# Nature of Spatial Data

## Outline

- "Spatial is special"
- Bad news: the pitfalls of spatial data
- Good news: the potentials of spatial data

## Spatial Is Special

- Are spatial data special?
  - Why spatial data require spatial analytic techniques, distinct from standard statistical analysis that might be applied to any ordinary data?
  - Bad news & good news

Number of cases of Lyme disease

## Bad News First

- Many of the standard techniques and methods documented in standard statistics textbooks have significant problems when we try to apply them directly to the analysis of the spatial distributions and phenomena

## Bad News First (Cont.)

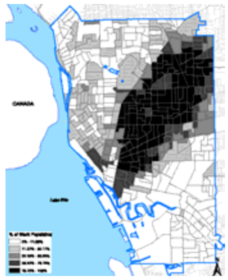
- The fundamental requirements of conventional statistical analysis
  - Identifiable study objects
  - Independence of cases
  - Normal distribution
  - Linearity
  - Stationary
  - Data accuracy

## Bad News First (Cont.)

- Spatial data always violate many of these fundamental requirements, due to:
  - Spatial autocorrelation
  - Modifiable areal unit problem
  - Ecology fallacy
  - Scale
  - Nonuniformity of space
  - Edge effect

## Spatial Autocorrelation

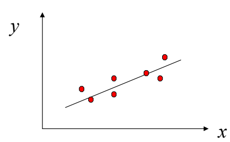
- Waldo Tobler's 1st Law of Geography (1970): '*Everything is related to everything else but nearby things are more related than distant things*'
  - Housing market
  - Elevation change
  - Air temperature



African American Population Concentration in Buffalo, NY

## Spatial Autocorrelation (Cont.)

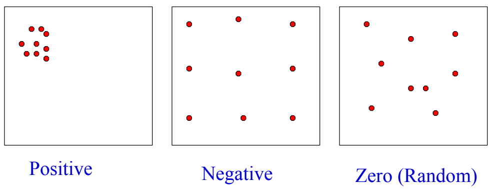
- Violation of independency assumption: The nonrandom distribution of phenomena in space → dependence among data in different locations→ violate the independence assumption
  - Data redundancy (affecting the calculation of confidence intervals
  - Biased parameter estimates




## Types of Spatial Autocorrelation

- **Positive** autocorrelation: nearby locations are likely to be similar from one another
- **Negative** autocorrelation: observations from nearby observations are likely to be different from one another
- **Zero** autocorrelation: no spatial effect is discernible, and observations seem to vary randomly through space
  - Standard statistical methods

## Types of Spatial Autocorrelation




## Detecting Spatial Autocorrelation

- Spatial autocorrelation diagnostic measures
  - Joins count statistics
  - Moran's I
  - Geary's C
  - Variogram cloud

## Why Spatial Autocorelation?

- **Types of Spatial Variation**
  - **First order** SV:
    - occurs when observations across a study region vary from space to space due to changes in the underlying properties of the local "environment"
  - **Second order** SV:
    - due to local interaction effects between observations
  - In practice, difficult to distinguish them

## Modifiable Areal Unit Problem (MAUP)

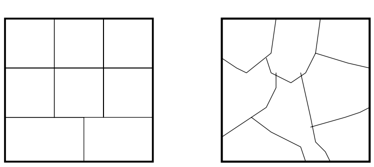
- Spatial analysis often relies on spatially aggregate data
  - e.g. census data: collected at the household level but reported for practical and privacy reasons at various levels of aggregation (block, block group, tract, county, state, etc.)
  - e.g. traffic analysis zone (TAZ): relies on census data to predict future traffic demand
- Potential problems in almost every field that utilizes spatial data → violation of identifiable study objects

## MAUP (cont.)

- MAUP: the aggregation units used are often arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way
  - If the spatial units in a particular study were specified differently, very different patterns and relationships might be observed
  - Many standard statistical analysis are sensitive to the analysis units

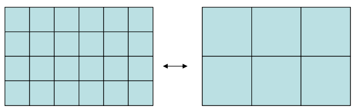
## Understanding MAUP

- **Modifiable Area**: Units are arbitrarily defined → different organization of the units may create different analytical results

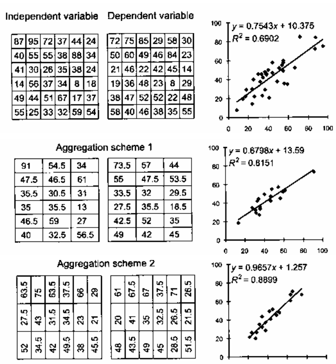


# Understanding MAUP

- **Scale issue**: involves the aggregation of smaller units into larger ones.
  - Generally speaking, the larger the units, the stronger the relationship among variables



# Illustration



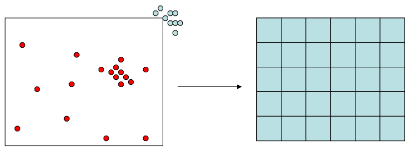

Openshaw and Taylor (1979) showed that with the same underlying data it is possible to aggregate units together in ways that can produce correlations anywhere between -1.0 to +1.0

# Special Consideration with MAUP

- Reasons?
  - Problems of data?
  - Problems of spatial units?
- Solutions?
  - Using the most disaggregated data
  - Produce a optimal zoning system
  - Others?

# Edge Effects

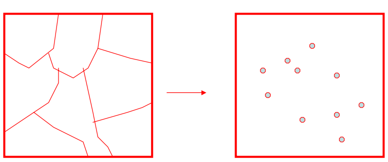
- Edge effects arise where an artificial boundary is imposed on a study, often just to keep it manageable
  - Entities at the edge of study area only neighbors in one direction (towards the middle)



Spatial interpolation

# Ecological Fallacy

- A situation that can occur when people makes an inference about an individual based on aggregate data for a group



(Reference: http://jratcliffe.net/research/ecolfallacy.htm)

# Ecological Fallacy

- We might observe a strong relationship between income and crime at the county level, with lower-income counties being associated with higher crime rate.
- Which conclusion?
  - Lower-income persons are more likely to commit crime
  - Lower-income counties tend to experience higher crime rates

# Ecological Fallacy

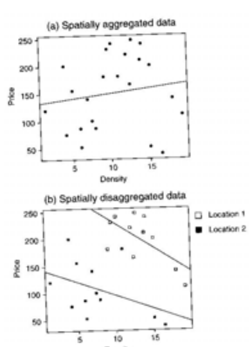
- We might observe a strong relationship between income and crime at the county level, with lower-income counties being associated with higher crime rate.
- Which conclusion?
  - Lower-income persons are more likely to commit crime
  - Lower-income counties tend to experience higher crime rates

# Ecological Fallacy

- Two related issues:
  - Identifying associations between aggregate figures is defective?
  - Inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are wrong?
- What should we do?
  - Be aware of the process of aggregating or disaggregating data may conceal the variations that are not visible at the larger aggregate level (**Simpson's Paradox**)

# Ecological Fallacy

- Relationship between and **MAUP**?



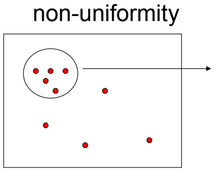

Simpson's Paradox

## Scale

- Besides MAUP, the geographical scale at which we examine a phenomenon can also affect the observations we make and must always be considered prior to spatial analysis
- Multiple Representation Problem
  - Is there an optimal scale?

## Non-uniformity of Space

- Non-uniformity: space is not uniform: First Order SV
    - e.g. population is not evenly distributed.
  - In most case, we are indeed interested in this non-uniformity



Crime locations

## Why are we interested in spatial data then?

- We make an assumption that location matters.
  - Spatial pattern in data is of interest: cluster of crime events
  - Statistical distribution is of interest: different from population at risk?
  - The relationships btw them is what counts: Why? Any other factors?

# Finally, Good News

- Potential insights provided by the examination of locational attributes of data
  - Distance
  - Adjacency
  - Interaction
  - Neighborhood
  - Proximity polygon

# Distance

- Distance between the spatial entities of interest can be calculated with spatial data

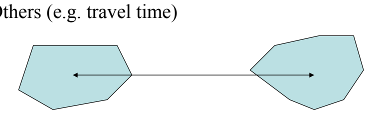


Euclidean distance: Pythagoras Theorem $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$

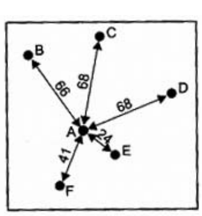
Manhattan distance $d_{ij} = |x_i - x_j| + |y_i - y_j|$

Network distance

Others (e.g. travel time)



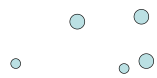
# Distance Matrix



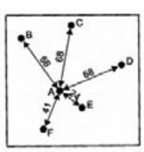

$$\longrightarrow \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0 & 66 & 68 & 68 & 24 & 41 \\ 66 & 0 & 51 & 110 & 99 & 101 \\ 68 & 51 & 0 & 67 & 91 & 116 \\ 68 & 110 & 67 & 0 & 60 & 108 \\ 24 & 99 & 91 & 60 & 0 & 45 \\ 41 & 101 & 116 & 108 & 45 & 0 \end{bmatrix} \end{array}$$

# Adjacency

- Adjacency can be thought of as the nominal, or binary, equivalent of distance. Two spatial entities are either adjacent or not
  - 1 or 0 → adjacency matrix
- Can be defined differently
  - Example 1: two entities are adjacent if they share a common boundary (e.g. Kentucky and Tennessee)
  - Example 2: two entities are adjacent if they are within a specified distance



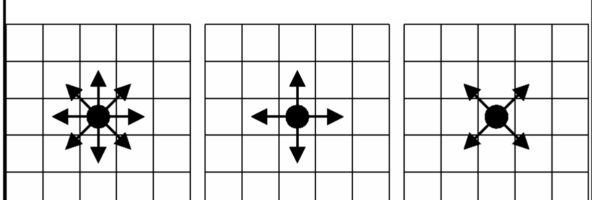
# Adjacency Matrix



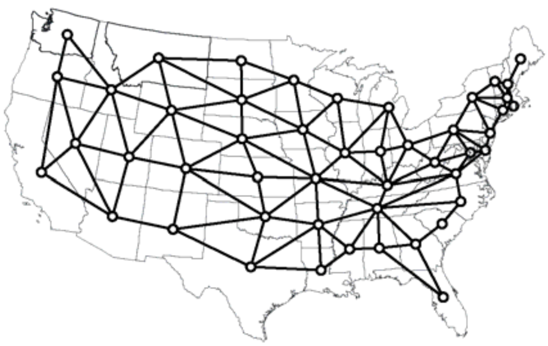

| | | | | | |
|---|---|---|---|---|---|
| 0 | 66 | 68 | 68 | 24 | 41 |
| 66 | 0 | 51 | 110 | 99 | 101 |
| 68 | 51 | 0 | 67 | 91 | 116 |
| 68 | 110 | 67 | 0 | 60 | 108 |
| 24 | 99 | 91 | 60 | 0 | 45 |
| 41 | 101 | 116 | 108 | 45 | 0 |

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | * | 0 | 0 | 0 | 1 | 1 |
| B | 0 | * | 0 | 0 | 0 | 0 |
| C | 0 | 0 | * | 0 | 0 | 0 |
| D | 0 | 0 | 0 | * | 0 | 0 |
| E | 1 | 0 | 0 | 0 | * | 1 |
| F | 1 | 0 | 0 | 0 | 1 | * |

# Queen vs Rook (occasionally bishop)

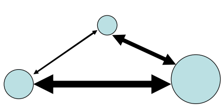


# A Simple Example



# Interaction

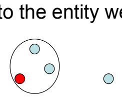
- Interaction may be considered as a combination of distance and adjacency and rests on the intuitively obvious idea that nearer things are "more related" than distant things,
  - *Waldo Tobler's 1st Law of Geography*
- Many geographic structures are indeed results of spatial interaction



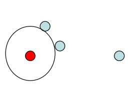
$$\omega_{ij} \propto \frac{p_i p_j}{d^k}$$

# Neighborhood

- Different definitions
  - Example 1: a particular spatial entity as the set of all others adjacent to the entity we are interested in

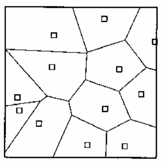


  - Example 2: a region of space associated with that entity and defined by distance from it



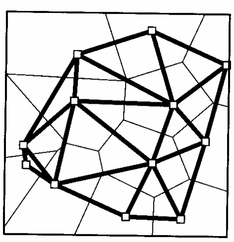
# Proximity Polygon

- The proximity polygon of any entity is that region of the space which is closer to the entity than it is to any other
  - Often called Thession Polygon



*Applications*:
Service area delineation (e.g. schools, hospital, supermarket, etc.)

# Dual Model: Delaunay Triangulation



**Delaunay Triangulation**

Potential applications:
1. TIN model
2. Others

# Review

- Bad News
  - Spatial autocorrelation
  - Modifiable areal unit problem
  - Ecology fallacy
  - Scale
  - Nonuniformity of space
  - Edge effect
- Good News
  - Distance
  - Adjacency
  - Interaction
  - Neighborhood

**Spatial Is Special!!!**