Lab 2: Linear geostatistical model

April 19, 2021

Summary

In this lab session, you will

- explore spatial correlation by varying parameters of different correlation functions;
- use the Galicia lead concentration data to
  - explore spatial correlation of a linear outcome variable with and without covariates;
  - interpret the parameter of spatial correlation;
  - fit a linear geostatistical model with and without the nugget effect and understand the differences.

1 Understanding spatial correlation

1. Open the variogram app in R. You may do this by typing the following into the R console:

   ```
   shiny::runGitHub(repo="variogramApp", username= "olatunjijohnson",
   ref="master", subdir = "inst/variogramApp")
   ```

2. By Visualising variograms and the simulated surfaces of the Gaussian process for specified parameters, investigate the following.

   (a) How do the variogram and surface change when you increase/decrease the variance and scale parameters?

   (b) How do the variogram and surface change when you increase/decrease the nugget effect?

   (c) Varying the correlation function and the covariance parameters. Do you see any differences in the variogram and surface?

   (d) How do the variogram and surface change when you increase/decrease the smoothness parameter of the Matérn?

   (e) Do the variograms give some indications of spatial correlation? How do you know?

2 Exploring the Galicia Lead Concentration Data

1. Open the MBG app in R by typing the following into the R console:

   ```
   shiny::runGitHub(repo="MBGapp", username= "olatunjijohnson",
    ref="main", subdir = "inst/MBGapp")
   ```

2. Using the **Explore** tab, investigate possible relationships between `loglead`, the log transform of `lead`, and `pm10`, `long` and `lat`.

   (a) Do you see substantial association between `loglead` and any of the covariates?

   (b) If you were to include any of the covariates in the your model, which ones would you include, and what transformation of those covariates might make the relationship with `loglead` approximately linear?

3 Investigating Residual Spatial Correlation in the Galicia Lead Concentration Data

1. Using the **Variogram** tab of the MBG app, explore possible spatial correlations in the data.

   (a) Does the addition and removal of covariates change the residual spatial correlation?

   (b) Why is this the case?

2. Carry out the test for spatial independence based on the empirical variogram. Is there evidence of residual spatial correlation in the data?

4 Geostatistical Modelling of the Galicia Lead Concentration Data

1. Using the `Estimation` tab of the MBG app, fit linear geostatistical models to the log transformed lead concentration as follows.

   (a) Do not including any covariate, and do not include the nugget effect i.e.

   $$\log(Y_i) = \beta_0 + S(x_i), \tag{1}$$

   where $S(x_i)$ is a stationary Gaussian process with exponential correlation function.

   (b) Do not including any covariate, but include the nugget effect i.e.

   $$\log(Y_i) = \beta_0 + S(x_i) + Z_i, \tag{2}$$

   where and Z_i are i.i.d. Gaussian random variables.

   (Hint click the `show table` button if the result summary does not show)

   What does the estimate of the nugget effect tell you about its importance in this model?

   (c) Fit another model including the east-west ordinates (of the geo-coordinates as a covariate on the linear scale, but excluding the nugget effect.

   $$\log(Y_i) = \beta_0 + \beta_1 x_{i,1} + S(x_i). \tag{3}$$

   How did adding $x_{i,1}$ affect the estimates of the covariance parameters?

2. Which of the three models you have just fitted, (1) , (2) and (3), do you consider the best model? Why?

3. What do the estimates of the parameters of your best model mean?

5 Exercise 1: Malaria-height relationship

1. In the MBG app, use the `Liberia_malaria_height_data` to build a geostatistical model to investigate the relationship between height-for-age z-scores and malaria incidence at population level in Liberia. Make sure to check residual spatial correlation.

   Data description: The HAZ data were obtained from the DHS programme. The variable names in the data are self-explanatory. The data were originally at individual level, but for the purpose of this exercise, they have been aggregated to cluster (location) level. Caution should therefore be taken when interpreting the results, because of the so called ecological falacy, and because a single pooled height-for-age z-scores to represent all children at a given location is problematic.

6 Exercise 2: `R` Code Challenge

1. The `R` code `Lab2.R` performs the tasks in Sections 4 above. Run the code to carry out the analysis and make sure you understand how each step of the analyses has been implemented.

2. Using the codes `Lab1.R` and `Lab2.R` as guide, write your own `R` to carry out the task in Exercise 1 without using the MBG app.

   NB: Should you find it helpful to include a variable x as a piecewise-linear trend, you can do this by including in your linear predictor $\beta_m \min\{x, k\} + \beta_n \min\{x - k, 0\}$ where β_m and β_n are the slopes of the first and second pieces, respectively.