ISCAMI 2012
Malenovice

Geostatistical Methods in R

Adéla Volfová, Martin Šmejkal

Czech Technical University in Prague, Faculty of Civil Engineering
Thakurova 7, 166 29 Praha 6, Czech Republic
adelavolfova@gmail.com, martin.smejkal@fsv.cvut.cz

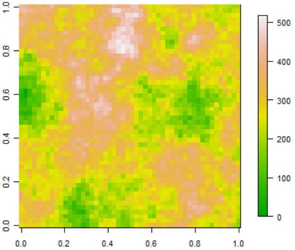
Geostatistics

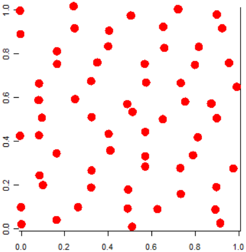
Geostatistics is a branch of statistics that deals with spatial data. It is a tool for:

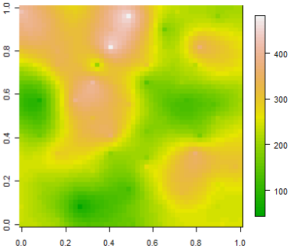
- analyzing spatial data,
- describing spatial continuity,
- predicting:
  - ordinary kriging,
  - ordinary cokriging.

Spatial data carry information of a natural phenomenon including location.

Prediction – Goal of Geostatistics







Project R

R is a language and an environment for statistical computations and creating graphics.

- Free Software (GNU General Public License),
- multi–platform,
- easy to learn,
- with a huge amount of additional packages.

There are tens of packages for geostatistics – the best known are `geoR`, `gstat`, `sp`.

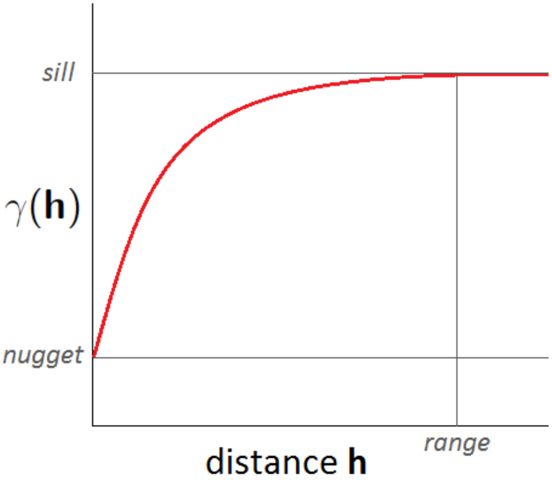
Univariate Geostatistics

- Data exploration,
- finding variogram model (spatial dependence),
- prediction (ordinary kriging).

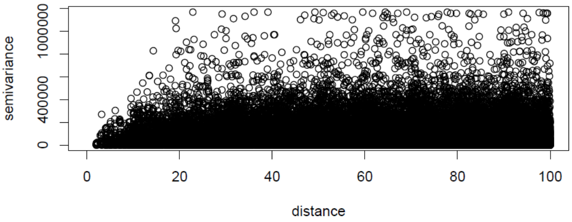
Variogram is a geostatistical tool describing the spatial dependency.

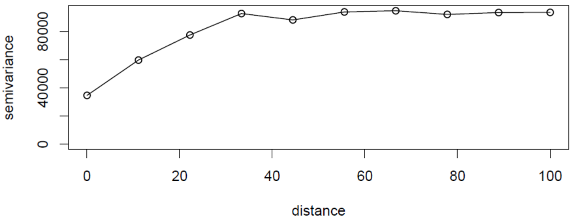
$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{n} [Z(x_i) - Z(x_i + h)]^2,$$

Variogram

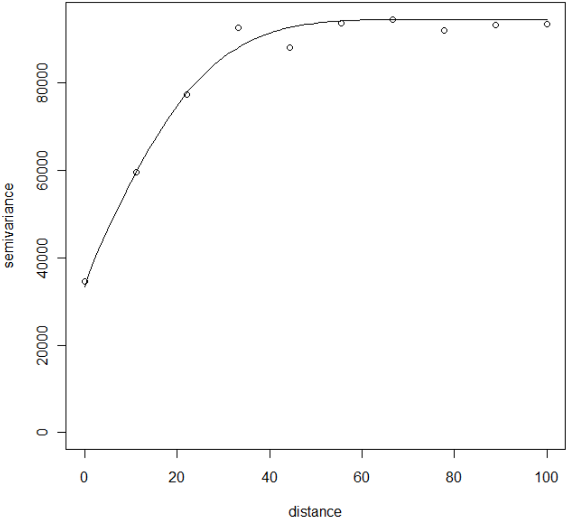


Empirical Variogram





Theoretical Variogram



Ordinary Kriging

Unsampled location – x_0.

$$Z^*(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha),$$

where λ_α is a weight for value $Z(x_\alpha)$ at x_α.
Ordinary kriging is aliased BLUP (best linear unbiased predictor) and therefore:

- a sum of weights is equal to 1 (guarantees the unbiasedness of the prediction),
- a variance of estimation errors is minimal.

Ordinary Kriging

The vector of weights:

$$\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} C_{10} \\ \vdots \\ C_{n0} \\ 1 \end{bmatrix},$$

where μ is a Lagrange parameter and C_{ij} is a covariance between $Z(x_i)$ and $Z(x_j)$.
A relationship between a covariance and a variogram:

$$C_{ij} = Cov(Z(x_i), Z(x_j)) = C(0) - \gamma(x_i - x_j),$$

where $C(0)$ is the *sill* of the variogram model.

Multivariate Geostatistics

Secondary variable enhances prediction of primary variable.
Linear model of coregionalization – spatial dependency of two or more variables.
Cross-variogram:

$$\gamma_{12}(h) = \frac{1}{2}E[(Z_1(x+h) - Z_1(x))(Z_2(x+h) - Z_2(x))].$$

Pseudo cross-variogram:

$$\psi_{12}(h) = \frac{1}{2}E[(Z_1(x+h) - Z_2(x))^2].$$

Ordinary Cokriging

$$Z^*(x_0) = \sum_{S_1} \lambda_{1\alpha} Z_1(x_\alpha) + \sum_{S_2} \lambda_{2\alpha} Z_2(x_\alpha),$$

where S_1 and S_2 are sets of samples for the primary and secondary variables respectively.
The following hold:

- the sum of weights $\lambda_{1\alpha}$ is equal to 1 and the sum of weights $\lambda_{2\alpha}$ is equal to 0 (guarantees the unbiasedness of the prediction),
- a variance of estimation errors is minimal.

Ordinary Cokriging

A relationship between a cross–variogram and a cross–covariance is:

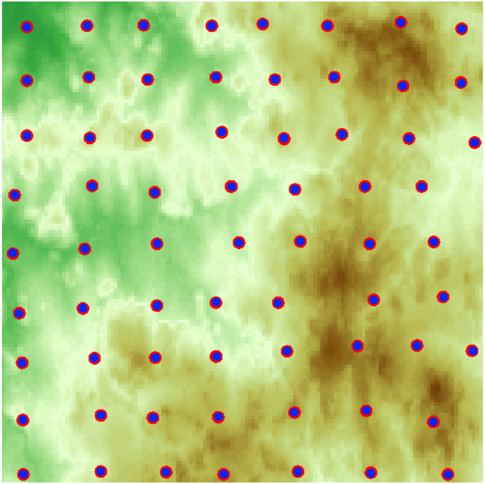
$$\gamma_{12}(h) = C_{12}(0) - \frac{C_{12}(h) + C_{21}(h)}{2},$$

Then, ordinary cokriging system in matrix form is given as:

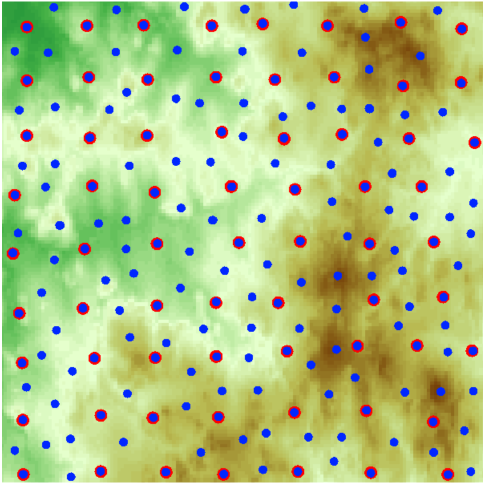
$$\begin{bmatrix} C_{11} & C_{12} & 1 & 0 \\ C_{21} & C_{22} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} C_{01} \\ C_{02} \\ 1 \\ 0 \end{bmatrix},$$

where C_{11} and C_{22} are covariance matrices of primary and secondary variables respectively, and C_{12} is a cross–covariance matrix.

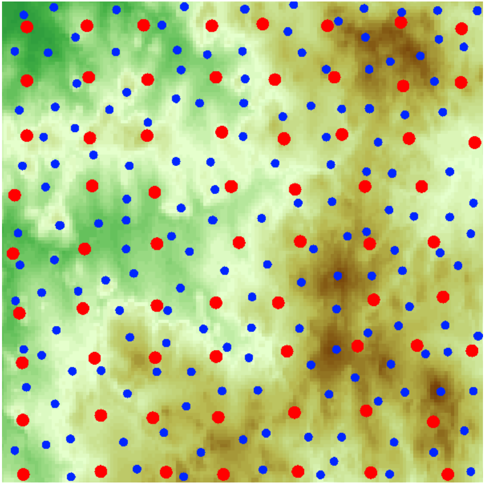
Sample Locations of Covariables



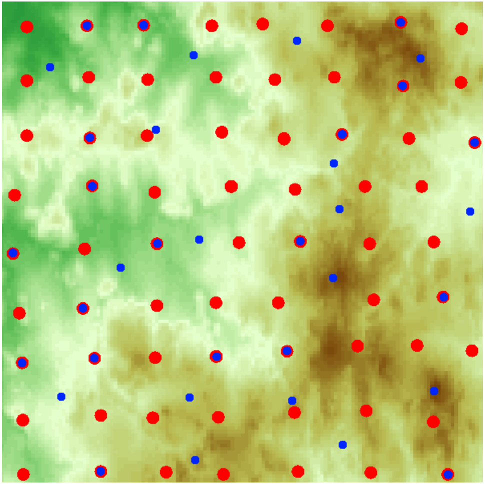
Sample Locations of Covariables



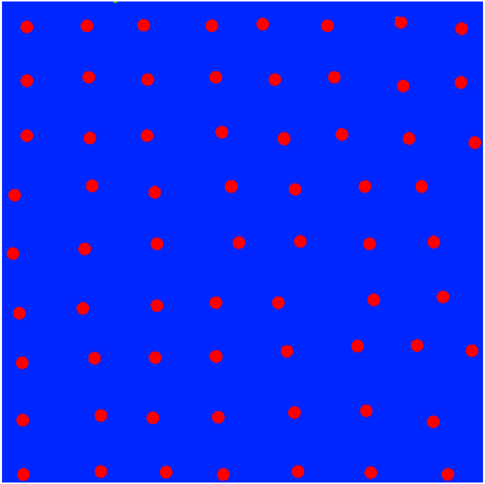
Sample Locations of Covariables



Sample Locations of Covariables

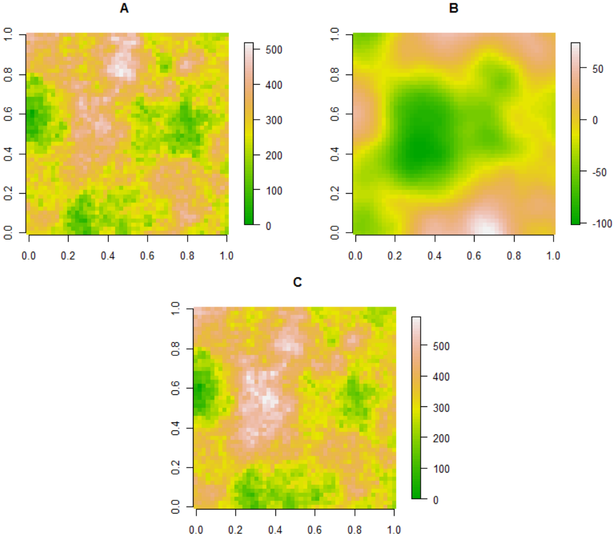


Sample Locations of Covariables

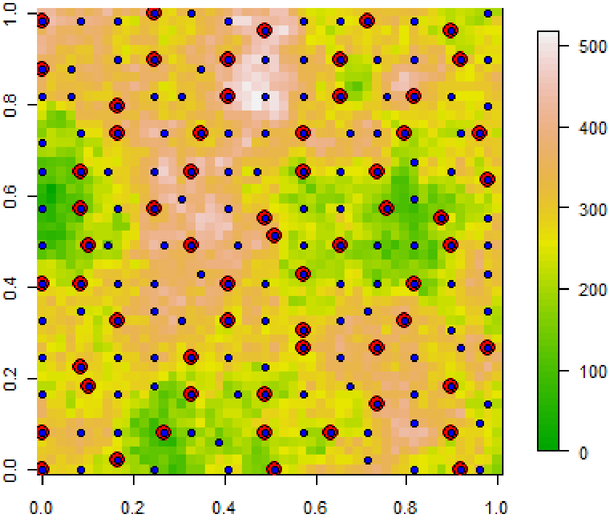


Kriging with external drift.

Gauss Random Fields



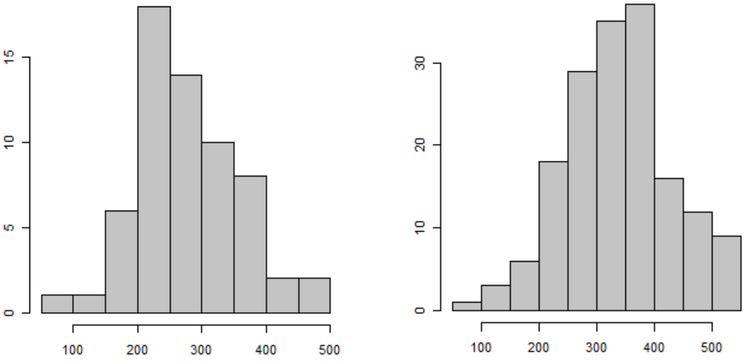
Measurement



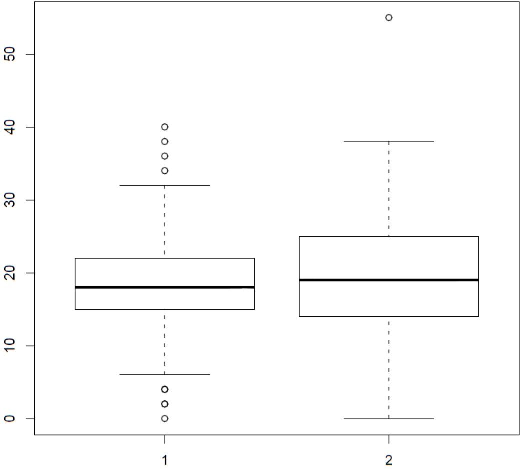
Summary

```
> summary(C)
Number of data points: 166

Coordinates summary
        Coord1 Coord2
min 0.0000000      0
max 0.9795918      1

Distance summary
       min        max
0.04081633 1.39985720

Data summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60.98  280.80  334.10  337.30  388.90  538.00
```

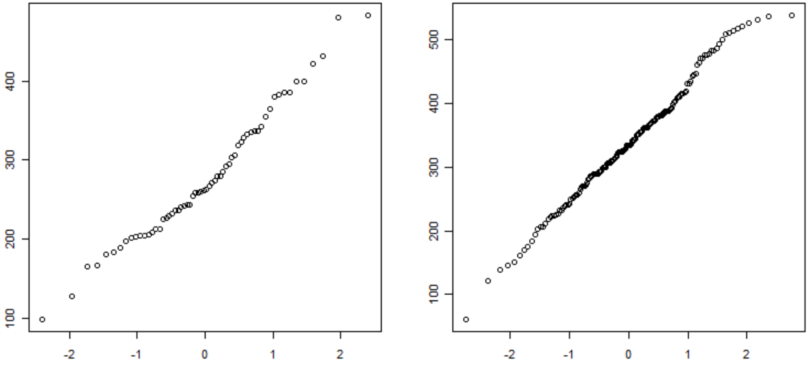
Histogram



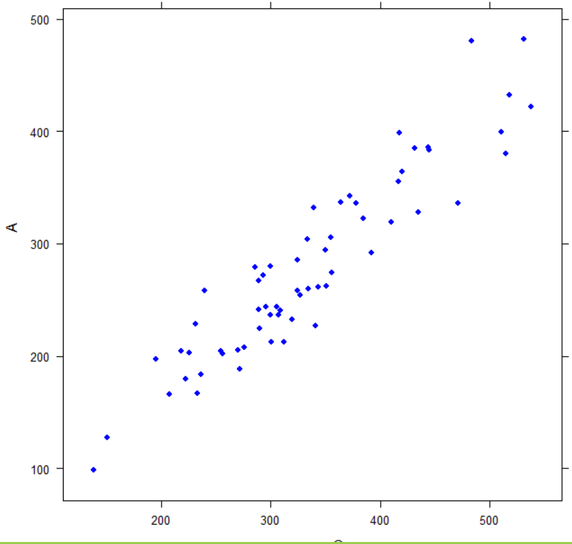
Box-and-whisker Plot



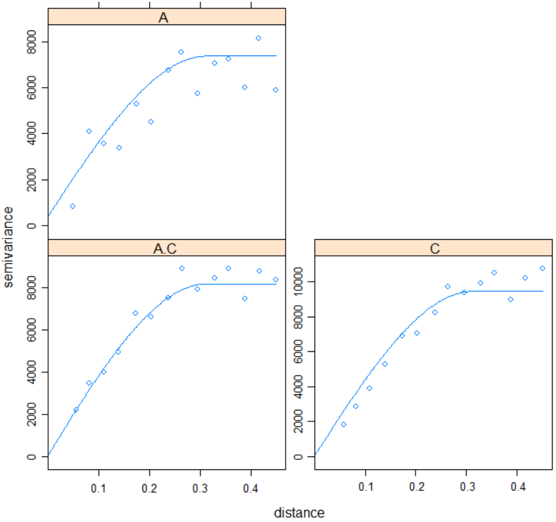
Q-Q Plot



Scatterplot



Variogram and Cross-variogram



Ordinary Kriging

```
# create a grid for the prediction
gr = data.frame(
    Coord1=A$coords[,"x"],Coord2=A$coords[,"y"])
gridded(gr) = ~Coord1+Coord2

# assign coordinates to variable A
coordinates(dataFrameA) = ~Coord1+Coord2

# variogram model
vm = variogram(data~1,dataFrameA)
vm.fit = fit.variogram(
    vm, vgm(6500, "Sph", 0.3, 50))

# prediction using ordinary kriging
OK_A = krige(data~1,dataFrameA,gr,vm.fit)
```

Linear Model of Coregionalization

```
g<-gstat(
  NULL,id="A",form=data~1,data=dataFrameA)
g<-gstat(
  g,id="C",form=data~1,data=dataFrameC)

#empirical variogram and cross-variogram
v.cross<-variogram(g)

#vmA_fit is previously created variogram model
g<-gstat(g,id="A",model=vmA_fit,fill.all=T)

#create linear model of coregionalization
g<-fit.lmc(v.cross,g)
```

Ordinary Cokriging

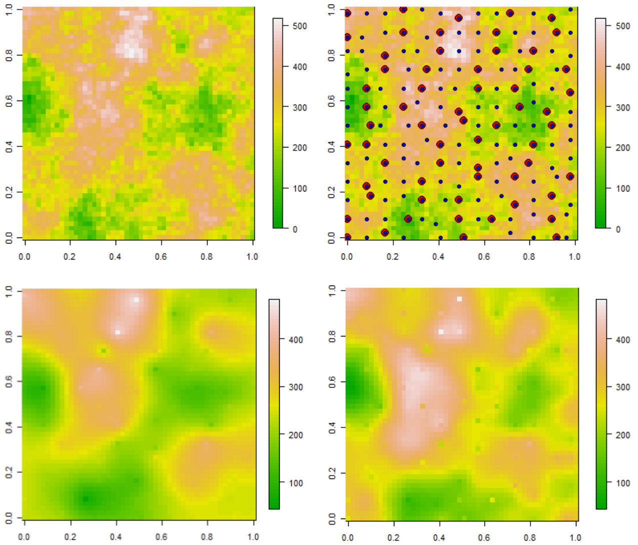
```
CK <- predict.gstat(g, gr)
```

Statistics Comparison

Data	Min.	Med.	Mean	Max.	Mean of var.pred.	Max.var. of pred.
A real	0.0	289.8	286.1	517.0	–	–
OK *A*	98.2	277.4	280.8	482.4	2804	5329
CK *A*, *C*	42.3	279.6	281.0	482.4	1617	3293
Data	**Min. diff.**	**Mean diff.**	**Max. diff.**	**Med. of diff.**	**Med. of abs(diff.)**	**RMSE**
OK *A*	-153.5	-5.1	179.1	-4.1	32.2	49.1
CK *A*, *C*	-177.0	-5.1	166.4	-3.0	30.8	46.8

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z^*(x_i) - Z(x_i))^2}.$$

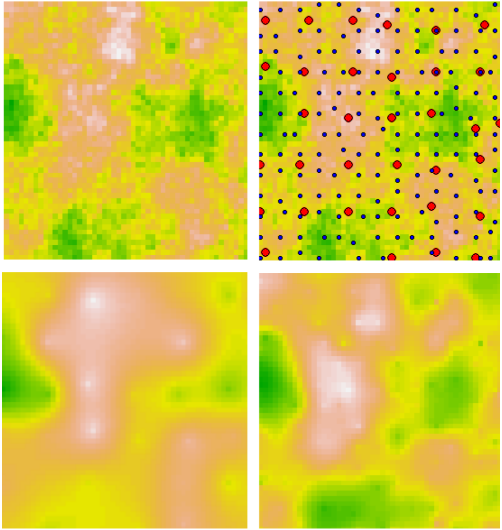
Graphical Output



Conclusion

- Secondary variable with strong spatial dependency enhances the prediction.
- Sufficient samples of the primary variable with a proper layout – good ordinary kriging prediction.
- Consider the cost of obtaining the primary and the secondary variable.

Insufficient Number of Observations



The End

Thank you for your attention!