GIS and Spatial Analysis for Public Health

Spatial Statistics for Advanced Practitioners and Researchers

Levente Juhász

Assistant Professor of Geospatial Analytics
Geospatial **A**nalytics, **T**echnology and **O**pen **R**esearch (GATOR) Lab
School of Forest, Fisheries and Geomatics Sciences
University of Florida
https://gatorlab.io

FIU-RCMI, Miami, FL
September 18, 2026

Hello from the workshop host

- 6 years at FIU (2019 - 2025)
  - Research Assistant Professor
    - Research Associate Professor (2025)
  - Director (interim), GIS Center (2022 - 2023)
  - Assistant Director of GIScience (2023 - 2025)
  - Research Associate Professor (2025)
- University of Florida (2026 -)
  - Assistant Professor of Geospatial Analytics
  - Director, Geospatial Analytics, Technology and Open Research (GATOR) Lab
  - Located in South Florida (Davie)

GATOR
GEOSPATIAL ANALYTICS, TECHNOLOGY AND OPEN RESEARCH

Housekeeping

Time	Block
8:30 - 9:00	Breakfast
9:00 - 12:00	**Morning Session** • GIS Principles • Spatial Statistics • Regression
12:00 - 13:00	Lunch
13:00 - 14:00	Afternoon Session

Goals for today

Review and deepen understanding of *spatial autocorrelation* and *local clustering*

Diagnose *spatial bias* in regression models using residual analysis

Apply advanced solutions, such as *Geographically Weighted Regression (GWR)* and *Spatial Eigenvector Filtering*

Build a full spatial analysis pipeline in R on real public health data

❗ ❗ ❗ **Interactive style** - please interrupt and ask questions anytime ❗ ❗ ❗

Overview

- **Morning session**
  - Recap of Workshops #1 and #2
    - Fundamentals of GIS
    - Introductory spatial statistics
    - Regression models in spatial contexts
  - Diagnosing spatial problems in regression models
- **Afternoon session**
  - Advanced Spatial Modeling Solutions
  - Hands-on Spatial Statistics with R

Fundamentals of GIS (recap)

1. What is GIS?
2. GIS software
3. GIS data models
4. Basic GIS operations and analysis methods

What is GIS

- **GIS** means Geographic Information Systems, but sometimes refers to GIScience
- The former refers to the **systems**. The latter focuses on the **theories and methodologies** that support the functionalities of the systems.
- GIS
  - **(computer) system** that collect, manage, model, and analyze geospatial data to understand relationships, patterns, and trends in the data.
- GIScience
  - **scientific discipline** that studies geospatial data structures and computational techniques to capture, model, process, and analyze geographic information

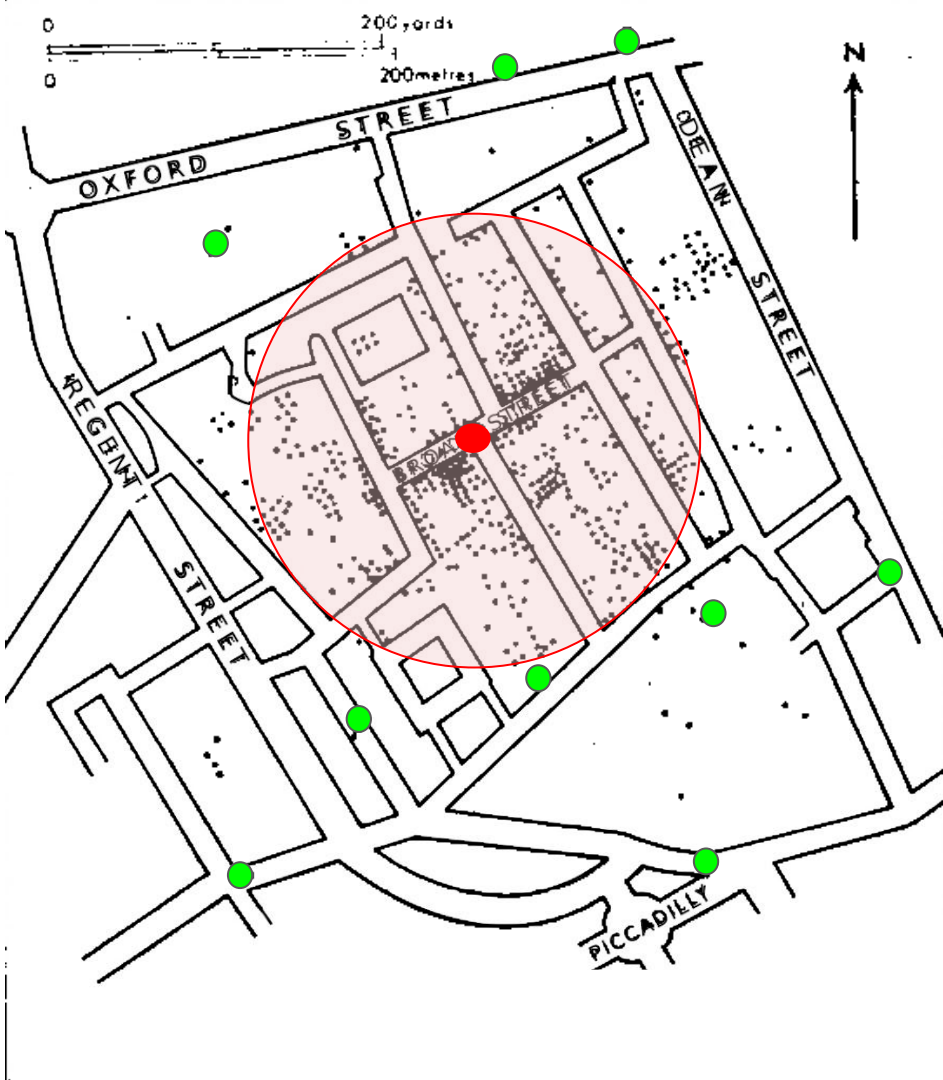
What is GIS?

- How many people live within one mile of this hospital?
- Where are the counties that have had high COVID-19 transmission rates in 2020?
- Which neighborhoods are in a mandatory evacuation zone?
- What is the fastest route to get to an accident?

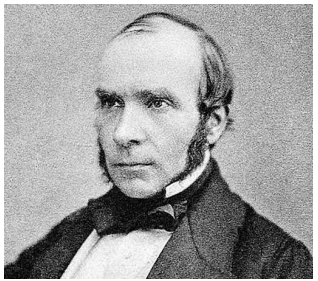
History of GIS

12 Broad St.
46 Oxford St.
13 Broad St.
1 Ocean St
15 Broad St.
35 Ocean St.

Let's put these on a map!



In 1854, John Snow, an English physician, depicted a cholera outbreak in London using points to represent the locations of individual victim through a map,possibly the earliest use of "spatial analysis" in epidemiology.



Basic GIS concepts and terminology

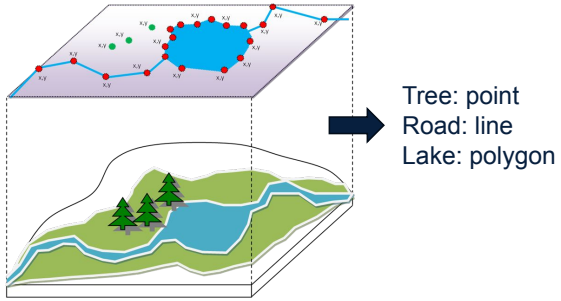
- Representation of space
- Data type
  a. Vector data
     i. Geometry
     ii. Attribute
  b. Raster data
- Geoprocessing operations
- Types of maps

Representation of space

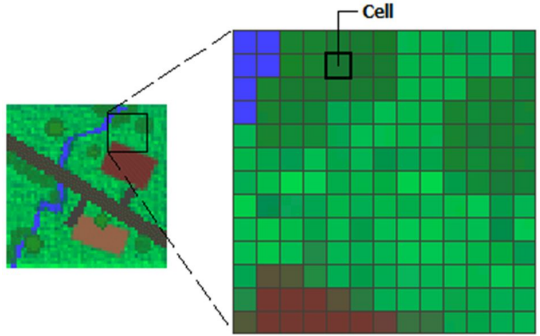
- Two fundamental ways of representing geographic space
  - Discrete objects
  - Continuous surfaces
- Discrete object
  - The world is “empty”, except where it is occupied by objects with well-defined boundaries
- Continuous field
  - The world is filled with values of one or multiple variables. The variable has values at every position, and do not necessarily have well defined boundaries
- In GIS, discrete objects are most commonly represented in the **vector data model**, while continuous fields are most commonly described with the **raster model.**

GIS data types

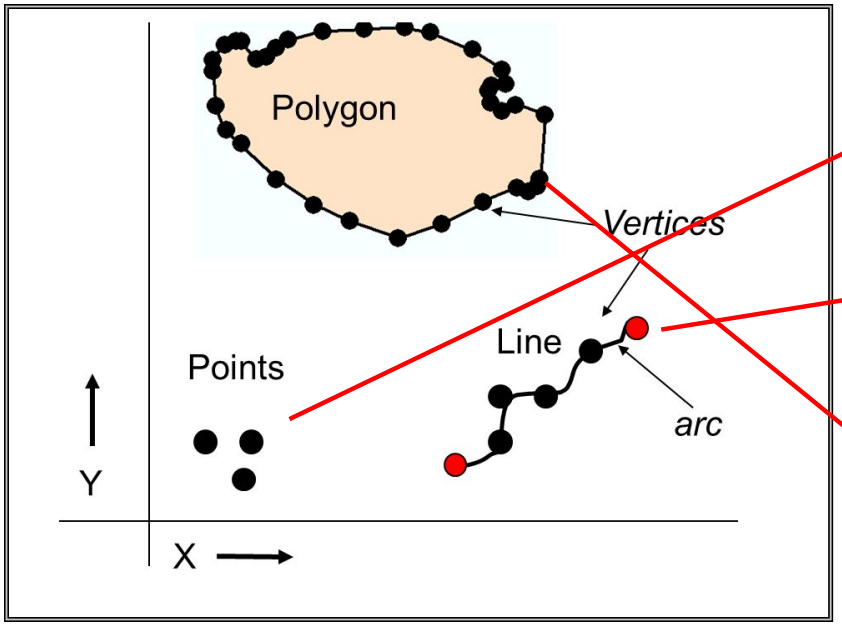
- **Vector data**
  - Vector data use points, lines and polygons to represent spatial features as discrete objects.
  - Good at representing spatial features with clear boundary and even interior.



- **Raster data**
  - Raster data define the world as a regular set of cells in a grid pattern.
  - Each cell is associated with a value that represent the attribute (e.g., elevation, land cover, housing density…)



Geometry of vector data



Point 1: (X1, Y1)
Point 2: (X2, Y2)
Point 3: (X3, Y3)

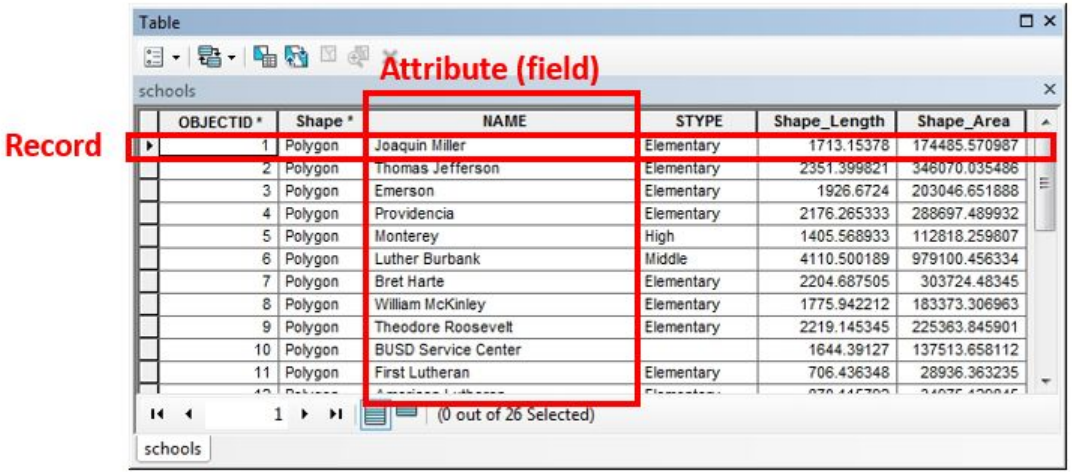
Line: [(X1, Y1), (X2, Y2),…, (Xn, Yn)]

Polygon: [(X1, Y1), (X2, Y2),…, (Xn, Yn)]

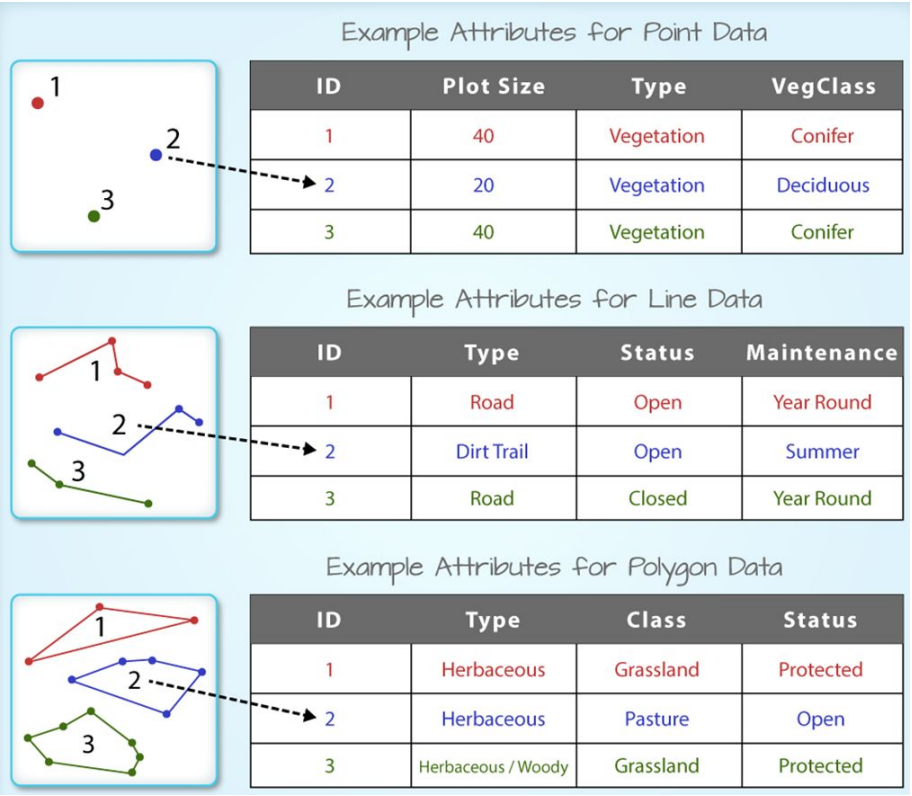
- Point: a pair of coordinates (x, y) in a 2D space
- Line: a sequence of points
- Polygon: the area inside a closed boundary (line)

Tables

- Data are stored in tables (also known as relations or relational tables)
- Columns (also called fields) contain attributes that describe features.
- Each rows (record) represent an individual record/feature.



Attributes of vector data



ID	Plot Size	Type	VegClass
1	40	Vegetation	Conifer
2	20	Vegetation	Deciduous
3	40	Vegetation	Conifer

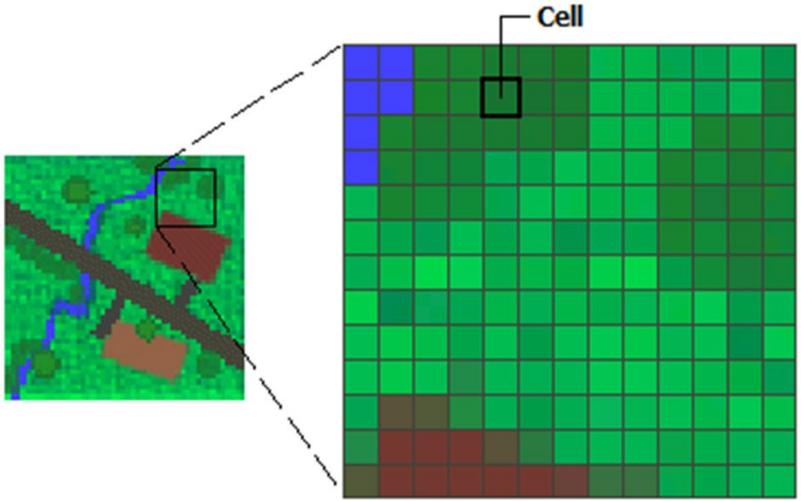
ID	Type	Status	Maintenance
1	Road	Open	Year Round
2	Dirt Trail	Open	Summer
3	Road	Closed	Year Round

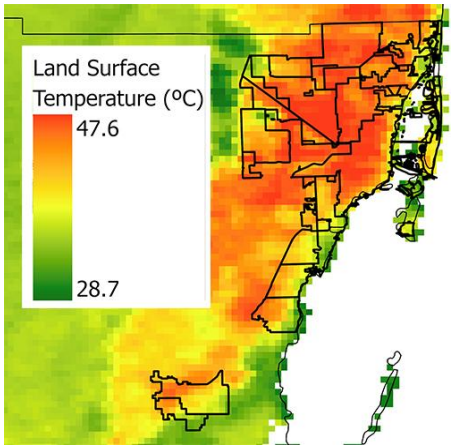
ID	Type	Class	Status
1	Herbaceous	Grassland	Protected
2	Herbaceous	Pasture	Open
3	Herbaceous / Woody	Grassland	Protected

- Attributes are non-spatial characteristics associated with the geometries (locations)
- Attributes are stored in a table (attribute table).
- Geometries are linked with records in attribute table by feature IDs.

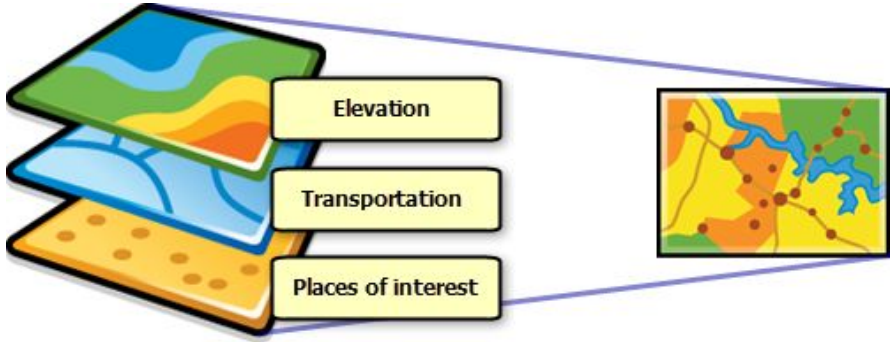
Raster data

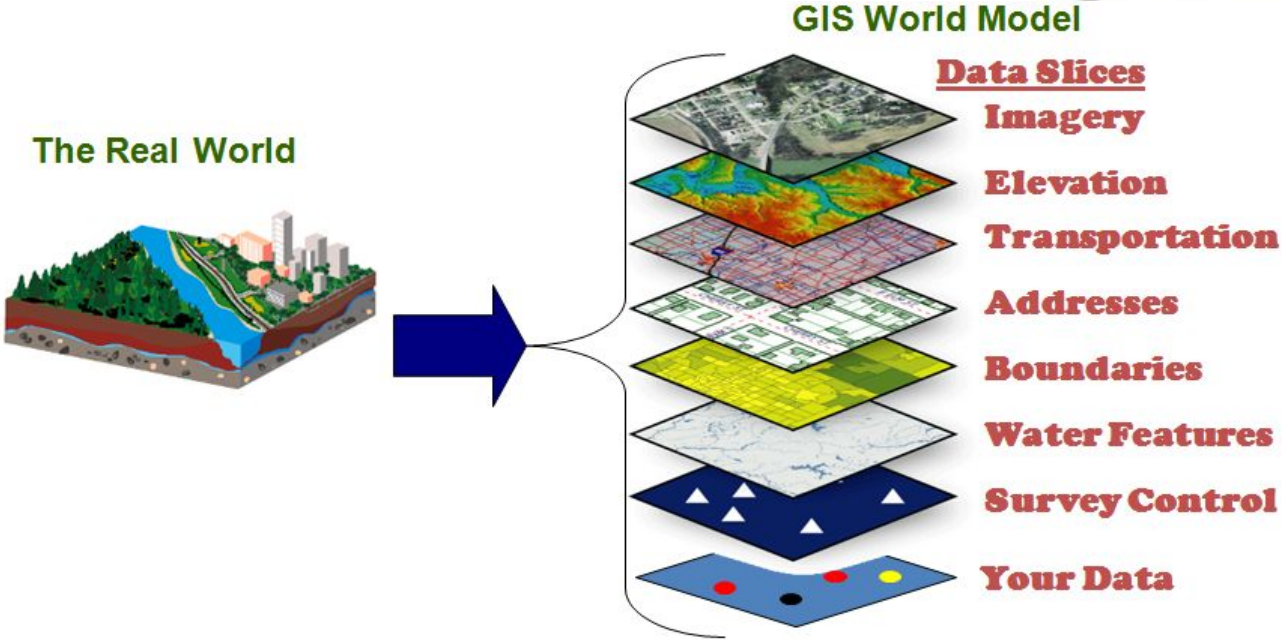
- Raster data define the world as a regular set of cells in a grid pattern
- The cells are typically square and arranged in X, Y directions
- Each cell is associated with a value represent the attribute (e.g. elevation, land cover, housing density…)





GIS organizes data into layers



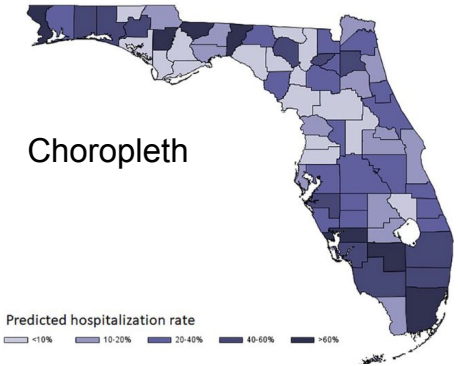


Thematic maps

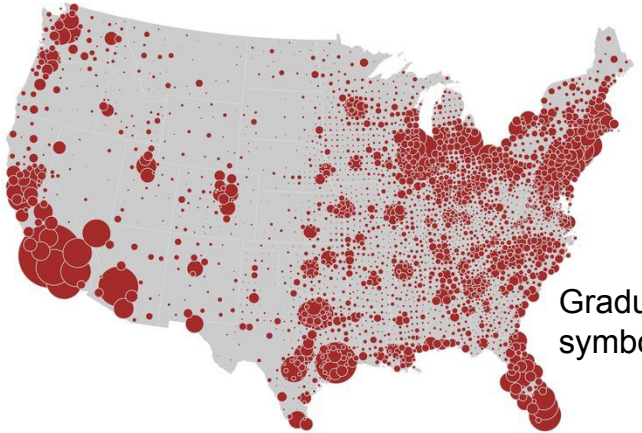
Presentation of maps matters!

- **Choropleth map:** areas are shaded or patterned in proportion to a statistical variable, such as population density or income levels. Each color represents a specific range of values.
- **Dot density map:** uses dots to represent a quantity of a variable in a given area. Each dot represents a specific number of occurrences, providing a visual sense of distribution or density.
- **Graduated symbol map:** symbols of varying sizes (e.g., circles) are used to represent data values. Larger symbols indicate greater values, making it easy to compare magnitudes.
- **Heat map:** shows the intensity of data using a gradient of colors.
- Etc...

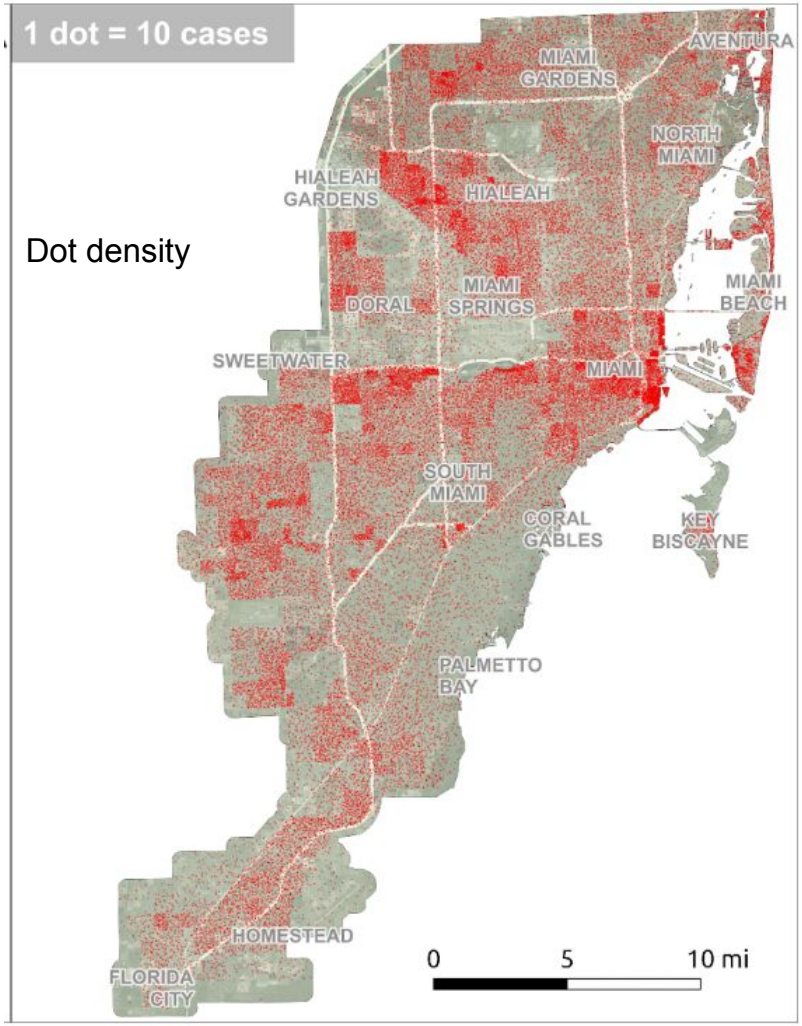
Thematic maps



Choropleth



Graduated symbol



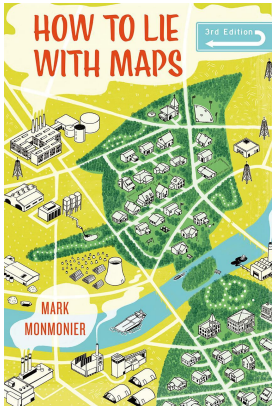
Dot density

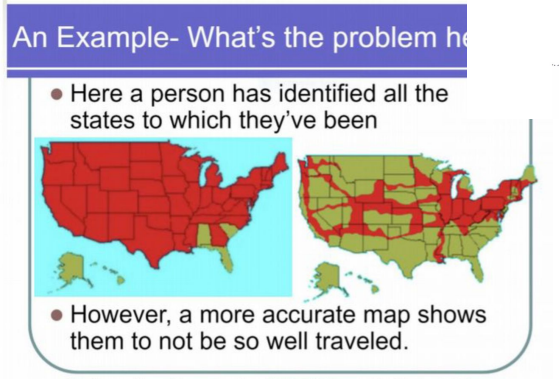
Map classification

"...classification introduces the risk of a mapped pattern that distorts spatial trends. ***Arbitrary selection of breaks between categories might mask a clear coherent trend with a needlessly fragmented map or oversimplify a meaningful intricate pattern*** *with an excessively smoothed view."*

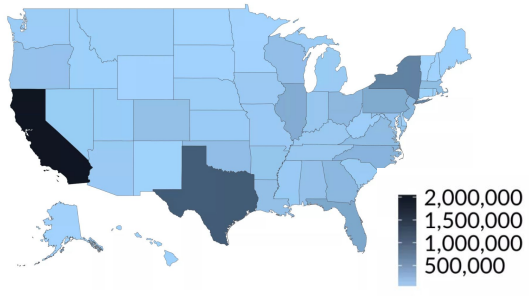
-

Mark Monmonier

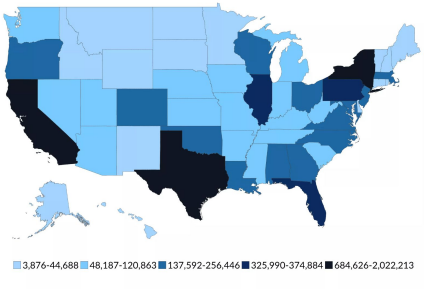




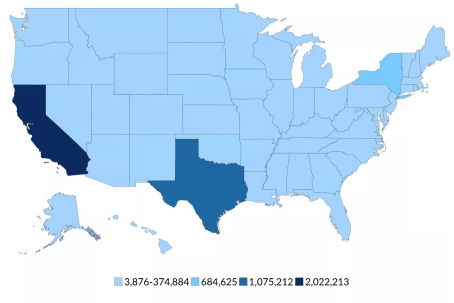
State CHIP Enrollment: No Bins Color Model



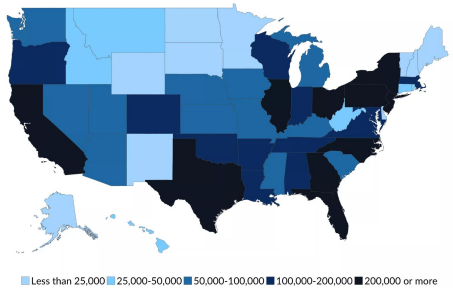
State CHIP Enrollment: KFF Map with 5 Custom Groups



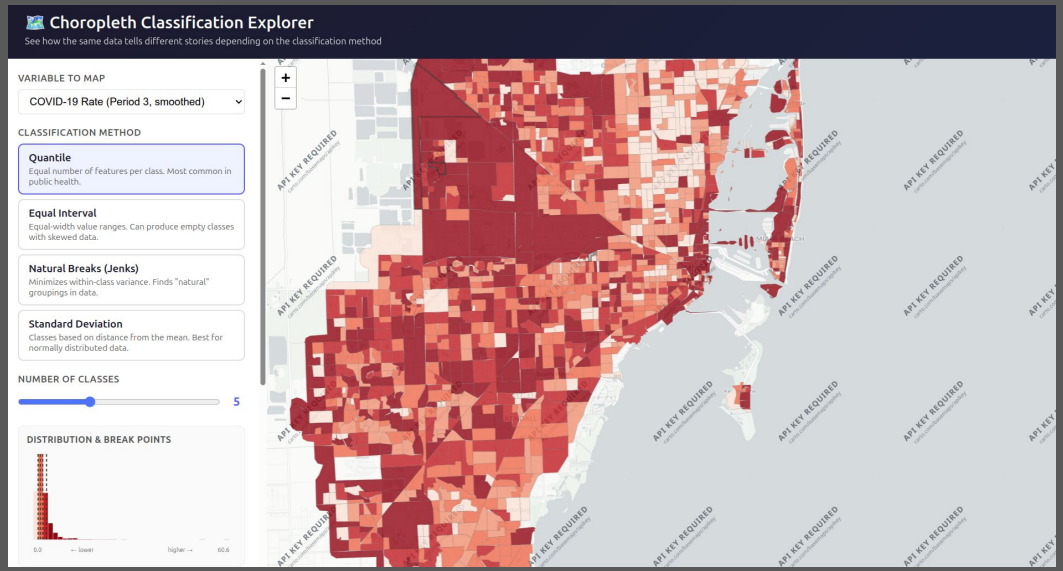
State CHIP Enrollment: KFF Map with 5 Equal Intervals



State CHIP Enrollment: KFF Map with 5 Round Groups



Interactive example - Choropleth Classification



Workshop materials: https://gatorlab.io/presentations/rcmi_advanced_2026/

Map classification example: https://gatorlab.io/presentations/rcmi_advanced_2026/classification_explorer.html

Exploring GIS basics in ArcGIS Pro (optional)

- Exercise for previous RCMI GIS workshop (intermediate concepts)

- **Step 1:** Download Tract 2020 from the Miami-Dade Open Data Hub
- **Step 2:** Add Data in ArcGIS Pro
- **Step 3:** Open Attribute table, create new fields, use the Field Calculator
- **Step 4:** Customize Symbology

Instructions/Guide:
https://maps.fiu.edu/gis/news/fiu-rcmi-gis-intermediate-workshop-exercise-1/

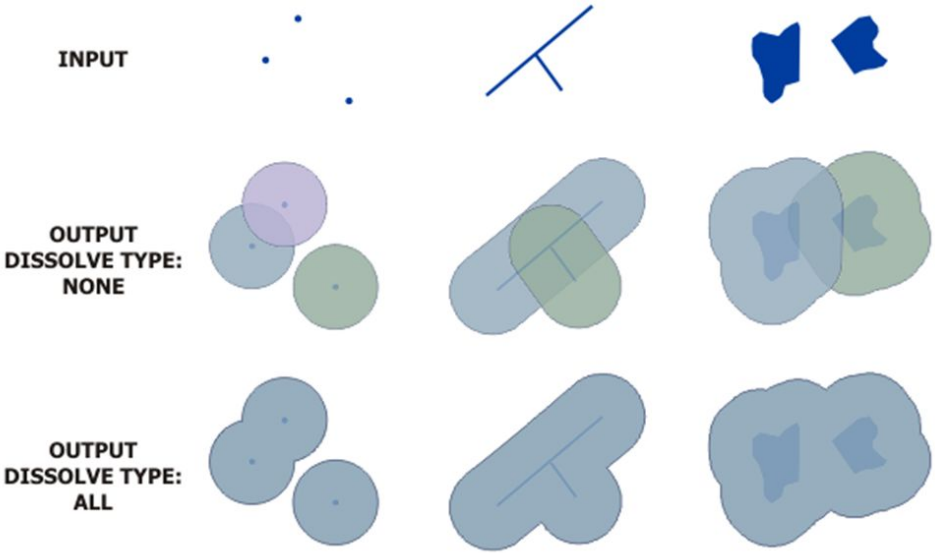
Geoprocessing operations

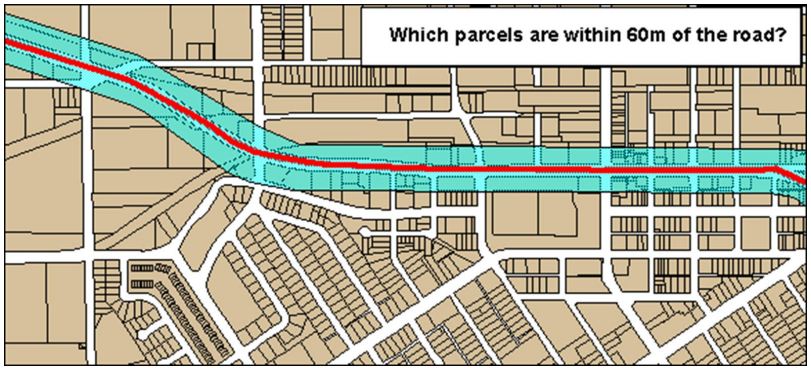
- Executes analytical operations on existing spatial data.
- Utilizes feature geometry and attributes to generate new, derived datasets.

Public Health Application: Quantifying spatial relationships (e.g., delineating exposure zones or intersecting demographic data with hazard areas).

Buffer

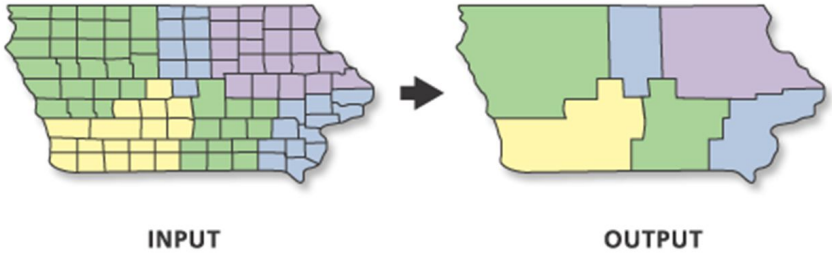
- Creates a zone within a specified distance of a vector feature





Dissolve

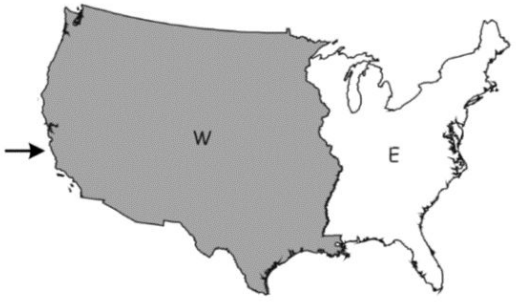
- Combines vector features based on common attributes



Dissolve operation

Dissolve table

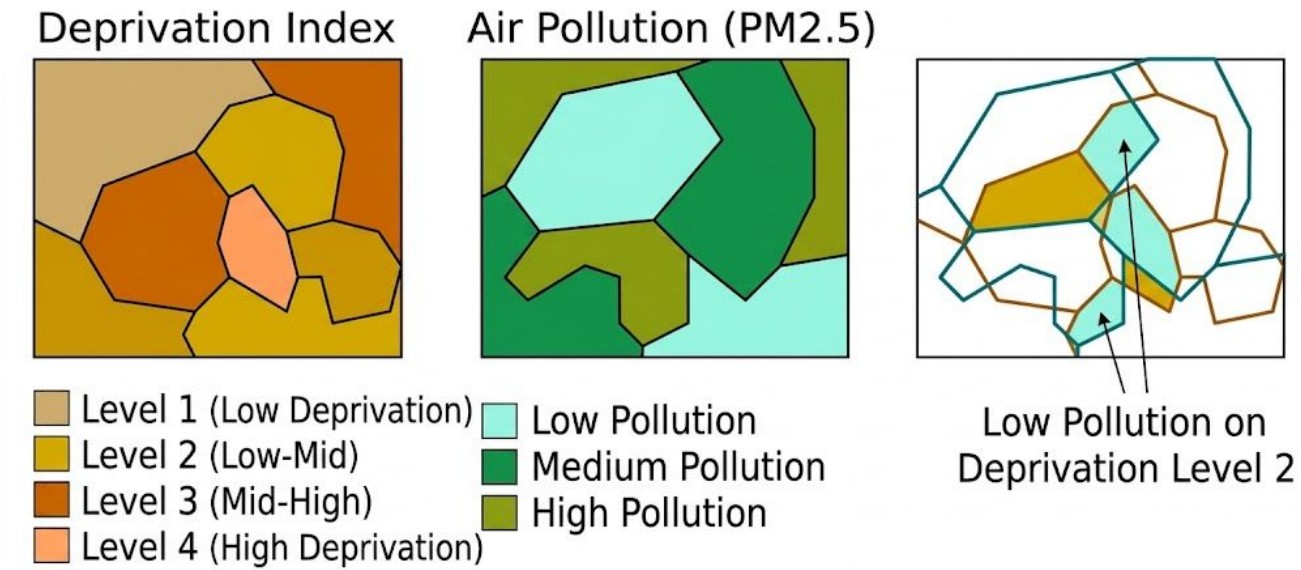
state name	is_west	dissolve value
Alabama	0	E
Arizona	1	W
Arkansas	1	W
Colorado	1	W
Connecticut	0	E
....	....	....
Wyoming	1	W



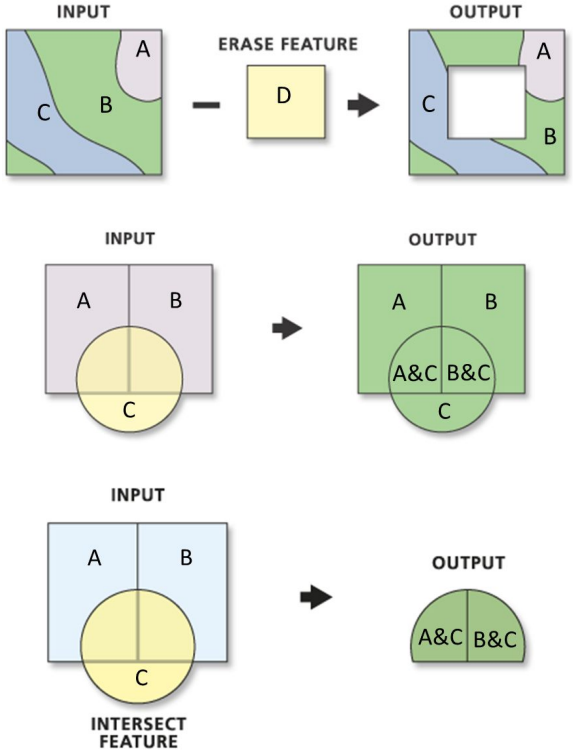
Spatial overlay

Spatial overlay of vector data include operations on **geometry and attribute**.

Common types of spatial overlay are **Erase, Union, and Intersect.**



Erase, Union and Intersect



Erase: erases an area from the input data

Union: Combines two inputs and their attributes

Intersect: Combines two inputs and their attributes, and keeps the **mutual parts**

Introduction to spatial statistics

- Why spatial statistics?
- Key concepts in spatial statistics:
  - Point pattern analysis
  - Spatial dependence
  - Spatial autocorrelation
    - Global Moran’s I
    - General G statistic

Why spatial statistics?

- Traditional statistics often **assume independence between observations**
- **Spatial data violates this** assumption due to spatial dependence
- Spatial statistics
  - Helps identify patterns, relationship and clusters in geospatia ldata
  - Next step after “looking at data through maps”

Example applications:

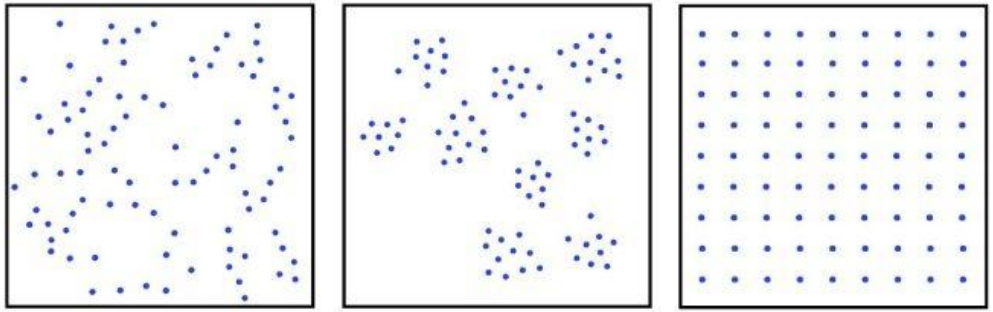
- Detecting disease hot spots
- Identifying underserved areas
- Analyzing environmental exposure

Key concepts in Spatial Statistics

- **Point pattern analysis**
  - Analyzes the distribution of points (e.g. disease cases) to determine randomness, clustering or uniformity
- **Spatial dependence**
  - Tobler's First Law of Geography:
    - "Everything is related to everything else, but near things are more related than distant things"
- **Spatial heterogeneity**
  - Relationships can vary across space
- **Spatial autocorrelation**
  - Measures how similar or dissimilar values are in space

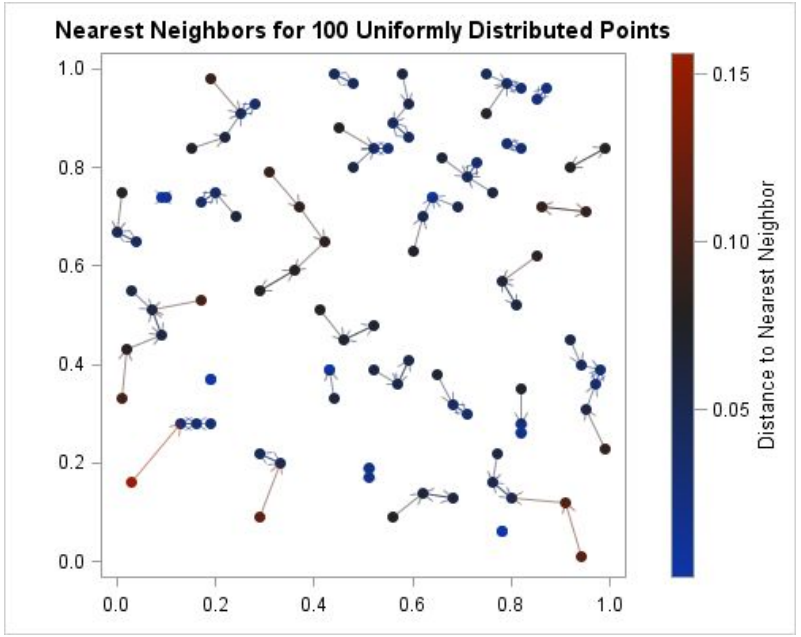
Point pattern analysis

- Study of the spatial arrangements of points in space
- **Objective:** Understand whether points (e.g. disease cases, facilities) are distributed randomly, or are clustered or dispersed
- **Common methods:**
  - Nearest neighbor analysis
  - Quadrat analysis
  - Ripley's K function



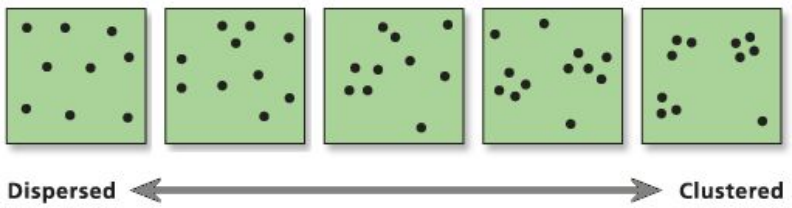
Nearest Neighbor Analysis

1. Measure the distance between each feature and their nearest neighbor
2. Average all these nearest neighbor distances
3. Answer question: Are points are randomly distributed?
    - Calculate Nearest Neighbor Index (NNI) using observed distances and expected distances



- *NNI* = Observed Avg distance / Expected Avg distance

- If *NNI* > 1, pattern tends to be dispersed
- If *NNI* < 1, pattern exhibits clustering

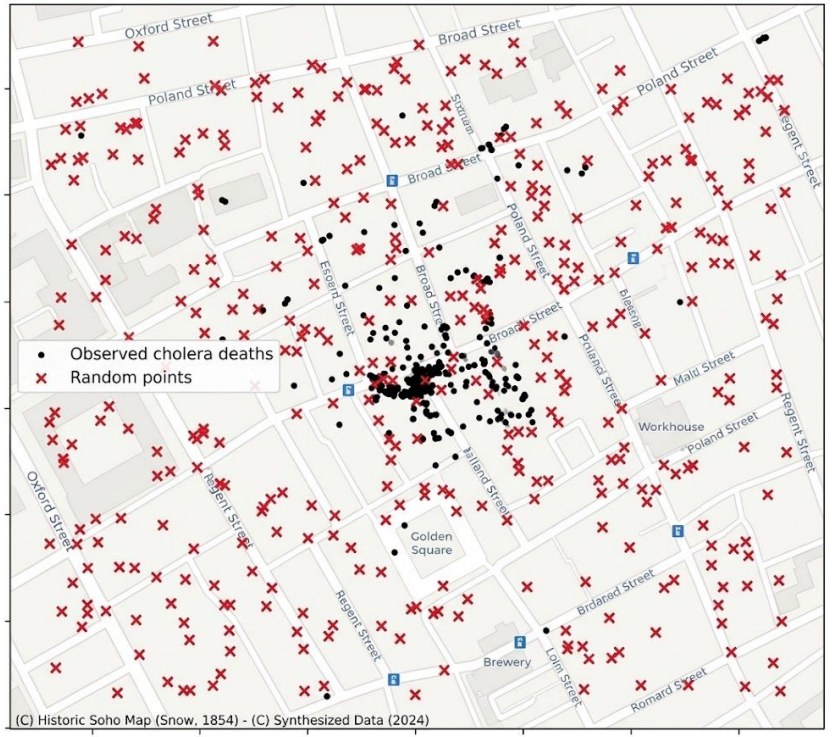


Nearest Neighbor Analysis

How do we characterize cholera deaths in London in 1860?

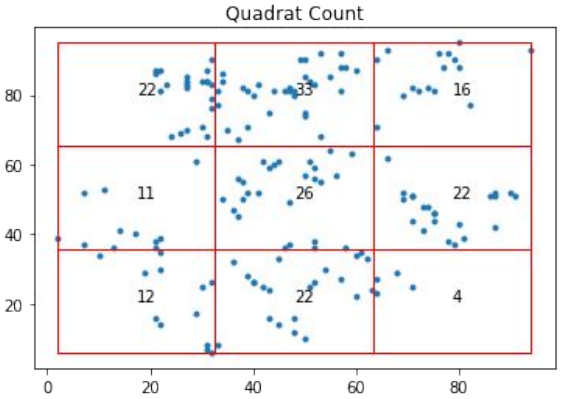
1. Measure average nearest distance between recorded deaths (**black dots**)
2. Measure average nearest distance between randomly generated points (**red x**)
   - Theoretical value can also be calculated based on area and number of points
3. Calculate Nearest Neighbor Index (*NNI)*

***NNI* < 1**



This statistic supports Dr. Snow's empirical finding.

Quadrat analysis



- Divide area into equal-sized grids (quadrats)
- Count points in each quadrat
- Calculate expected counts
- Use the **Chi-squared statistic** to measure the deviation of observed counts from expected counts under the null hypothesis of complete spatial randomness
  - Total points: 168
  - E = 168 / 9 = 18.67
  - df = Number of Quadrats – 1 = 8
  - χ^2=34.2. We reject the null hypothesis, i.e. the point pattern is not random

$$\chi^2 = \sum \frac{(O_i - E)^2}{E}$$

- Variance-to-Mean Ratio (VMR) is used to determine clustering or dispersion
  - VMR = Variance of Observed Counts / Mean of Observed Counts
  - If VMR >1, pattern is clustered
  - If VMR<1, pattern is dispersed

Ripley's *K* function

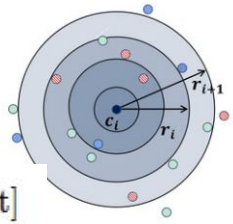
- **Many point patterns show a combination of effect**
  - Clustering at large scales
  - But regularity at small scales
  - We will call this **spatial heterogeneity**

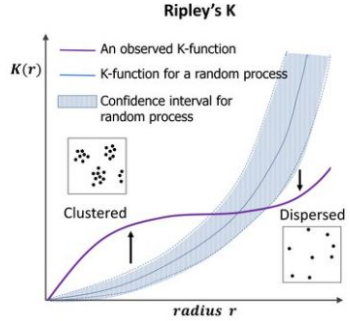
Example: Walmart locations

- At the **local level** (e.g. Miami-Dade County), they tend to be evenly distributed across the urban area
- At the **state level**, they tend to cluster around population centers

Ripley's *K* function helps quantify and detect this spatial heterogeneity by measuring point distributions at different scales

Ripley's *K* function





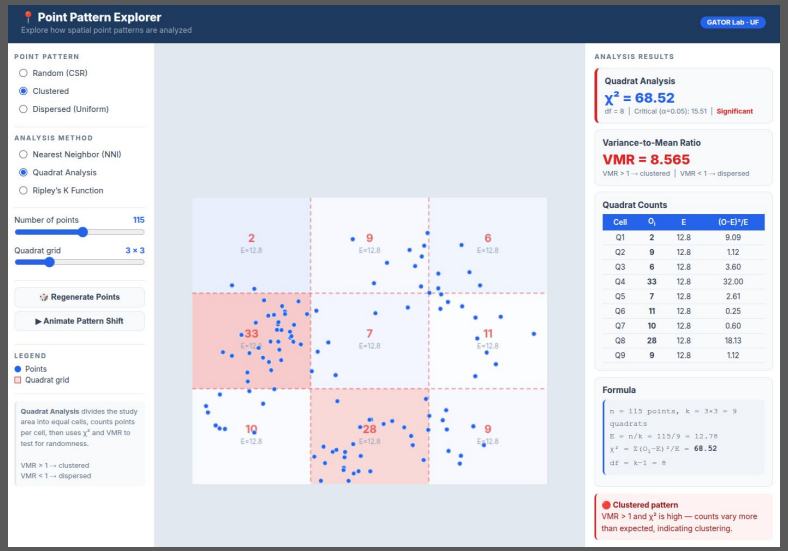
- Insert circles with increasing radius around points

$$K(r) = \lambda^{-1}\mathbb{E}[\text{number of extra events within distance } r \text{ of a randomly chosen event}]$$

- Compare the observed counts against expected counts under complete spatial randomness
- **Interpretation**
  - Above the random envelope indicates clustering
  - Below the random envelope indicates dispersion
- **Problem:** K(r) under CSR = π r², which is a steep curve, and hard to visually judge
- **Solution:** Besag's L transformation (1977):
  - Under CSR, K(r) = πr², so L(r) = √(πr²/π) − r = r − r = 0 (flat line)

$$L(r) = \sqrt{\frac{K(r)}{\pi}} - r$$

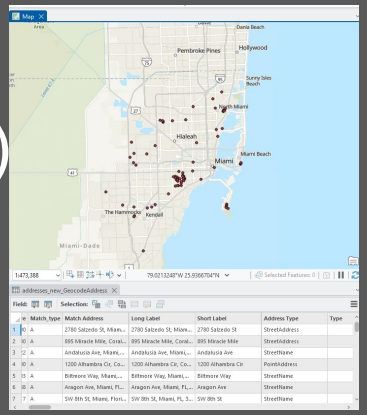
Interactive tool - Point Pattern Explorer



Workshop materials: https://gatorlab.io/presentations/rcmi_advanced_2026/

Map classification example: https://gatorlab.io/presentations/rcmi_advanced_2026/pointpattern-explorer.html

Analyzing Point Patterns in ArcGIS Pro (optional)



- Exercise for previous RCMI GIS workshop (intermediate concepts)

- **Exercise:** Analyzing Disease Infection Patterns Using Geocoded Data
- **Step 1:** Download the provided CSV file containing addresses of individuals diagnosed with a disease
- **Step 2:** Add Data in ArcGIS Pro
- **Step 3:** Use the Geocode Addresses tool to create point geometries for these addresses
- **Step 4:** Perform spatial analysis

Instructions/Guide:
https://maps.fiu.edu/gis/news/fiu-rcmi-gis-intermediate-workshop-exercise-2/

Spatial dependence and Tobler's First Law

"Everything is related to everything else, but near things are more related than distant things." - Waldo Tobler

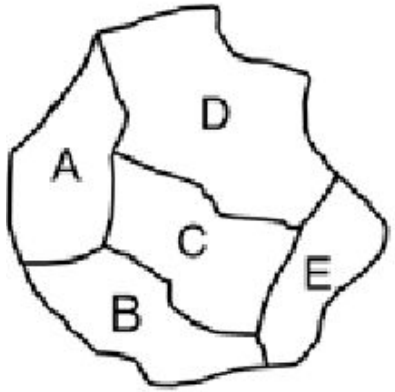
Two types of spatial dependence:

- **Endogenous:** the value at a location depends on values at neighboring locations. *Example: infectious disease spreading neighborhood to neighborhood*
- **Exogenous:** values are similar because of a shared underlying cause (spatially correlated covariates). *Example: high poverty AND high disease rates in same areas because of shared structural factors*

Distinguishing these matters for model choice. Endogenous = spatial lag model. Exogenous = spatial error model or eigenvector filtering.

Determining neighbors: Spatial Weights Matrix

- The foundation of all spatial statistics
- A spatial weights matrix ***W*** is an N x N matrix, where w_{ij} = 1 if unit *i* and *j* are neighbors, 0 otherwise
- **Row standardization**: divide each row by its row sum so weights sum to 1.



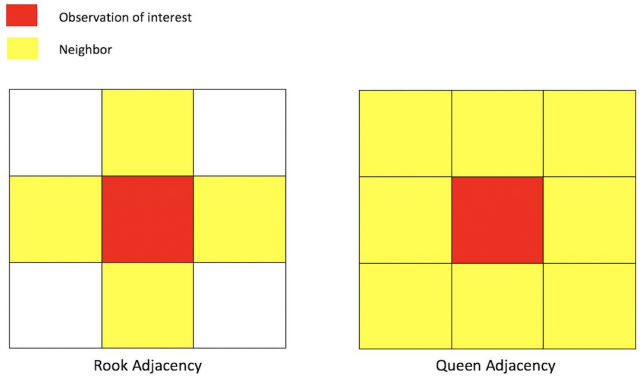
	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	1
C	1	1	0	1	1
D	1	0	1	0	1
E	0	1	1	1	0

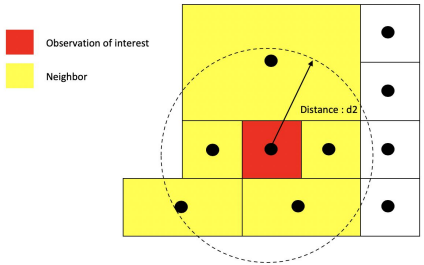
Determining neighbors

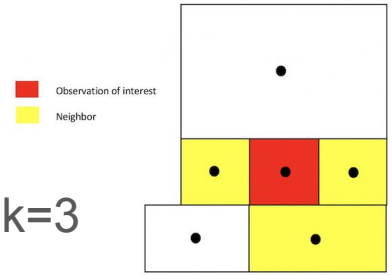
- Contiguity or adjacency
  - **Queen contiguity:** Any shared point or edge. More inclusive. Standard for polygon data
  - **Rook contiguity:** Shared edges only. More conservative.
- **Distance based:** based on geographic distance
- **K-Nearest Neighbor:** Based on k nearest neighbors

Distance-based and KNN are useful when polygon sizes vary greatly.

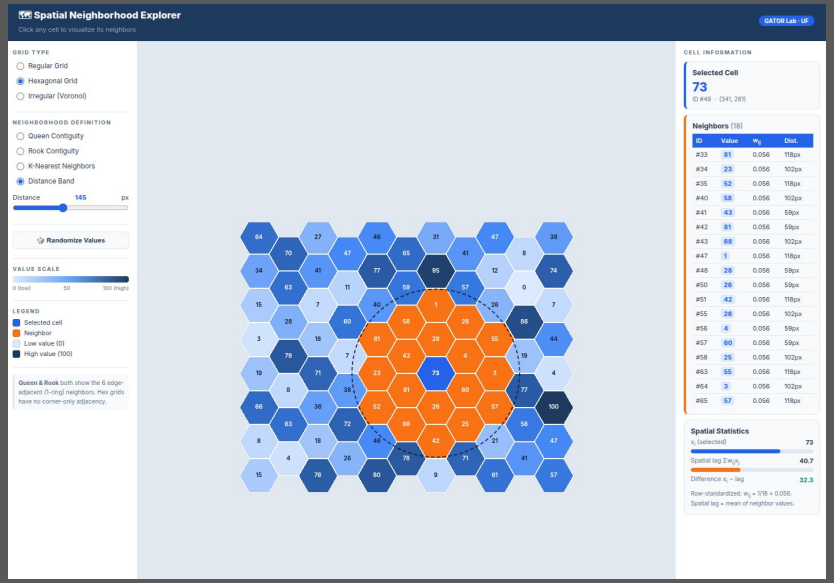
Every spatial statistic you run depends entirely on this matrix. Different choices may produce different results.







Interactive tool - Neighborhood Explorer



Workshop materials: https://gatorlab.io/presentations/rcmi_advanced_2026/

Map classification example: http://localhost:3326/presentations/rcmi_advanced_2026/neighborhood-explorer.html

Spatial weights: real-world considerations

Should two census block groups separated by I-95 be considered neighbors?

The weights matrix assumes contiguity = connectivity, but real barriers break that assumption.

Approaches to address barriers:

- Network-based weights: use road network travel time instead of Euclidean distance
- Binary barrier adjustment: set $w_{ij} = 0$ if a major barrier separates *i* and *j*.

Spatial autocorrelation

- The **degree of similarity** between objects (observations) that are located near each other
  - Can be seen as the extension of temporal autocorrelation in higher dimensions, with complex shapes, across distance and proximity metrics

In simple terms, it describes the spatial patterns of observations

- Clustered, Random, Dispersed

Spatial Autocorrelation can be measured and quantified, both for an entire region (global), or for smaller areas within the region (local)

Pat Moran

Global Moran's *I* statistic

Common measure of (global) spatial autocorrelation. It measure of the overall clustering of spatial data

Are values that are near each other in space more similar than we would expect by chance?

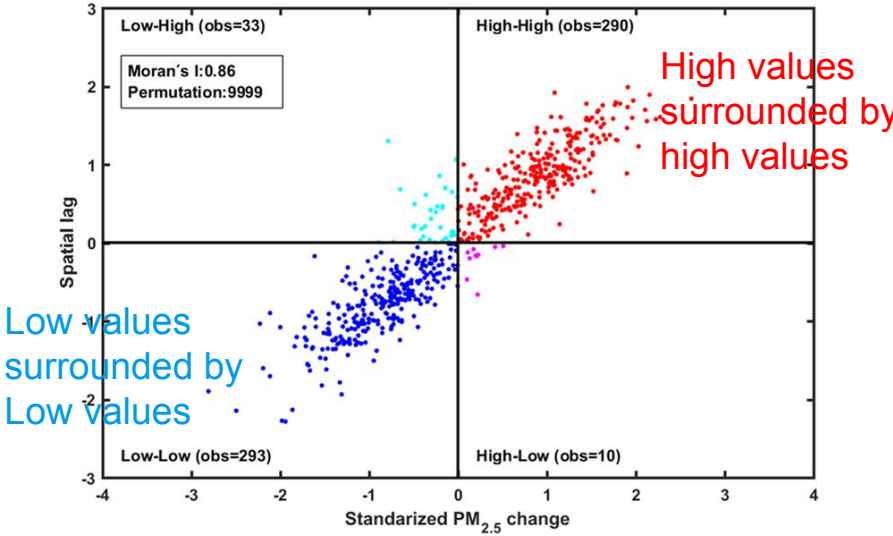
Value	Interpretation
I → +1	Perfect positive clustering
I ≈ 0	Spatial randomness (CSR)
I → -1	Perfect dispersion

Range: -1 to +1

Moran scatterplot:

- X-axis: standardized value at location i
- Y-axis: spatially lagged value (mean of neighbors)
- Slope of regression: Moran's *I*

Four quadrants: HH (upper right), LL (lower left), HL (upper-left), LH (lower-right)



Global Moran's *I*: formula

- We need observations (i.e. values) and locations
- We need to define "near each other". That is, we build a **spatial weight matrix**

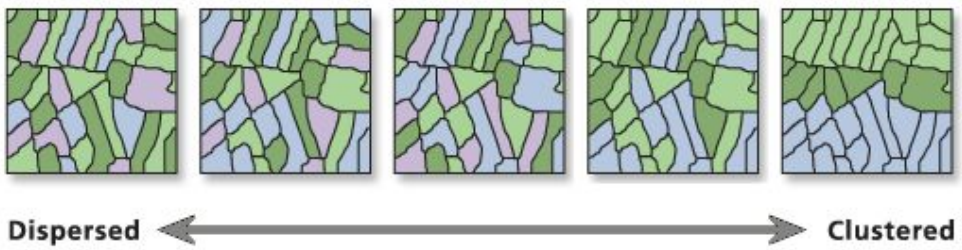
$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

N: number of objects (e.g. census tracts)
x: variable of interest (e.g. infection rate)
$w_{i,j}$: elements of a spatial weights matrix
W: sum of $w_{i,j}$

- **Numerator:** cross-products of deviations from the mean, weighted by spatial proximity — high when nearby locations have similar values
- **Denominator:** variance normalization
- **Result:** if similar values cluster together → I > 0; if similar values are spread apart → I < 0

A significant positive Moran's *I* tells you clustering exists somewhere. It **does NOT** tell you where. That's what local statistics are for.

Moran's I

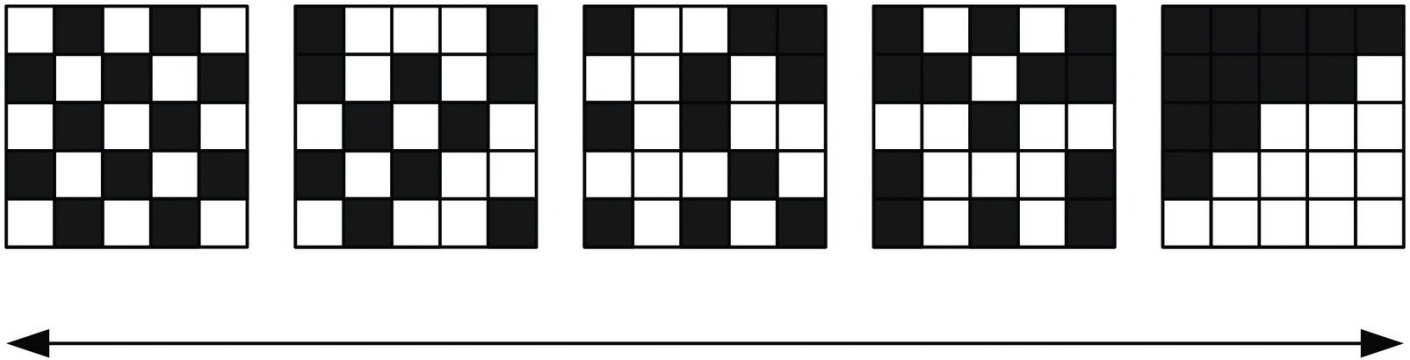


I = -1
Perfect dispersion

I = 0
Complete Spatial Randomness (CSR)

I = 1
Perfect clustered pattern

Moran's *I*



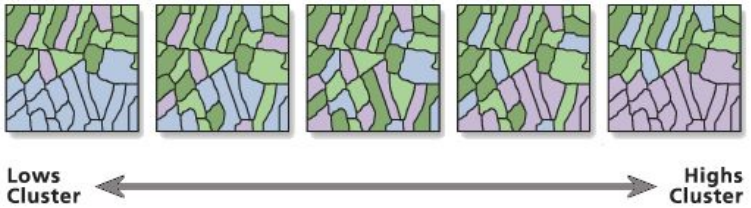
Getis-Ord General G statistic


J. Keith Ord


Arthur Getis

Limitations of Moran's I: detects clustering but does not distinguish between high-value clusters and low-value clusters. General G statistic addresses this.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \quad \forall j \neq i$$



where x_i and x_j are attribute values for features i and j, and $w_{i,j}$ is the spatial weight between feature i and j. n is the number of features in the dataset and $\forall j \neq i$ indicates that features i and j cannot be the same feature.

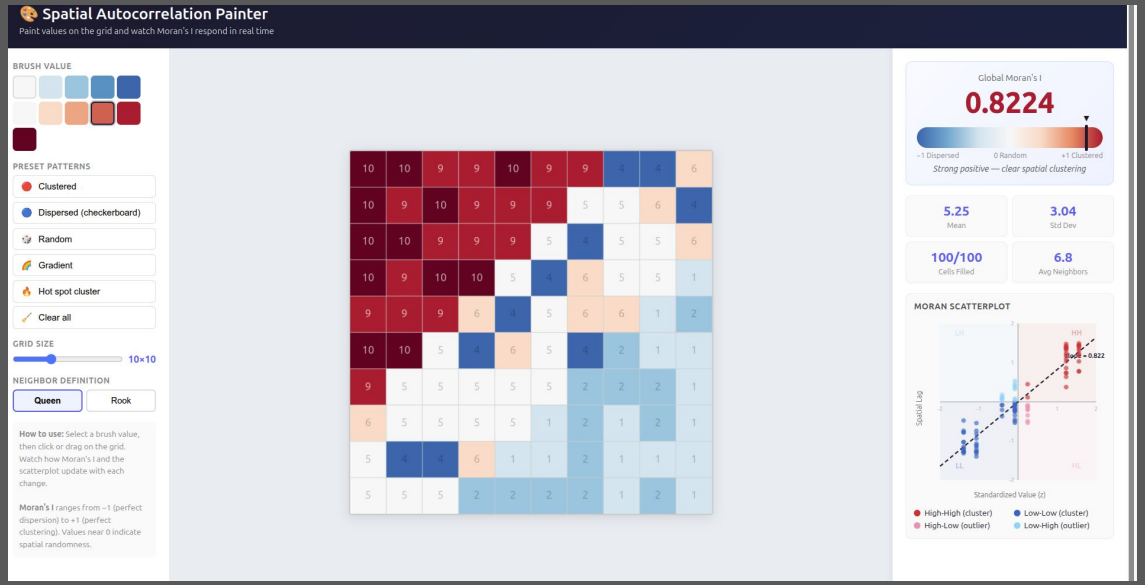
- **High G** (significant positive z-score): high values cluster together → concentration of high burden
- **Low G** (significant negative z-score): low values cluster together → concentration of low burden

Example interpretation

- Moran's I = 0.45, p < 0.001 → **significant clustering exists**
- G z-score = +3.2, p < 0.001 → i**t's a HIGH-value cluster** (elevated disease rates)
- G z-score = -2.8, p < 0.01 → **it's a LOW-value cluster** (suppressed rates)

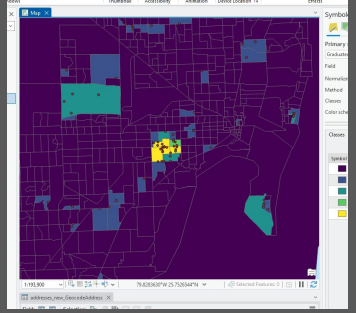
Use Moran's *I* to detect clustering.
Use General G to characterize the TYPE of clustering.

Interactive tool - Spatial Autocorrelation Painter



Workshop materials: https://gatorlab.io/presentations/rcmi_advanced_2026/

Map classification example: http://localhost:3326/presentations/rcmi_advanced_2026/spatial_autocorrelation_painter.html



Spatial Analysis in ArcGIS Pro

Objective:

- Learn how to aggregate point data into polygons, calculate spatial statistics (Global Moran's I and General G), and interpret spatial patterns in public health data

Exercise: Analyzing Spatial Patterns of Disease Cases Using Aggregation and Global Statistics

- **Step 1:** Load the Tracts and Geocoded locations from previous steps
- **Step 2:** Count number of occurrences within each tract
  - Spatial Join
  - Summarize Within

Step 3: Calculate Global Moran's I and General G statistics

Instructions/Guide:
https://maps.fiu.edu/gis/news/fiu-rcmi-gis-intermediate-workshop-exercise-3/

Introduction to spatial statistics

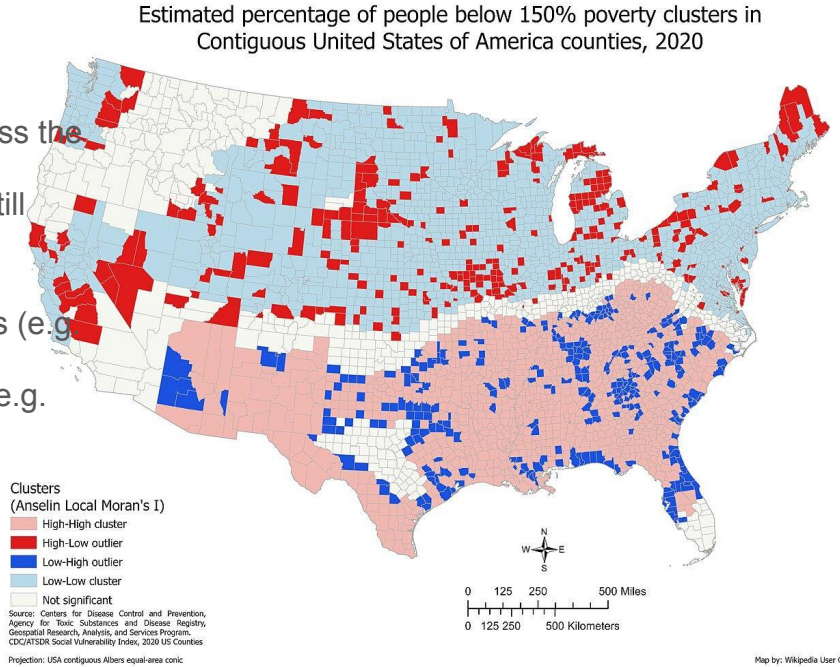
- More spatial statistics techniques
- Local Indicators of Spatial Autocorrelation (LISA)
  - Local Moran’s I
  - Getis-Ord Gi*
- Regression analysis
  - OLS regression in a spatial context
  - Geographically Weighted Regression (GWR)

Local Indicators of Spatial Autocorrelation (LISA)

- Localized measures of spatial autocorrelation to identify clusters and outliers in spatial data
- **Purpose:**
  - Detect spatial heterogeneity within the dataset
  - Highlight areas of significant clustering or dispersion
- **Methods**
  - Local Moran’s I
    - Identifies local clusters and outliers
  - Local Getis-Ord Gi*
    - Detects hot spots and cold spots

Local Moran's I

- **Purpose:** Measure spatial autocorrelation at the local level
- **Why?**
  - Global Moran's I assumes spatial homogeneity, that is, the relationship between neighboring values is consistent across the study area. Most processes are not spatially stationary.
  - Even if there is no global spatial autocorrelation, we may still find clusters at the local level
- **Output:** Clusters of similar values and outliers
  - High-High Clusters: High values surrounded by high values (e.g. disease hot spot)
  - Low-Low Clusters: Low values surrounded by low values (e.g. areas with minimal cases)
  - High-Low Outliers: High values surrounded by low values
  - Low-High Outliers: Low values surrounded by high values



Local Moran's I

- I_i: local Moran's I value for a location *i*
- X_i: value of the variable (e.g. disease rate) at a location
- *X_hat*: mean of all values
- *w*: spatial weights matrix indicating relationships between locations
- m_2: the variance of the dataset

$$I_i = \frac{x_i - \bar{x}}{m_2} \sum_{j=1}^{N} w_{ij}(x_j - \bar{x})$$

where:

$$m_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

then,

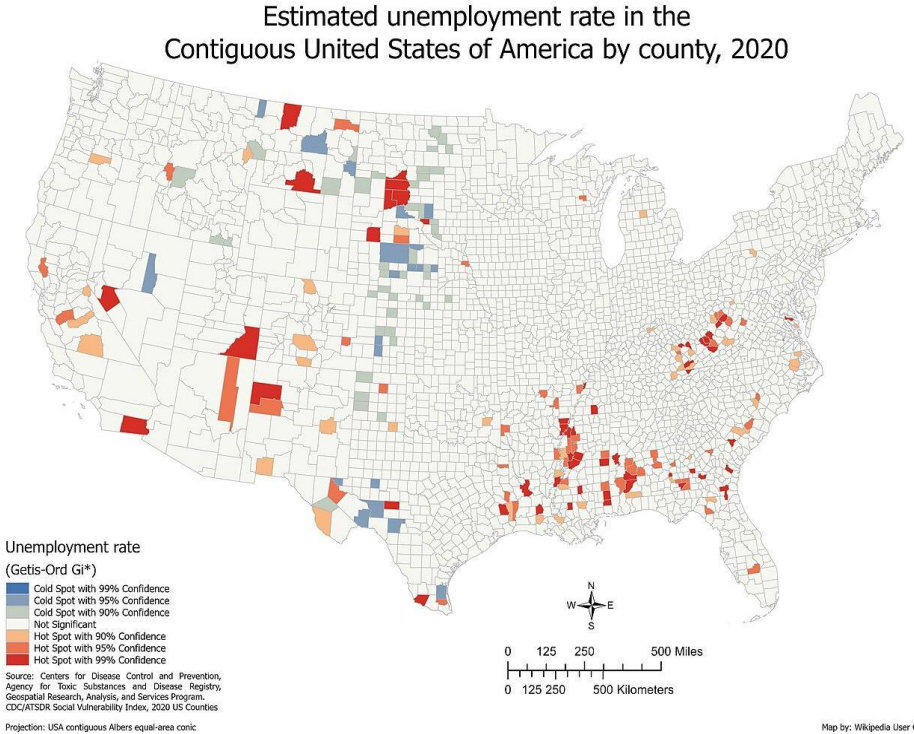
$$I = \sum_{i=1}^{N} \frac{I_i}{N}$$

- **Applications of Local Moran's I:**
  - Identifying disease hot spots and areas needing intervention
  - Highlighting neighborhoods with high or low socioeconomic indicators
  - Detection of pollution clusters of ecological anomalies

Getis-Ord Gi*

- A spatial statistical method that identifies areas of high (hot spots) and low (cold spots) values in a dataset.
- **Purpose:**
  - Detect local concentrations of high or low values
  - Highlight statistically significant clusters of spatial data.

- Explicitly focuses on identifying concentrated hot spots (high values) and cold spots (low values) vs. Local Moran's I, which included clusters and outliers



Getis-Ord Gi*

- X_j: value at location *j*
- $w_{i,j:}$ spatial weights between locations *i* and *j*

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

- Positive G*i** indicates a hot spot: High values location *i* and its neighbors
- Negative G*i** indicates a cold spot: Low values at location *i* and its neighbors

- The magnitude of G*i** values indicate the strength of clustering

Local Moran’s I vs. Getis-Ord Gi*

Aspect	Local Moran's I	Getis-Ord Gi*
Purpose	Detects clusters (hot spots/cold spots) and spatial outliers.	Detects concentration of high or low values (hot and cold spots).
Statistic meaning	Measures how similar or dissimilar a location is compared to neighboring values, identifying clusters or outliers (high-high, low-low, high-low, low-high).	Measures local clustering by evaluating if nearby features collectively have high or low values compared to the entire dataset.
Interpretation	Identifies clusters and spatial outliers explicitly; suitable for spotting anomalous locations	Identifies significant hotspots (high-value clusters) and cold spots (low-value clusters).
Output	Classified as clusters/outliers (HH, LL, HL, LH).	Identifies statistically significant hotspots (high values) and cold spots (low values).
Typical application	Identifying localized patterns and spatial outliers in data, e.g., disease hotspots and anomalies.	Detecting localized areas of significantly high or low concentration, e.g., crime hotspots or economic clusters.

Regression analysis

- A statistical method to describe the relationship between variables
- Typically involves one dependent (response) variable and one or more independent (explanatory, predictor) variables
- **Purpose:**
  - To explain variation in the dependent variable
  - To predict values of the dependent variable
- A typical linear regression models the linear relationship between response and predictor variables, and might be of the form of

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

Error term
(residuals)

Linear regression

- **Assumptions of linear regressions:**
  - **Linearity:** Relationship between response and predictors is linear
  - **Normality:** Residuals are approximately normally distributed
  - **Homoscedascitiy:** constant variance of residuals
  - **Independence:** Observations are independent of each other

Spatial data almost always violates the independence assumption. When it does, OLS standard errors are wrong.

- **Interpretation of linear regressions**
  - **Coefficient estimates:** magnitude and direction of relationship
  - **R^2:** Proportion of variance explained by the model
  - **p-value**: significance of the relationship
  - **Residuals:** Difference between observed and predicted values

Mapping residuals

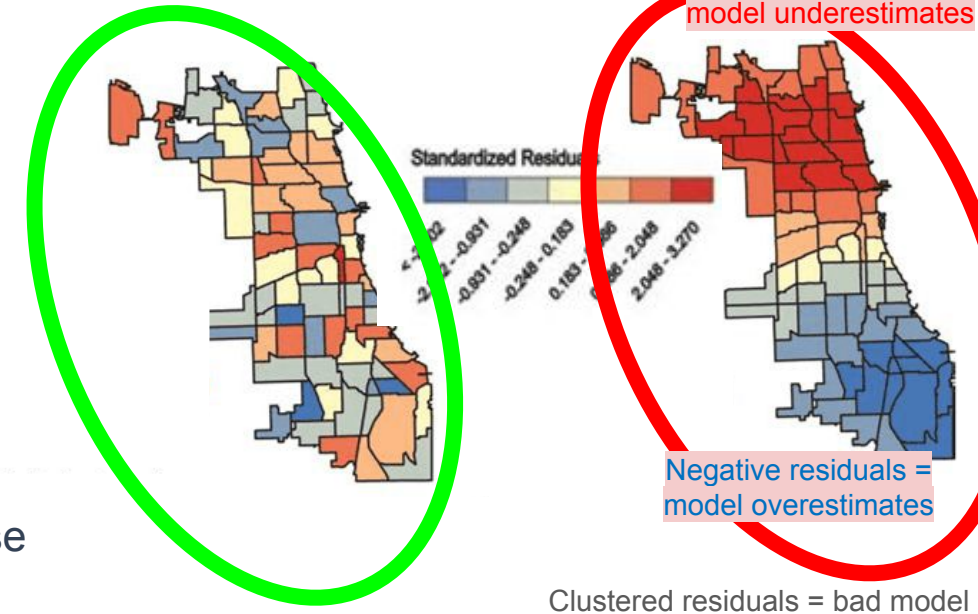
Moran's I diagnosis: Calculate Global Moran's I on the residuals. If significant (e.g., I = 0.6, p < 0.01) → spatial autocorrelation in residuals → model needs correction.

Residuals are the difference between observed and predicted values. In a well-specified model, **residuals should be random noise**, that is, no geographic pattern.

Let's plot the residuals for each area on a map. This is Exploratory Spatial Data Analysis.

If residuals cluster spatially:

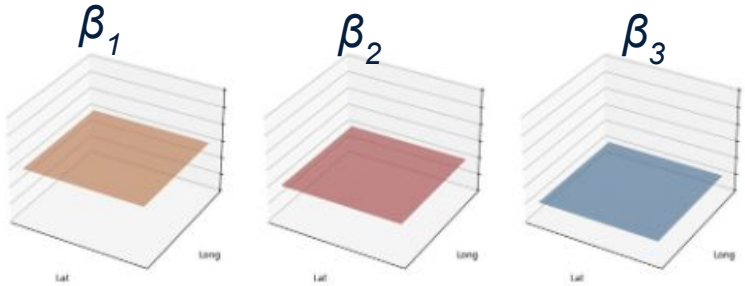
1. the model is missing spatially-structured predictors
2. the independence assumption is violated
3. OLS standard errors are too small → false positives



Random residuals = good model

Clustered residuals = bad model

Linear regression

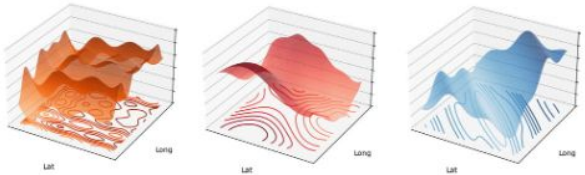


- Let's stop and think about the OLS regression
  - $Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + e$

- Coefficients (β) are global constants across the study area
- They are the same for each spatial unit

What are our options?

- Ignore the observed spatial autocorrelation and accept OLS as is
- Allow the coefficients to vary with location
  - We can use Geographically Weighted Regression (GWR)
- We can use spatial lag or spatial error models



Spatial Error and Spatial Lag models

- **Spatial Error Model (SEM):**
  - Addresses residual spatial autocorrelation. For SEM, spatial dependence is viewed as a nuisance parameter.

$$y = x\beta + u, u = \lambda W u + \epsilon$$

The residual term (u) is modeled by a different regression equation that predicts the residual using a spatial autoregressive parameter λ and a spatial weights matrix (W), along with its own residual term (ε).

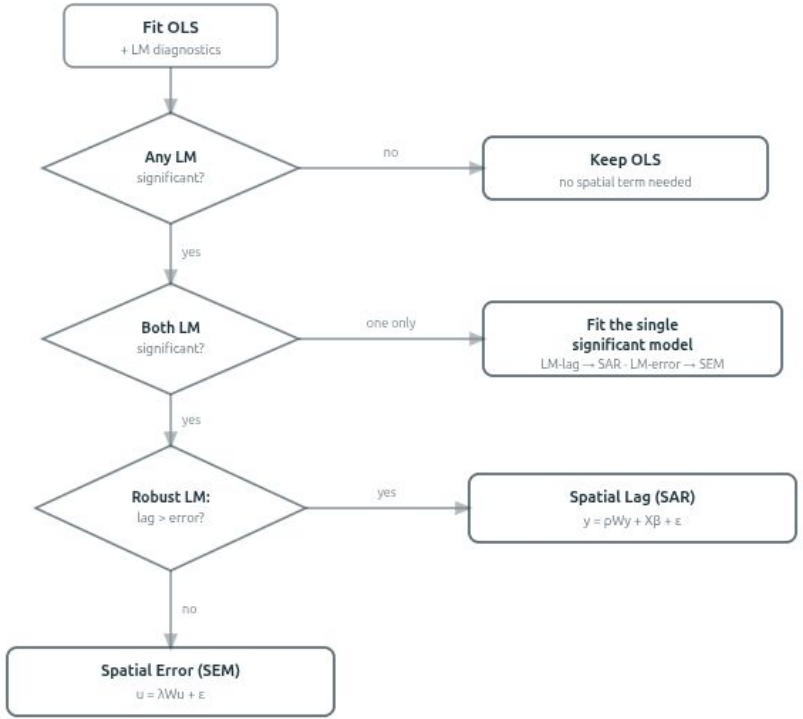
- **Spatial Lag Model (SLM):**
  - Incorporates the spatial dependence as an explanatory variable. SLM is used when the dependent variable has a large amount of spatial autocorrelation and exhibits a spatial spillover effect (i.e. changes in one area elicit changes in neighboring areas).

$$y = \rho W y + X\beta + u$$

Wy: spatial lag (i.e. weighted average from neighboring locations)
ρ: autoregressive parameter (strength of neighbor's influence)

Spatial Error and Spatial Lag Models

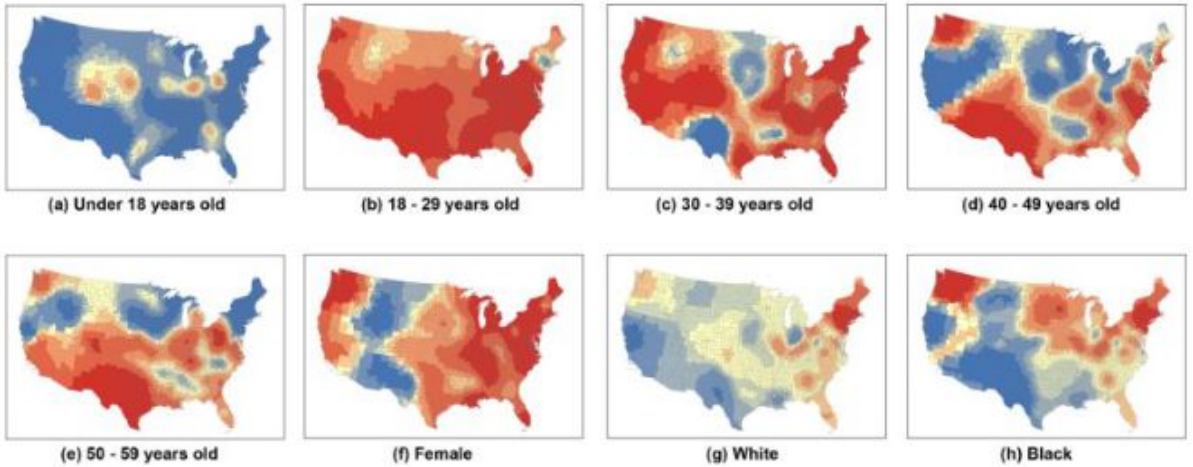
- SEM or SLM?
  - Choice is formalized by Anselin and Rey (2014)
  - Series of statistical test called a Lagrange Multiplier (LM) test



Anselin, L., and Sergio J. Rey. 2014. "Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL." ISBN 9780986342103.

Geographically Weighted Regression

- Local regression method that **allows model parameters to vary spatially**
- **Captures spatial heterogeneity** by producing location-specific regression coefficients
  - Practically, GWR generates a local model for each county, considering the observation values from neighboring counties



Geographically Weighted Regression

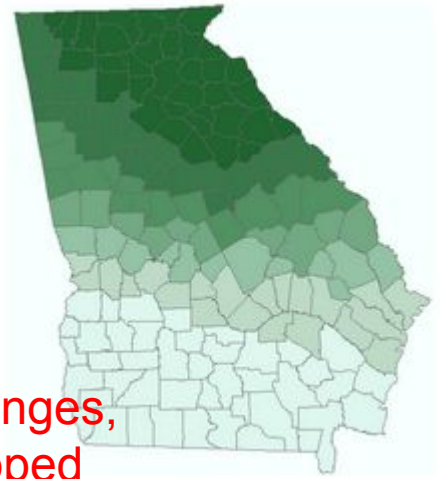
- **Bandwidth selection**
  - Determines the size of the neighborhood influencing each local regression
  - Small bandwidth: rapid distance decay
  - Large bandwidth: smoother weighted scheme
- Bandwidth can be defined manually or determined with adaptive methods
- **Interpreting GWR**
  - For each explanatory variable, GWR generates as many estimates as features included in the model (e.g. 1 value for every county)
  - Residuals and coefficients for each explanatory variable can be mapped

- Georgia Educational Attainment example (https://gwr.maynoothuniversity.ie/wp-content/uploads/2016/01/GWR_Tutorial.pdf)
  -

Comparing OLS and GWR

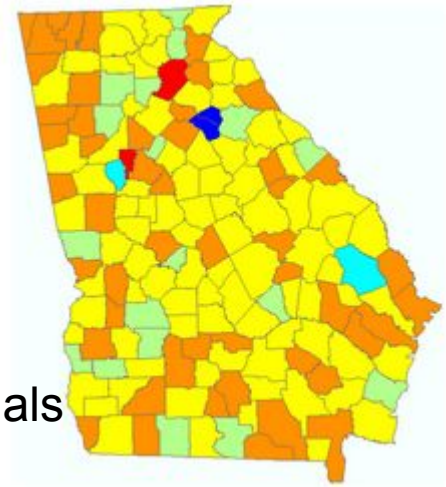
- **Predicting educational attainment in Georgia**

Predictor variables	Global parameter estimate	GWR parameter estimates
Total population, β_1	0.24 x10^{-4}	0.14 to 0.28 x10^{-4}
% rural, β_2	-0.044	-0.06 to -0.03
% elderly, β_3	-0.06 (not signif.)	-0.26 to -0.06
% foreign born, β_4	1.26	0.51 to 2.42
% poverty, β_5	-0.15	-0.20 to –0.00
% black, β_6	0.022 (not signif.)	-0.04 to 0.08
Intercept, β_0	14.78	12.62 to 16.49
Diagnostics		
Residual SS	1816	1506
Adjusted R^2	0.63	0.68
AICc	855.4	839.2



Coefficients are expressed as ranges, and can be mapped

% foreign-born ranges from 0.51 to 2.42. The global coefficient of 1.26 hides enormous geographic variation.



Mapped residuals

Advantages and limitations of GWR

- GWR provides a framework to model spatially varying relationships, offering deeper insights into spatial data

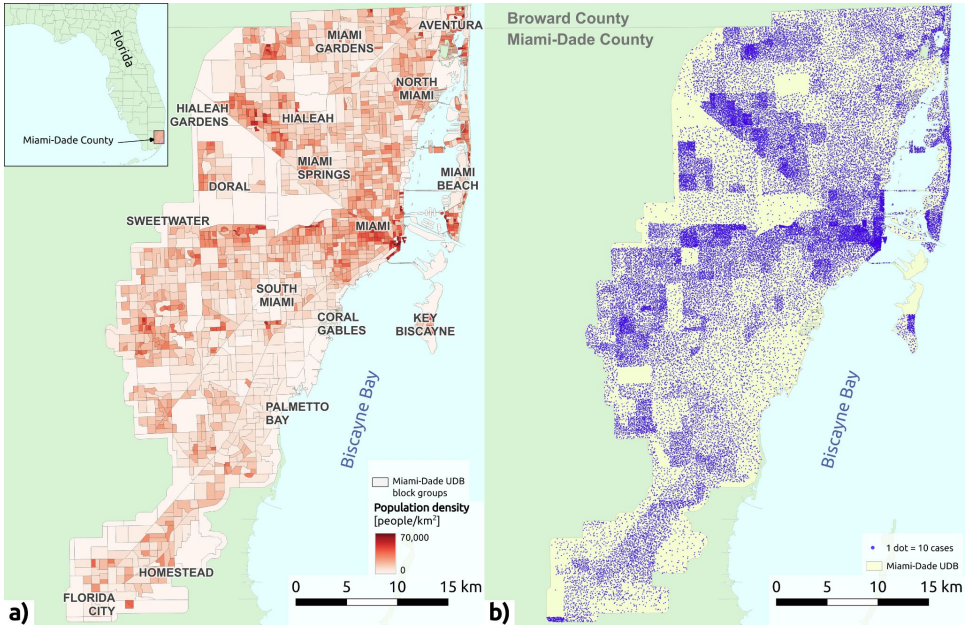
- **Advantages**
  - Identifies spatial heterogeneity (non-stationarity) in relationships
  - Provides localized insights for targeted interventions
- **Limitations**
  - Computationally more intensive, especially with large datasets
  - Risk of overfitting due to multiple local models

Spatial Statistics in Practice

- Analyzing factors associated with COVID-19 cases in Miami-Dade county
  - Count data -> Negative Binomial Regression
  - Diagnosing the NBM
  - Spatial Eigenvector Filtering
- Hands-on Session with R on Google Colab
  - Implementing the COVID-19 case study
  - Exploring clustering in Miami-Dade

Case Study: Factors associated with COVID-19 cases

- **Study area:** MDC
  - ~2.7M population
  - 1,831 census block groups
- **Demographics:**
  - 66% hispanic or Latino
  - 18% Black or African American
  - 54% foreign born
  - Median HH income: $58K
- **Data sources:**
  - 491,769 geocoded COVID-19 cases (count data)
  - Socioeconomic variables (ACS)



- **RQ:** Which neighborhood-level socio-economic and demographic characteristics predicts areas of elevated or suppressed COVID-19 case rates?

Why not OLS for count data?

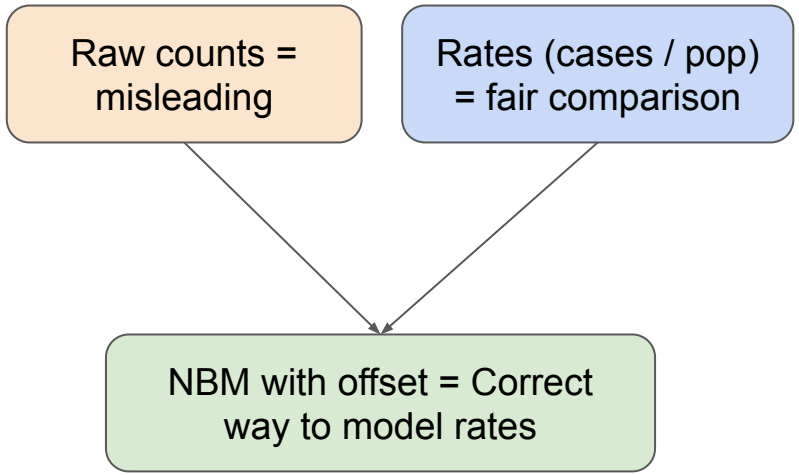
- Disease surveillance data = counts (cases per area)
- OLS assumes continuous, normally distributed residuals
  - Can predict negative case counts
  - Doesn't handle mean-variance relationship in counts
- Solution: Generalized Linear Models (GLMs)
  - **Link function: log(μ)** = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$
  - Ensures predicted counts are always ≥ 0

Model	When to use
Poisson	Counts where mean ≈ variance
Negative Binomial	Counts where variance >> mean (overdispersion)

In practice: health data is almost always overdispersed -> use **Negative Binomial**

Negative Binomial Regression (NBM)

- Adds a dispersion parameter (θ) to handle overdispersion
- As $\theta \to \infty$, NB converges to Poisson



The offset term

- Instead of
  log(cases) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$
- We write
  log(cases) = log(population) + $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$

- The log(population) is the offset. It's not estimated, it's fixed.
- This effectively models rates (cases/population) while keeping cases as the response
- Essential when areas have different population sizes

Interpreting NBM - Incidence Rate Ratios

- NB coefficients (β) are on the log scale, hard to interpret directly
- Exponentiate to get **IRR = exp(β)**

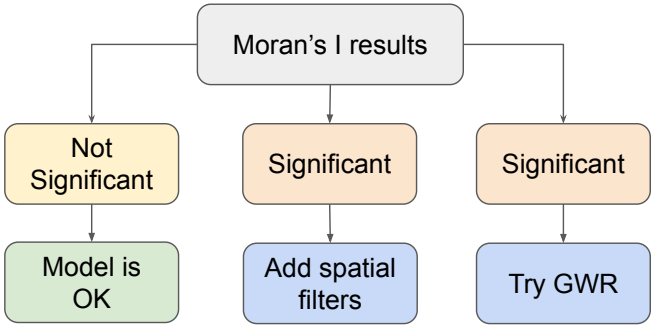
IRR	Interpretation
IRR = 1.0	No effect
IRR = 1.15	15% increase in expected rate per unit increase in X
IRR = 0.85	15% decrease
IRR = 0.995	0.5% decrease per unit increase in X

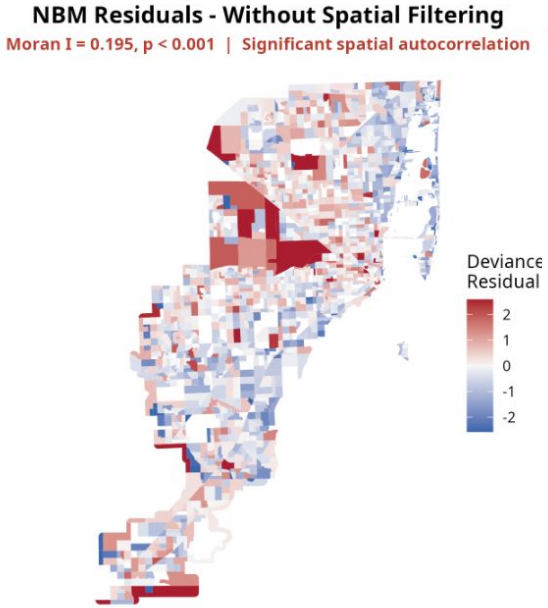
Example from Miami-Dade Covid-19 study:

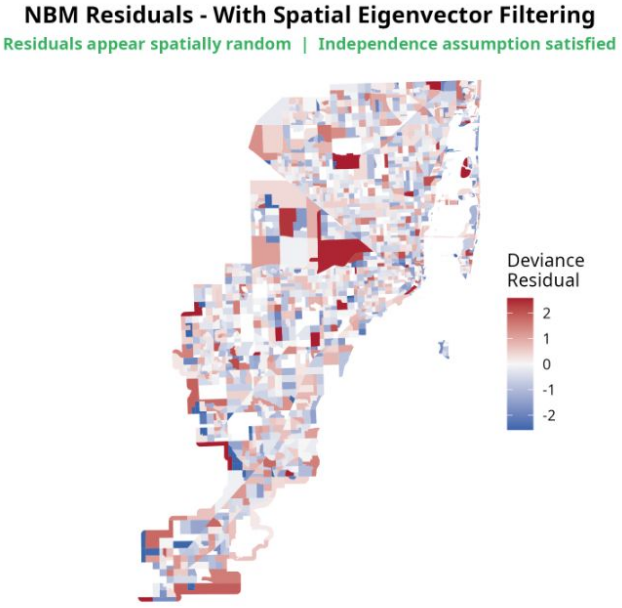
- % of black population: IRR = 0.991 - 0.995 ($p<0.001$)
  - Each 1% increase in Black population -> 0.5 = 0.9% fewer reported cases.
  - Counterintuitive. Lower reported cases likely represented testing barriers, not lower transmission

After fitting NBM - Check your Residuals

1. Fit NBM -> extract residuals
2. Map the residuals. Do you see spatial patterns?
3. Calculate Moran's I on residuals

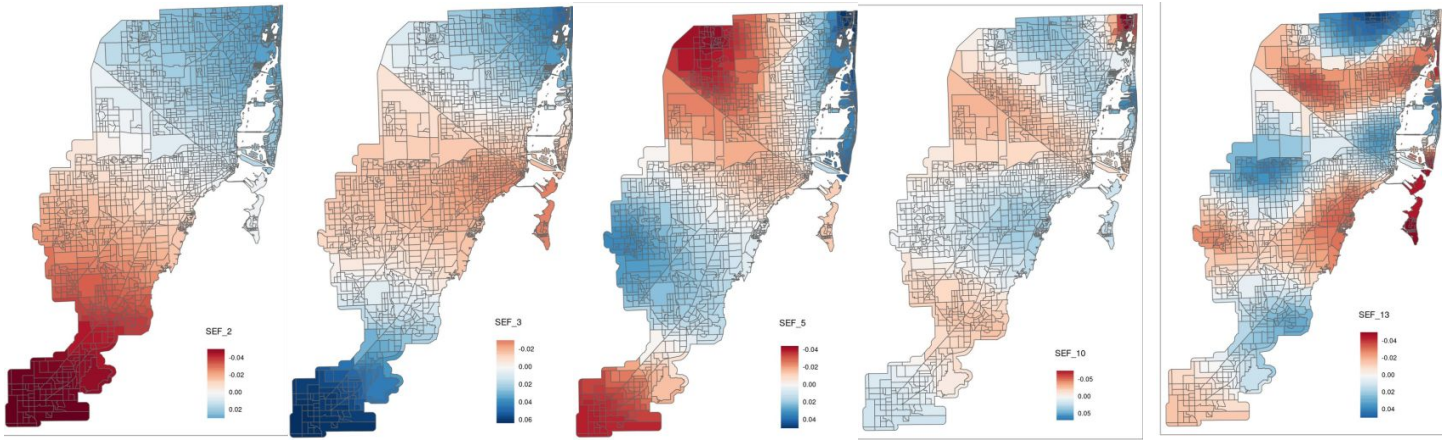






Spatial Eigenvector Filtering (SEF)

- **The problem:** Residuals show spatial autocorrelation -> violates independence assumption
- The idea:
  - Extract spatial patterns directly from the spatial weights matrix
  - Decompose *W* into **eigenvectors** - each one represents a different spatial pattern
  - Add selected eigenvectors as **extra predictors**
  - **They “absorb” the spatial autocorrelation** -> residuals become independent



Think of it as: The eigenvectors are "spatial basis functions." The model picks whichever spatial scales are needed to remove autocorrelation.

Spatial Eigenvector Filtering - How it works in practice

1. Build spatial weights matrix (e.g. queen contiguity)
2. Compute eigenvectors from the weights matrix
3. Fit baseline model (e.g. NBM without spatial filters)
4. Test residuals for spatial autocorrelation (Moran's I)
5. Add eigenvectors as predictors - e.g. use stepwise selection (AIC)
6. Final model retains key socioeconomic predictors + a subset of eigenvectors

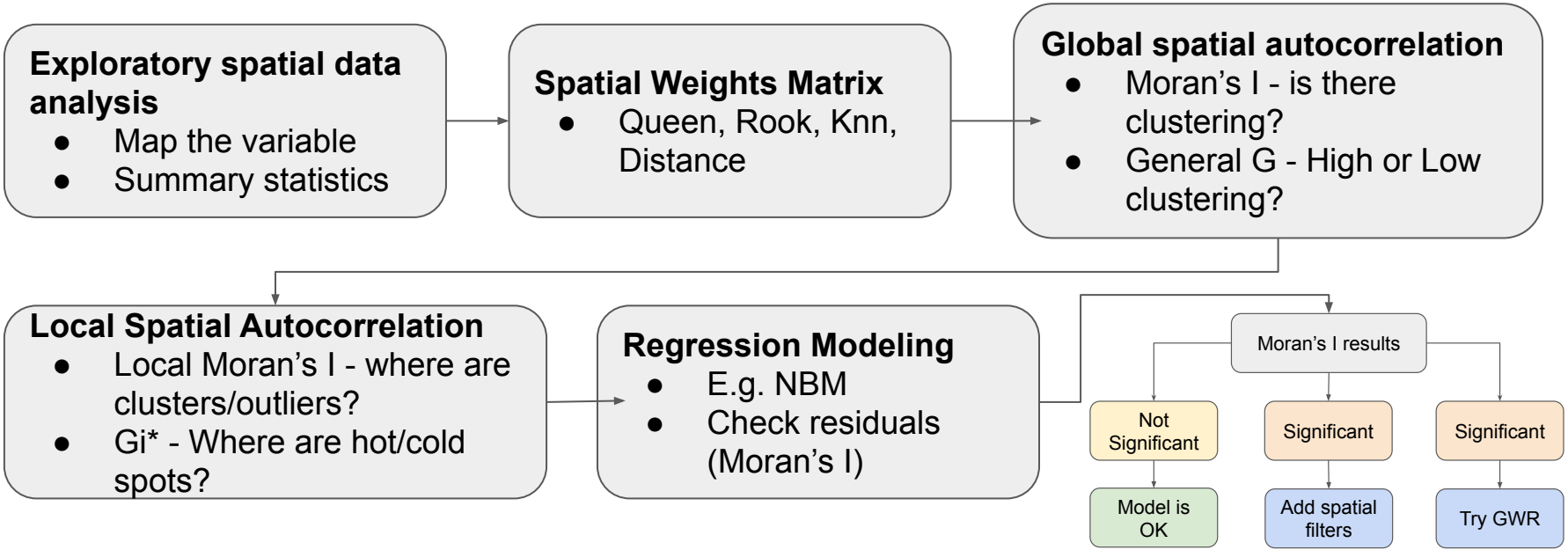
Advantages

- Works with any regression type (OLS, Poisson, NBM, logistic)
- No need to choose between SEM vs. SLM
- Captures spatial patterns at multiple scales
- Interpretable

Limitations

- Can be computationally intensive with many spatial units
- Risk of absorbing substantive spatial variation (not just nuisance autocorrelation)

The spatial analysis process



SEF vs. GWR

	SEF	GWR
Primary goal	Fix spatial autocorrelation in residuals	Explore spatial heterogeneity in coefficients
Output	One set of globally corrected coefficients	One coefficient per location per predictor
Model family	Any GLM (NBM, Poisson, logistic, OLS)	OLS or Gaussian (GWmodel)
Interpretability	Standard IRRs / regression coefficients	Coefficient maps
Computational cost	Low–moderate	Moderate–high
Handles autocorrelation?	Yes (explicitly)	Partially (large bandwidth reduces it)
When to use	Inferential global model with unbiased CIs	Exploratory spatial variation in relationships

Hands-on session

- Analyzing factors associated with COVID-19 cases in Miami-Dade county
  - Count data -> Negative Binomial Regression
  - Diagnosing the NBM
  - Spatial Eigenvector Filtering
- Hands-on Session with R on Google Colab
  - Implementing the COVID-19 case study
  - Exploring clustering in Miami-Dade

Why R for spatial analysis?

- Opensource, free, reproducible
- Rich ecosystem for spatial data (sf, sfdep, tmap, GWmodel, etc.)
- Used in academic research

Package	Purpose
`sf`	Read, write, manipulate spatial data (vector)
`spdep` / `sfdep`	Spatial weights, Moran's I, spatial regression
`MASS`	Negative Binomial regression (`glm.nb`)
`GWmodel`	Geographically Weighted Regression
`tmap`	Thematic mapping (choropleth, hot spots)
`ggplot2`	General-purpose plots and charts

R Basics

- Variables and assignment:
```
x <- 5
```
- Reading data:
```
st_read("file.geojson")
```
- Piping:
```
data |> filter(...) |> mutate(...)
```
- Fitting a model:
```
glm.nb(y ~ x1 + x2, data = df)
```
- Plotting:
```
tm_shape(data) + tm_polygons("variable")
```

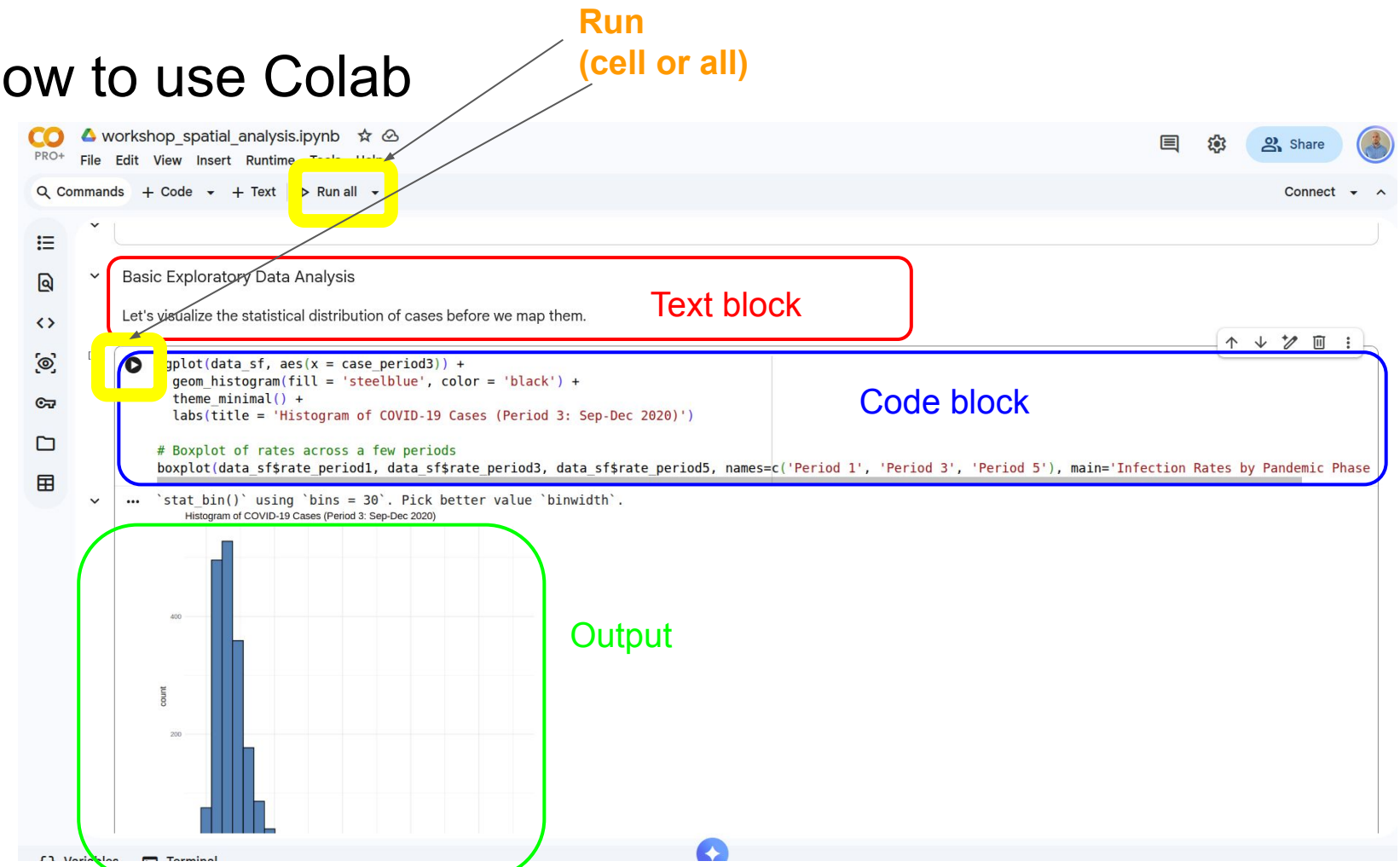
You don't need to memorize these. The notebook(s) has everything pre-written.

You will just run cells and interpret results.

What is Google Colab?

- Free, cloud-based Jupyter Notebook environment by Google
- No installation required - runs in your browser
- Supports R (not just Python)
- Some pre-installed libraries (no spatial)
- GPU access, if needed
- Perfect for workshops because everyone has the same environment

How to use Colab



Let’s get started

- Hands-on Session with R on Google Colab
  - Implementing the COVID-19 case study
  - Exploring clustering in Miami-Dade

1. Follow the QR code or go to https://gatorlab.io/presentations/rcmi_advanced_2026/ Scroll down and find the link to the Google Colab Notebook

2. Click File -> Save in Drive before modifying and running it.

Key takeaways

- Geography matters
  - Spatial dependence violates the independence assumption of standard models
- Always define your spatial weights matrix thoughtfully
- Always check residuals for spatial autocorrelation
  - Map them, Moran's I, decide next step
- Three tools we briefly covered today
  - **Getis-Ord Gi* / Local Moran's I** - You need to find **where** cluster are
  - **Spatial Eigenvector Filtering** - You want a **global** model but need to fix spatial bias in residuals
  - **Geographically Weighted Regression** - You suspect relationships **vary across space**

UF UNIVERSITY of FLORIDA

UF IFAS UNIVERSITY of FLORIDA

Thank you!

Questions?

Workshop materials: https://gatorlab.io/presentations/rcmi_advanced_2026/

https://gatorlab.io

Levente Juhász
levente.juhasz@ufl.edu

GATOR
GEOSPATIAL ANALYTICS, TECHNOLOGY AND OPEN RESEARCH