Tutorial Sheet 4

1 Areal processes

The figure below shows coloured dissolved organic matter (CDOM) across a lake along with the corresponding Moran's I plot (Figure 1).

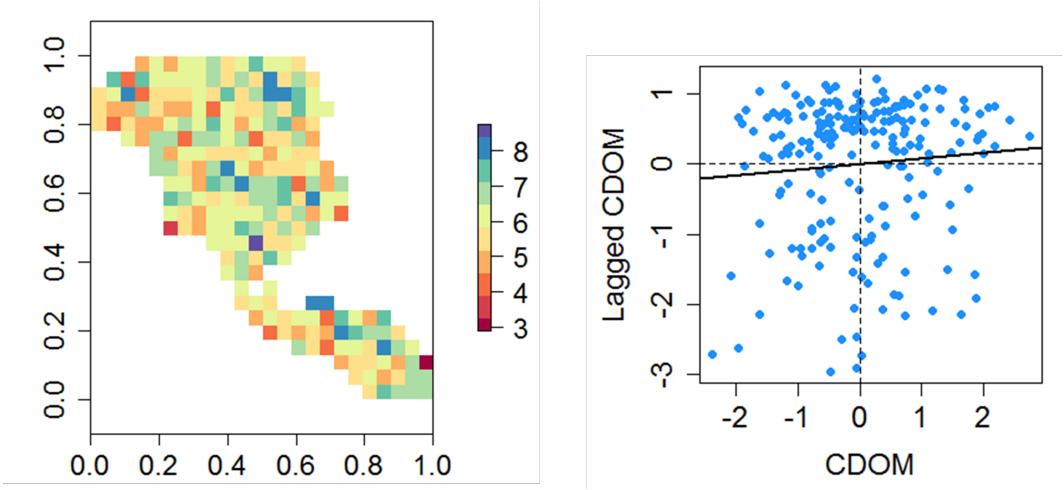

Figure 1: Lake CDOM plotted over space (left) and the corresponding Moran's I plot (right) based on Queen's distance

The estimated Global Moran's I is 0.02 with a variance of 0.02.

Interpreting Moran's I plot

A **Moran's I scatter plot** visualizes spatial autocorrelation by plotting the values of observations in different areas against their **spatially lagged values**, which represent the weighted average of neighboring values. A positive correlation in the plot suggests **spatial clustering** (similar values are located near each other), while a negative correlation indicates **spatial dispersion** (neighboring values tend to be dissimilar). The slope of the regression line through the points provides an estimate of **Moran's I**, a statistic that quantifies the degree of spatial autocorrelation.

Task 1

Define what is meant by the term neighbourhood matrix and discuss two approaches for defining it, giving a drawback of each
Solution

1. **Binary:** Set the *ij*th element w_{ij} of W equal to 1 if areas (i; j) share a common border and equal to zero otherwise.
    - Disadvantages: For small geographical units this can lead to two units being close together but not treated as being spatially close. Length of shared border is not accounted for.
    - Advantage: Don't need to specify a distance (but there are options of what constitutes a common border if regular shaped grid data, like pixels

are used)

2. **Distance:** Set $w_{ij} = 1$ if the distance between the central points of each unit is less than a pre-specified threshold d, and $w_{ij} = 0$ otherwise.
    - Disadvantages; Threshold d is arbitrary and needs to be chosen by the user, how do we choose d?
    - Advantages: small geographical units can be neighbours if close together but do not share a common border.

Task 2

Comment on the spatial variability in CDOM for this lake, with specific reference to the Moran's I plot (Figure 1).
Solution
No evidence of spatial correlation in the dataset using the Moran's I plot. There are points in each quadrant of the plot, and the slope of the line, which represents the Moran's I value, is very flat (close to 0).

Task 3

In your own words what does the Modifiable Areal Unit Problem (MAUP) refers to?
Solution
MAUP is a statistical bias that occurs when data are aggregated into different areal units. The MAUP can lead to different results depending on the scale of the analysis. For example, if you aggregate data at the county level, you may get different results than if you aggregate data at the state level.

Task 4

Figure 2 shows the number of lung cancer cases in Pennsylvania per county, in 2002.

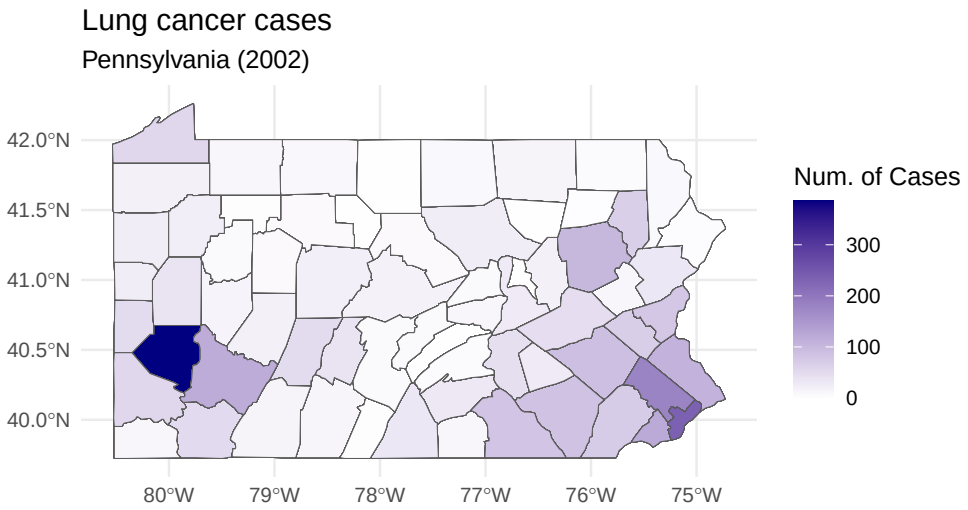

Figure 2: Number of Lung Cancer cases in Pennsylavnia counties in 2002

1. Interpret the map , what does this tells you about the number of lung cancer cases?

2. In addition to the number of cases,,, the data set contains county-level information of the following variables:

- Expected number of cases E_i computed as $E_i = \sum_{j=1}^{m} r_j \times n_j$ (r_j is rate of disease and n_j the population in stratum j)
- Overall county smoking rate

Propose a reasonable spatially explicit LGM to assess the relationship between lung cancer risk and smoking.
Solution

1. You can see that a few counties—especially one in the southwest—have much higher numbers (dark blue), while most counties have relatively low to moderate cases (lighter shades). Overall, lung cancer cases are unevenly distributed, with certain areas experiencing significantly higher concentrations than others.

2. A possible LGM for disease risk modelling could be a BYM with the following structure:

- **observational model**: $Y_i \sim$ Poisson$(\theta_i \times E_i)$, where Y_i represent the number of cases, θ_i is the relative risk, and E_i the expected number of cases
- **the latent field** where the latent components of our model's linear predictor are linked to the RR as $\log(\theta_i) = \alpha_0 + \beta_1 \text{smoking rate}_i + u_i + v_i$ where $\exp(\alpha)$

is the baseline lung cancer relative risk, β_1 is the smoking effect , v_i represents an unstructured spatial effects and u_i is an iCAR component representing the spatially structured variability.

- the **hyperparameters**: the precision (or variance) parameters of the unstructured and structured spatial effects τ_u, τ_v

2 Geostatistics

Task 5

Suppose we have a geostatistical process, $\{Z(s); s \in D\},\ D \subset \mathbb{R}^2$ which is stationary with mean, μ_z and covariance Cov(h). Define what is meant by:

- weakly stationary
- isotropic

Then, write down an expression for the autocorrelation function $\rho_z(h)$ in terms of the covariance function.
Solution
Stationarity implies mean is constant across spatial domain D. Mathematically, a geostatistical process $\{Z(s); s \in D\},\ D \subset \mathbb{R}^2$ is weakly stationary if:

- $E[Z(s)] = \mu_z(s) = \mu_z$ for some finite constant μ_z which does not depend on s.
- Cov$[Z(s), Z(s+h)] = C_z(s, s+h) = C_z(h)$, a finite constant that can depend on *h* but not on *s*.

An isotropic geostatistical process is a stationary process (i.e., the mean is constant, and the covariance only depends on the spatial lag and not the values themselves), but where the covariance between points only depends on the distance between the points, and not the direction. Mathematically, a weakly stationary geostatistical process $\{Z(s); s \in D\}$ is isotropic if the covariance function $C_z(h)$ can be further simplified to $C_z(h) = C_Z(||h||)$; where $h = ||h||$ denotes the length of the lag vector h.
Assuming $Z(s)$ is stationary at lag h we have

$$\begin{aligned}\rho_z(h) &= \frac{\text{Cov}[Z(s), Z(s+h)]}{\sqrt{\text{Cov}[Z(s)]\text{Cov}(Z(s+h)]}} \\ &= \frac{C_z(h)}{\sqrt{C_z(0)C_z(0)}} \\ &= \frac{C_z(h)}{C_z(0)}\end{aligned}$$

Figure 3 below shows summary plots for measurements of a geostatistical process, namely levels of Nitrogen measured at different locations across Chesapeake Bay, USA.

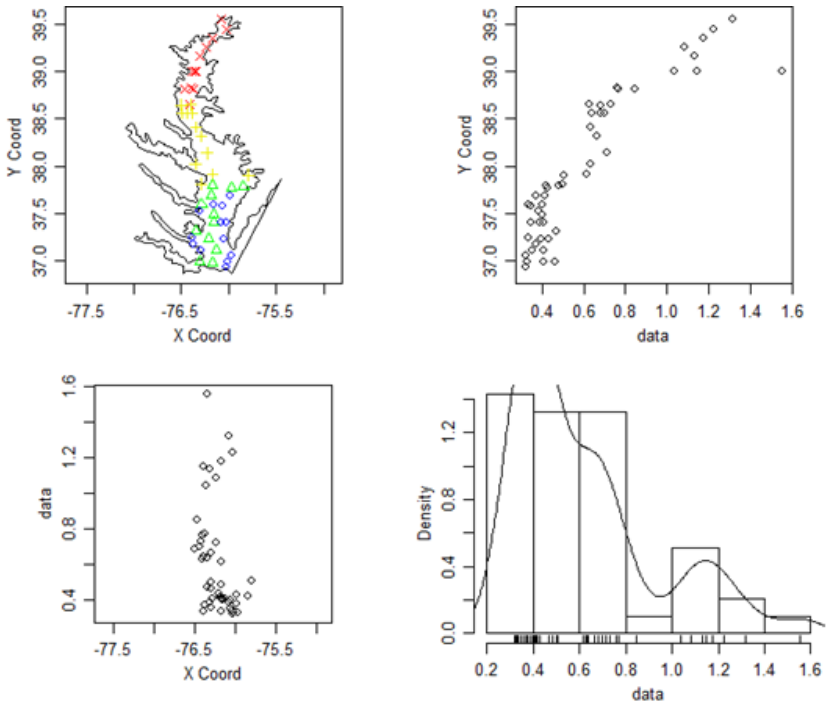

Figure 3: Summary plots for nitrogen levels across Chesapeake bay. Top left plot colour scale red to blue (high to low values).

Task 6

Subjectively, comment on what each of these plots tells you about the data
Solution
Histogram does not look bell shaped & symmetric, there appears to be some evidence of bimodality. It is possible a transformation could be applied to stabilize the variance (e.g., a log transformation). It should be noted however that histograms are difficult to interpret and how they appear can change substantially when the number of bins used is changed.

Task 7

Describe how you would check for spatial correlation in the data and model it if it were present.
Solution
This a geostatistical data set. To assess spatial correlation we could:

- Compute semi-variance for nitrogen levels for all pairs of points and plot (variogram cloud)
- Bin the variogram cloud to obtain empirical variogram
- Construct null envelops under the assumption of no spatial autocorrelation and check for deviations from this assumption.

A possible geostatistical model that accounts for spatial autocorrelation could take the form

$$\log(\text{nitrogen})_i \sim \text{Normal}(\mu_i, \sigma_e^2)$$
$$\mu_i = \beta_0 + Z(s)$$

Here, the log concentrations are assumed to follow a Gaussian distribution with an

observational error σ_e^2 and mean μ_i which is linked to the latent components β_0 ,representing the mean nitrogen concentration levels (on the log scale), and $Z(s)$ - an mean-zero isotropic and stationary Gaussian Field with a variance-covariance matrix (e.g., a Matérn covariance) that accounts for the spatial correlation. $Z(s)$ can be then approximated as a continuously indexed a Gaussian Markov Random field using the stochastic partial differential equation approach (SPDE).

Task 8

An empirical semi-variogram with a MC envelope is shown below (Figure 4). Comment on the plot with regards to the presence of spatial correlation.

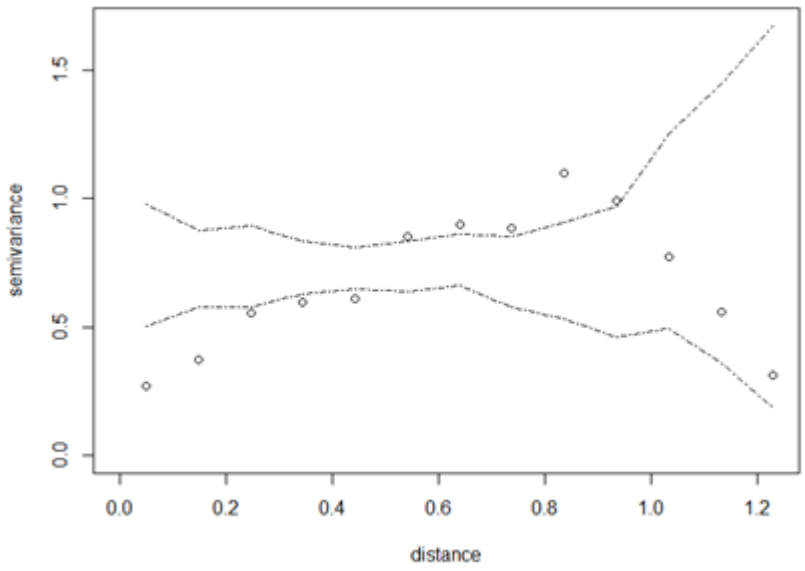

Figure 4: Empirical variogram and MC simulation envelope

Solution
The dashed lines represent a simulated Monte Carlo envelope corresponding to no spatial correlation. This envelope corresponds to the range of plausible semi-variograms that could be produced if the data contained no spatial correlation. Therefore, if the empirical semi-variogram that is calculated from the real data lies completely within the envelopes, then there is no evidence of spatial correlation. If they lie outside at some point then there is evidence of correlation.

3 Point Processes

Task 9

Explain what a spatial point pattern is, define complete spatial randomness, and how it is assessed.
Solution
A Spatial point process is a set of locations, irregularly distributed within a designated region and presumed to have been generated by some form of stochastic (random) mechanism - A realisation from a spatial point process is termed a spatial point pattern – a countable collection of points x_i.

Given any spatial region A, CSR asserts that

i. Conditional on the total number of events in A, the events are uniformly distributed over $|A|$

ii. The random variable *total number of events in* A follows a Poisson distribution with mean $\lambda|A|$.

In (ii) above, λ is termed the intensity, or the expected number of events per unit of area.
A process satisfying (i) and (ii) is called a Spatial Poisson process (with intensity λ).
$K(t) = (1/\lambda) \times E(\text{number of events within a distance t of an arbitrary event})$

- For the case of a Poisson process, $K_{CRS}(t) = \pi t^2$.
- For the case of clustered patterns, we would expect for short distances t that $K(t) > \pi \times t^2$. For regular patterns, we would expect that for short distances t that $K(t) < \pi \times t^2$.

Task 10

Discuss two limitations of using Ripleys' K for assessing CSR in spatial point patterns.
Solution
Ripley's K function grows quadratically with distance (t) because it is based on area (πt^2) This makes interpretation difficult since the expected value under Complete Spatial Randomness (CSR) also follows this quadratic relationship.
Sensitivity to Edge Effects; Points near the boundaries of the study area have fewer neighboring points within distance (t), leading to underestimation of $K(t)$.

Task 11

Figure 5 (a, left) shows the K function for the locations of 500 trees of a particular species within an area of tropical rainforest while Figure 5 (b, right) shows the K function for the Lansing wood tree species. For each of these plots comment on the spatial pattern of the data that generated the K functions with respect to complete spatial randomness.

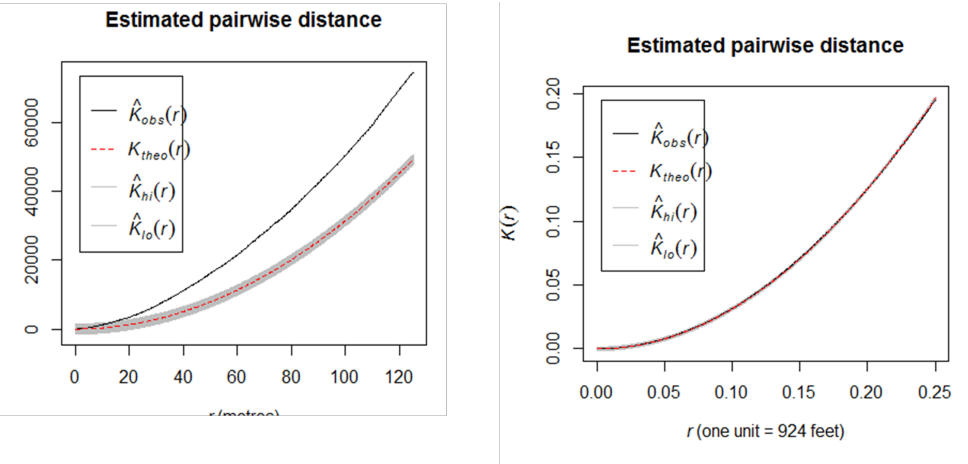

Figure 5: (a) K function for location of a particular species of tree in a tropical rainforest and (b) the K function for the distribution of trees in Lansing Wood.

Solution

- A: The observed K function is above the theoretical K function (under CSR) and outwith the grey confidence region. Hence we can say that the location of trees in the tropical rainforest follow a clustered spatial pattern.
- B: The observed K function is difficult to distinguish from the K function under CSR and falls entirely within the grey confidence region. Hence we can say that the locations of trees in Lansing wood follow a homogeneous spatial Poisson process.