Spatial Statistics
GEOG 419: Lembo

• Global methods to analyze point patterns across entire study region (or a map)
  – Quantitative tools for examining a spatial arrangement of point locations on the landscape
• Two common types of analysis
  – spacing of individual points – nearest neighbor analysis
    • Ex. fire stations locations – random or dispersed
    – Goal: equitable service throughout region
    – Design new configuration (e.g., relocating, new stations)
    – More or less dispersed than original configuration
  – nature of overall point pattern – are locations dispersed or clustered
    • Ex. diseased trees in a national forest
    – Widespread aerial spraying versus concentrated ground treatment

Point Pattern Analysis

Center Point

Euclidean (straight-line) distance
• Total distance from all other points is lowest
Center Point

Mean Center

- mean center – average location of a set of points
  - Center of gravity of point pattern (spatial distribution)
  - average X, Y values
  - equal weights

where:

\[ \bar{X} = \frac{\sum X_i}{n} \quad \text{and} \quad \bar{Y} = \frac{\sum Y_i}{n} \]

Outliers……
- add point (15, 13)
- Average location but…
Mean Center

- geographic “center of population” – point where a rigid map of the country would balance if equal weights (i.e., location of each person) were situated over it.

Weighted mean center

- Unequal weights applied to points
  - Ex. retail store volume, city populations, etc.
  - Weights analogous to frequencies

\[
X_{\text{wm}} = \frac{\sum f_i X_i}{\sum f_i} \quad \text{and} \quad Y_{\text{wm}} = \frac{\sum f_i Y_i}{\sum f_i}
\]

where

- \(X_{\text{wm}}\) = weighted mean center of \(X\)
- \(Y_{\text{wm}}\) = weighted mean center of \(Y\)
- \(f_i\) = frequency (weight) of point \(i\)

### Table 4.4

<table>
<thead>
<tr>
<th>Place</th>
<th>X</th>
<th>Y</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>145</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>175</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>5</td>
<td>189</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>197</td>
</tr>
</tbody>
</table>

\[
X_{\text{wm}} = \frac{145 \cdot 1 + \cdots + 197 \cdot 5}{145 + \cdots + 197} = 3.10
\]

\[
Y_{\text{wm}} = \frac{145 \cdot 1 + \cdots + 197 \cdot 5}{145 + \cdots + 197} = 2.86
\]

Weighted mean center coordinates: (3.10, 2.86)
Spatial measures of dispersion

- **standard distance** – measures the amount of absolute dispersion in a point distribution
  - spatial equivalent to standard deviation
  - calculate Euclidean distance from each point to mean center

\[
S_p = \sqrt{\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n}}
\]

\[
S_d = \sqrt{\frac{\sum X_i^2}{n} - \bar{X}^2} + \left(\frac{\sum Y_i^2}{n} - \bar{Y}^2\right)
\]

---

**Standard distance**

![Standard Distance Diagram](image)

---

**Weighted standard distance**

- Used with weighted mean center

\[
S_{wp} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{w_i}}
\]

\[
S_{wd} = \sqrt{\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{w_i}}
\]

- Difference
  - 1.54 vs. 1.70

---

![Weighted Standard Distance Table](image)

---

© Arthur J. Lembo, Jr.
Salisbury University
**Standard Deviational Ellipse**

- Extends standard distance to include orientation of the point pattern
  - Calculated separately for X and Y
  - Average distance points vary from mean center on X and average distance points vary
    \[
    \sigma_D = \sqrt{\frac{\sum (x - \mu_x)^2}{n}} \quad \text{and} \quad \sigma_D = \sqrt{\frac{\sum (y - \mu_y)^2}{n}} \tag{1.16}
    \]

**Standard Deviational Ellipse**

- Cross of dispersion
- Trigonometric function – angle of rotation
  - Rotated about mean center to minimize distance between both arms and points

**Nearest Neighbor Analysis – (NNA)**

- Distance of each point to its nearest neighbor measured and mean distance for all points is determined
  - Objective: describe the pattern of points in a study region and make inferences about the underlying process
Nearest Neighbor analysis – (NNA)

- Compare calculated value from point data to theoretical point distributions
  - Outcomes: random, clustered, dispersed
  - average nearest neighbor distance is an absolute index
    - Dependent on distance measure (ex. miles, km, meters, etc.)
    - Minimum = 0 (clustered), maximum is function of point density
  - standardized nearest neighbor index (R) is often used
    - Comparison of data to random

\[ R = \frac{\text{NND}}{\text{NND}_E} \]

(NNA)

\[ \text{NND}_E = \frac{1}{2D} \]

\[ \text{Density} = \frac{\text{number of points}}{\text{Area}} \]

\[ R = \frac{2.67}{1.38} \]

(NNA – R values)

- Continuum...
  - Result?
  - Descriptive test.

\[ R = \frac{\text{NND}}{\text{NND}_E} \]

\[ R = \frac{2.67}{1.38} \]
Nearest neighbor analysis (nna)

- A difference test can be used to determine if the observed nearest neighbor index (NNA) differs significantly from the theoretical norm (NNA_R).

\[ Z = \frac{\text{NNA} - \text{NNA}_R}{\sigma_{\text{NNA}}} \]  

where \( \sigma_{\text{NNA}} \) is standard error of the mean nearest neighbor distance.

The standard error for the nearest neighbor test can be estimated with the following formula:

\[ \sigma_{\text{NNA}} = \frac{2n^3}{3(Density)^2} \]  

where \( n \) = number of points, \( Density = \frac{\text{number of points}}{\text{Area}} \).

Nearest neighbor analysis (nna)
Example: Community Services in Toronto

- Emergency services: fire and police
  Seek dispersion to provide services equally.
- Nonemergency services: polling sites and elementary schools
  Seek clustering—why?
Nearest neighbor analysis (nna) Example: Community Services in Toronto

- Result?

Nearest neighbor analysis (nna) Example: Community Services in Toronto

- Issues to consider...
  - Study area boundaries – political boundary or research delimited
    - Doesn’t impact NNA distances but does impact area (point density function)
  - Nearest feature – may be outside study area!
    - Problem with using political boundaries
  - More advanced techniques available – Ripley’s K
    - Evaluates more than one nearest neighbor
    - Can define distances – How many police stations within 1km? 2km?
Geographers are interested in spatial patterns produced by physical or cultural processes.

- Explain patterns of points and areas
  - "global" overall arrangement
    - Random vs. Nonrandom spatial processes
  - "local" concentrations or absences
    - Clusters – points or areas within larger area
      - Groups of high values – "hot spots"
      - Groups of low values – "low spots"

General Issues in Inferential Spatial Statistics

- Compare existing pattern to theoretical pattern
  - Clustered
    - Density of points varies significantly from one part of study area to another
      - Points: retail locations near highway interchange
      - Areas: registered majority political party affiliation
    - Patterns result from nonrandom factors
      - Accessibility, income, race, etc.
  - Dispersed
    - Uniformly distributed across study area
      - Suggests systematic spatial process
      - Area example: Central Place Theory
        - Settlements are uniformly distributed across landscape to best serve needs of a dispersed rural population
Types of Spatial Patterns

- Random
  - No dominant trend toward clustering or dispersion
  - Suggests spatially random process (Poisson)
  - Ex. lightning strikes

- Geographic problems
  - Patterns typically appear as some combination of these three patterns
  - Along continuum...

Random – No dominant trend toward clustering or dispersion

Suggests spatially random process (Poisson)

Ex. lightning strikes

Geographic problems

Patterns typically appear as some combination of these three patterns

Along continuum...

Spatial Autocorrelation

- Tobler’s Law – “Everything is related to everything else but near things are more related than distant things”
- Spatial autocorrelation: measures the degree to which a geographic variable is correlated with itself through space
  - Positive, negative or non-existent
    - Positive spatial autocorrelation: objects near one another tend to be similar
      - Features with high values are near other features with high values, features with medium values are near other features with medium values, etc.
    - Negative spatial autocorrelation: objects near one another tend to have sharply contrasting values
      - Features with high values near features with low values
  - Most geographic phenomena exhibit positive spatial autocorrelation
    - Examples: rainfall amounts, home values, etc.

Variogram

- Visualization of spatial autocorrelation
- Variogram: scatterplot that display the differences in values between geographic locations against the differences in distances between the geographic locations
  - Y-axis: average variance (really half the variance) in values for a set of geographic objects
  - X-axis: distance between objects
  - Use plot to determine average difference in values at specific distances
  - Ex. 100 miles, 500 miles

Geographic locations near one another tend to have smaller differences than geographic locations at greater distances (positive autocorrelation)!
Variogram

- Displayed as best-fitting curve (function)
  - Differences in values with distance noted and then diminishes
    - range – distance at which the difference in values are no longer correlated
    - sill – average difference in value where there is no relationship between location and value
    - nugget – degree of uncertainty when measuring values for geographic locations that are very close to each other
      - Effect of sampling, measurement error, etc.
      - Unlikely that two samples near each other will have the exact same value
- Variation becomes less similar with distance

Variogram Example: Last Spring Frost IN SE United States

- Two nearby stations, LSF dates should be similar
  - 0 to 400 miles: distances between stations are large, dates are different
  - Beyond 400 miles, no longer spatially autocorrelated

Spatial Autocorrelation: Importance in Geographic Research

- GIS – push of a button
  - Calculates relationship for any distances...
  - Is the test appropriate for any distance?
- Presence of spatial autocorrelation
  - Inferential statistics assume independent observations
  - Impact: sample locations close together, just like taking the same sample
    - Sample size impacts size of standard error
    - Smaller standard error than warranted
    - Standard deviation calculation impacted
      - Even smaller standard error
- Global or local measurement
  - global – examine a distribution of subset (ex. ethnic group) across entire area (ex. city)
    - One group more clustered, dispersed or random than another
  - local – compares each geographic object (ex. all group members) with its surrounding neighbors
    - Is area (ex. neighborhood) more clustered, dispersed or random than another
Spatial Autocorrelation: Neighbor Definitions

- Measure of interaction between geographic features
  - Defining neighbor:
    - adjacency: share common border
      - Binary: yes or no
    - distance threshold: cut-off distance
      - Salisbury, MD – neighbor definition 60 miles...Easton, Wilmington, DE
    - inverse-distance: strength of “neighborliness” between two objects as a function of distance separating them
      - 1/distance
      - New York City and Boston: 1/189 miles or .005
      - NYC and LA: 1/2588 miles or .0004
      - Interaction measure (“neighborliness”) is 12 times stronger between NYC and Boston versus NYC and LA
  - In equations/modeling, takes the form of weights
    - \( w_{ij} \): weight between geographic object \( i \) and \( j \)
      - Binary: 0 or 1
      - Inverse-distance: continuous value ...

Spatial Autocorrelation

- First law of geography: “everything is related to everything else, but near things are more related than distant things” – Waldo Tobler
- Many geographers would say “I don’t understand spatial autocorrelation” Actually, they don’t understand the mechanics, they do understand the concept.

Spatial Autocorrelation

- Spatial Autocorrelation – correlation of a variable with itself through space.
  - If there is any systematic pattern in the spatial distribution of a variable, it is said to be spatially autocorrelated
  - If nearby or neighboring areas are more alike, this is positive spatial autocorrelation
  - Negative autocorrelation describes patterns in which neighboring areas are unlike
  - Random patterns exhibit no spatial autocorrelation
Why spatial autocorrelation is important

• Most statistics are based on the assumption that the values of observations in each sample are independent of one another.
• Positive spatial autocorrelation may violate this, if the samples were taken from nearby areas.
• Goals of spatial autocorrelation
  – Measure the strength of spatial autocorrelation in a map.
  – Test the assumption of independence or randomness.

Spatial Autocorrelation

• Spatial Autocorrelation is, conceptually as well as empirically, the two-dimensional equivalent of redundancy.
• It measures the extent to which the occurrence of an event in an areal unit constrains, or makes more probable, the occurrence of an event in a neighboring areal unit.

Spatial Autocorrelation

• Non-spatial independence suggests many statistical tools and inferences are inappropriate.
  – Correlation coefficients or ordinary least squares regressions (OLS) to predict a consequence assumes that the observations have been selected randomly.
  – If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise.
  – They are biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed.
Indices of Spatial Autocorrelation

- Moran’s I
- Geary’s C
- Ripley’s K

Moran’s I Index (Global)

- Popular technique for quantifying level of spatial autocorrelation in a set of geographic areas
- Moran’s I Index takes into account geographic locations (points or areas) as well as attribute values (ordinal or interval/ratio) to determine if areas are clustered, randomly located or dispersed
  - Positive: clustered – nearby locations have similar attribute values
  - Negative: dispersed – nearby locations have dissimilar attribute values
  - Near zero: attribute values are randomly dispersed throughout study area

Moran’s I Index (Global) Weighted cross-products: deviation values for contiguous pairs multiplied together and summed

- Positive: neighboring areas with similar attribute values either large or small (clustered)
  - Larger deviation from mean, greater magnitude
- Negative: neighboring areas with dissimilar attribute values contiguous (dispersed)
  - Larger deviation from mean, greater magnitude
- Near zero: random…

- I ranges from -1.00 to 1.00
Moran’s I Index (Global):

**Significance test**

The null hypothesis is that the area values are arranged in a completely random spatial pattern, and that spatial autocorrelation is not present within the study area. The test is computed as:

\[ \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

where: \( x_i \) = value for the i-th area

Similar to the first and non-stationary sampling techniques mentioned with point count analysis and binary data, the variance or formula I may be calculated under the assumption of a second-order spatial dependence or second-order spatial dependence as in the variance of a variable given a normally distributed population, where the standard deviation is in a second-order spatial dependence.

- **H\(_0\)**: No spatial autocorrelation in the data (Values of areas are completely random)
- **H\(_1\)**: Spatial autocorrelation in the data (Values of areas are not completely random)

If the p-value is not significant, then you should not reject the null hypothesis.

- The observed pattern is not different from complete spatial randomness.

- p-value significant and Z-score positive
  - clustering

- p-value significant and Z-score negative
  - dispersed

Result?

**Example: Cleveland Census Block Groups**
Moran’s $I$ Index (Global)

**Moran’s Index**

- **Primary Objective:** Identify significant spatial patterns within a study area.

- **Requirements and Assumptions:**
  1. Minimum of 20 geographic features.
  2. Attribute values measured on an ordinal or interval/ratio scale.

- **Hypotheses:**
  - $H_0$: Attribute values are randomly distributed across features in the study area.
  - $H_1$: Attribute values are not randomly distributed across features in the study area.

- **Test Statistic:**
  
  \[ I = \frac{2C(x) - P(x)}{P(x)(n-1)} \]

- **Interpretation:**
  - $I < 0$: (observed pattern is dispersed)
  - $I = 0$: (observed pattern is random)
  - $I > 0$: (observed pattern is clustered)
Global spatial autocorrelation (Moran’s I) may indicate a lack of spatial autocorrelation.

- Local pockets may exist – hotspots
- LISA – Local Indicators of Spatial Association
  - Quantify similarity of each geographic observation with an identified group of geographic neighbors
  - Identifies local clusters – geographic locations where adjacent or nearby areas have similar values
  - Spatial outliers – geographic locations that are different from adjacent or nearby areas
- Each geographic area receives individual measure

Moran’s I Index (local): Example: Obesity in PA

\[
I = \frac{\sum_{i,j} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i,j} w_{ij}}
\]

where: 
- \( x_i \) = the value for a particular geographic entity
- \( x_j \) = the value for the neighboring geographic entity
- \( \bar{x} \) = the average of all attributes
- \( w_{ij} \) = the spatial weight.

The overall computation of the local Moran index is beyond the scope of this book, but a short example will help illustrate its usefulness.

- Global Moran’s I = .69, p-value = .25
- Local Moran’s I for each county...

Positive values: similar levels in adjacent counties (clustering)

Example of Moran’s I – Per Capita Income in Monroe County

Using Polygons:
Moran’s I: 66
P: < .001

Using Points:
E: 12
Z: 65
Example of Moran’s I –
Random Variable

Using Polygons:
Moran’s I: .012
p: .515
Using Points:
Moran’s I: .0091
Z: 1.36