Applied Spatial Statistics (Stat 4121)

Chapter 2-Spatial Data Analysis

Bedilu A. Ejigu

Department of Statistics
Addis Ababa University, Ethiopia

Outline

- Basics on point pattern data (clustering, spatial point processes)
- Lattice data and analysis (spatial weight construction)
- Spatial autocorrelation measures
- Areal data visualization
- Spatial auto-regressive models
- Practical lab examples

Spatial point pattern data

- A **spatial point process** generates realizations that are finite or countably infinite sets of points in the plane.
- Goal: describe, model, and infer structure in the **locations** (and optionally **marks**) of events/objects.
- Observed locations $\{x_i\}_{i=1}^{n} \subset W \subset \mathbb{R}^2$ of events/objects.

What to know

- **Characterizations** of point-process distributions (e.g., via product densities / generating functionals).
- **Campbell theorem** & **moment measures** (mean/variance; higher-order moments).
- **Palm theory**: interior/exterior conditioning (distribution seen from a typical point).

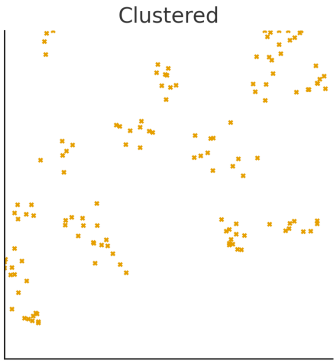
Motivating example (forest stand)

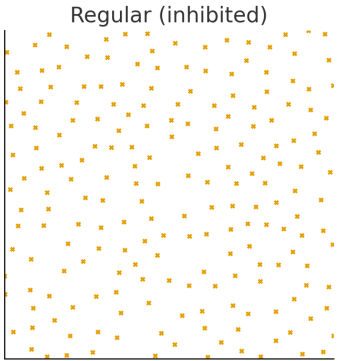
- **Data**: 270 trees in a 100×100 m plot.
- **Phenomena to capture**:
  - **Inhibition** among large trees (competition for space/nutrients).
  - **Clustering** of seedlings near adults (recruitment).
  - **Environmental heterogeneity** (e.g., soil fertility) driving apparent clusters.
- Visual evidence of **clustering** (short-range attraction) and **inhibition** (preferred spacing).
- Large (older) trees exhibit **preferred inter-tree distance** (e.g., ≈ 4 m) $\Rightarrow$ potential design/management signal.

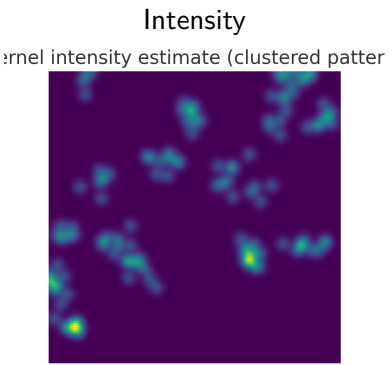
Marked point patterns (adding attributes)

- Each point carries a **mark** (e.g., diameter, species, health status): (x_i, m_i).
- Subpatterns (small vs large trees) reveal structure: small trees often fill gaps around large trees.
- Practical question: are marks **independent** of locations or do we see **mark–location** interaction?

Examples: Simulated point process data







Outline

Spatial influence

- **Goal**: quantify how geographic objects affect each other ("spatial influence").
- Encoded via **adjacency** or **(inverse) distance**; commonly stored in a spatial weights matrix W
- True influence is complex and unobservable $\implies$ we use practical proxies.
- Examples for polygons (e.g., countries):
  - Binary adjacency (share a border = 1, else 0)
  - Centroid distance ("as-the-crow-flies")
  - Border length (longer shared boundary $\implies$ stronger influence)

SEDA

- How to explore spatial relationships with the variable of interest?
- spatial weight construction
- Global and local spatial autocorrelation Versus variogram
- How to handle spatial data in R

Adjacency

- Objects are adjacent if they "touch" (e.g., neighboring countries) or if they are within a distance threshold (common for point data).
- Build an adjacency matrix A from a distance matrix D:
  - Threshold rule: $A_{ij} = \mathbb{I}\{D_{ij} \leq 50\}$ (units as in D).
  - Diagonal convention: set $A_{ii} =$ NA (or 0); no self-adjacency.
  - Convert TRUE/FALSE to 1/0 as needed (e.g., multiply by 1).

Two nearest neighbours (k-NN adjacency)

- Build a k-nearest neighbours adjacency (e.g., $k = 2, 3, 4$):
  - For each row i, rank columns j by increasing D_{ij}.
  - Mark the k smallest non-self distances as neighbours:

$$A_{ij}^{(k)} = \begin{cases} 1, & \text{if } j \text{ is among the } k \text{ nearest to } i, \\ 0, & \text{otherwise.} \end{cases}$$

  - Useful under varying point density: ensures each feature has k neighbours.

Weights matrix

- Use **continuous** weights instead of binary adjacency:
  - Inverse distance: $w_{ij} = D_{ij}^{-\alpha}$ with $\alpha > 0$.
  - Kernel decay (e.g., Gaussian), or weights based on shared border length.
- **Row-normalize** so each row sums to 1: $\tilde{w}_{ij} = w_{ij} \Big/ \sum_j w_{ij}$.
- Handle **Inf/NA** carefully:
  - Self-distances $D_{ii} = 0 \Rightarrow 1/0 = \infty$; set diagonals to 0 or NA before computing W.
  - Replace remaining ∞ with NA, then normalize over valid entries only.

Basic activities to do SEDA

- Type of spatial data and research question
  - Areal/lattice, geostatistics, point pattern
- Assessment of spatial dependency
  - Autocorrelation: Occur in either space(spatial autocorrelation) or time(temporal autocorrelation)
  - spatio-temporal autocorrelation
- Stationarity, Non-stationarity
  - **Stationarity**: process is not changing with respect to either time or space.
  - **Non-Stationarity**: the process changes with respect to time or space.
- Global and Local Statistics
  - Global: one statistic is used to adequately summarize the data
  - Local:useful when the process you are studying varies over space, i.e. different areas have different local values that might cluster together to form a local deviation from the overall mean
- Neighborhoods:who is next to who
  - How the weights constructed (distance, queen, rock,)

- Understand what **spatial autocorrelation** is and why it matters for spatial statistical analysis.
- Understand how a **spatial weights matrix** is constructed.
- Understand the difference between **global** and **local** spatial autocorrelation methods.
- Know the differences among **Local Moran's** I, **Local Geary's** C, and **Getis–Ord** G_i / G_i^*.
- Perform local spatial autocorrelation and hotspot analysis on imported data and **interpret** the results.

Spatial Autocorrelation

- Spatial autocorrelation measures quantify the correlation of the spatial random field $Y(l)$ with itself at different locations.
  - Different statistics have been developed to test for the presence and magnitude of spatial association among areal units
    1. Moran's I, and Geary's C (see Cressie 1993, Banerjee 2015, for discussion).
    2. B-statistics (Ejigu 2020)
- Summary of numeric scales for Moran's I, Geary's C and *B-statistic* indices

Spatial pattern	Geary's C	Moran's I	B (spatio-env)
Clustered, similar $(+\rho)$	$0 < C < 1$	$I > \mathbb{E}[I]$	$B > \mathbb{E}[B]$
Random, independent $(\rho = 0)$	$C = 1$	$I \approx \mathbb{E}[I]$	$B \approx \mathbb{E}[B]$
Dissimilar, contrasting $(-\rho)$	$C > 1$	$I < \mathbb{E}[I]$	$B < \mathbb{E}[B]$

- Main difference between Moran's I and Geary's C statistics are:
  - Moran's index is based on populations, while Geary's index is based on samples, and
  - The numerical range values of both index are different.

Spatial Autocorrelation

- Question: Is the value of an outcome at a location influenced by its **neighbours**?
- Examples: disease status, housing prices, population density.
- Visual assessment alone may be insufficient:
  - How confident are we that a *significant* spatial pattern exists?
  - How can we **quantify** the spatial pattern?
- Compare x_i (at location i) with neighbours x_j and with the overall dataset.
- Different measures address distinct questions:
  - **Moran's** I: are deviations from the mean similar across neighbours? $(x_i - \bar{x})$ vs $(x_j - \bar{x})$
  - **Geary's** C: are *pairwise differences* $(x_i - x_j)$ small/large relative to variance?
  - **Getis–Ord** G/G^*: are local *products/sums* of values large or small compared to global?
- Neighbour definition is essential $\Rightarrow$ weights.

Global vs Local Spatial Autocorrelation

- **Global:** combines information into a *single* statistic for the dataset.
  - Question: Is there overall spatial autocorrelation?
- **Local:** computes a statistic *for each location*.
  - Components/"decomposition" of global measures.
  - Multiple values per dataset (one per feature).
  - Focuses on **location–neighbour** interactions.
  - Identifies **where** clusters are located.
  - Distinguishes **cluster type**:
    - High–High (*hotspots*)
    - Low–Low (*coldspots*)
    - High–Low and Low–High (*outliers*)
- Use the **same** spatial weights matrix in both settings.

Moran's I statistic and spatial weighting matrix

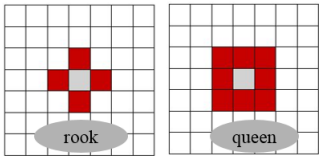
- Moran's I statistic

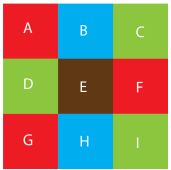
$$I = \frac{n \sum_i \sum_j w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2} \tag{1}$$

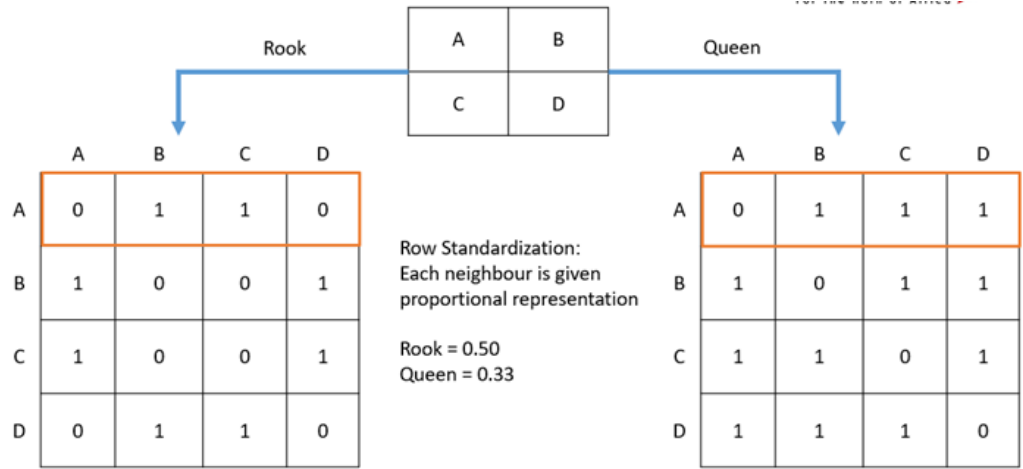
- The choice of a weighting matrix is a central component of Moran's I as it assumes **prior structure of spatial dependence**.
- Specification of spatial weights matrix starts by identifying neighborhood structure of each cell.
- Spatial weighting matrix is created based on the concept of 1st law of geography.
  1. Rook Contiguity
  2. Queen Contiguity
  3. q-nearest neighbors of cases
  4. Exponential distance weights

Weighting Matrix Construction Approaches

- Typically, a spatial weighting matrix is created based on the concept of geographical distance/neighborhood.
- Common type of weighting matrices (Waller & Gotway, 2004; Getis, 2009):
  1. Rook Contiguity $w_{ij} = \begin{cases} 1, & \textit{if location } i \textit{ and } j \textit{ share boundaries}, \\ 0, & \textit{otherwise}. \end{cases}$
  2. Queen Contiguity: each neighboring cell in all directions are given the value 1.







- **Row standardization**: each neighbour receives proportional weight so row sums are 1 (e.g., Rook $\rightarrow$ 0.50, Queen $\rightarrow$ 0.33 per neighbour in this toy example).

Weighting Matrix Construction Approaches

3 q-nearest neighbors of cases:

$$w_{ij} = \begin{cases} 1 \text{ , } \textit{if site } i \textit{ is among } q \textit{ nearest neighbors of site } j, \\ 0 \text{ , } \textit{otherwise}. \end{cases}$$

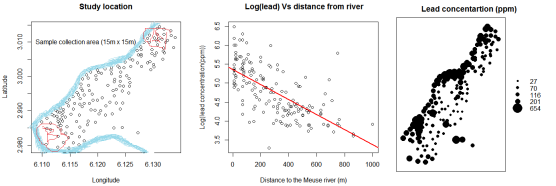
4. Exponential distance weights:

$w_{ij} = \exp(-pd_{ij})$, where d_{ij} represents Euclidean distance

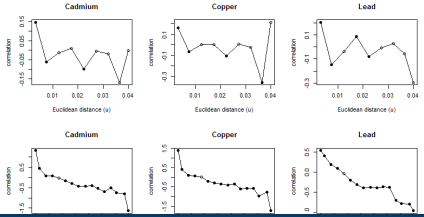
- Limitations of distance based-functions to create weighting matrices in spatial statistics is not well explored (Earnest *et al.* 2007, Wang *et al.* 2013).
- Methods of constructing weights should take into account how the outcome variable is generated over-spatial units under consideration.
- **None of these tools directly represents dynamic aspects of environmental effects on the occurrence of the outcome variable.**

Motivation to consider 3^{rd} LG

- Prevalence of environmentally-mediated diseases highly linked with environment $\Longrightarrow$ Environmental configuration
- **Meuse River Dataset** - Heavy Metal Concentration
  - Distance from the river has impact on lead concentration



  - Correlation decreases as distance from the river increases.

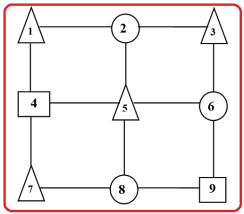


$\Leftarrow$ Correlogram plot using 1^{st} law of geography

$\Leftarrow$ Correlogram plot using 3^{rd} law of geography.

Other weighting matrix construction approaches

- Two locations close geographically but separated by other barriers may not be considered as near neighbors (Ejigu & Wencheko 2020).
- To create a meaningful weighting matrix:
  1. decide which relationship between observation are to be given non-zero weights
  2. assign weights to identified neighbor links:
     a) environmental contiguity, b) using meaningful functions



- Similar symbol represents similar environmental condition.
- Ejigu *et al* 2020 weighting matrix

$$m_{ij} = \exp(-(\alpha u_e + (1-\alpha)u_s)), \quad (2)$$

where $u_e = |e_i - e_i'|$ is the absolute difference in the environmental covariate between two locations, and $u_s = ||s - s'||$ is the euclidean distance between two locations

- Environmentally, for site 1 neighbors are not 2 and 4, rather 3,5,7.

$$N = \text{number of observations}, \quad \bar{x} = \frac{1}{N}\sum_i x_i, \quad W = \sum_i \sum_j w_{ij}.$$

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}\,(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}.$$

Toy grid ($N = 4$), mean $\bar{x} = 27$, Rook weights with $W = 8$.

Values and deviations

Cell	x	$x - \bar{x}$
A	36	9
B	22	-5
C	26	-1
D	24	-3

Cross-deviations summary

- $\sum_{i,j} w_{ij}(x_i - \bar{x})(x_j - \bar{x}) = -72$
- $\sum_i (x_i - \bar{x})^2 = 116$
- $I = \frac{4}{8} \cdot \frac{-72}{116} \approx -0.31$

Let x_i be the value at location i, $\bar{x}$ the mean, N be the number of observations, $\bar{x}$ the mean, w_{ij} spatial weights ($w_{ii} = 0$), and $W = \sum_i \sum_j w_{ij}$. For location i,

$$I_i = \frac{N}{W} \frac{(x_i - \bar{x}) \sum_j w_{ij} (x_j - \bar{x})}{\sum_k (x_k - \bar{x})^2}.$$

- **Inputs:** x (outcome), indices i, j, weights w_{ij}, dataset mean $\bar{x}$, total weight W.
- **Interpretation (range $\approx [-1, 1]$):** $I_i > 0$ cluster [suggests a local cluster (high with high, or low with low)], $I_i = 0$ none, $I_i < 0$ outlier[suggests a local spatial outlier (high with low, or low with high)].

Toy 2×2 grid with values:

$$\begin{matrix} 36 & 22 \\ 26 & 24 \end{matrix} \quad \Rightarrow \quad \bar{x} = 27,\ N = 4,\ \text{Rook weights with } W = 8.$$

Cross-deviations (per cell):

$A : (36-27)(22-27) + (36-27)(26-27) = -45 + (-9) = -54$

$B : -45 + 15 = -30$

$C : -9 + 3 = -6$

$D : 15 + 3 = 18$

Denominator $\sum (x - \bar{x})^2 = 116$

Local indices:

$$I_A = \frac{4}{8} \cdot \frac{-54}{116} = -0.23\,, \quad I_B = \frac{4}{8} \cdot \frac{-30}{116} = -0.13\,,$$

$$I_C = \frac{4}{8} \cdot \frac{-6}{116} = -0.03\,, \quad I_D = \frac{4}{8} \cdot \frac{18}{116} = 0.08\,.$$

Type labels: High–High, Low–Low (clusters); High–Low, Low–High (outliers).

Interpreting Results

- Use **z-scores** and **p-values** based on observed vs. expected and variance.
- If $p > \alpha$: no evidence of autocorrelation.
- If $p < \alpha$:
  - Positive z: positive spatial autocorrelation.
  - Negative z: negative spatial autocorrelation.
- Moran scatter plot: **Moran's I equals the slope** of the regression line.

Interpreting Results (Quadrant Plot)

- **Top Right:** High–High cluster (hotspot).
- **Bottom Left:** Low–Low cluster (coldspot).
- **Top Left:** Low–High outlier.
- **Bottom Right:** High–Low outlier.

Geary's C

Geary's C test statistics

$$C = \frac{N-1}{2W} \frac{\sum_i \sum_j w_{ij} (x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2}.$$

- **Interpretation**: $C < 1$ positive autocorrelation; $C = 1$ none; $C > 1$ negative autocorrelation.
- Geary's C worked example (Using Rook weights, $N - 1 = 3$, $2W = 16$.)
- $\sum_{i,j} w_{ij}(x_i - x_j)^2 = 608$, $\sum_i (x_i - \bar{x})^2 = 116$
- $C = \frac{3}{16} \cdot \frac{608}{116} \approx 0.98$

Local Geary's C, for location i,

$$C_i = \frac{N-1}{2W} \frac{\sum_j w_{ij} (x_i - x_j)^2}{\sum_k (x_k - \bar{x})^2}.$$

- **Interpretation (range $[0, 2]$):** $C_i < 1$ cluster, $C_i = 1$ none, $C_i > 1$ outlier.

Getis–Ord G and G^*

$$G = \frac{\sum_i \sum_j w_{ij}\, x_i x_j}{\sum_i \sum_j x_i x_j}, \qquad \text{(for } G^*\text{, include } i = j \text{ in the numerator).}$$

- No centering by $\bar{x}$; not advised to use row standardization.
- Cannot identify *negative* autocorrelation (focus on hot/cold spots).
- Interpret relative to expected value $\mathbb{E}[G] = \frac{W}{N(N-1)}$ (for simple binary w_{ij}, $i \neq j$).

Getis–Ord G example with Rook weights $W = 8$, $N = 4$.

- Numerator $\sum_{i,j} w_{ij} x_i x_j = 5760$
- Denominator $\sum_{i,j} x_i x_j = 4316$
- $G = 5760/4316 \approx 1.33$
- Expected $G = \dfrac{W}{N(N-1)} = \dfrac{8}{4 \cdot 3} \approx 0.67$
- $G > \mathbb{E}[G] \Rightarrow$ potential hotspots.

Local Geary's C (worked example)

Same toy grid, Rook weights; $N-1=3$, $2W=16$.

$$\begin{aligned}
&\text{Differences (A)} : 14^2 + 10^2 = 296,\\
&\text{Differences (B)} : 14^2 + (-2)^2 = 200,\\
&\text{Differences (C)} : 10^2 + 2^2 = 104,\\
&\text{Differences (D)} : 2^2 + (-2)^2 = 8.
\end{aligned} \qquad \sum (x - \bar{x})^2 = 116.$$

Local indices:

$$C_A = \frac{3}{16} \cdot \frac{296}{116} = 0.48,\ C_B = \frac{3}{16} \cdot \frac{200}{116} = 0.32,\ C_C = \frac{3}{16} \cdot \frac{104}{116} = 0.17,\ C_D = \frac{3}{16} \cdot \frac{8}{116} = 0.01.$$

Type labels: $C_i < 1$ cluster (HH/LL); $C_i > 1$ outlier.

Getis–Ord G_i / G_i^* (definition)

Let x be the outcome, w_{ij} spatial weights. No adjustment by N and no centering by $\bar{x}$. Not advised to row-standardize.

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}, \quad \text{(for } G_i\text{, exclude } j = i \text{ in the numerator)}.$$

- Interpreted relative to **expected value** (for simple binary w_{ij}, $i \neq j$):

$$\mathbb{E}[G] \;=\; \frac{W}{N(N-1)}.$$

- Detects **hot/cold spots**; does *not* diagnose negative autocorrelation.

Getis–Ord G_i^* (worked example)

Toy grid with Rook weights: $N = 4$, $W = 8$.

$$\sum x = 36 + 22 + 26 + 24 = 108.$$

Neighbour sums (including i for G_i^*):

$$A : 36 + 22 + 26 = 84 \Rightarrow G_A^* = \frac{84}{108} = 0.78,$$
$$B : 22 + 36 + 24 = 82 \Rightarrow G_B^* = \frac{82}{108} = 0.76,$$
$$C : 26 + 36 + 24 = 86 \Rightarrow G_C^* = \frac{86}{108} = 0.80,$$
$$D : 24 + 22 + 26 = 72 \Rightarrow G_D^* = \frac{72}{108} = 0.67.$$

Expected value:

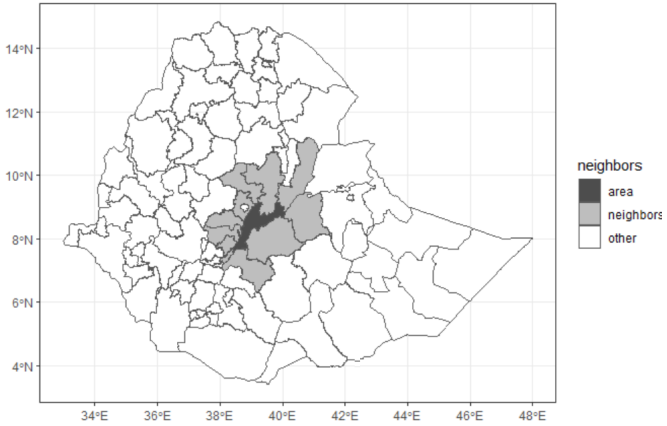
$$\mathbb{E}[G] = \frac{W}{N(N-1)} = \frac{8}{4 \cdot 3} \approx 0.67.$$

Rule of thumb: $G^* > \mathbb{E}[G]$ potential hotspots; $G^* < \mathbb{E}[G]$ potential coldspots.

Constructing neighborhoods for areal data using R

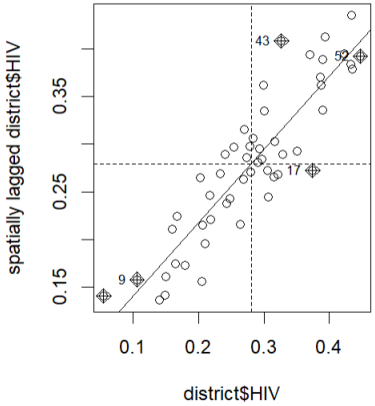
- The **spdep** package has a number of function to construct neighborhoodness based on different criteria.
  - **poly2nb**: to construct the list of neighbors based on areas with contiguous boundary.
  - knearneigh : to obtain matrix of indices of points belonging to the set of the k nearest neighbors.
  - knn2nb: to convert the list obtained using knearneigh function into a neighbor list.
  - dnearneigh: to get list of neighbors based on a distance between specific limits.
  - nblab & nblag: to get neighbors of order k based on contiguity.
  - **nb2listw**: to create a spatial neighborhood matrix containing the spatial weights corresponding to a neighbors list.
- Review the RMarkdown output "Spatial EDA.html" file for further details and practical examples using R.

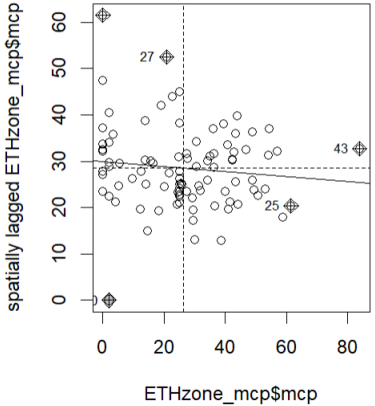
- Consider the admin2 boundaries (Zones) of Ethiopia, and
  1. Visualize which zone is a neighbor of others?
  2. Identify zones with maximum and minimum number of neighbors



Moran I scatter plot

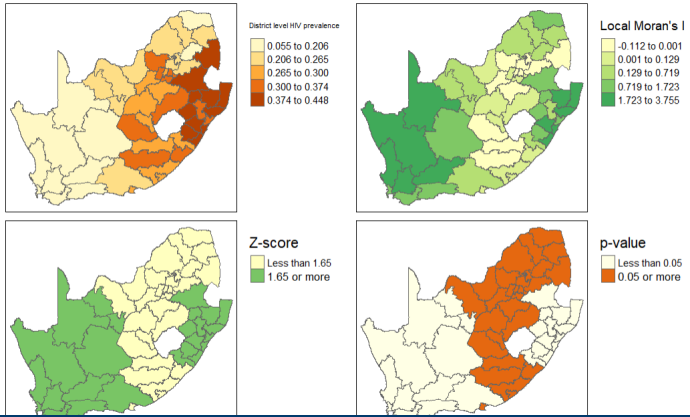
- Moran's *I* scatter plot: A plot of spatial data against its spatially lagged values.
- What the two Moran's *I* scatter plot tells us?
  - In which dataset spatial dependency demonstrated?





Hot-spot or Cluster identification

- Local Moran's I used to identify clusters of the following types:
  - High-High: areas of high values with neighbors of high values,
  - High-Low: areas of high values with neighbors of low values,
  - Low-High: areas of low values with neighbors of high values,
  - Low-Low: areas of low values with neighbors of low values.



Practical Session

- In this practical session, you will learn the following:
  1. How to define/construct spatial neighborhood using *poly2nb*function;
  2. How to convert spatial neighborhood into spatial weights;
  3. How to get global and local Moran's *I* statistics to check spatial autocorrelation;
  4. Cluster identification (if any).
- Instructions for the lab practice
  - Load "*mCP*2019.*csv*" and "*SA_SDP_HIV.csv*" datasets and be familiar with the variables in the data.
  - Check the columns of the data and ensure that you understand what each column means.
  - Get and visualize the shapefiles of Ethiopia nd South-Africa
  - Construct spatial neighborhood matrix and compute Moran's I statistics to check the presence of spatial autocorrelation.
- For detailed support and get all codes, refer the Rmarkdown output "Spatial EDA.html" file.

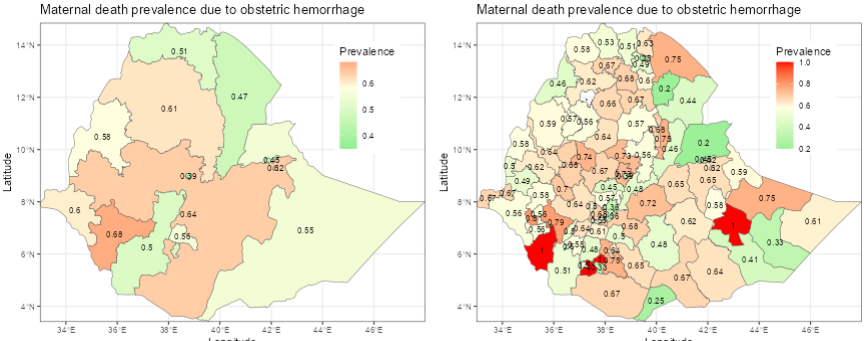
Areal Data Visualization

Areal data: what & why

- **Areal data** = attributes attached to *polygons* (e.g., Zones, sub-cities, woredas, districts).
- Typical tasks: **choropleths**, **comparisons across areas**, **hotspot screens**,
- Key ingredients:
  - **Geometry:** polygon boundaries (topology, gaps/overlaps, dissolved levels).
  - **Attributes:** counts, rates, indices, categories.
  - **Support scale and direction indicator:** areal unit size/shape; North-east indicator.
  - **Projection/CRS:** choose an appropriate local projection for distance/area fidelity.

Areal data: what & why

- **Areal data** = attributes attached to *polygons* (e.g., Zones, sub-cities, woredas, districts).
- Typical tasks: **choropleths**, **comparisons across areas**, **hotspot screens**,
- Key ingredients:
  - **Geometry:** polygon boundaries (topology, gaps/overlaps, dissolved levels).
  - **Attributes:** counts, rates, indices, categories.
  - **Support scale and direction indicator:** areal unit size/shape; North-east indicator.
  - **Projection/CRS:** choose an appropriate local projection for distance/area fidelity.



Choropleth fundamentals

A choropleth map is a thematic map that uses color or patterns to represent statistical data across predefined geographic areas.

- **Map the right quantity:** use *rates/ratios* or *intensive* variables (not raw counts).
- **Palettes:**
  - **Sequential** (low→high) for unipolar data (e.g., prevalence).
  - **Diverging** (low↔high around a meaningful center) for anomalies/changes.
  - Prefer colorblind-safe schemes; avoid misleading saturation for tiny areas.
- **Classing:** quantiles, equal interval, natural breaks, Keep small classes.
- **Legend & NA:** label units & breaks clearly; show *No data*/*Zero* distinctly.
- **Design tips:** thin boundaries, hierarchy (main vs internal borders), contextual basemap kept light.

Examples

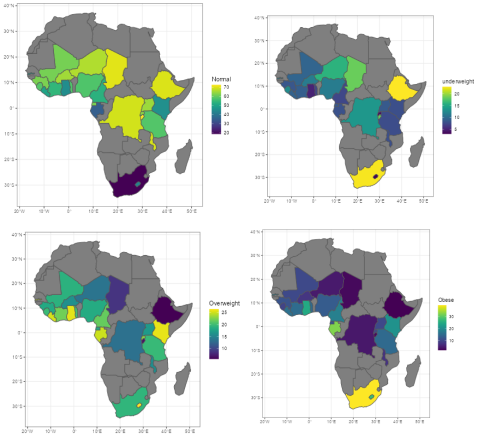


Figure 1: Malnutrition

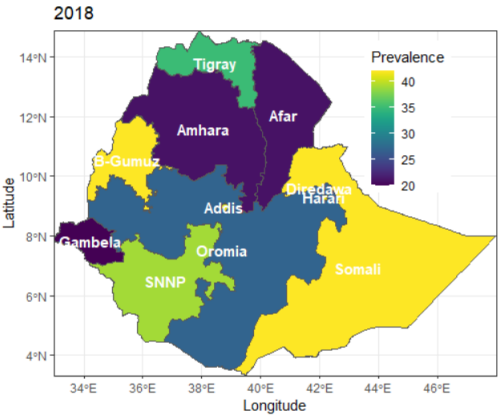


Figure 2: Quality of family planning counseling

Preparing areal attributes

- **Normalize:** rate or proportion

$$r_i = \frac{y_i}{n_i} \times c \quad \text{(e.g., } c = 1000 \text{ for per-1,000)}$$

- **Standardize:** age/sex standardization for fair comparisons across areas.
- **Transform:** log or asinh for skew; winsorize extreme outliers cautiously.
- **Stabilize small-n noise:** empirical Bayes (EB) smoothing for rates

$$\hat{r}_i^{\mathrm{EB}} \;=\; w_i\, r_i \;+\; (1 - w_i)\, \bar{r}, \quad w_i = \frac{n_i}{n_i + \alpha}$$

  where n_i is the area population/exposure and α a prior strength (conceptually).
- **Uncertainty:** map confidence/credible intervals or flag low-reliability areas.

Quality checks & common pitfalls

- **Areal/Ecological fallacy:** results depend on zoning/scale; avoid individual-level claims from area data.
- **Projection/CRS:** pick one suited to your region; document it on the figure or caption.
- **Break sensitivity:** conclusions can flip with different classing—report method and test robustness.
- **Sparse data:** unreliable rates for tiny populations—flag, smooth, or aggregate.
- **Perception:** large rural polygons dominate area—consider cartograms or symbol overlays.
- **Reproducibility:** save code, seeds, and the exact breakpoints/palette for auditability.

Areal/Lattice data Modeling
Spatial Autoregressive Models

Background

- Inferences relies not only on the quality and completeness of the collected data, but also chosen statistical model validity.
- Models that can adequately represent the true underlying process by which the data were created, reveal data structure, provide performance of parameter estimators, evidence robustness of statistical inferences, and allow evaluation of the underlying assumptions that support interpreting analytic findings.
- Depending on how the data generated and the source of uncertainty, statisticians have different modeling approaches.
  - Categorical data $\Rightarrow$ Logistic Regression,GEE, **GLMM**
  - Count data $\Rightarrow$ Poission Regression
  - Continuous data $\Rightarrow$ Linear regression, LMM
  - Time-to-event data $\Rightarrow$ Survival modeling approaches
  - Geo-referenced data $\Rightarrow$ **Spatial Modeling** (*our focus*)

Recap: Model Building

1. Clearly state the main research question
2. Perform exploratory data analysis
   - assess the relationship between the outcome and covariates
   - test for residual spatial correlation.
3. Decide the **purpose of modeling**
   1. **Explanatory modeling**, where greater emphasis is placed on understanding the relationships between the health outcome and risk factors.
      - model-building process focus on the selection of candidate covariates which should be informed by context-specific scientific knowledge of the underlying disease process.
   2. **Predictive modeling**, whose main focus is on maximizing the predictive accuracy of the model.
      - effort is directed primarily towards the development of a model that can predict as accurately as possible future data generated by the same underlying process.
4. **Model validation**: Assess assumptions and model compatibility

The issue of over-dispersion in count spatial data

- What is overdispersion?
  - Occurs when the variance of count data exceeds the mean.
  - Violates the Poisson assumption:
$$Var(Y) = E(Y)$$
- Why does it matter?
  - Standard models (e.g., Poisson regression) underestimate uncertainty.
  - Leads to overly optimistic confidence intervals and p-values.
- Example
  - Consider $Y = \sum_i^n X_i$, where X_i are correlated binary variables.
  - If X_i are independent, Y follows a Binomial distribution: $E(Y) = np$, $Var(Y) = np(1-p)$.
  - If X_i are correlated, the variance increases:
$$Var(Y) = np(1-p) + \sum_{i \neq j} Cov(X_i, X_i).$$
  - This leads to overdispersion.

Interest lies on handling spatial dependency.

1. **Spatial autoregressive model** - Areal/lattice data
   - the response at each study location l_i is a function of not only the possible exploratory variables at l_i , but of the values of the response at neighbouring location(s)

$$Y = X'\beta + \rho WY + \epsilon, \tag{3}$$

2. **Geostatistical model** (Diggle & Ribeiro, 2007) -
   - spatial dependence in the response variable is modelled by directly specifying a parametric covariance structure.

$$Y_i = d(x_i)^\top \beta + S(x_i) + Z_i. \tag{4}$$

where $d(x_i)$ is a vector of spatially-referenced covariates, $\{S(x) : x \in \mathbf{R}^2\}$ is a stationary and isotropic Gaussian process with zero mean, variance σ^2, and $\text{Corr}\{S(x), S(x')\} = \rho(x, x') = \exp(-d_{ij}/\phi)$, $Z_i \sim N(0, \tau^2)$.

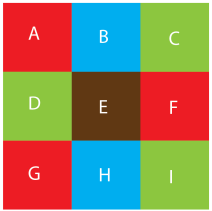
Areal/lattice data modeling

- Spatial neighborhood matrices (spatial weights) play a crucial role in areal/lattice type of data modeling.
- How to construct neighborhood matrices
  1. Neighbors based on geographical contiguity
  2. Neighbors based on ***k*** nearest neighbors
  3. Neighbors based on distance
  4. Neighbors based on distance and covariates (3^{rd} law of geography)
- Spatial weights
  1. binary neighbor list
  2. inverse distance values

Spatial Autoregressive Models (SAM)

- The response at each location i is a function of not only the exploratory variable at i, but also of values of the response at neighboring locations j as well (Cressie, 1993).

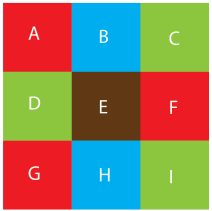
$$Y = X'\beta + \rho WY + \epsilon, \qquad (5)$$
$$Var(Y) = (I - \rho W)^{-1}\sigma^2 I(I - \rho W')^{-1}$$



Spatial Autoregressive Models (SAM)

- The response at each location i is a function of not only the exploratory variable at i, but also of values of the response at neighboring locations j as well (Cressie, 1993).

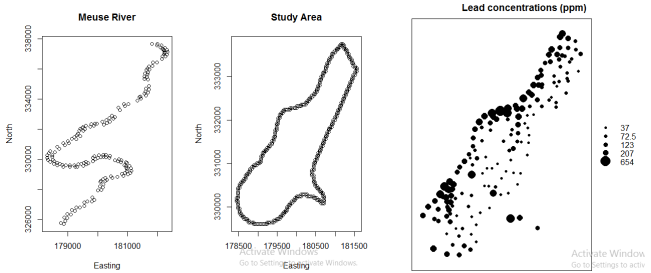
$$Y = X'\beta + \rho WY + \epsilon, \qquad (5)$$
$$Var(Y) = (I - \rho W)^{-1}\sigma^2 I(I - \rho W')^{-1}$$



- The choice of a weighting matrix is a central component of SAM as it assumes **prior structure of spatial dependence**.
- Specification of spatial weights matrix starts by identifying neighborhood structure of each cell.

Application to Meuse Heavy Metals Data

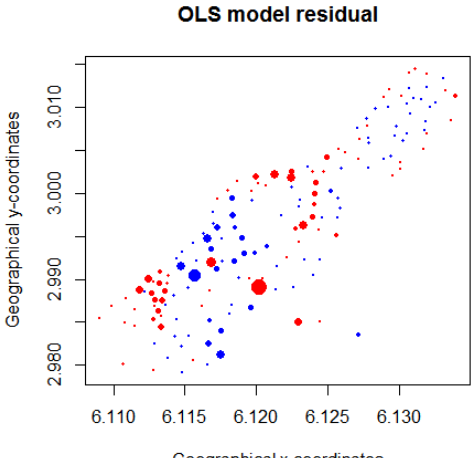
- Meuse dataset is one of the classical datasets in spatial statistics to demonstrate different theories in the field.



- This dataset is available in **gstat** and **sp** packages.
- Moran's I and B statistics reveal that, there is significant spatial and environmental autocorrelation.

Cont'd

- Residuals based on the classical modeling approach show spatial pattern (Moran's I_{res}, P-value=0.0001).
- Our modeling approach should address such autocorrelation.



Model Result

- Models fitted based on standard and proposed weighting matrix approach
- Models B, C and D, the weighting matrix W is computed using both geographical separation distance and distance from the river (at $\alpha = 0.3, 0.6, 0.9$).

		Traditional weights	Covariate Dependent weights		
		Model A	Model B	Model C	Model D
		$\hat{\beta}(se(\hat{\beta}))$	$\hat{\beta}(se(\hat{\beta}))$	$\hat{\beta}(se(\hat{\beta}))$	$\hat{\beta}(se(\hat{\beta}))$
log (lead)	β_0	5.303 (0.302)*	5.319 (0.309)*	6.428 (0.779)*	6.954 (0.593)*
	Elevation	0.003 (0.019)	0.009 (0.026)	−0.181 (0.046)*	−0.217 (0.056)*
	Distance from river	0.002 (0.0002)*	−0.002 (0.0002)*	−0.001 (0.0002)*	−0.001 (0.0002)*
	λ	0.856*	0.856*	0.963*	0.923*
	σ^2	0.228	0.233	0.194	0.184
	$pseudo - R^2$	37.679	41.287	42.026	41.161
	AIC	236.211	239.503	212.698	205.193
	RMSE	0.477	0.484	0.44	0.43
log (copper)	β_0	4.026 (0.169)*	4.030 (0.237)*	4.877 (0.643)*	5.155 (0.439)*
	Elevation	0.008 (0.015)	0.017 (0.0198)	−0.131 (0.035)*	−0.154 (0.043)*
	Distance from river	−0.001 (0.0002)*	−0.001 (0.0001)*	−0.001 (0.0001)*	−0.001 (0.0002)*
	λ	0.755*	0.856*	0.956*	0.917*
	σ^2	0.139	0.137	0.116	0.111
	$pseudo - R^2$	35.974	39.498	40.511	38.329
	AIC	154.503	151.93	128.2	121.03
	RMSE	0.373	0.37	0.34	0.333

Spatial weights W: defining neighbors on areal units

- **Contiguity:** *rook* (shared edge), *queen* (edge or vertex).
- **Distance / kNN:** guarantee k neighbors per area (useful if units vary in size).
- **Row-standardization:** $\tilde{W}$ with rows summing to 1; simplifies interpretation of spatial lag.
- **Choice matters:** results can be sensitive; test robustness across plausible W.

$$W_{ij} = \begin{cases} 1 & \text{if } j \text{ is a neighbor of } i \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad \tilde{W}_{ij} = \frac{W_{ij}}{\sum_j W_{ij}}$$

Get the data and prepare the data for analysis.

```
    library(gstat)
    library(sp)
    data("meuse.all")
    data=meuse.all
    # Transform coordinates (UTM to geographical coordinates)
    sps  <- SpatialPoints(data[, c("x", "y")], proj4string = CRS("+proj=utm +
zone=32"))
    spst <- spTransform(sps, CRS("+proj=longlat +datum=WGS84"))
    # Adding long and lat to the dataset after conversion
    data[,c("long", "lat")] <- coordinates(spst)
    cords=cbind(data$long,data$lat) #Geographical cordinates
    cord=cbind(data$dist.m,data$dist.m) #Cords based on distance from the
river
    cord2=cbind(data$elev,data$elev)# Cords based on relative  elevation
```

Weighting matrix construction

Standard distance based weighting matrices

```
library(spdep)
#a. Exponential
sep.dist<- as.matrix(dist(cords))
w_Ex <- exp(-sep.dist); diag(w_Ex) <- 0
rs_Ex <- rowSums(w_Ex);
w_Ex<- apply(w_Ex, 2, function(q) q/rs_Ex)
# b. Inverse distance
Inv_w=as.matrix(1/dist(cords));diag(Inv_w) <- 0
rs_inv <- rowSums(Inv_w); # row sum
Inv_w <- apply(Inv_w, 2, function(q) q/rs_inv)
```

Weighting matrix construction (2)

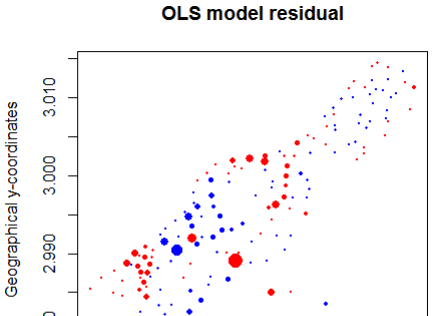
Both Geographical proximity and covariate dependent weights

```
                #alpha={0.2, 0.5,0.8} # Controls the strength of covariate effects
    on W
                alpha1=0.2
                pw1=as.matrix(exp(-(alpha1*dist(cord)+(1-alpha1)*dist(cords))))
                diag(pw1) <- 0; rs_pw1 <- rowSums(pw1);
                pw1 <- apply(pw1, 2, function(q) q/rs_pw1)
                alpha2=0.5
                pw2=as.matrix(exp(-(alpha2*dist(cord)+(1-alpha2)*dist(cords))))
                diag(pw2) <- 0; rs_pw2 <- rowSums(pw2);
                pw2 <- apply(pw2, 2, function(q) q/rs_pw2)
                alpha3=0.8
                pw3=as.matrix(exp(-(alpha3*dist(cord)+(1-alpha3)*dist(cords))))
                diag(pw3) <- 0 ;rs_pw3 <- rowSums(pw3);
                pw3 <- apply(pw3, 2, function(q) q/rs_pw3)
```

Model Fitting

Classical linear model

```
ols=lm(log(lead) ~elev + dist.m, data=data)
plot(data$long, data$lat, col=c("blue",
"red")[sign(resid(ols))/2+1.5], pch=19,
cex=abs(resid(ols))/max(resid(ols))*2, xlab="geographical
xcoordinates", ylab="geographical y-coordinates")
```



Model fitting using tradational weighting matrices

Null model and With covariates

```
library(spdep)
library(spatialreg)
summary(St2_null<-spautolm(log(copper) ~1,family="SAR", listw =mat2listw(w_Ex,
    style ='W'), data=data))
summary(St3_null<-spautolm(log(copper) ~1,family="SAR", listw =mat2listw(Inv_w ,
    style ='W'), data=data))

# Model A: Including  distance from the river only in W
summary(PW1<-spautolm(log(copper) ~ elev,family="SAR", listw =mat2listw(pw1, style
     ='W'), data=data))
summary(PW2<-spautolm(log(copper) ~ elev,family="SAR", listw =mat2listw(pw2, style
     ='W'), data=data))
summary(PW3<-spautolm(log(copper) ~ elev,family="SAR", listw =mat2listw(pw3, style
     ='W'), data=data))
```

Model fitting (2)

Including covarite both in the mean structure and W

```
    summary(pw1_cov<-spautolm(log(copper) ~ elev+dist.m,family="SAR", listw =
mat2listw(pw1, style ='W'), data=data))
    summary(pw2_cov<-spautolm(log(copper) ~ elev+dist.m,family="SAR", listw =
mat2listw(pw2, style ='W'), data=data))
    summary(pw3_cov<-spautolm(log(copper) ~ elev+dist.m,family="SAR", listw =
mat2listw(pw3, style ='W'), data=data))
```

Model Assessment

- R^2
- AIC
- Mean square error
- Moran residual test and plots

```
            cat(paste ("Standard-Model 2-R2:", round((1-exp(-(2/nrow(data))*(
logLik(St2_cov)-logLik(St2_null ))))*100,3), "\n"));
            cat(paste ("Standard-Model 3-R2:", round((1-exp(-(2/nrow(data))*(
logLik(St3_cov)-logLik(St3_null ))))*100,3), "\n"))
            # Using Proposed weights
            cat(paste ("Proposed-Model 1-R2:", round((1-exp(-(2/nrow(data))*(
logLik(pw1_cov)-logLik(pw1_null))))*100,3), "\n"));
            cat(paste ("Proposed-Model 2-R2:", round((1-exp(-(2/nrow(data))*(
logLik(pw2_cov)-logLik(pw2_null))))*100,3), "\n"));
            cat(paste ("Proposed-Model 3-R2:", round((1-exp(-(2/nrow(data))*(
logLik(pw3_cov)-logLik(pw3_null))))*100,3), "\n"))
```