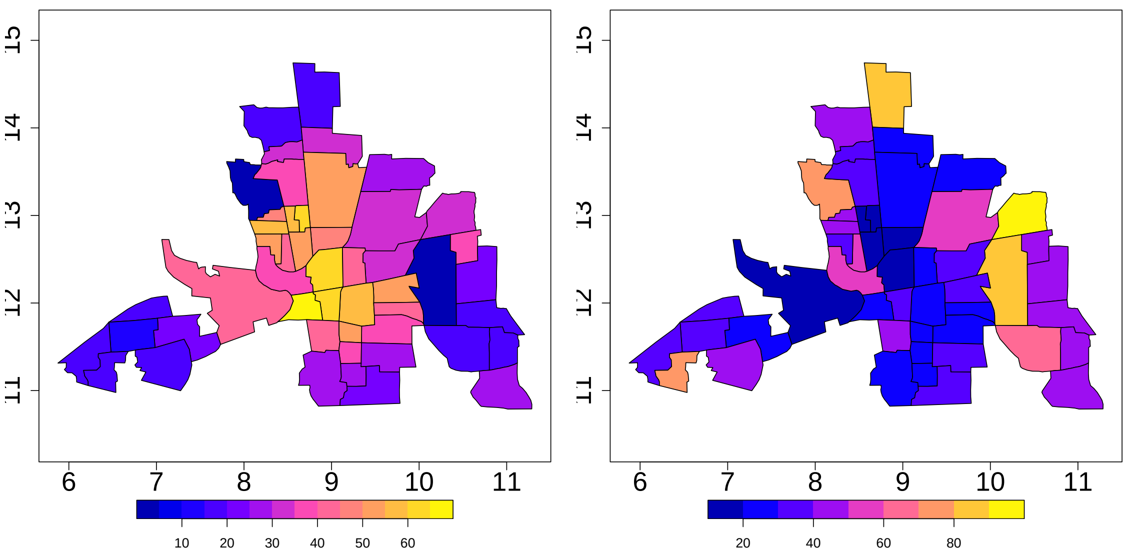
Spatio-temporal statistics (MATH4341) Michaelmas term, 2024

Exercise sheet

Lecturer: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Part 1. Types of spatial data

($\star$)(Columbus Columbus OH data set) Figure 1a shows the Property crime (number per thousand households) in 49 districts in Columbus in 1980, as well as the average value of the house in USD. Figure 1b presents the corresponding average house value. This is the R dataset `columbus{spdep}`. Interest may lie to find whether high rates of crime are clusterd in a particular areas, and if yes, perhaps what is the association of it with the value of the houses in the area. To which principal spatial statistical are would you associate this problem?



(A) Property crime (number per thousand households)

(B) Average value of the house in USD

FIGURE 1. Columbus Columbus OH spatial analysis dataset

Solution. Aerial unit data / spatial data on lattices

Exercise 1. ($\star$)(Soil chemistry properties data set.) It contains measurements of various chemical properties of soil samples collected at different locations in a field. These properties include: the acidity or alkalinity of the soil (PH), the salt concentration in the soil (Salinity), and others. It is the R dataset `soil250{geoR}`. Figure 2 presents the locations these measurements are taken. The data (measurements) are in fixed locations at a regular grid of points. The domain scientist would be interested in the nutrient levels and pH to assess soil fertility and make recommendations for agricultural practices. The statistician could (i.) estimate/predict values of soil properties at unsampled locations based on measurements at sampled locations; and (ii.) assess the spatial variability of soil properties (nutrient levels and pH) to identify regions with high or low variability. To which principal spatial statistical are would you associate this problem?

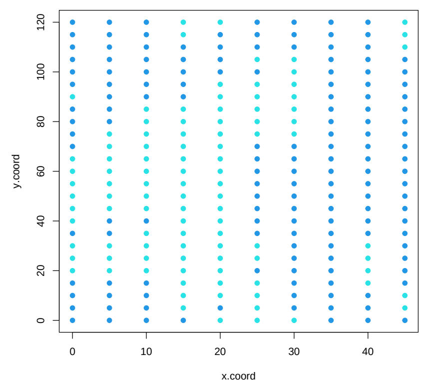


FIGURE 2. Soil chemistry data set

Solution. Point referenced data, or geostatistcal data

Exercise 2. ($\star$)(Scallop abundance data) The scallop data is based on a 1990 survey cruise in the Atlantic continental shelf off Long Island, New York, U.S.A. They are available from R as `scallop {SemiPar}`. Figure 3 presents 148 locations (degrees of longitude & latitude) in the Atlantic waters off the coasts of New Jersey and Long Island New York as coordinates and the size of scallop catch at the corresponding location as the dot size. The sites are at fixed locations within an irregular grid of points. Sustainable scallop abundance is critical for the long-term economic viability of the fishing industry. A healthy and stable scallop population supports a consistent source of income for fishermen and related businesses.

(1) To which principal spatial statistical are would you associate this problem?
(2) Can you suggest a Bayesian hierarchical model for this? Justify your suggestion.

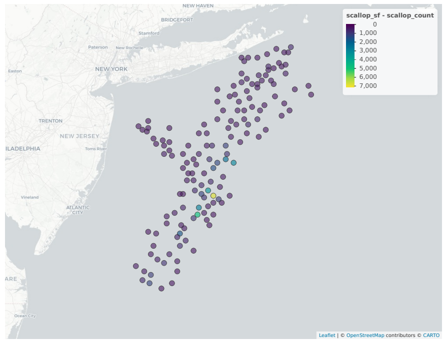


FIGURE 3. Scallop abundance data

Solution.

(1) Point referenced data, or geostatistcal data

(2) It contains $n = 148$ observations $\{(Z_i, s_i)\}_{i=1}^{n}$ where the i-th observation contains the observed scallop catch size Z_i at location specified by coordinate $s_i = (s_{1,i}, s_{2,i})^\top$ that is the degrees latitude north of the Equator $s_{2,i}$, and degrees longitude west of Greenwich $s_{1,i}$.

This is definitely a geostatistics problem. Here, we will present a naive way to model it by reflecting what we discussed earlier.

Data model: One may consider that the observations $\{Z_i\}$ at each location are a realization of the actual mean (hence unknown) scallop catch abundance $Y_i$and as it is a count its distribution can be represented as a Poisson distribution, i.e.

$$Z_i|Y_i, \sigma^2 \overset{\text{iid}}{\sim} \text{Poi}\left(Y_i\right), \quad i = 1, ..., n$$

Spatial process model: One may consider that the mean scallop catch abundance $Y(s)$ (over the spatial domain $s \in S$)in log scale is a function where at each finite set of locations $\{s_i\}_{i=1}^{n}$ follows a Normal distribution with a mean μ with $[\mu]_i = \mu(s_i)$ parametrized as

$$\mu(s) = \exp\left(\beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{12} s_1 s_2\right), \text{ at a location s=}(s_1, s_2)^\top$$

with unknown parameter β, and covariance matrix $[C]_{i,j} = c(s_i, s_j)$ parametrized with covariance function

$$c\left(s, s'\right) = \sigma^2 \exp\left(-\phi \left\|s - s'\right\|\right)$$

to impose that nearer locations cause stronger dependences in the model. Here β, ϕ, and σ^2 are unknown parameters.

Hierarchical model: To sum up, we have build the hierarchical model

$$\begin{cases} Z_i|Y_i, \sigma^2 \overset{\text{iid}}{\sim} & \text{Poi}\left(\mu\left(s_i\right) = \exp\left(Y_i\right)\right), \text{ data model} \\ Y|\sigma^2, \beta, \phi \sim & \text{N}_n\left(S\beta, C\right), \quad \text{spatial process model} \end{cases} \tag{1}$$

Figure **??** shows the hierarchical spatial model (1) for different values of $\theta = \left(\sigma^2, \beta, \phi\right)$; the surface corresponds to the spatial process $\left\{Y(s)\,; s \in \mathbb{R}^2\right\}$ and is presented at three different instances each of them with different values for (β, ϕ), while the dots correspond to the observations $\{(Z(s_i), s_i)\}_{i=1}^n$ and their deviation from the spatial process is controlled by σ^2 .

Bayesian hierarchical model: If we work on the fully Bayesian framework, we can complete the model with priors on $\theta = \left(\sigma^2, \beta, \phi\right)$ for instance $\sigma^2 \sim \mathrm{IG}(\kappa_\sigma, \lambda_\sigma)$, $\phi \sim \mathrm{IG}(\kappa_\phi, \lambda_\phi)$, and $\beta \sim \mathrm{N}_4(b, Iv)$, with some known hyper-parameters $\kappa_\sigma, \lambda_\sigma, \kappa_\phi, \lambda_\phi, b, v$. To sum up, we have build the Bayesian model

$$\begin{cases} Z_i | Y_i, \sigma^2 \overset{\text{iid}}{\sim} & \mathrm{Poi}\left(\mu(s_i) = \exp(Y_i)\right), \text{ data model} \\ Y | \sigma^2, \beta, \phi \sim & \mathrm{N}_n(S\beta, C), \text{ spatial process model} \\ \beta \sim & \mathrm{N}_4(b, Iv), \text{ hyper-parameter prior model} \\ \sigma^2 \sim & \mathrm{IG}(\kappa_\sigma, \lambda_\sigma), \text{ hyper-parameter prior model} \\ \phi \sim & \mathrm{IG}(\kappa_\phi, \lambda_\phi), \text{ hyper-parameter prior model} \end{cases}$$

Exercise 3. ($\star$)(Wolfcamp-aquifer data) Figure 4 presents locations and levels (in feet above sea level) of piezometric head for the aquifer; they are obtained by drilling a narrow pipe into the aquifer and letting the water find its own level in the pipe. After rigorous screening of unsuitable wells, 85 remained. There is interest to find where the radionuclide contamination would flow from the site in Deaf Smith County, Texas. Beneath Deaf Smith County is a deep brine aquifer known as the Wolfcamp aquifer, a potential pathway for any radionuclides leaking from the repository. The predicted direction of flow can be used to determine locations of downgradient and upgradient wells for a groundwater monitoring system. A first direction in analyzing this spatial data set is to draw a map of a predicted surface based on the (irregularly located) 85 data. To which principal spatial statistical are would you associate this problem?

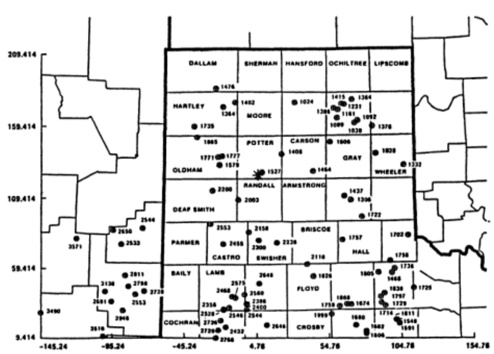


FIGURE 4. Wolfcamp-aquifer data. Piezometric-head levels (feet above sea level) vs coordinates.

Solution. Point referenced data, or geostatistcal data

Exercise 4. ($\star$)(Swiss rainfall data) Figure 5 presents the locations of the 100 locations in Switzerland as dots whose size and color indicates the amount of the corresponding rainfall measurements (in 10th of mm) taken on May 8, 1986. This is the R data set `SIC{geoR}`. Observation sites are irregularly spaced, and fixed. A scientific objective may be to analyzing rainfall patterns with purpose to optimize crop planting and irrigation schedules. A statistician is able to estimate rainfall values at unsampled locations based on available measurements, create maps that represent the spatial distribution of rainfall, or quantify the uncertainty associated with rainfall estimates and predictions, which are important for risk assessment and decision-making. To which principal spatial statistical are would you associate this problem?

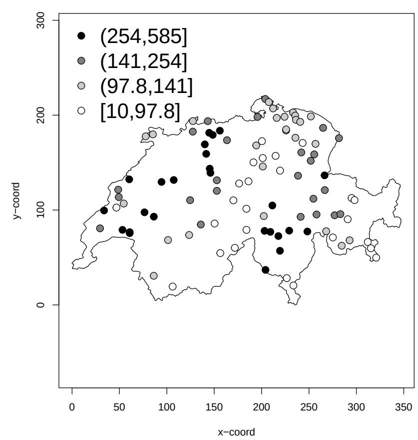


FIGURE 5. Swiss rainfall data

Solution. Point referenced data, or geostatistcal data

Exercise 5. ($\star$)(Air pollution in Piemonte.) Figure 6 presents the average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte region (Northern Italy). The data (measurements) are at fixed locations at irregular grid points. PM10 is one of the most troublesome pollutants in the area. Environmental agencies need models to predict PM10 at unmonitored sites in order to assess PM10 concentration over an entire region. A geostatistician can build a model which is satisfactory in terms of goodness of fit, interpretability, parsimony, prediction capability and computational costs with purpose to build reliable PM10 concentration maps, equipped with the corresponding uncertainty measure.To which principal spatial statistical are would you associate this problem?

Solution. Point reference data / geostatistical data

Exercise 6. ($\star$)An example of spatio-temporal data where aerial spatial data are time referenced is given in Figure 7 which shows a spatio-temporal dataset representing the population of the counties of Ohio, USA, from 1968 to 1988. The dataset is available from the SpatialEpiApp R package.

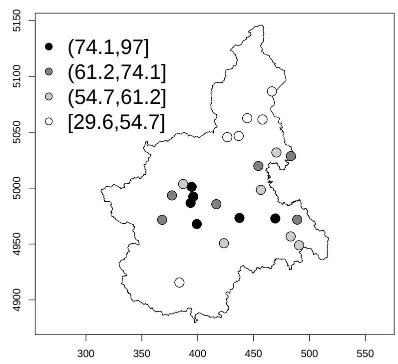


FIGURE 6. (Air pollution data) Average PM10 ($\mu g/m^3$) concentration during October 2005–March 2006 for the 24 monitoring stations in the Piemonte

Interest lies in not only how the population is only distributed over the spatial domain but also how it evolves during the time. To which principal spatial statistical are would you associate this problem?

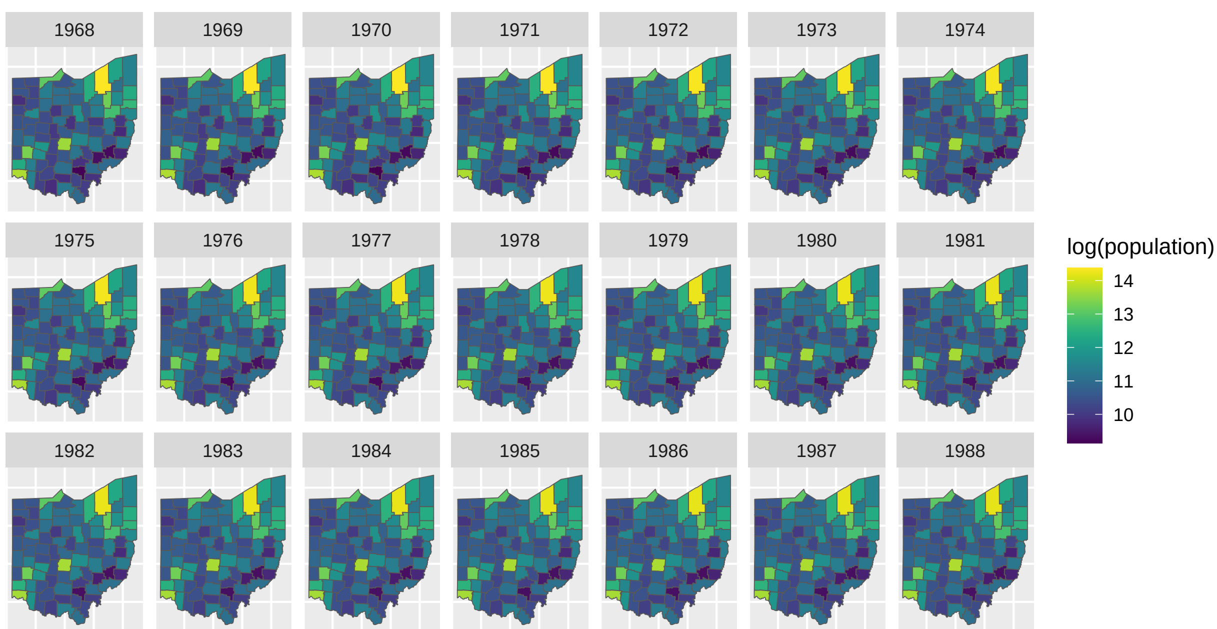


FIGURE 7. Population of the counties of Ohio, USA, from 1968 to 1988.

Solution. Aerial spatio-temporal data

Part 2. Point referenced data / Geostatistics

Part 3. Aerial unit data / spatial data on lattices