Spatial Data Analysis in R

Dealing With Spatial Dependence 2
Classes of Spatial Models

Eco 697DR – University of Massachusetts, Amherst – Spring 2022
Michael France Nelson

Areal Data

Areal data are data that exist at a polygon level (as opposed to point data).

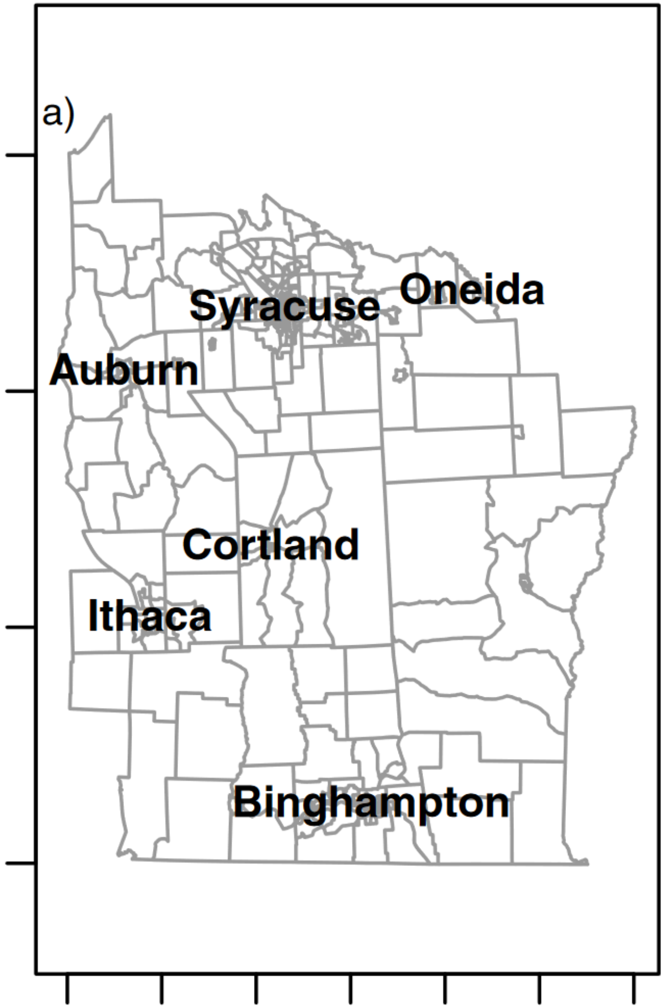
Aerial entities can be

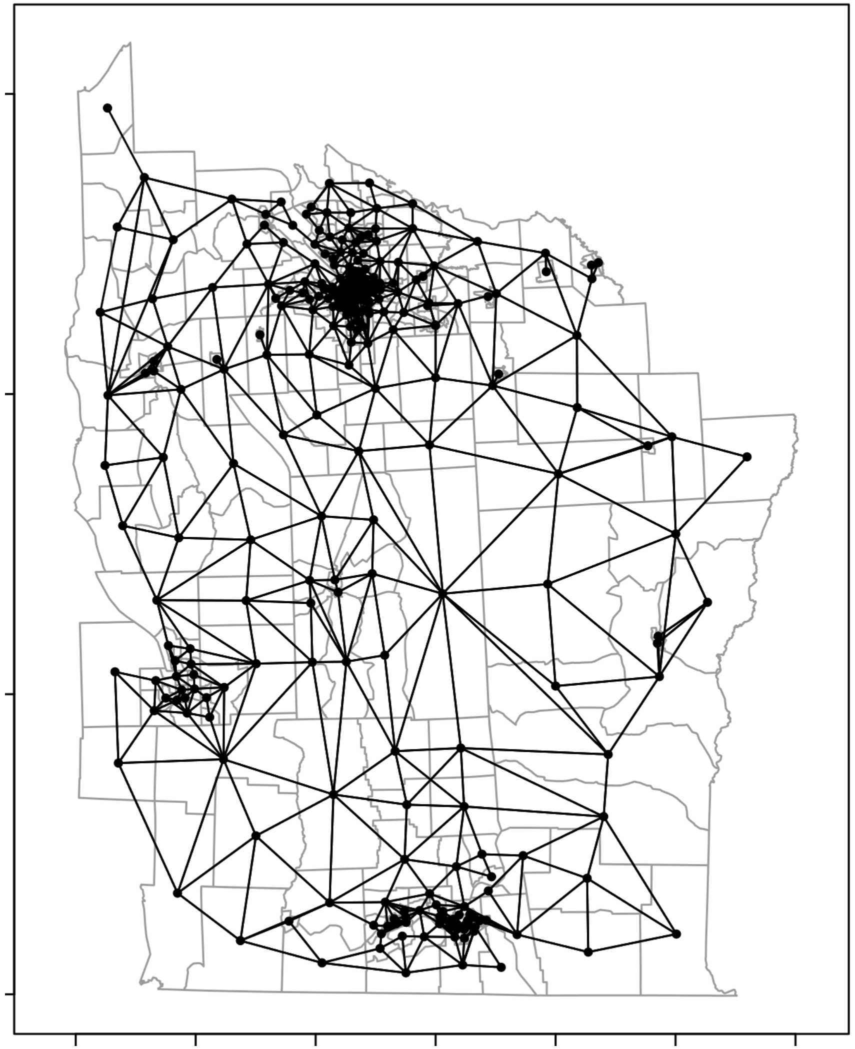
- Aggregations of point data
- Exhaustive tessellations of a region
- Political boundaries
- Area codes, postal codes

Modifiable unit area problem: modifying areal unit boundaries may lead to different outcomes:

- Gerrymandering.
- Polygons of different sizes: can be normalized for comparison

Areal Neighborhoods





Neighbors share an edge

Neighborhoods can be represented as a graph, a.k.a. and network.

From Bivand et al., 2008

Neighbors and Weights: 4-Neighbor Rule

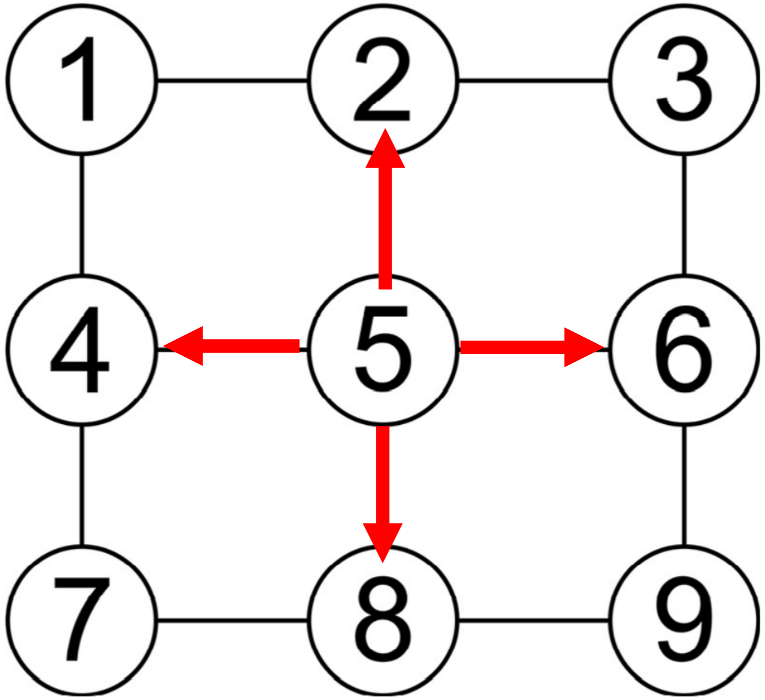


Fig. 1. Spatial arrangement of sites in a simple 3 × 3 gri
where the numbers label each site.

$$\mathbf{W} = \begin{array}{ccccccccc}
 & 2 & & 4 & & 6 & & 8 & \\
\end{array}
\left(\begin{array}{ccccccccc}
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & \boxed{1} & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & \boxed{1} & 0 & 1 & 0 & 0 \\
0 & \boxed{1} & 0 & \boxed{1} & 0 & \boxed{1} & 0 & \boxed{1} & 0 \\
0 & 0 & 1 & 0 & \boxed{1} & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & \boxed{1} & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{array}\right)$$

Point Data

Neighborhoods are usually constructed from:

- Distance: points within a radius, d, are in the neighborhood.
  - If d is small, you can have points with zero neighbors.
  - How could you get around this issue?
  - Usually, neighborhoods are not the same size.
- Count: k nearest points are within the neighborhood.
  - Neighborhoods are always the same size
  - Neighbors may not be within consistent distances

Tradeoffs between methods

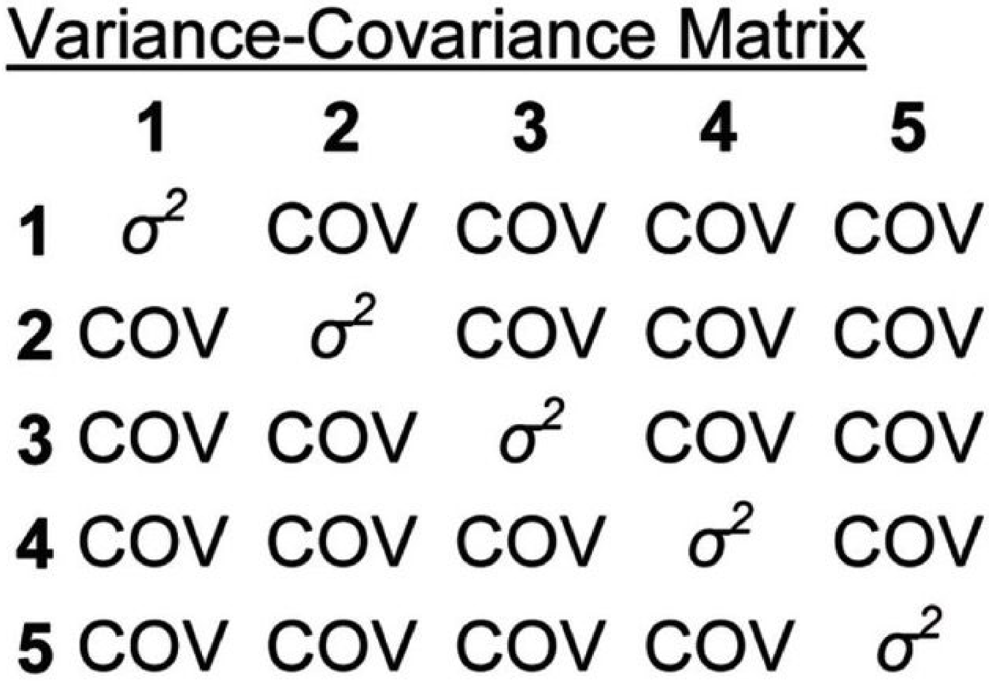
Variance/covariance structures

Simple linear models consider independence of errors. This is almost never the case…

What’s a variance/covariance matrix?

The simple linear model equation hides the variance/covariance matrix:

- SLR uses an identity matrix
- All off-diagonals are zero



Variance-Covariance Matrix

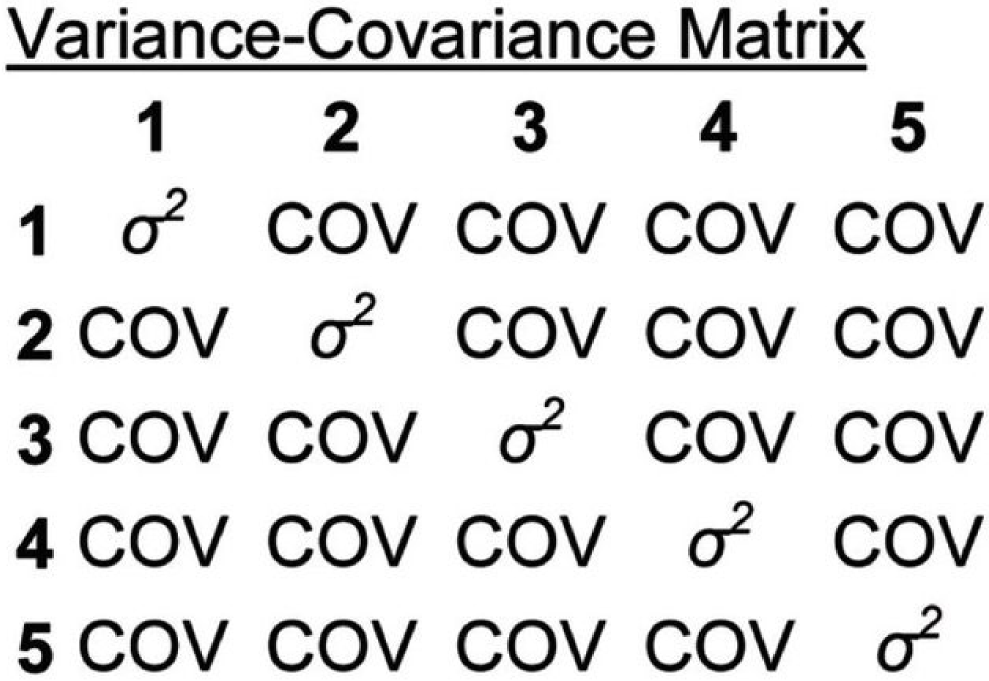
	1	2	3	4	5
1	σ^2	COV	COV	COV	COV
2	COV	σ^2	COV	COV	COV
3	COV	COV	σ^2	COV	COV
4	COV	COV	COV	σ^2	COV
5	COV	COV	COV	COV	σ^2

From Goedert et al. 2013

Variance/covariance structures

Generalized Least Squares (GLS) allows to modify the variance/covariance matrix directly

- We can populate the off-diagonals of the matrix to specify covariance structures.



Variance-Covariance Matrix

	1	2	3	4	5
1	σ^2	COV	COV	COV	COV
2	COV	σ^2	COV	COV	COV
3	COV	COV	σ^2	COV	COV
4	COV	COV	COV	σ^2	COV
5	COV	COV	COV	COV	σ^2

From Goedert et al. 2013

Types of Spatial Models

Spatially-Aware Regression

Types of spatially-aware models include:

- Spatial trend
- Spatial filter
- Spatial lag models: Autocovarites
- Autoregressive models
  - Simultaneous Autoregressive Models (SAR)
  - Conditionally Autoregressive Models (CAR)

Trend surfaces

Model a continuous dependence surface using x-y coordinates.

Coordinate-based. Capture patterns in exogenous factors, great for large-scale, smooth patterns.

Fitting a mathematical **function** to describe to environmental data based on location.

- Polynomial regression
- Smoothing: local regressions

- Doesn’t explicitly model autocorrelation in predictors or response
- Can help remove autocorrelation from residuals.

Trend surfaces

The environmental trend surface model can be incorporated into an overall regression to ‘even out’ spatial autocorrelation in residuals.

This can help flatten out some of the spatial structure in our response variable of interest.

A simple trend surface may not be enough, though….

Why not?

Trend Surface Analysis

$$\underset{\substack{\downarrow\\ \text{Response}}}{y_i} = \underbrace{lat^2 + lon^2}_{\substack{\text{Coordinate-based terms}\\ \text{Often polynomial}}} + \underbrace{\alpha + x_i\beta}_{\text{Coefficients + predictors}} + \underset{\substack{\searrow\\ \text{Error}\\ \text{(single term)}}}{e_i}$$

Spatial Filtering Models

What is spatial filtering?

- A technique to ‘filter out’ spatial autocorrelation.
- Separates an autocorrelated variable into correlated and noncorrelated parts
- The original variable is the sum of the correlated and noncorrelated parts.
- Attempts to remove autocorrelation from predictors or response.

What is an eigenvector?

- Property of a matrix, it gives us information about happens when we multiply things by a matrix.
- A vector whose direction doesn’t change, but that can be stretched under a linear transformation.
- Kind of like roots of a polynomial
- Not super important to know the details

* Disclaimer: I’m not a linear algebra expert!

Eigenvector Filtering

- “Eigenvector spatial filtering (ESF) is proposed to account for spatial effects. It selects a subset of eigenvectors of a spatially [sic] weight matrix and adds them to the original regression model as new independent variables. The linear combination of these eigenvectors filters the spatial autocorrelation out of the observations, thus enabling model processes to proceed as if the observations were independent…”
- From Zhang et al, 2018

Spatial Lag: Autocovariates

Diffusion or contagious processes

Include a **lag** term to account for nearby values

- Each observation has a potentially different value for the term

Spatial lag: distance or neighborhood which influences an observation.

Neighborhood concepts:

- Distance based
- K nearest-neighbor based
- Polygon continuity
- Critical distance

What techniques have we seen to determine an appropriate spatial lag?

Spatial Lag Models: Autocovariate

$$\underset{\underset{\text{Response}}{\swarrow}}{y_i} = \underbrace{\rho WY}_{\text{Neighborhood-based lag component}} + \underbrace{\alpha + x_i\beta}_{\text{Coefficients + predictors}} + \underset{\underset{\text{Error (single term)}}{\searrow}}{e_i}$$

Generalized Lest Squares: Autoregressive Models

$$y_i = \underbrace{\alpha + x_i\beta}_{\text{Coefficients + predictors}} + \underbrace{E}_{\text{Error}}$$

Response

Coefficients + predictors

Error
Variance/Covariance
Matrix.
The ‘lag’ or
autoregressive part
happens here

SAR and CAR

- Autoregressive models of non-independent residuals
  - Variance/covariance structure
- Simultaneous Autoregressive Models: lag term calculated from all data
- Conditionally Autoregressive Models: lag term calculated from neighbors
- Distinction is fuzzy, it is possible to view a SAR as a type of CAR and vice-versa.

Autoregressive models:

- Model the variance/covariance matrix
- GLS: CAR and SAR

Autocovariate models (spatial lag)

- Models the lag as covariates
- Uses a simple error term

Hierarchical Models: Random Effects

Random Effects:

- Random effects are always categorical
- Think of them as a grouping factor
- Observations within a group are autocorrelated
- Observations among groups [are assumed to be] independent
- Random factor levels effects are assumed to be normally distributed
  - There should usually be at least 5 levels

Hierarchical Model

$$\underset{\underset{\text{Response}}{\swarrow}}{y_i} = \underbrace{\gamma_j}_{\substack{\text{Random Effect} \\ \text{For group j}}} + \underbrace{\alpha + x_i\beta}_{\text{Coefficients + predictors}} + \underset{\underset{\substack{\text{Error} \\ \text{(single term)}}}{\searrow}}{e_i}$$

You can [approximately] think of random effects as samples from separate populations with [hopefully] similar properties, but whose mean values may differ.

- This is an extreme oversimplification

Examples

- Observations on transects: individual transects are the random effect
- Observations in schools: individual schools are the random effect

From Arnqvist, 2019

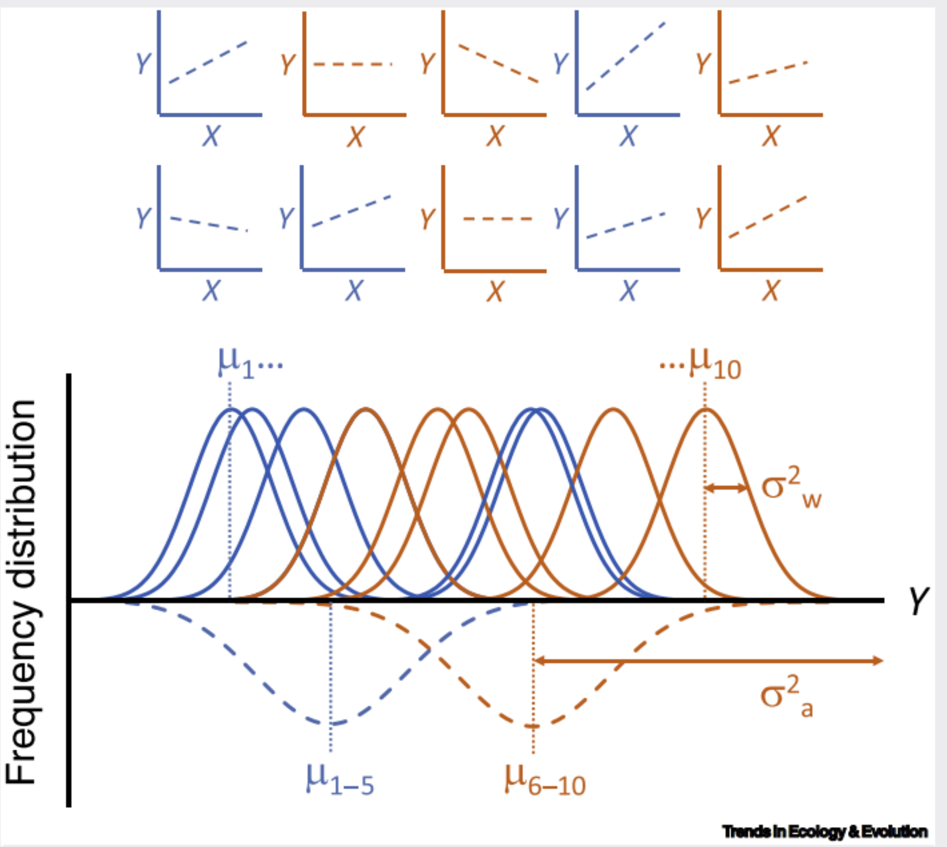


Figure I. A Schematic Illustration of Data from 10 Different Populations. Populations belong to either a blue or an orange category, that differ in the mean value (μ) of some response variable Y that shows a common within-population variance (σ^2_w). Each category of five populations also shows a grand mean value (e.g., μ_{1-5}) that shows a common among-population variance (σ^2_a). The 10 populations also differ in the manner in which a covariate (X) relates to the response variable, as illustrated at the top.

Spatially-Aware Regression Workflow

A general workflow for spatially-aware regression

1. Data import, prep, exploration, cleaning
2. Examine autocorrelation in predictors and response.
3. Fit nonspatial model, examine autocorrelation in residuals
4. Propose spatially-aware options
    - non-spatial covariates such as:
        - random effects, polynomial terms, additional predictors
    - spatial covariates such as:
        - terrain/landscape covariates, spatial lag terms
    - variance/covariance structures
5. Fit nonspatial model, examine autocorrelation in residuals
6. Iterate 4 and 5 until you are satisfied

Model Type Summary

- Spatial Filtering: ‘filters out’ autocorrelation component in variables, adds autocorrelated predictors.
- Trend Surface: Fits a model of the spatial coordinates
- Spatial Lag – Autocovariate: fits new neighborhood-based predictors, which are functions of the response at nearby locations.
- Spatial Autoregressive: Does not fit new covariates, but rather works with the variance/covariance structure
- Hierarchical Models: Specify the hierarchical structure in the experiment with random effect terms (covariates). Can accommodate complicated experimental structures