ACM Analytics, LLP: Informational Paper

The Relevance of Neighborhood (Spillover) Effects on Data Analysis

Paulo Macedo, Ph.D., Founder and Chief Analytics Officer
www.acmanalytic.com

Spatial Data

The analysis of spatial data presumes the availability of geographic information associated with summary numbers of the phenomenon being analyzed such as incidents of measles in a geographic area. The relative position of the spatial units has a known location structure that is considered in the analysis – be areal units such as counties or geographic points of interest.

When geographical information is available, it is interesting to investigate if the distribution of the data displays a spatial pattern. The spatial distribution of the data may or may not follow a random pattern. Non-randomness may be due to the presence of either positive or negative spatial autocorrelation. Positive autocorrelation has clusters (e.g; cluster of counties) distributed in either a high-high ("hot-spot") or low-low ("cold spot") geographic layout. For example, a high concentration of crimes in neighboring counties characterizes a "hot-spot." Negative autocorrelation has the units of analysis displaying values in significant contrast to the neighboring units characterizing a high-low or low-high spatial arrangement; for example, if measles outbreak occurs in isolated counties only but not in neighboring counties.

According to Luc Anselin (1992), "*An attention to location, spatial interaction, spatial structure and spatial processes lies at the heart of research in several sub-disciplines in the social sciences. Empirical studies in these fields routinely employ data for which locational attributes (the where) are an important source of information. Such data typically consist of one or a few cross-sections of observations for either micro-units, such as households, store sites, settlements, or for aggregate spatial units, such as electoral districts, counties, states or even countries. Observations such as these, for which the absolute location and/or relative positioning (spatial arrangement) is taken into account are referred to as spatial data.*"[1]

Connectivity in Space

The geographic neighborhood effects that may be associated with social-economic (or other types of) data are not as straightforward as the intuitive neighborhood effects related to time-series data. For example, the growth in the gross domestic product (GDP) of the country in the current year will most likely include spillover effects of the GDP growth in the previous year, unless a sharp trend break happens precisely at the end of the previous year. Time-series data are recorded in well-established periods of time and people usually do not think about the arbitrariness implicit in the choice of those standard time units. The arrow-of-time nature of time-series data implies that autocorrelation is unidirectional. On the other hand, spatial data is not unidirectional; for example, if groups of counties have high measles outbreak, the spread of the disease may go back and forth among the counties.

[1] *Spatial Data Analysis with GIS: An Introduction to Application in the Social Science*. Luc Anselin (1992), Technical Report 82-10, University of California, Santa Barbara, http://www.ncgia.ucsb.edu/technical-reports/PDF/92-10.pdf

In the case of areal units such as counties the relevant neighborhood can be defined as contiguous units and those units sharing a common border with the reference areal unit. The information of sharing or not a border is summarized by a binary indicator. This notion of contiguity assumes the existence of a map from which the boundaries can be discerned.

In the case of geographical points such as housing units, the idea of relevant neighborhood is frequently associated with the distance between the data points. A standard approach is to define a threshold distance allowing the classification of every unit within the threshold distance as neighboring with respect to the reference observation. Another method often used ranks neighborhood effects as related to the inverse distance between the data points – for any reference unit the strength of those effects will then be inversely proportional to the distance from the other units in the data. At least two sources of autocorrelation across spatial units are worth mentioning:

- **The administrative boundaries:** administrative boundaries (country boundaries, counties, zip-code tabulation areas) often do not contain the phenomena summarized by their areal aggregate measures. For example, space plays a significant role in international trade flows (phenomenon) as transportation costs increase with distance that traded goods must travel (geographic information). A country that undergoes sustained economic growth strengthens the trade flows with neighboring countries as a result positively affecting their economic performance.
- **The existence of spatial processes with contagion:** these are processes in which entities influence each other by contagion – such as the adoption of similar policies by neighbors in coping with the spread of infectious diseases. Counties that experience significant increase in the incidence of measles are likely to adopt the practices of the most successful neighboring counties in dealing with the disease outbreak.

Observed spatial autocorrelation may occur by a combination of these two sources of neighborhood effects making the identification of their relative importance a task worth pursuing.

Spatial Data Structures

- *Points*: The neighborhood effects can have their strength proportional to the inverse of the distance between points.
- *Polygons*: The neighborhood effects can be represented by the clusters of areal units sharing common borders and having close measurements in a targeted metric.
- *Grids:* The neighborhood effects can be represented by the clusters of square cells in a grid, referenced by their row and column numbers, and having close measurements in a targeted metric.

Examples of Spatial Statistics Applications

- Modeling:
  - The spread of contagious diseases.
  - The adoption of an innovation.
  - The atmospheric/oceanic phenomena.
- Identification of crime spots.
- Predicting the housing prices.

An Example: Neighborhood Effects Analysis on House Prices

The following analysis uses a dataset containing 60 housing units from a number of zip-code areas in the city of Alexandria, Virginia including information on sale prices and a limited set of house attributes – square footage of the unit, the number of bathrooms (half-bathrooms are assigned value 0.5) and an indicator whether the house is detached or not. The following two questions will be addressed:

- Is the limited set of attributes enough for a standard multiple linear regression to do a good job in finding a statistically significant pattern between the prices of the houses and their attributes?

- Does the distribution of the data display a pattern of neighborhood effects so that the specification of a spatial lag model may improve the predictive power of the analysis by explicitly considering the spatial autocorrelation across the units?

Despite the small set of attributes utilized in this example, the standard regression method validated the assumption that they are all correlated with the price of the units – in this case 64% of the price variation across the housing units could be ascribed to those attributes. The existence of neighborhood effects requires a threshold distance that characterizes the relevant neighborhood for the analysis. This identification of the threshold distance was done on a what-if basis by setting several threshold distances to classify every unit within the threshold distances as neighboring unit with respect to the reference observation. The threshold distances were tested to identify the existence of a non-random spatial pattern. The results indicated that:

- When the neighborhood boundary was set to a threshold distance of 2.18 miles the spatial lag model identified significant neighborhood effects of the type high-low (or low-high) – in which adjacent observations tend to have very contrasting values with respect to the reference observation – negative spatial autocorrelation.

- The spatial lag model (SLM) displayed a better predictive power than the standard regression model (SRM) as reported by a number of indicators, for example, the average difference between predicted and actual prices of the units and the standard deviation of these differences was lower for SLM and the SLM predicted prices were closer to the actual price in 62% of the units in the data.

Summary Statistics of the Data

- ✓ *Average price: $930,447.50*
- ✓ *Maximum price: $1,680,000.00*
- ✓ *Average (geodesic) distance: 2.79 miles*
- ✓ *Max distance: 9.45 miles*

Data Limitation: What If More Data Points Were Available?

In what conditions would the analysis above lead to the identification positive autocorrelation? Analyzing a larger number of units, for example 500 instead of 60 data points would probably lead to the identification of clusters of similar values characterizing the neighborhood effects as positive autocorrelation.