assume $\mu(\boldsymbol{s}) = \mu$ which implies, using (1), that the probability of success is constant $p(\boldsymbol{s}) = p$. In addition we assume a constant number of trials, $n(\mathbf{s}) = n$. This implies that we are assuming a constant mean and variance for the binomial random field. To obtain a simulation from the binomial random field we need to specify the (constant) mean, a parametric correlation model $\rho(\boldsymbol{h})$ for the underlying Gaussian random field and the (fixed) number of trials n.

For the correlation function $\rho(\boldsymbol{h})$ of the "parent" Gaussian random field G we assume an isotropic Matérn model (Matérn, 1986):

$$\rho_{\alpha,\gamma}(\boldsymbol{h}) = \frac{2^{1-\gamma}}{\Gamma(\gamma)} \left(\frac{\|\boldsymbol{h}\|}{\alpha} \right)^{\gamma} \mathcal{K}_{\gamma} \left(\frac{\|\boldsymbol{h}\|}{\alpha} \right), \qquad \|\boldsymbol{h}\| \geq 0. \tag{4}$$

where $\mathcal{K}_{\gamma}$ is a modified Bessel function of the second kind of order γ, $\gamma > 0$ is the smoothness parameter and $\alpha > 0$ the spatial scale parameter. Then, we set the parameter associated to this correlation model:

```
corrmodel = "Matern";
scale = 0.25/3;
smooth =0.5;
nugget =0;
```

We need to fix the number of trials n.

```
n=10
```

Finally we set the mean parameter, $\mu = 0.5$. This implies a marginal (constant) probability of success $p = \Phi(0.5) = 0.691$. We are now ready to simulate a realization of the Binomial random field using the function *GeoSim*:

```
mean = 0.5 #  mean parameter
param=list(nugget=nugget,mean=mean, scale=scale, smooth=smooth);
data_s <- GeoSim(coordx=coords ,corrmodel=corrmodel ,
        param=param ,model="Binomial",n=n)$data;
```

Note that empirical mean (variance) is very close to the theoretical mean (variance) as expected:

```
mean(data_s);var (data_s)
[1] 6.985
[1] 2.229004
 p=pnorm(mean)
 n*p;n*p*(1-p)
[1] 6.914625
[1] 2.133421
```

The following Figure shows the histogram of the data and the associated coloured map

```
par(mfrow=c(1,3))
plot(table(data_s),ylab = "Frequency")
quilt.plot(coords,data_s,nlevel=n+1,zlim=c(0,n))
```

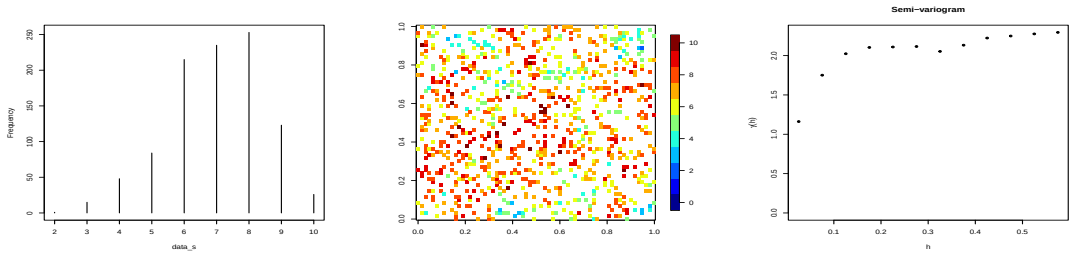

Figure 1: Histograms and coloured maps of a realization from a stationary binomial random field. On the right the associated empirical semivariogram estimation.

A graphical representation of the empirical semivariogram estimation can be obtained using the *GeoVariogram* function (see Figure 2 right part).

```
fit = GeoVariogram(coordx=coords,data=data_s,maxdist=0.6)
plot(fit, xlab='h', ylab=expression(gamma(h)),
     ylim=c(0, max(fit$variograms)), pch=20,
     main="Semi-variogram")
```

Non-stationary Binomial random fields

A non stationary binomial random field can be specified by assuming that $\mu(\boldsymbol{s})$ is not constant and/or $n(\mathbf{s})$ is not constant. We assume a linear specification $\mu(\boldsymbol{s}) = X(\boldsymbol{s})^T\boldsymbol{\beta}$ where $X(\boldsymbol{s})$ is a k-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is a k-dimensional vector of (unknown) parameters. In this tutorial we assume $k = 2$.

Thus, in order to obtain a realization from a non stationary binomial random field we need to specify the mean parameters and a parametric correlation model $\rho(\boldsymbol{h})$ for the "parent" Gaussian random field. For the correlation function we use, as in the stationary case, the Matern model in Equation (4) with the same parameters as in the stationary case.

Finally we set the mean parameters and the regression matrix:

```
mean = 0.3 # regression paramteres
mean1= -0.25
set.seed(29)
a0=rep(1,N);a1=runif(N)
X=cbind(a0,a1); ## regression matrix
```

and the vector of (non-constants) trials

```
n=sample(10:20,nrow(coords),replace=TRUE)
head(n)
[1] 15 19 18 10 18 19
```

We are ready to simulate a non stationary binomial random field using the function *GeoSim*:

```
param=list(nugget=nugget,mean=mean,mean1=mean1, scale=scale,
           smooth=smooth);
data_ns<- GeoSim(coordx=coords ,corrmodel=corrmodel , param=param ,n=n,
                 X=X,model="Binomial")$data;
```

Estimation of binomial random fields

Let y_j denotes a realization of the random variable $Y_{n(\mathbf{s}_j)}(\mathbf{s}_j)$. Hereafter, for notation simplicity we set $n_j = n(\mathbf{s}_j)$ and $p_j = p(\mathbf{s}_j)$, $\mu_j = \mu(\mathbf{s}_j)$ and $p_{11} = p_{11}(\mathbf{h})$. Given a realization $\boldsymbol{y}_n = (y_1, \ldots, y_l)^T$ of the binomial random field, observed at $(\mathbf{s}_1, \ldots, \mathbf{s}_l)$ location sites, the estimation of the regression and correlation parameters can be performed using composite likelihood estimation based on pairs.

Let $\Pr(Y_{n_i}(\mathbf{s}_i) = y_i, Y_{n_j}(\mathbf{s}_j) = y_j)$ the probability density of the bivariate random vector $(Y_n(\boldsymbol{s}_i), Y_n(\boldsymbol{s}_j))^T$, then the conditional pairwise likelihood function is defined as:

$$pl(\boldsymbol{\theta}) = \sum_{i=1}^{l} \sum_{j=1, j\neq i}^{l} c_{ij}[log(\Pr(Y_{n_i}(\mathbf{s}_i) = y_i, Y_{n_j}(\mathbf{s}_j) = y_j) - log(\Pr(Y_{n_j}(\mathbf{s}_i) = y_j))], \qquad (5)$$

where in this case $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)^T$ and c_{ij} are non-negative symmetric weights. Using results in Marshall and Olkin (1985) the bivariate distribution is given by:

$$\Pr(Y_{n_i}(\mathbf{s}_i) = y_i, Y_{n_j}(\mathbf{s}_j) = y_j) \quad = \quad \begin{cases} \sum_{k=0}^{n_i - n_j} \binom{n_i - n_j}{k} p_i^k (1-p_i)^{n_i - n_j - k} [b(y_i - k, y_j; p_i, p_j; n_j)] & n_i > n_j; \\ b(y_i, y_j; p_i, p_j, n) & n = n_i = n_j; \\ \sum_{k=0}^{n_j - n_i} \binom{n_j - n_i}{k} p_j^k (1-p_j)^{n_j - n_i - k} [b(y_j - k, y_i; p_j, p_i; n_i)] & n_j > n_i; \end{cases}$$

where

$$b(x_1, x_2; v_1, v_2; N) = \quad = \quad \sum_{a=max(0, x_1+x_2-N)}^{min(x_1, x_2)} \binom{N}{a, x_1 - a, x_2 - a, N - x_1 - x_2 + a} p_{11}^a (v_1 - p_{11})^{x_1 - a} (v_2 - p_{11})^{x_2 - a} (1 + p_{11} - v_1 - v_2)^{N - x_1 - x_2 + a}.$$

An efficient way to specify the weights from computational and efficient viewpoint is based on neighborhoods:

$$c_{ij}(k) = \begin{cases} 1 & \mathbf{s}_i \in E_k(\mathbf{s}_j) \\ 0 & \text{otherwise} \end{cases} \qquad (6)$$

Here $E_k(\mathbf{s}_l)$ the set of the neighbors of order $k = 1, 2, \ldots$ of the point $\mathbf{s}_l$.

The conditional 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 can be estimated. Conditional pairwise likelihood estimation can be performed using the function *GeoFit2*. The maximization can be performed using different algorithms. In this case we consider the option *nlminb* that allows for box constrained optimization using PORT routines.

As a consequence we need to specify the lower and upper bound of the parametric space of the parameters to be estimated.

```
fixed2<-list(nugget=0,smooth=smooth) # fixed parameters
start2<-list(mean=mean,mean1=mean1,scale=scale) # starting paramters
lower<-list(mean=-5,mean1=-5,scale=0);
upper<-list(mean=5,mean1=5,scale=10);

fit1_ns<- GeoFit2(data=data_ns,coordx=coords,corrmodel=corrmodel,
  likelihood="Conditional",type="Pairwise",n=n, X=X, neighb=4,
  sensitivity=TRUE,optimizer="nlminb",lower=lower,upper=upper,
  start=start2,fixed=fixed2, model="Binomial");
```

Note that the option *neighb=4* set the neighborhood order of the weight function (6) i.e. $k = 4$.

The object `fit1_ns` includes informations about the conditional pairwise estimation including the estimates:

```
##################################################################
##################################################################
Maximum  Composite-Likelihood Fitting of Binomial Random Fields
Setting: Conditional Composite-Likelihood
Model: Binomial
Type of the likelihood objects: Pairwise
Covariance model: Matern
Optimizer: nlminb
Number of spatial coordinates: 1000
Number of dependent temporal realisations: 1
Type of the random field: univariate
Number of estimated parameters: 3
Type of convergence: Successful
Maximum log-Composite-Likelihood value: -7668.58
Estimated parameters:
    mean     mean1     scale
 0.31756  -0.23810   0.07267
##################################################################
```

Standard error estimation can be performed using parametric bootstrap using the function *GeoVarestbootstrap*. The procedure is simulation based and can be time consuming. Note

that, the function *GeoVarestbootstrap* return the same object, updated with the standard error informations.

```
fit1_ns_sd=GeoVarestbootstrap(fit1_ns,K=100,
                    optimizer="nlminb",lower=lower, upper=upper)
 mean       mean1       scale
0.03429359 0.02470289 0.01110608
```

Prediction of binomial random fields

For a given spatial location $\boldsymbol{s}_0$ with associated covariates $X(\boldsymbol{s}_0)$ and trial number n_0, the optimal linear prediction of a binomial random field is given by:

$$\widehat{Y_{n_0}(\mathbf{s}_0)} = \mathbb{E}(Y_{n_0}(\mathbf{s}_0)) + \boldsymbol{c}^T\Sigma^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}) \tag{7}$$

where $\boldsymbol{\mu} = (\mathbb{E}(Y_{n_1}(\mathbf{s}_1)), \ldots, \mathbb{E}(Y_{n_l}(\mathbf{s}_l))^T$ with $\mathbb{E}(Y_{n_i}(\mathbf{s}_i)) = n_i p_i$, and $\boldsymbol{c} = [Cov(Y_{n_0}(\mathbf{s}_0), Y_{n_i}(\mathbf{s}_l)]_{i=1}^{l}$ and $\Sigma = [Cov(Y_{n_j}(\mathbf{s}_i), Y_{n_j}(\mathbf{s}_j))]_{i=j=1}^{l}$. The covariance matrix Σ and the covariance vector $\boldsymbol{c}$ can be computed using (3).

The associated mean squared error is given by:

$$MSE(\widehat{Y_n(\mathbf{s}_0)}) = \text{var}(Y_{n_0}(\mathbf{s}_0)) - \boldsymbol{c}^T\Sigma^{-1}\boldsymbol{c}. \tag{8}$$

Kriging and associated MSE can be obtained using the `GeoKrig` function using the estimated parameters. We first need to specify the spatial locations (and the associated values of the covariates) to predict and, in this example, we consider a spatial regular grid:

```
xx=seq(0,1,0.025)
loc_to_pred=as.matrix(expand.grid(xx,xx))
NN=nrow(loc_to_pred)
a0=rep(1,NN);a1=runif(NN)
Xloc=cbind(a0,a1);;
```

In addition, we need the number of trials associated with the location to predict.

```
nloc=sample(10:20,nrow(loc_to_pred),replace=TRUE)
```

Then the optimal linear prediction (7), using the estimated parameters, can be performed using the `GeoKrig` function for both models:

```

pr=GeoKrig(fit1_ns,loc=loc_to_pred,Xloc=Xloc,,nloc=nloc,mse=TRUE)
```

Finally, a kriging map with associate mean square error (Figure 2) can be obtained with the following code for the binomial case:

```
par(mfrow=c(1,3))
#### map of  data
nn=max(n)
quilt.plot(coords, data_ns,nlevel=nn+1,zlim=c(0,nn))
# map predictions
map=matrix(pr$pred,ncol=length(xx))
image.plot(xx,xx,map,xlab="",zlim=c(0,nn),ylab="",main="Kriging")
#map MSE
map_mse=matrix(pr$mse,ncol=length(xx))
image.plot(xx,xx,map_mse,xlab="",ylab="",main="MSE")
```

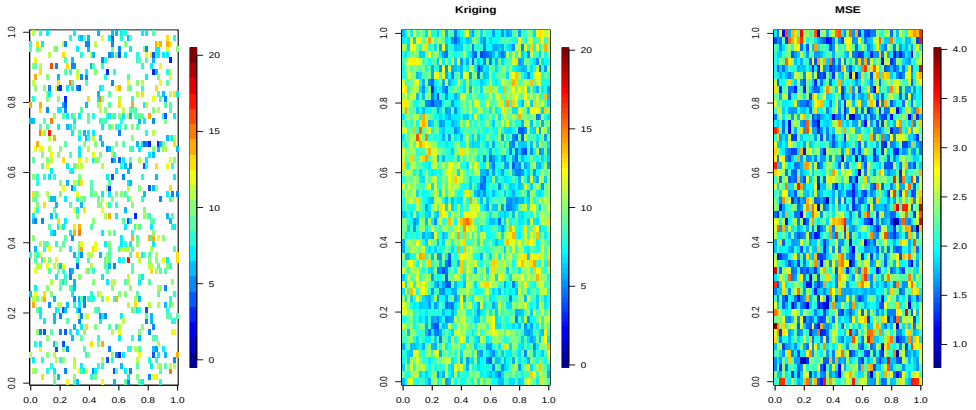

Figure 2: From left to right: observed spatial data, associated kriging map and mean square error map for a nonstationary binomial random field.

References

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.

Marshall, A. W. and I. Olkin (1985). A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association 80*(390), 332–338.

Matérn, B. (1986). *Spatial Variation: Stochastic Models and their Applications to Some Problems in Forest Surveys and Other Sampling Investigations* (2nd ed.). Heidelberg: Springer.