Spatial Data Analysis

Week 7: Areal Data II

Meredith Franklin

Department of Statistical Sciences and School of the Environment

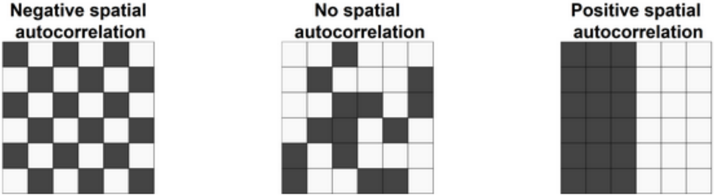
October 17th, 2025

Areal Data II

- More on Global Indexes of Spatial Autocorrelation
- Local Indexes of Spatial Autocorrelation

Global Indexes of Spatial Autocorrelation

- The goal of global indexes of spatial autocorrelation is to summarize the degree to which similar observations tend to occur near each other
- Global indexes are summaries over the entire study area, akin to testing clustering rather than a test to detect individual clusters
- Indexes share a common structure: calculate the similarity of values at locations i and j then weight the similarity by the proximity of locations i and j
- High similarities with high weight indicate similar values that are close together; low similarities with high weight indicate dissimilar values that are close together



Examples of configurations of areas showing different types of spatial autocorrelation. From Moraga, Paula. (2023) Spatial Statistics for Data Science: Theory and Practice with R.

Global Indexes of Spatial Autocorrelation

Indexes of spatial autocorrelation

- We want to summarize similarity between nearby areal units
- Spatial autocorrelation is the the correlation of the same measurement taken at different areal units
- The similarity of values at locations B_i and B_j are weighted by the proximity of i and j
- The weight w_{ij} defines proximity
- In general the extent of similarity is represented by the weighted average of similarity between areal units: indexes of spatial autocorrelation are built on this basic form:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} sim_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

Moran's I

- Moran's I (1950) measures the similarity between values at neighboring areal units.
- Similarity between areas i and j: $sim_{ij} = (y_i - \bar{y})(y_j - \bar{y})$
- Mean value: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- The standardized form of Moran's I:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

- This normalization matches the implementation in `spdep::moran.test()`.

Interpreting Moran's I

Value of I	**Spatial Pattern**
$I > 0$	Positive spatial autocorrelation (clustering)
$I < 0$	Negative spatial autocorrelation (dispersion)
$I \approx 0$	Random spatial pattern (no association)

Moran's I

- I is a random variable having a distribution defined by the distributions of and interactions between the y_i
- When neighboring regions have similar values (pattern is clustered), I will be positive
- When neighboring regions have different values, I will be negative
- When there is no correlation between neighbouring values: $E(I) = -\frac{1}{n-1}$
- When $n \to \infty$, $E(I) \to 0$
- I is asymptotically normally distributed where $\frac{I+\frac{1}{n-1}}{\sqrt{Var(I)}} \sim N(0,1)$

Moran's I

- Moran's I is similar to Pearson's correlation but it is not bounded on [-1,1] because of the spatial weights.
- Null hypothesis: NO spatial association, i.e. y_i iid.
- Compare the z-score to a standard normal distribution.
- The z-score that we compare to the standard normal is $z = \frac{I - E(I)}{\sqrt{Var(I)}}$ where $E(I) = -\frac{1}{n-1}$ and $V(I)$ is shown in the next slides.

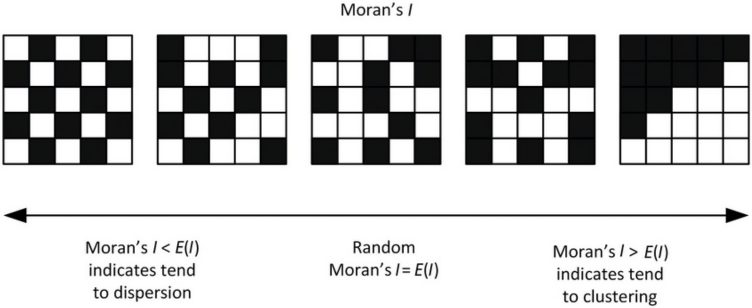


Illustration of how map clustering affects Moran's I score. Image adapted from (Kirkegaard 2015).

Variance for Moran's I

- $E(I) = \frac{-1}{(n-1)}$ under null hypothesis of no autocorrelation
- $V(I)$ is dependent upon the weight matrix:

$$V(I) = \frac{ns_1 - s_2 + 3(\sum_i \sum_j w_{ij})^2}{(n-1)(n+1)(\sum_i \sum_j w_{ij})^2} - E(I)^2$$

Where:

$$s_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} (w_{ij} + w_{ji})^2 \text{ sum of sqared spatial weights}$$

$$s_2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} (w_{ij} + w_{ji}) \right)^2 \text{ sum of sqared row sums of spatial weights}$$

$$\sum_i \sum_j w_{ij}^2 \text{ total sum of the spatial weights matrix}$$

Here the variance of Moran's I is calculated under the assumption of normality and reflects how Moran's I would behave under this assumption.

Moran's I Variance: Analytical vs Randomization

- Two common ways to compute the variance of Moran's I:
  - **Analytical (Normality assumption):** Assumes y_i are normally distributed. Variance derived from weight matrix structure.
  - **Randomization (Permutation-based):** Observations are shuffled among areal units B_i many times. Variance and p-value are estimated empirically.
- Both test H_0: no spatial association (spatial randomness).
- Randomization avoids strong distributional assumptions and provides a more robust test.

Moran's I via Monte Carlo

- Monte Carlo approach repeats randomization of the observations into the B_i a large number of times (e.g. $N_{sim} \sim 999$). It is basically a bootstrap permutation.
- For each bootstrap permutation Moran's I is calculated.
- Across the simulations the mean (equivalent to $E(I)$) and the variance (equivalent to $V(I)$) is calculated.
- Compare the observed Moran's I to the randomization set through difference: observed-expected/s.d(expected).
- The p-value is calculated as the proportion of values as extreme or more extreme than the statistic observed in the direction of the alternative hypothesis.
- If the actual I falls at the 5th/95th percentile (or smaller/greater) then it is significant at $\alpha = 0.05$.

Moran's I via Monte Carlo

Benefits of Randomization in Moran's I

- Randomization allows for a formal test to determine if the observed spatial autocorrelation is statistically significant.
- Permutation-based randomization doesn't rely on strict distributional assumptions (i.e. standard normal), making it more robust.
- The randomized distribution is used to compute p-values, providing a way to assess the likelihood of the observed Moran's I arising by chance.
- It tests the null hypothesis of no spatial autocorrelation (random spatial distribution), providing a baseline for evaluating spatial patterns.

Monte Carlo Moran's I in R

- Monte Carlo simulation of global Moran's I in spdep:

R output

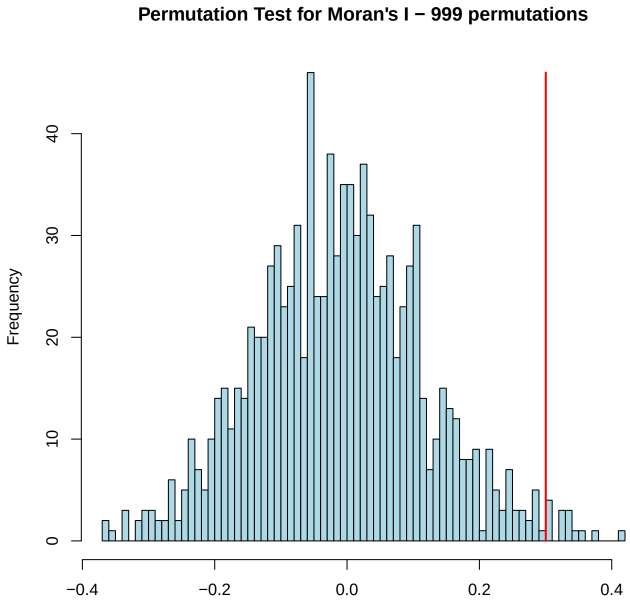
```
moranSIDS_mc <- moran.mc(SID79/BIR79, sids_kn1_w, nsim = 999)

Monte-Carlo simulation of Moran's I
data:  SID79_BIR79
weights: sids_kn1_w
number of simulations + 1: 1000

statistic = 0.3003, observed rank = 989, p-value = 0.011
alternative hypothesis: greater
```

- The observed Moran's $I = 0.30$ lies in the upper tail $\Rightarrow$ reject H_0 (positive spatial autocorrelation).

Global Indexes of Spatial Autocorrelation



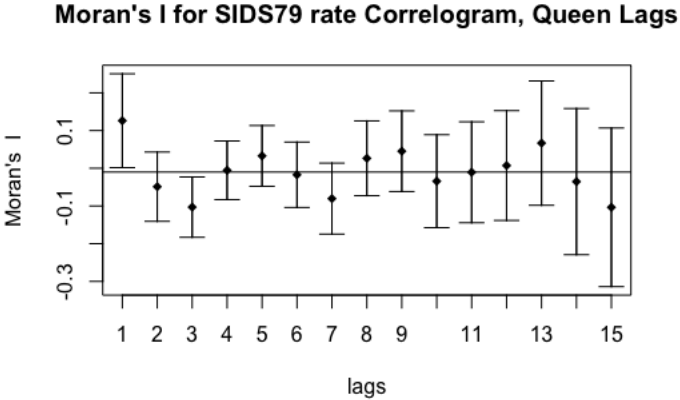
Issues with Global Indexes of Spatial Autocorrelation

- Global tests assume **stationarity —** any trend has been removed.
- Centering the mean $(y_i - \bar{y})$ assumes a constant mean model.
- Removing trend may be difficult without covariate data.
- Spatial weights may be **misspecified**: too few or too many neighbors can bias the test.
- Require roughly $n \geq 20$ for asymptotic (z-score) results.

Determining Neighbors for Moran's I

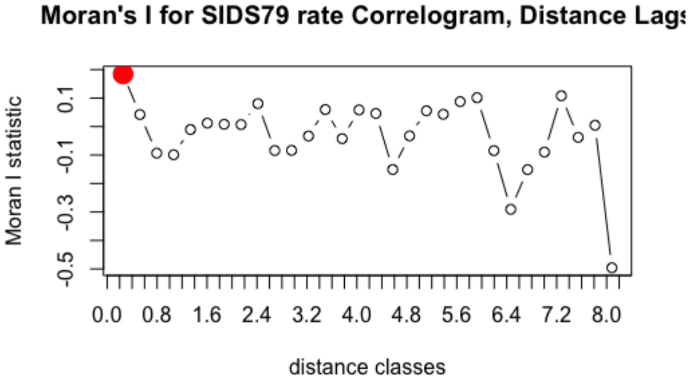
- Use **connectivity diagnostics** first: ensure no islands and a single connected component.
- Explore **knn graphs**: compute I for $k \in \{1, \ldots, 8\}$ (or more) and check stability.
- Explore **contiguity graphs**: queen vs. rook; compare resulting I and connectivity.
- Explore **distance bands**: choose d so every area has ≥ 1 neighbor (use the max of each area's nearest neighbor).
- For interpretation on the Moran scatterplot, prefer **row-standardized** weights (`style="W"`).
- Always conduct a **sensitivity check** across several k or d values.

Correlogram of Moran's I



Correlogram: $I(h) = \frac{n}{S_0} \frac{\mathbf{z}^\top \mathbf{W}_h \mathbf{z}}{\mathbf{z}^\top \mathbf{z}}$, $\mathbf{z}$ is standardized y, $\mathbf{W}_h$ encodes h-order neighbors, $S_0 = \sum_{ij} w_{ij}$.

Distance Lags of Moran's I



Distance correlogram: $I(d) = \frac{n}{S_0(d)} \frac{\mathbf{z}^\top \mathbf{W}(d)\mathbf{z}}{\mathbf{z}^\top \mathbf{z}}$, where $\mathbf{W}(d)$ links pairs with d_{ij} in a chosen distance class (e.g., $(d_k, d_{k+1}]$).

Quasi Local Indexes of Spatial Autocorrelation

- A **Moran scatterplot** visualizes how each area relates to its neighbors.
- Standardize variable values:

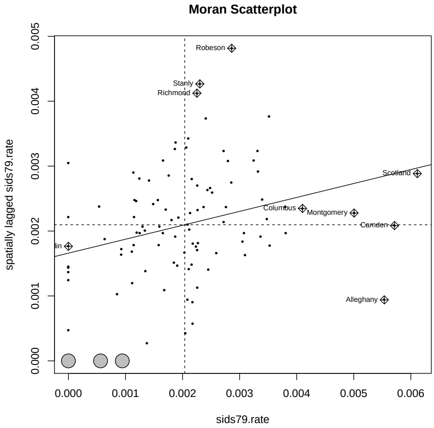
$$y_i = \frac{Y_i - \bar{Y}}{s_Y}, \qquad (Wy)_i = \sum_j w_{ij} y_j$$

  where W is row-standardized.
- Plot $(Wy)_i$ on the y-axis against y_i on the x-axis.
- The **slope** of the least squares regression line

$$(Wy)_i = \alpha + \beta y_i + \varepsilon_i$$

  is equal to the **global Moran's** I.

Quasi Local Indexes of Spatial Autocorrelation



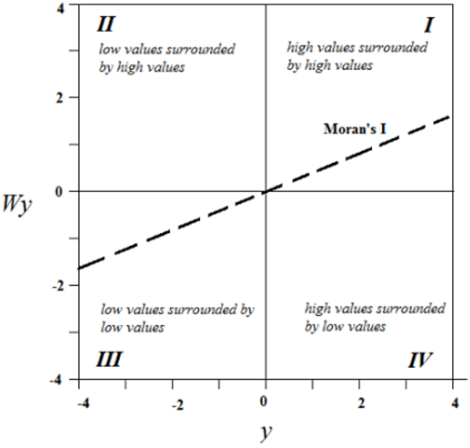
Moran scatterplot: slope = Moran's I when x is standardized and $\mathbf{W}$ is row-standardized.

Moran Scatterplot and Quadrants

- Standardize y: $z_i = \dfrac{y_i - \bar{y}}{s_y}$; use row-standardized W.
- Plot spatial lag vs. value: $(Wy)_i$ vs. y_i (or $(Wz)_i$ vs. z_i).
- Slope of dashed line = global Moran's I.

I	$y_i > 0,\ (Wy)_i > 0$	High–High (cluster)
II	$y_i < 0,\ (Wy)_i > 0$	Low–High (outlier)
III	$y_i < 0,\ (Wy)_i < 0$	Low–Low (cluster)
IV	$y_i > 0,\ (Wy)_i < 0$	High–Low (outlier)

HH/LL $\Rightarrow$ positive local association;
HL/LH $\Rightarrow$ negative local association.



Axes: y (x-axis), Wy (y-axis). Quadrants at $y = 0$, $Wy = 0$.

Quasi Local Indexes: Influence Diagnostics

- The Moran scatterplot regression:

$$(Wy)_i = \alpha + \beta y_i + \varepsilon_i$$

  where $\beta =$ Moran's I.
- Each point (areal unit) influences β differently:
  - **Leverage** (h_{ii}): how far y_i is from the mean

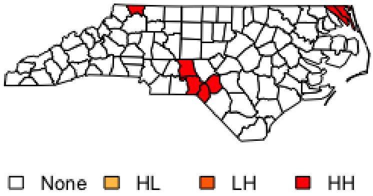
$$h_{ii} = \frac{1}{n} + \frac{(y_i - \bar{y})^2}{\sum_j (y_j - \bar{y})^2}$$

  - **Cook's distance**: combines leverage and residual size

$$D_i = \frac{(e_i^2 / p\,\hat{\sigma}^2)\,h_{ii}}{(1 - h_{ii})^2}$$

- High leverage or large $D_i \Rightarrow$ influential points. Often occur in Quadrants II (Low–High) and IV (High–Low) $\rightarrow$ potential spatial outliers.
- Diagnostic plots or maps help identify which regions drive global Moran's I.

Quasi Local Indexes of Spatial Autocorrelation



Mapping Moran scatterplot quadrants: HH/LL = cluster cores (positive association); HL/LH = spatial outliers (negative association).

Local Indexes of Spatial Autocorrelation

- Global measures (Moran's I or Geary's c) are a single value that apply to the entire study area
- The same pattern or process occurs over the entire geographic area
- Global statistic suggests that there is clustering but does not identify areas of particular clusters
- Global test is often used first to determine if there is evidence of spatial association
- Want to detect local areas of similar values, need a local statistic
- LISAs are decompositions of global indicators into the contribution of each individual observation (i.e. $B_i \in D$)
- As a result the sum of LISAs is proportional to the equivalent global indicator
- Local Moran's I, Getis-Ord G*

From Global to Local Moran's I_i

- Global Moran's I summarizes spatial autocorrelation **across the whole region**.
- To identify **where** clusters or outliers occur, decompose I into **local contributions**.
- Each observation i has its own local statistic:

$$I_i = z_i \sum_j w_{ij} z_j,$$

  where $z_i = \frac{y_i - \bar{y}}{s_y}$ and $\mathbf{W}$ is row-standardized.
- $\Rightarrow$ Local Moran's I_i measures the similarity between i and its neighbors relative to the global mean.

Interpreting Local Moran's I_i

- Sign and magnitude of I_i reveal local spatial association:

Pattern		**Interpretation**
$z_i > 0$, neighbors > 0	(HH)	High surrounded by high values (cluster core)
$z_i < 0$, neighbors < 0	(LL)	Low surrounded by low values (cluster core)
$z_i > 0$, neighbors < 0	(HL)	High outlier amid lows (negative association)
$z_i < 0$, neighbors > 0	(LH)	Low outlier amid highs (negative association)

- $I_i > 0 \Rightarrow$ local clustering (HH or LL)
  $I_i < 0 \Rightarrow$ local spatial outlier (HL or LH)
- The global $I = \frac{1}{n}\sum_i I_i$ (up to a scaling constant).

Significance Testing for Local Moran's I_i

- The null hypothesis H_0: no local spatial association (random spatial arrangement).
- For each area i, randomize y_i among all locations and recompute I_i many times (`nsim` ≈ 999).
- Compare observed I_i to the simulated distribution:

$$p_i = \frac{\#(|I_i^{sim}| \geq |I_i^{obs}|) + 1}{N_{sim} + 1}$$

- Adjust for multiple testing using False Discovery Rate (FDR) or Bonferroni correction.
- Map significant I_i values, colored by quadrant type (HH, LL, HL, LH) $\Rightarrow$ **LISA cluster map**.

Local Indexes of Spatial Autocorrelation (LISA)

- **LISA = Local Indicators of Spatial Association**
- Capture the degree of spatial autocorrelation **around each areal unit**.
- Two main types:
  - **Local Moran's** I_i — identifies **local clusters** (similar neighbors)
  - **Local Getis–Ord** G_i^* — identifies **hotspots and coldspots**
- LISAs decompose the global statistic into contributions for each location:

$$I = \frac{1}{n}\sum_{i=1}^{n} I_i$$

Moran's I vs. Local Moran's I_i

Global Moran's I:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

Local Moran's I_i:

$$I_i = z_i \sum_j w_{ij} z_j, \quad \text{where } z_i = \frac{y_i - \bar{y}}{s_y}$$

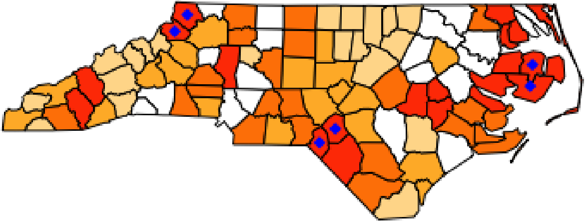
- I_i gives each area's contribution to overall spatial autocorrelation.
- $I_i > 0$: clustering of similar values (HH or LL) $I_i < 0$: spatial outlier (HL or LH)

Local Moran's I_i: Interpretation and Significance

$$I_i = z_i \sum_j w_{ij} z_j$$

- A positive I_i indicates similar values among neighbors (clusters). Negative I_i indicates dissimilarity (spatial outliers).
- To test significance:
  - Randomize values across locations many times (~ 999 permutations).
  - Compute empirical p_i as the fraction of simulated I_i as extreme as observed.
- Because multiple I_i are tested simultaneously:
  - Adjust p_i (e.g. Bonferroni or False Discovery Rate correction).
- Significant I_i values are mapped by quadrant type (HH, LL, HL, LH).

Local Moran's I_i: Significant Clusters



Statistically significant Local Moran's I_i (blue points). High–High (HH) and Low–Low (LL) indicate clustering; High–Low (HL) and Low–High (LH) indicate outliers.

Getis–Ord G vs. Local Getis–Ord G_i^*

- Getis–Ord statistics identify **hotspots (high values)** and **coldspots (low values)**.

Global G:

$$G = \frac{\sum_i \sum_j w_{ij} y_i y_j}{\sum_i \sum_j y_i y_j}$$

Local G_i^*:

$$G_i^* = \frac{\sum_j w_{ij} y_j}{\sum_j y_j}$$

- High positive G_i^* $\Rightarrow$ high-value cluster (hotspot)
- Low negative G_i^* $\Rightarrow$ low-value cluster (coldspot)

Testing Getis–Ord G and G_i^*

- Compute z-scores from randomization or analytical expectation:

$$z = \frac{G - E(G)}{\sqrt{Var(G)}}$$

- Expected value:

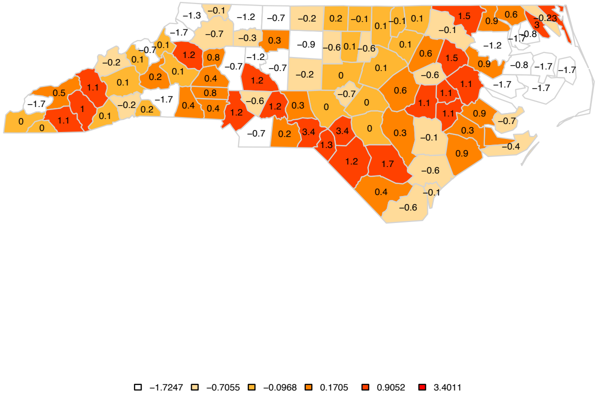
$$E(G) = \frac{\sum_i \sum_j w_{ij}}{n(n-1)}$$

- Variance (simplified form):

$$Var(G) = \frac{s_y^2}{\bar{y}^2} \frac{\sum_i \sum_j w_{ij}(n - \sum_i \sum_j w_{ij})}{n-1}$$

- Interpretation:
  - $z > 0$: high values cluster together (hotspot)
  - $z < 0$: low values cluster together (coldspot)

Local Getis–Ord G_i^* Example



Local Getis–Ord G_i^* for SIDS79 rates. Red = hotspots (high-value clusters); blue = coldspots (low-value clusters).

From Autocorrelation Tests to Spatial Models

- Both global and local tests assume that mean trends are removed (e.g., income effects when analyzing SIDS rates).
- A common workflow:
  1. Fit a regression model: `lm(rate ~ covariates)`
  2. Test residuals for spatial autocorrelation: `lm.morantest()`
- If residuals still show spatial structure $\Rightarrow$ use
  - Spatial Lag Model (SLM)
  - Spatial Error Model (SEM)
- These explicitly model spatial dependence instead of treating it as noise.