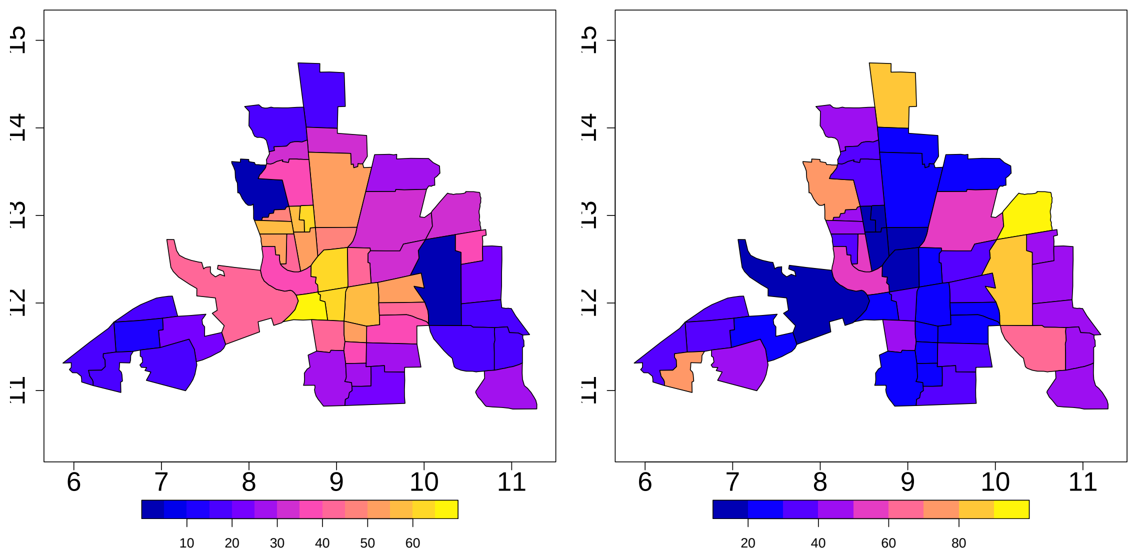
Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Exercise sheet

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

Part 1. Types of spatial data

Exercise 1. ($\star$)(Columbus Columbus OH data set) Figure 2a 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 2b 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?



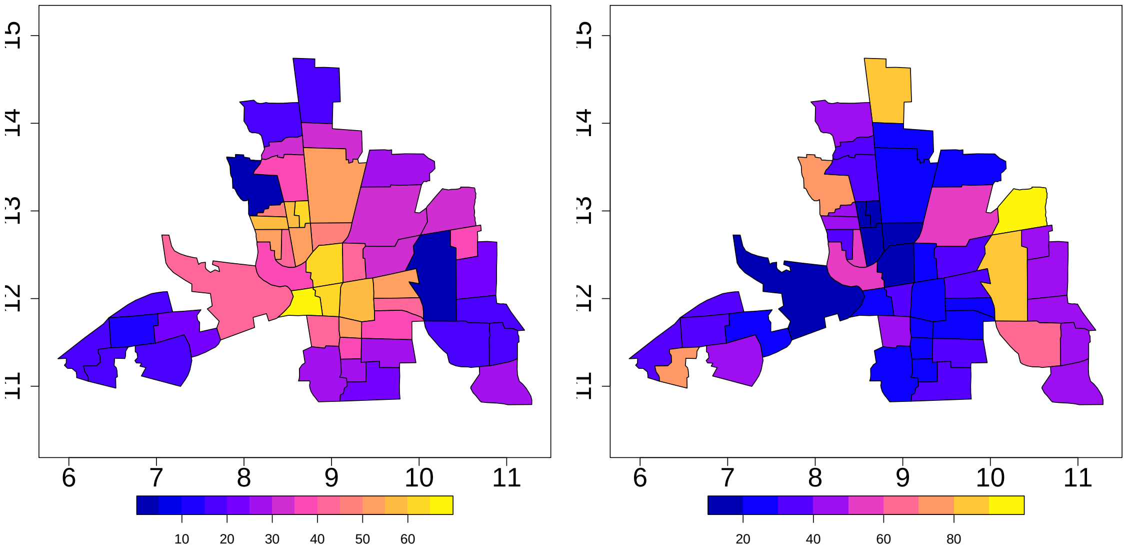
(A) Property crime (number per thousand households)

(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 2. ($\star$)(Columbus Columbus OH data set) Figure 2a 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 2b 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 2. Columbus Columbus OH spatial analysis dataset

Solution. Aerial unit data / spatial data on lattices

Exercise 3. (⋆)(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 3 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?

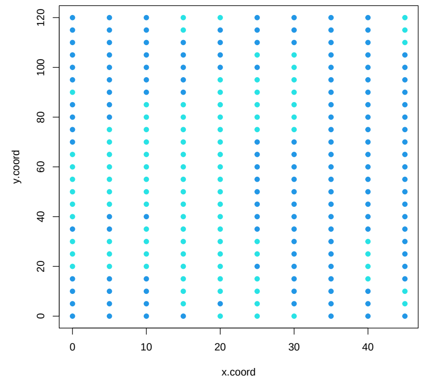


FIGURE 3. Soil chemistry data set

Solution. Point referenced data, or geostatistcal data

Exercise 4. ($\star$)(Scallop abundance data) Figure 4 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. To which principal spatial statistical are would you associate this problem?

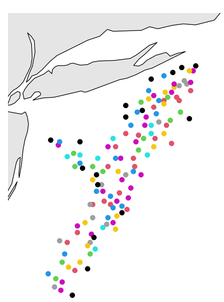


FIGURE 4. Scallop abundance data

Solution. Point referenced data, or geostatistcal data

Exercise 5. ($\star$)(Wolfcamp-aquifer data) Figure 5 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?

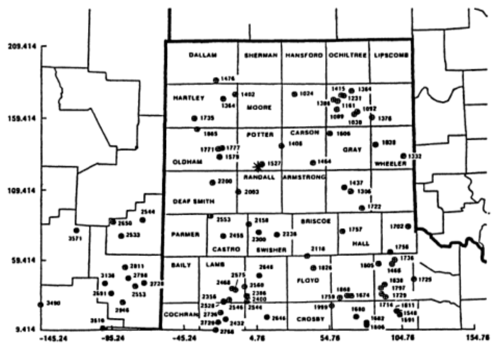


FIGURE 5. Wolfcamp-aquifer data. Piezometric-head levels (feet above sea level) vs coordinates.

Solution. Point referenced data, or geostatistcal data

Exercise 6. ($\star$)(Swiss rainfall data) Figure 6 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?

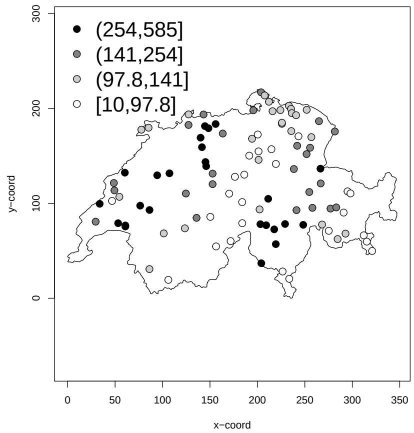


FIGURE 6. Swiss rainfall data

Solution. Point referenced data, or geostatistcal data

Part 2. INLA

Exercise 7. ($\star$)Consider the model

$$\begin{cases} z_i|\eta_i \sim \text{Poisson}\left(\exp\left(\eta_i\right)\right) \quad i = 1, ..., n \\ \eta_i = \beta_0 + \beta_1 w_i + u_{j(i)} \\ u \sim \text{N}_m\left(0, I\tau^{-1}\right) \end{cases}$$

where $\{w_i\}$ are covariates, $j(i)$ is a known mapping from $1 : n$ to $1 : m$ (given below in the dataset as idx).

For training use the following data set $\{(z_i, w_i)\}_{i=1}^{n}$ by running

```
rm(list=ls())
# generate the dataset
set.seed(123456L)
n = 50;
m = 10
w = rnorm(n, sd = 1/3)
u = rnorm(m, sd = 1/4)
intercept = 0;
beta = 1
idx = sample(1:m, n, replace = TRUE)
z = rpois(n, lambda = exp(intercept + beta * w + u[idx]))
table(z, dnn=NULL)
```

Do the following, by using R-INLA

(1) Run inla{INLA} in order to train the above model, and generate an inla object (that you will call it out.inla). For the function inla{INLA} specify the formula, data, and family

arguments. To approximate the conditional pdf of latent variables of the GMRF use the Gaussian approximation. For the rest parameters just use the default R-INLA options.

(2) Print a summary of the marginal posteriors

(3) Produce and print the marginal posterior pdf of $\Pr(\beta_1|z)$.

Solution.

(1)
```
my.data = data.frame(z, w, idx)
formula = z ~ 1 + w + f(idx, model="iid")
out.inla = inla(formula, data = my.data,
                family = "poisson",
                control.inla = list(strategy = "gaussian")
                )
```

(2)
```
summary(out.inla)
```

```
> res.predict$summary.linear.predictor[7,]
                 mean        sd 0.025quant 0.5quant 0.975quant     mode          kld
Predictor.07 3.021652 0.1847223   2.639797 3.029161   3.362469 3.045581 1.224947e-07
```

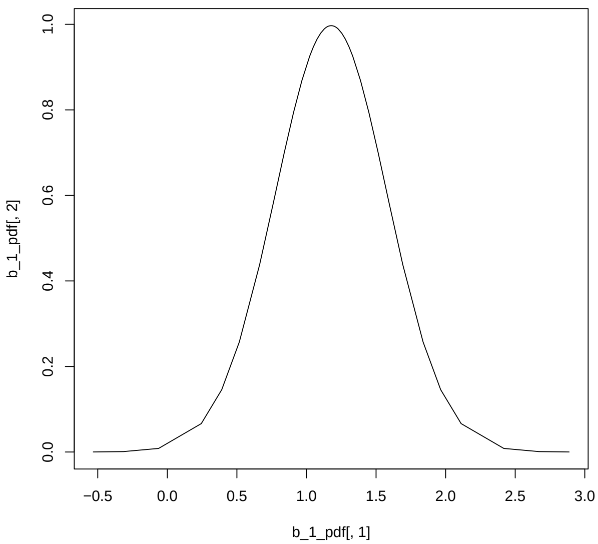
```
Call:
   c("inla.core(formula = formula, family = family, contrasts = contrasts, ", " data =
   data, quantiles = quantiles, E = E, offset = offset, ", " scale = scale, weights =
   weights, Ntrials = Ntrials, strata = strata, ", " lp.scale = lp.scale, link.covariates
   = link.covariates, verbose = verbose, ", " lincomb = lincomb, selection = selection,
   control.compute = control.compute, ", " control.predictor = control.predictor,
   control.family = control.family, ", " control.inla = control.inla, control.fixed =
   control.fixed, ", " control.mode = control.mode, control.expert = control.expert, ", "
   control.hazard = control.hazard, control.lincomb = control.lincomb, ", " control.update
   = control.update, control.lp.scale = control.lp.scale, ", " control.pardiso =
   control.pardiso, only.hyperparam = only.hyperparam, ", " inla.call = inla.call,
   inla.arg = inla.arg, num.threads = num.threads, ", " blas.num.threads =
   blas.num.threads, keep = keep, working.directory = working.directory, ", " silent =
   silent, inla.mode = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
   .parent.frame)")
Time used:
    Pre = 0.763, Running = 0.224, Post = 0.0168, Total = 1
Fixed effects:
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept) -0.069 0.153     -0.370   -0.069      0.232 -0.069   0
w            1.178 0.401      0.391    1.178      1.964  1.178   0

Random effects:
  Name    Model
    idx IID model

Model hyperparameters:
                      mean       sd 0.025quant 0.5quant 0.975quant   mode
Precision for idx 19980.67 19912.46     599.82 13939.51   74289.61 214.74

Marginal log-Likelihood:  -69.62
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

(3) b_1_pdf = out.inla$marginals.fixed$w
plot(b_1_pdf[,1], b_1_pdf[,2], type ="l")



Part 3. **Point referenced data / Geostatistics**

Exercise 8. ($\star$)If $c:\mathbb{R}^d\rightarrow\mathbb{R}$ is the covariogram of a weakly stationary random field $Z=(Z_s)_{s\in\mathbb{R}^d}$ then $c(\cdot)$ is semi-positive definite; i.e. for all $n\in\mathbb{N}$, $a\in\mathbb{R}^n$, and $\{s_1,...,s_n\}\subseteq S$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jc(s_i-s_j)\geq 0$$

Solution. To show that $c(\cdot)$ is semi-positive definite, I need to show that $\forall a\in\mathbb{R}^n-\{0\}$ it is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jc(s_i-s_j)\geq 0$$

Well it is

$$\begin{aligned}0\leq \text{Var}\left(\sum_{i=1}^{n}a_iZ(s_i)\right)&=\text{Cov}\left(\sum_{i=1}^{n}a_iZ(s_i),\sum_{j=1}^{n}a_jZ(s_j)\right)\\ &=\sum_{i=1}^{n}a_i\sum_{j=1}^{n}a_j\text{Cov}(Z(s_i),Z(s_j))\\ &=\sum_{i=1}^{n}a_i\sum_{j=1}^{n}a_ja_jc(s_i,s_j)=\sum_{i=1}^{n}a_i\sum_{j=1}^{n}a_jc(s_i-s_j)\end{aligned}$$

Exercise 9. ($\star$) Show that if $c_1(\cdot,\cdot)$ and $c_2(\cdot,\cdot)$ are are covariance functions (are non-negative definite) then so are $c_3(\cdot,\cdot)=bc_1(\cdot,\cdot)+dc_2(\cdot,\cdot)$ and $c_4(\cdot,\cdot)=c_1(\cdot,\cdot)\,c_2(\cdot,\cdot)$.

Solution. For all $n \in \mathbb{N}$ and $a_1, ...a_n$

$$\begin{aligned}\sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j c_3 (s_i, s_j) &= \sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j (bc_1 (s_i, s_j) + dc_2 (s_i, s_j)) \\ &= \underbrace{\sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j bc_1 (s_i, s_j)}_{\geq 0} + \underbrace{\sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j dc_2 (s_i, s_j)}_{\geq 0} \\ &\geq 0\end{aligned}$$

Regarding c_4, assume independent stochastic processes $(Y_s)_{s\in S}$ and $(X_s)_{s\in S}$ with mean zero and covariance functions $c_1(\cdot,\cdot)$ and $c_2(\cdot,\cdot)$ correspondingly. Let stochastic processes $(Z_s)_{s\in S}$ with $Z_s = Y_s X_s$. Then

$$\begin{aligned}\mathrm{Cov}(Z_s, Z_t) &= \mathrm{Cov}(Y_s X_s, Y_t X_t) \\ &= \mathrm{E}(Y_s X_s Y_t X_t) \\ &= \mathrm{E}(Y_s Y_t X_s X_t), \quad \text{but} \quad Y_s \perp X_s \\ &= \mathrm{E}(X_s X_t)\, \mathrm{E}(X_s X_t) \\ &= \mathrm{Cov}(X_s, X_t)\, \mathrm{Cov}(Y_s, Y_t) \\ &= c_1(s,t)\, c_2(s,t) = c_4(s,t)\end{aligned}$$

that is $c_4(\cdot,\cdot)$ is a covariance function of a stochastic processes.

Exercise 10. ($\star$) Consider the Gaussian c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_2^2\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Compute the spectral density from Bochner's theorem

Solution. It is

$$\begin{aligned}f(\omega) &= \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp\left(-i\omega^\top h\right) \sigma^2 \exp\left(-\beta \|h\|_2^2\right) \mathrm{d}h \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^{d} \int_{\mathbb{R}} \exp\left(-i\omega_j h_j - \beta h_j^2\right) \mathrm{d}h \\ &= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^{d} \int_{\mathbb{R}} \exp\left(-\beta\left(h_j - (-i\omega/(2\beta))\right)^2\right) \mathrm{d}h_j \\ &= \sigma^2 \left(\frac{1}{4\pi\beta}\right)^{d/2} \exp\left(-\|\omega\|_2^2 / (4\beta)\right)\end{aligned}$$

Exercise 11. ($\star$) Consider the Exponential c.f. $c(h) = \sigma^2 \exp\left(-\beta \|h\|_1^1\right)$ for $\sigma^2, \beta > 0$ and $h \in \mathbb{R}^d$. Compute the spectral density from Bochner's theorem

Solution. It is

$$\begin{aligned}
f(\omega) &= \left(\frac{1}{2\pi}\right)^d \int_{\mathbb{R}^d} \exp\left(-i\omega^\top h\right) \sigma^2 \exp\left(-\beta \|h\|_1^1\right) \mathrm{d}h \\
&= \sigma^2 \left(\frac{1}{2\pi}\right)^d \prod_{j=1}^d \int_{\mathbb{R}} \exp\left(-i\omega_j h_j - \beta |h_j|\right) \mathrm{d}h_j
\end{aligned}$$

where

$$\begin{aligned}
\int_{\mathbb{R}} \exp\left(-i\omega_j h_j - \beta |h_j|\right) &= \int_{-\infty}^0 \exp\left(-i\omega_j h_j - \beta |h_j|\right) \mathrm{d}h_j + \int_0^\infty \exp\left(-i\omega_j h_j - \beta |h_j|\right) \mathrm{d}h_j \\
&= \int_{-\infty}^0 \exp\left(-i\omega_j h_j + \beta h_j\right) \mathrm{d}h_j + \int_0^\infty \exp\left(-i\omega_j h_j - \beta h_j\right) \mathrm{d}h_j \\
&= \int_{-\infty}^0 \exp\left(-\left(i\omega_j - \beta\right) h_j\right) \mathrm{d}h_j + \int_0^\infty \exp\left(-\left(i\omega_j + \beta\right) h_j\right) \mathrm{d}h_j \\
&= \int_0^\infty \exp\left(-\left(\beta - i\omega_j\right) h_j\right) \mathrm{d}h_j + \int_0^\infty \exp\left(-\left(i\omega_j + \beta\right) h_j\right) \mathrm{d}h_j \\
&= \frac{1}{(\beta - i\omega_j)} + \frac{1}{(\beta + i\omega_j)} = \frac{2\beta}{\beta^2 + \omega_j^2}
\end{aligned}$$

hence

$$f(\omega) = \sigma^2 \left(\frac{\beta}{\pi}\right)^d \prod_{j=1}^d \frac{1}{\beta^2 + \omega_j^2}$$
