# Module 12: Geostatistics - Lesson 5
# Mapping exercise

18 June 2018

## 1 Aims

The aim of this exercise is use to the concepts and skills developed in the first week of the module to perform a simple "end-to-end" geostatistical mapping. You will work on the exercise in your group and submit ONE assignment.

## 2 Overview

Today's exercise builds on the concepts and skills that you developed during the first week. You will need to refer back to your lecture notes and to the notes you made from the reading. The exercise requires that you use the R skills developed during the last week. We give you some simple preliminary instructions; however, to develop this further you will need to refer to the exercises completed over the last week. Feedback will be quickly provided (before the beginning of Week 3).

This exercise uses a dataset of air quality observations for the European Economic Area (EEA), which are made available via AirBase (*Air* quality data*Base*)[1]. This provides quality-assured data for various air quality components at a range of time scales. The 24-hour PM10 concentrations were extracted for 7 April 2009 for central Europe (Austria, Belgium, Czech Republic, Denmark, Germany, Luxembourg, Netherlands, Switzerland) for rural background locations. PM10 is particulate matter less than 10 $\mu$m in diameter. These data were used in a recently published study (Hamm et al. 2015).

You are required to write a short report (5 pages $\pm 1$, including figures) which addresses the six issues listed below. Please add your own observations and comments as necessary.

## 3 Assessment and submission

You only need to complete ONE assignment per group. Please name the document clearly as `Group???_mapping.doc`. Each group member should submit the same assignment using the Blackboard assignment tool. We will check and grade your answers. Please ensure that you include your group on every page of the document. References should be used as appropriate.

You should submit your report by midnight (24:00) on Thursday 21 June, 2016; however, I strongly recommend that you complete the assignment sooner.

[1] http://acm.eionet.europa.eu/databases/airbase/ (accessed 1 June 2016)

# 4 Instructions

1. Import the data into R and check what is in the dataset

```
> library(gstat)
> library(maptools)
> library(rgdal)
> load("AirBasePM10_070409.RData")
> head(abpm10)
```

```
   mNo      lat      lon  easting northing height country pm10.obs
23  23 46.60833 14.39861 4658.091 2620.540    430      AT     26.7
24  24 48.39167 13.67111 4592.866 2815.451    525      AT     45.7
27  27 48.10611 15.91945 4761.515 2794.613    581      AT     34.9
42  42 48.23667 16.63695 4813.484 2813.615    150      AT     71.2
51  51 48.08611 16.43333 4799.839 2795.595    172      AT     61.5
56  56 47.77000 16.76640 4827.679 2762.801    117      AT     44.2
```

   Both geographic and projected coordinates are provided. The projection used is the European Terrestrial Reference System 1989 (ETRS) Lambert azimuthal equal-area (LAEA) projection which gives a coordinate reference system for the whole of Europe. Distance units are kilometres. We should use these projected coordinates because they allow the calculation of Euclidean distances. We should convert the data frame to a spatial object and assign the map projection:

```
> coordinates(abpm10) <- ~easting + northing
> proj4string(abpm10) <- CRS("+init=epsg:3035 +units=km")
```

   For the other columns, `mNo` is an index, `height` is altitude, `country` is the 2-letter ISO code for each country and `pm10.obs` is the observed PM10 concentration in $\mu$g m$^{-3}$.

2. Plot the data

```
> plot(abpm10, axes=T)
> bubble(abpm10, zcol="pm10.obs", scales=list(draw=T), main="Sensible title")
```

   Refer to your notes from the last week to improve this plot.

3. Estimate the empirical variogram

```
> abpm10.ev <- variogram(pm10.obs~1, data=abpm10)
> plot(abpm10.ev)
> abpm10.ev
```

```
   np      dist     gamma dir.hor dir.ver   id
1  63  23.62018  88.68138       0       0 var1
2 173  53.53947 119.78784       0       0 var1
3 217  87.68357 122.05095       0       0 var1
4 241 120.46024 127.45085       0       0 var1
5 325 156.21037 152.44536       0       0 var1
6 333 191.68783 184.83536       0       0 var1
7 413 226.32728 163.47725       0       0 var1
8 398 260.58345 161.39955       0       0 var1
```

```
9  409 296.10671 205.38441       0       0 var1
10 423 330.07574 197.95906       0       0 var1
11 433 365.14507 214.52983       0       0 var1
12 388 399.09620 190.67092       0       0 var1
13 388 435.43082 203.05083       0       0 var1
14 399 468.76597 218.77180       0       0 var1
15 374 504.71392 219.08644       0       0 var1
```

(Hint: check `np`, the number of points used to calculate the variogram at each lag. Is this sufficient? Does the sample variogram flatten out and reach an apparent sill?)

4. Fit a model to the empirical variogram

```
> abpm10.mv <- fit.variogram(abpm10.ev,vgm(150, "Exp", 300, 50))
> abpm10.mv
```

```
  model     psill    range
1   Nug  75.91735   0.0000
2   Exp 152.35795 216.7027
```

```
> plot(abpm10.ev, model=abpm10.mv)
```

5. Use the modelled variogram for interpolation

```
> xy  <- expand.grid(x=seq(3800, 5000, by=50), y=seq(2500,3600, by=50))
> xys <- SpatialPoints(xy)
> gridded(xys) <- TRUE
> proj4string(xys) <- CRS("+init=epsg:3035 +units=km")
> plot(xys, axes=T)
> points(as.data.frame(abpm10)[,4:5], col=2, pch=19)
> pmk <- krige(pm10.obs~1, abpm10, xys, model=abpm10.mv)
```

```
[using ordinary kriging]
```

Experiment with a denser grid for prediction.

6. Display a map with the interpolated values

```
> spplot(pmk, "var1.pred", sp.layout=list("sp.points", pch="+", abpm10),
+          scales=list(draw=TRUE), main="Sensible title")
```

`sp.layout` adds an extra layer/s to the map. See the help for `spplot`.

7. Display a map with the prediction error variance (kriging variance).

```
> spplot(pmk, "var1.var", sp.layout=list("sp.points", pch="+", abpm10),
+          scales=list(draw=TRUE), main="Sensible title")
```

You should edit the above two maps to add labels to the x- and y-axes.

8. Add the contry boundaries to the above maps. First we use the `maptools` library to import the shape file. Before doing this you should download `ceShape.zip` from Blackboard and unzip it into the current directory.

```
> ceShape.shp <- readShapePoly("./ceShape/CentralEuropeShape.shp")
> proj4string(ceShape.shp) <- CRS("+init=epsg:3035 +units=km")
```

We can now add both the country boundaries and the observation locations to the map, for example

```
> spl1 <- list("sp.polygons", ceShape.shp, first = FALSE)
> spl2 <- list("sp.points", pch="+", abpm10, col=2, cex=2)
> spl <- list(spl1, spl2)
> spplot(pmk, "var1.pred", sp.layout = spl, scales=list(draw=TRUE), main="Kriged predic
```

Note that here we add two layers to the map. Repeat this for both the kriged predictions and the kriging variance.

9. You should calculate the statistics for cross validation. An example command for this is

```
> pmk.exp.cv = krige.cv(pm10.obs~1, abpm10, model=abpm10.mv)
```

Plot the histogram of the residuals and examine this. Calculate the mean error (ME) and RMSE (RMSE). The command for ME is as follows, you need to write the code to get RMSE yourself:

```
> ME.sph.cv = sum(pmk.exp.cv$residual)/length(pmk.exp.cv$residual)
```

10. Interpret the two maps, also in comparison with the map obtained during step 2.

The above instructions give a basic procedure. Critical assessment of these steps together with experimenting and retrying is essential to improve and fine-tune the analysis. You should consider the following six points:

(a) Take a careful look at the data (e.g., using boxplots and histograms). Try log-transforming them. Examine this this and, if you consider it appropriate, re-do the analysis on the log transformed data. Explain your choice. (10%)

(b) Go back to step 3. Here you took the default options for calculation of the sample variogram. You should check "np" in the variogram output. Is the number of pairs of points at each lag sufficient to obtain a robust estimate for the sample variogram? Does the sample variogram approach a sill? Try chainging the `cutoff` and bin `width`. Does the resulting sample variogram differ from the previous one? Pick one variogram *model* and examine the effect of the different sample variograms on the estimated parameter values and cross-validation values. Finalize your choice of sample variogram. Explain your choice. (20%)

(c) Go back to step 4. Keeep your choice of sample variogram (see (b) above) fixed. Take a different variogram model, and examine the effect on the estimated parameters, the accuracy of the variogram fit and the ME and RMSE obtained from cross-validation. Which model do you consider to be most appropriate? Explain your choice. (20%)

(d) Interpret your sample variogram and modelled variograms. What do these tell you about the spatial structure in the data? Pay particular attention to the nugget, sill and range. How is this information used for interpolation (kriging)? (15%)

(e) Examine the maps of the kriged predictions and kriging variance. Where are highest predicted concentrations of PM10 found? Where is the kriging variance (prediction error variance) small? Where is it large? Explain these results. (15%)

(f) In this mapping exercise you have implemented ordinary kriging (OK). Considering that you have now learnt about regression kriging (RK) and co-kriging (CK) please reflect on whether OK is the best option for this dataset. Do you consider OK to be most appropriate, or would RK or CK be worth considering? Explain your answer. You do not have to do extra analysis but you should provide clear and convincing arguments. (10%)

Your report should address these six issues listed above (a-d). The mark distribution is indicated. Addtionally 10% is allocated for presentation, including the quality of the tables and figures.

## References

Hamm, N. A. S., Finley, A. O., Schaap, M. & Stein, A. (2015), 'A spatially varying coefficient model for mapping air quality at the European scale', *Atmospheric Environment* **102**, 393–405.