MA22019 2025 - Problem Sheet 7

Semi-variogram and principal component analysis

Overview

This week's problem sheet focuses on the methods for analyzing spatial dependence introduced in Sections 4.3 and 4.4 in the lecture notes. Exercises 1-2 help you with revising the content of the lecture in Week 8.

Tutorial Questions 1 and 2 each consider one of the two techniques we covered in Week 8.

Your answer to the Homework Question can be submitted on Moodle to your tutor for feedback. The submission deadline is 17:00 on Thursday 3 April 2025. You should submit a single PDF or Word file that provides your R code, any created R output and all your comments.

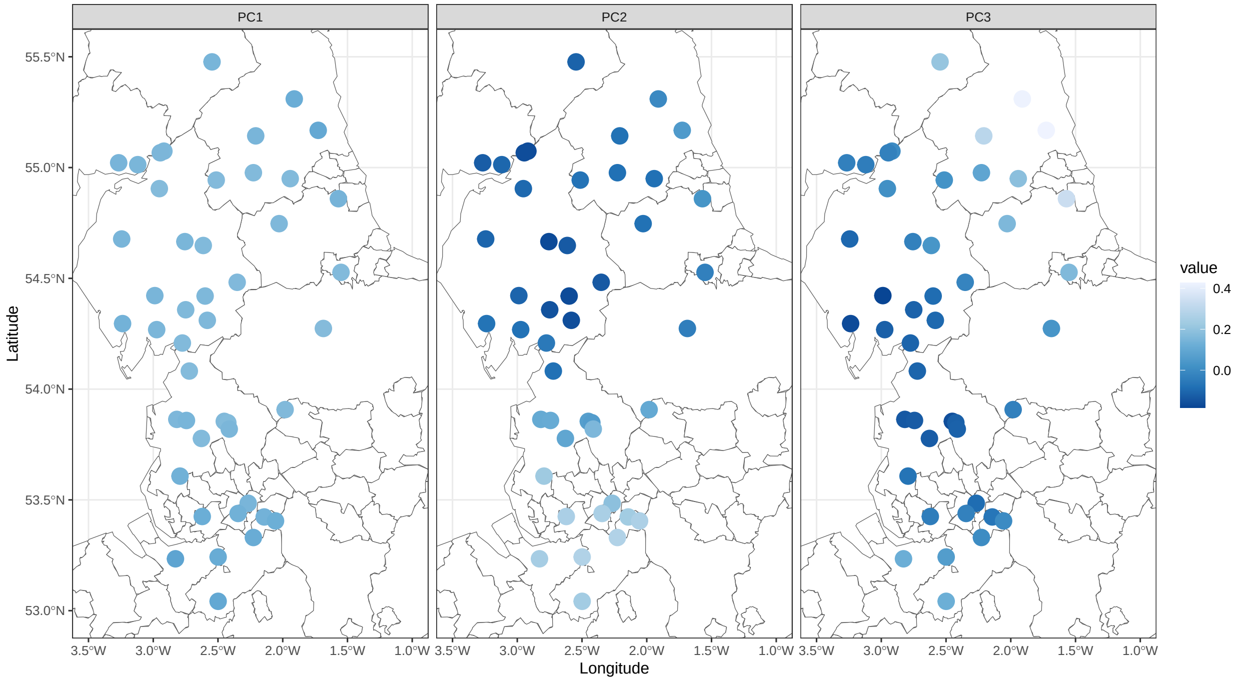
You may want to load the following packages before starting the exercise:

```r
library( dplyr )
library( lubridate )
library( ggplot2 )
library( tidyr )
library( sf )
library( ggspatial )
library( prettymapr )
library( sp )
library( gstat )
```

When working on a University PC, you will have to first install some of these packages.

Exercise 1 - Exploring spatial structure in river flow data

A data scientist used principal component analysis to explore the spatial structure / dependence in the daily maximum river flow values across 45 gauges in Northern England and the South of Scotland. They identified that the first 3 components explain more than 90% of the variation in the data and created the following plots representing the first three eigenvectors:



Provide an interpretation for the three plots and then go to Moodle and complete the quiz.

Exercise 2 - Concentrations of zinc in the top soil

The file “Meuse.csv” gives topsoil zinc concentrations (in mg/kg) collected across 155 locations in a flood plain of the river Meuse. Specifically, we are given

- Lon,Lat - Longitude and latitude coordinate of the locations
- Zinc - Zinc concentration on logarithmic scale

Perform the following steps to explore the spatial structure in the data:

a) Visualize the zinc concentration on logarithmic scale using the ggspatial package. What do you conclude?

b) Explore the spatial dependence in the zinc concentration on logarithmic scale by estimating the semi-variogram for the data. Go to Moodle and answer the quiz.

Tutorial Question 1 - Spatial dependence in soil moisture

The file “SoilMoisture.csv” contains soil moisture measurements for 1 July 2022 for the Iberian peninsula (i.e. Spain and Portugal). Data were collected using satellites and each point corresponds to the soil moisture recorded for an area of 0.25ř × 0.25ř. We want to visualize the data and estimate the semi-variogram:

a) Load the data into R and create a spatial plot which shows the soil moisture for each point, that is, the x and y axis should correspond to longitude and latitude respectively, with colour being used as a visual cue representing soil moisture. What do you conclude?

b) Remove any grid cells with missing data from the data set using the function na.omit(). Why is this step necessary for estimating the semi-variogram?

c) Estimate the semi-variogram using the function variogram() from the gstat R package. Visualize your estimate. What do you conclude?

d) Discuss whether it is reasonable to assume that the value of the semi-variogram is fully specified by the distance between spatial sites, which is the key assumption we make when using the variogram() function.

Tutorial Question 2 - Maximum monthly temperature across Germany

The file "Temperature Germany.csv" contains maximum monthly temperatures recorded for 20 sites in the northern half of Germany for 2000-2023. The longitude and latitude coordinates for the sites can be found in the file "Sites Germany.csv". In the following we will visualize the data and analyse its spatial structure.

a) For all sites calculate their median monthly maximum temperature for June across the period 2000-2023. Create a plot to visualize the values you obtained. What do you conclude?

b) We now want to use principal component analysis to analyse the spatial structure in the data. One difference to the example from the lecture notes is that we standardize the data per month and site, rather than just per site. Formally, we standardize the data using

$$\tilde{x}_{i,j,t} = (x_{i,j,t} - \bar{x}_{i,j})/\hat{\sigma}_{i,j},$$

where $x_{i,j,t}$ is the maximum temperature for site i in month j and year t. We will have to perform this standardization "by hand" and can no longer rely on prcomp() to do it for us. Perform the following steps to perform PCA in this case and interpret your outputs:

1. Standardize the data per site and month using the equation above.
2. Use the pivot_wider() function to convert the data to a data frame such that each column contains the standardized observations for one side. Use the function na.omit() to remove any rows with missing data.
3. Apply the prcomp() function using the values obtained in Step 2.
4. Using the eigenvalues, decide on the number m of eigenvectors we should study.
5. Visualize the first m eigenvectors, where m is the number you decided on in part Step 4. What do you conclude?

Homework Question - Spatial dependence in temperatures across Brazil

Let's again consider the temperature data from Brazil considered in Problem Sheet 6. The data are provided in the file "Brazil Temperature.csv". In the following we want to explore the spatial dependence in the data.

a) Estimate the semi-variogram for the data. What do you conclude about the spatial dependence in the observed temperatures?

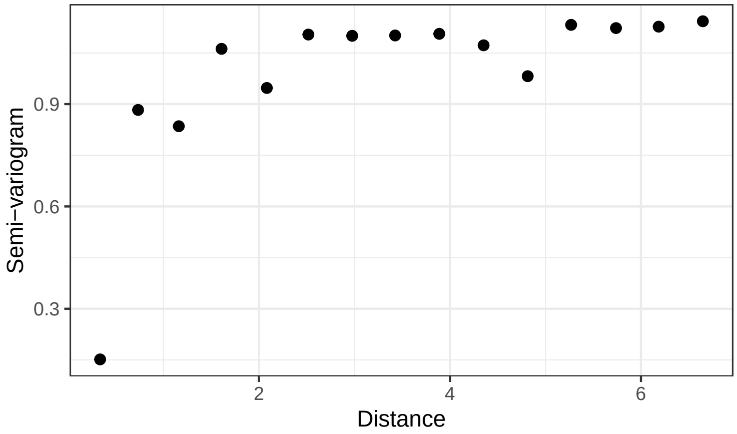
b) Discuss whether it is reasonable to assume that the spatial random process has a constant mean and that the value of the semi-variogram is fully specified by the distance between spatial sites.

c) **Bonus:** One may argue that the constant mean assumption may not hold for the spatial process $\{X(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$. The variogram() function allows us to define a model of the form

$$X(\mathbf{s}) = \beta_0 + \beta_1 \text{Longitude}(\mathbf{s}) + \beta_2 \text{Latitude}(\mathbf{s}) + Z(\mathbf{s}), \tag{1}$$

where $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ is termed the residual random process. The semi-variogram estimated by the variogram() function then corresponds to that for the residual random process. In R, we derive and visualize this estimate using

```
Brazil <- read.csv( "Brazil Temperature.csv" )
coordinates( Brazil ) <- ~Lon+Lat
gamma_hat <- variogram( MeanTemp ~Lon+Lat, data = Brazil )
ggplot( gamma_hat, aes( x=dist, y=gamma/2 ) ) + geom_point( size=2 ) +
  theme_bw() + labs( x="Distance", y="Semi-variogram" )
```



Describe the similarities and differences between this estimate and yours in part a). Discuss whether the conclusions we draw regarding spatial dependence in the residual process are the same as for the process in part a).

Remark: You may learn more about the type of models in Equation (1) in Statistical Modelling & Data Analytics 3B.