Assignment 14. Spatial Autocorrelation


Conceptual Issues & Basic Measures

- Always check for possible data errors (e.g., sum “Race,” gender, and housing tenure categories to make sure that they amount to the appropriate totals).
- “Spatial statistics let you compare the spatial distribution of a set of features to a hypothetical random spatial distribution...” (Mitchell, 19). Alternatively, they compare local empirical distributions to average empirical distributions across the study area.
  - Null hypothesis: the features are evenly distributed across the study area.
  - “To the extent that your distribution differs from a random distribution, there is a trend or pattern in the data” (Mitchell, 19). Alternative hypothesis: there is a trend or pattern in the distribution of the features across the study area.
- A pattern may vary according to the geographic scale of the study area and its borders:
  - Dispersed within a small study area but clustered within a larger area, or vice versa; or it may vary in other ways.
  - Study area borders can bias results; perhaps use border buffer to exclude border points from analysis or give higher weight to border points (which tend to have fewer neighbors).
- How the data are represented influences the results of spatial statistics (Mitchell, pages 184-89):
  - Distance measures are straightforward for point data but not for line data (e.g., is distance measured at the midpoint, endpoint, or a randomly selected point on a line? is the line continuous or a series of short lines?).
  - For areas, centroid-to-centroid distance or distance between nearest locations on the boundaries are used, but centroids should be used only if areas are roughly the same size and shape and merged areas will alter distance calculations.
  - For raster data, smaller cell size creates many features (cells) with the same or very similar sizes and hence exaggerates spatial dependence.
  - Buffering a boundary area and excluding observations within it may compensate for the tendency of boundary observations to have fewer neighbors. Giving boundary observations higher weighting may also do the same.

1

- Measuring the pattern formed by the locations of features: quadrant analysis & nearest neighbor
- Quadrant analysis: assess a feature’s spatial distribution by overlaying a grid and measuring the density of features (i.e. the number of features per unit
area). E.g., for small tight clusters of events such as burglaries, emergency 911 calls, earthquakes, but not for features that have direct interaction with each other.

- Quadrant analysis does not consider the proximity or relationship among features.
- Geographic scale can influence results.

**Nearest neighbor index:** "measures the distance between each feature and its nearest neighbor, then calculates the average. Distributions that have a smaller average distance between features than a random distribution are considered clustered" (Mitchell, page 80). E.g., for analyzing pattern of incidents of HIV/AIDS.

- *K-function:* if the “average number of features found at a distance is greater than the average number of features throughout the study area, the distribution is considered clustered at that distance” (Mitchell, pages 80-81).
- Using the k-function "can provide a more accurate measure of nearest neighbor because features close to each other or at the same location can skew the mean distance to be smaller than it otherwise would be” (Mitchell, page 95).
- But the geographic scale can also matter because it is the basis of calculating the expected mean distance.
- Points near a study area’s edge may have fewer neighbors because some of them may be outside the study area.

II

**Measuring whether similar values are clustered or dispersed—spatial autocorrelation**

- What does correlation tell us and not tell us? What data characteristics can bias its results? How do we assess the data’s characteristics?

**Global Geary’s c or Global Moran’s I (the preferred option):** “compare values for neighboring features and then compare the differences in values between each pair of neighbors to the difference in values between all features in the study area” (Mitchell, page 118).

- Based on Tobler’s First Law of Geography.
- “Specify a neighborhood based on adjacency, a set distance, or the distance to all features in the dataset” (Mitchell, page 118).
- For Moran’s I, zero coefficient indicates random distribution of values in space; positive coefficient (up to 1) indicates that more pairs have similar values than not (i.e. similar values are clustered in space); and negative coefficient (down to −1) indicates that more pairs have dissimilar values than not (i.e. similar values are dispersed in space).
- Global methods calculate a single statistic that summarizes the pattern for a study area.

**Vulnerabilities:**

- Results depend on the geographic scale and are influenced by borders (Mitchell, page 184).
- Testing a small number of features (say, less than 30) makes the results especially vulnerable to outliers (Mitchell, page 174).
- Because Geary’s c and Moran’s I conduct so many hypothesis tests, a Bonferroni or other multiple-test adjustment should be
used (Mitchell, page 184) (but see Fotheringham et al., pages 167 and 249-50, on exploratory versus conservative approaches to t-values).

- **Local versions** of Geary’s c and Moran’s I:
  - Local versions examine how individual features contribute to a global pattern and identify spatial anomalies and outliers; they identify “hot” and “cold” spots.
  - That is, given a global pattern of spatial autocorrelation, what local features account for the global pattern?
  - E.g., Local Moran’s I compares each value in a local pair to the mean value for all the features in the study area (Mitchell, pages 165-66).
  - Spatial weighting (see below): weighting based on adjacent features (using row-standardized weighting) should be used (Mitchell, page 166).
  - Vulnerabilities: the same as mentioned above.

- **General G-statistic**: measures the concentration of high or low values (Mitchell, pages 127-32).
  - Define a neighborhood based on distance.
  - Higher than expected coefficient if there’s concentration, lower than expected if there’s dispersion.
  - Vulnerabilities:
    - Geographic scale and size of the features being analyzed can affect the results, as can the range of values of the features.
    - Testing fewer than about 30 features makes the results vulnerable to outliers.
    - Use a Bonferroni or other multiple-test adjustment.

- **Local G-statistic**: see Mitchell, pages 175-80.
  - Spatial weighting (see below): use either adjacent features or inverse distance (either Euclidean or, e.g., travel time).
  - Vulnerabilities: as discussed above.

### III

- **Defining spatial neighborhoods & weights:**
  - **Neighborhood**: to compare the value of a target’s features to the values of neighboring features, GIS requires that we “define the area surrounding each target feature within which feature values are compared” (Mitchell, page 135).
  - **Spatial weights**: weights must be assigned “to each feature pair to specify whether the two features are in each other’s neighborhoods, and to represent the spatial relationship between the features” (Mitchell, page 136).

- **How to define a neighborhood**: do so based on how features interact with each other.
  - E.g., do closer values (e.g., property prices, crime rates, disease rates, voting patterns) influence each other? Does movement between features (e.g., daily travel, migration, social linkages, trade/investment) cause the values of nearby features to be similar? Or do common structural or compositional conditions (e.g., poverty/wealth, rural/urban, ethnic composition) cause features in nearby zones to be similar?
Most common representation of such spatial relationships:
  o Adjacencies: contiguous areas or rasters that share a border.
    ▪ 1=adjacent, 0=not adjacent.
    ▪ Shared borders only: “Rook” contiguity.
      • Implies that direct interaction across borders matters most (e.g., disease contagion) (Mitchell, page 141).
      • Length of shared border may matter.
    ▪ Shared borders and corners: “Queen” contiguity.
      • Implies that proximity matters more than direct interaction (e.g., property values) (Mitchell, page 141).
      • Length of shared border may matter.
  o Euclidean (straight-line) distance: for discrete or contiguous features.
    ▪ E.g., less than 1/4 mile=1 (adjacent), more than 1/4 mile=0 (not adjacent) (e.g., influence of property values).
    ▪ Or specify rate of distance decay (using inverse distance so that closer distance has greater influence).
  Other options:
    o E.g., length of border that is shared.
    o E.g., travel time, travel cost, travel frequency, frequency of social interaction, importance of social interaction, magnitude of trade, magnitude of investment, frequency of conflicts, magnitude of conflicts.
    o Interaction of length or percentage of border that is shared with disease rate.
  Is the influence local or regional? First-order vs. higher-order features:
    o If there’s minimal influence by surrounding values or influence is nearby only, or the phenomenon is very localized, use an adjacency measure (Mitchell, page 138).
    o If distant features are influential, use a distance measure (e.g., influence decreasing with greater distance, using inverse distance).
      ▪ Point-to-point distance.
      ▪ Distance between centroids: if the features are approximately the same size and shape.
      ▪ Distance between nearest location on a line or boundary: if the size or shape of the features varies greatly.
  Row standardized weighting: typically used with “adjacency-based neighborhoods to create proportional weights when features have unequal numbers of neighbors” (Mitchell, page 144).

“When using inverse distance weight with discrete features, the results can be distorted if there are features that are very close or at the same location” (Mitchell, page 145), because close features will have a comparatively very large weight.
  o One solution: add 1 to the value of all distances before calculating the inverse.
IV

- Different GIS softwares commonly give significantly different spatial statistical results? **This is a major problem in the field today, so beware.**

- The combination of graphical and statistical display is essential. Why is this so? Consider also presenting more than one statistical indicator (along with more than one kind of graphical display), or a particular statistical indicator with different weighting schemes.

- Pay close attention to the possible consequences of outlier data.

V

Exercise: Data Exploration in GeoDa

- Refer to Anselin, “Exploratory Spatial Data Analysis in GeoDa” and “An Introduction to Spatial Autocorrelation Analysis with GeoDa”; and to Mitchell, vol. 2, chapter 1.

- Open GeoDa. Open the dataset *sids2* (key variable: FIPSNO).

- Click **Explore**:
  - Useful commands
    - Map (on the toolbar) > Map Reset.
    - Right-click area outside map boundaries to eliminate last data selection.
  - *Histogram* SIDR74; click graph data display to link to map display. Right-click graph and increase the number of intervals to provide a different display. Describe the characteristics of the distribution (overall pattern and striking deviations—i.e., shape, center, spread, including outliers).
  - *Boxplot* SIDR74. What does it reveal about the distribution? Click extreme outliers to link to map display.

- Click **Map (on the toolbar)**
  - Note: to display and compare multiple maps at the same time - Edit->New Map, right-clicking the new map and selecting a map type, doing so for each type of map that you want to display.
  - *Box Map* SIDR74. Link to map.
    - Why use a box map?
  - *Standard Deviation Map* SIDR74. Link to map.
    - Why use a standard deviation map?
  - *Percentile Map* SIDR74. Link to map.
    - Why use a percentile map? Why not use it?
  - *Quantile Map* SIDR74. Link to map.
    - Why use a quantile map? Why not use it?
  - *Cartogram* SIDR79. Link to map.
    - Why use a cartogram?

- What are your conclusions about the data distribution? What are the implications for how to manipulate the data?
GeoDa notes:
- If your file is in the form of points (which is appropriate for distance weighting; see below) but you want to use contiguity weighting:
  - Tools>Shape>Points to Polygons.
- See ‘Tools’ and ‘Options’ on the toolbar for various procedures (e.g., range selection, field calculations, adding/deleting a column, joining a table).
- Use ‘range selection’ to exclude outliers.
- Window>Tile Vertical or Tile Horizontal can allow the simultaneous display of multiple maps or graphs at the same time.

Exercise: Constructing Spatial Weights in GeoDa

- We will create contiguity weights (Rook, Queen) for sids2, which in GeoDa require projected coordinates.
  - Use ArcCatalog to check whether sids2’s current coordinates are projected.
  - If they are not projected, then use ArcCatalog to select projected coordinates (e.g.: Projected Coordinate Systems>State Plane>NAD 1983 State Plane North Carolina FIPS).
- Click Tools>Weights>Create
  - Input file: sids2. ID variable: FIPSNO. Save output as: sids2_rook (perhaps resaving a duplicate copy in a folder outside of GeoDa for ArcGIS use).
  - Create.
    - Note: use Euclidean Distance for projected data only, ArcDistance for unprojected data.
- Click Tools>Weights>Properties>Characteristics>sids2_rook>OK.
  - Enlarge>widen the histogram, which displays # features (bins) by # neighbor connections (see right-hand legend). Click each bin to link to map. Right-click graph to increase number of intervals to provide a different display.
    - Objectives: 1) inspect the spatial pattern of connectivity; and 2) if it makes substantive/theoretical sense, try to eliminate any ‘islands’ (i.e. features with zero connections).
    - An island has no spatial relationship with other features and thus should not be analyzed spatially.
    - If it makes substantive/theoretical sense, eliminate islands by either using distance weights or by editing contiguity weights.
- Repeat the above steps to create Queen weights for sids2 (sids2_queen).
- Compare the spatial pattern of connectivity for Rook weights and Queen weights.
  - Tools>Weights>Properties>sids2_rook [Rook histogram appears]
  - Tools>Weights>Properties>sids2_queen [Queen histogram appears]
  - Edit>Duplicate Map
  - Window>Tile>Vertical (to make an orderly display of both histograms and a map).
  - Compare histograms and click bins of each histogram to explore mapped pattern of connectivity.
    - Which type of weighting has the densest connectivity and why?
Exercise: Global & Local Morans Statistics in GeoDa

- **Important:** use *Univariate Morans for a raw measure* (such as SID74) but use *Morans I with EB rate for rate measure* (such as SIDR74) (see Anselin).
- Click *Space>Univariate Morans with EB Rate* (clicking SID74 as *Event Variable* and BR74 as *Base Variable*; this combination=SIDR74). OK.
  - Select weight: sids2_rook. OK.
  - Options>Randomization>999 Permutations> Run (to obtain simulated significance test).
  - Envelope on (to examine highlight outlier coefficients).
  - Exclude selected (using clicking or rectangular ‘brush’ to exclude one or more outlying observations).
    - Perhaps also use the toolbar’s Options>Range Selection to exclude one or more outlying observations.
- Repeat the above steps with Queen weighting, comparing the results to those of Rook weighting.
- Repeat the above steps for Univariate Local Morans I (*LISA*) with Rook and Queen weighting, comparing and saving the results.
  - First, save the results to the spreadsheet: Toolbar’s Options>Save Results>Save LISA Results (click the first three map options, inserting ‘R’ after each default filename for Rook weighting and inserting ‘Q’ after each default filename for Queen weighting).
  - Second, save the spreadsheet as a shapefile: Toolbar’s Options>Save to Shapefile as sids2_lisa, perhaps saving a duplicate copy to a folder outside GeoDa for ArcGIS use.
    - Note: to make the duplicate copy or to make a change in this shapefile (e.g., sids2_lisa2).
  - Do the following LISA-specific procedures:
    - Cluster map.
      - Note: *low-high and high-low clusters are spatial outliers*.
    - Box plot map (refers to LISA results).
    - Significance map.
    - Significance filter (setting to more stringent thresholds to simulate Bonferroni-type adjustments, but see Fotheringham et al. on exploratory versus conservative approaches).
  - Window>Tile Vertical or Tile Horizontal
- Repeat the relevant steps for Multivariate Moran, which is useful for examining change in a variable at two different points in time. Use SIDR79 (x-axis) and SIDR74 (y-axis), saving the results (see Anselin).
- Repeat the relevant steps for Multivariate LISA, saving the results (sids2_lisa).
- Open ArcGIS>ArcCatalog.
  - Import sids2_lisa; examine in ArcCatalog; set ArcMap’s coordinate system appropriately; set map properties to store relative path names; check the map display units.
  - **Add data:** sids2_lisa; and map the coefficients and significance values in ArcGIS.
**Final Questions:** What have you learned about the meaning of weighting matrices? What do indices of spatial autocorrelation tell us and not tell us? Why is it essential to combine graphic and statistical display, and to display variants of each? What problems might you have to address with regard to, e.g., Percent Rental Occupancy spatial autocorrelation for the Miami-Dade data?