3

This chapter outlines conceptual and mathematical underpinnings of local indicators of spatial association, or LISA statistics. An empirical application shows how LISAs can be used to detect localized spatial processes impacting students and campuses.

Higher Education Hot Spots: Using Local Indicators of Spatial Association

Austin Lyke

Quantifying the all-encompassing nature of space and its reciprocal relationship with social processes often involves collection and manipulation of massive amounts of data. Applied researchers turn to geographic information systems (GIS) and other spatial analysis techniques capable of leveraging big data to investigate of hypotheses and research questions with scalar components. For institutional researchers, many of the motivating questions driving use of big data stem from a desire to better understand micro-level processes on and between college campuses (Rios-Aguilar, 2015). Analysis of geographic or spatial interactions on campus and among college students using big data therefore may in fact be *small* in scale. How then can higher education researchers use big data to answer small questions about socio-spatial processes affecting students and campuses?

One technique that can provide college and university researchers with descriptive information about local spatial patterns of students and institutional processes existing within larger geographic networks is the aptly named local indicator of spatial association, or LISA statistic (1995). Applied researchers often use LISAs as a means of identifying hot spots, cold spots, and outliers of various socio-spatial phenomena (Ord & Getis, 1995). For higher education institutions, LISAs offer campuses a unique lens into how students interact with each other and with place-based structures without many of the data collection hurdles faced with qualitative and survey research methods. The following chapter is a practical introduction and guide for institutional researchers that situates LISAs in the broader GIS toolkit, outlines their mathematical and conceptual underpinnings, and provides an example of an empirical application using administrative data from a college in California. Future directions for application of LISAs in higher education and institutional research are discussed throughout.

New Directions for Institutional Research, no. 180 © 2019 Wiley Periodicals, Inc.
Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.20286

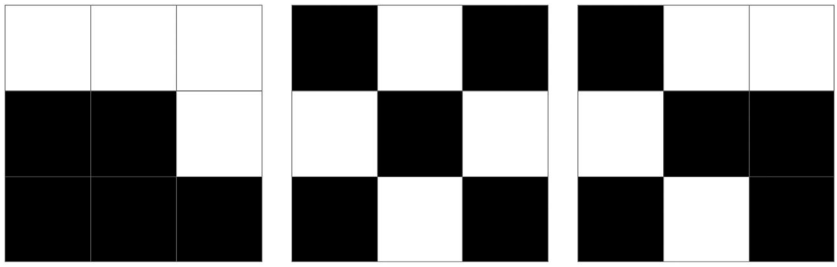
Indicators of Spatial Dependence

What Is Spatial Dependence? Tobler's (1970) famous "first law of geography" dictates that everything is related to everything else, a notion which complicates assumptions of many predictive models in social science research. What affects school graduation rates, for example, may in some way depend on graduation rates of neighboring schools. Graduation rates or other outcome variables are said to be *spatially dependent* or exhibit *spatial autocorrelation* when a variable at a focal location is jointly determined by that of its neighbors in a defined geographic region (Anselin & Rey, 1991). Another form of spatial autocorrelation stemming from diffusion or shocks across observations/spatial regions can affect error terms in a regression context, which impacts model efficiency if left unaddressed (LeSage & Pace, 2010). The first step for researchers responding to questions with spatial dimensions, therefore, is determining whether units of analysis are in fact ordered according to a distance-based pattern. In institutional research, for example, students living off-campus found to have lower GPAs may be subject to influence from fellow students with lower GPAs living nearby. As not accounting for those dependencies affecting GPA or other variables would yield biased (spatial dependence among dependent variables) and/or inefficient (spatial dependence in error terms) parameter estimates in a regression model, researchers can rely on various diagnostic tests to identify spatial autocorrelation prior to estimation of more sophisticated models.

Global Moran's *I* and Local Moran's I_i. One of the most common measures of spatial autocorrelation is the Moran's *I* statistic, which is used to test a null hypothesis of spatial independence of a given variable across an entire researcher-defined study space (Cliff & Ord, 1981). Ranging between -1 and 1, a Moran's *I* statistic that fails to reject the null hypothesis indicates that the variable tested among units in a specified neighborhood is randomly dispersed. Positive spatial autocorrelation, as determined from a positive Moran's *I* and a significant test statistic, indicates that the tested variable at a focal unit is dependent on those in a surrounding spatial region. Alternatively, negative spatial autocorrelation results from the tested variable being arranged in a checkerboard pattern (i.e., alternating high values near low values). Figure 3.1 depicts these relationships on a grid. In higher education research, scholars have identified positive spatial dependence in tuition variables (González Canché, 2014) and college spending (Titus, Vamosiu, & McClure, 2017) at the institution level using a series of Moran's *I* tests. In both cases, spatial regression models were used to adjust for bias and efficiency of parameter estimates.

Spatial dependence is initially teased out through definition of a positive $n \times n$ spatial weights matrix, *W*, where row-by-column elements, w_{ij}, represent the strength of the relationship between two units of analysis, *i* and *j*, and diagonals are zero. Researchers define this matrix ex ante, typically according to radial distance, a contiguity-based pattern, or a

Figure 3.1. Types of spatial autocorrelation



(a) Positive Autocorrelation (b) Negative Autocorrelation (c) No Autocorrelation

near-neighbor algorithm. Returning to the example above, this could mean measuring the influence of GPA at k nearest j addresses on i or addresses within d distance of i (encompassing j neighbors). The Moran's I statistic then is a function of the divergence of a given variable at each spatially weighted individual unit from that of the mean across the entire region under investigation, that is, a global indicator of spatial association. Another commonly used indicator of spatial association is the Geary's C, which is similar but distinct from Moran's I (see Sokal & Oden, 1978). Seeking to allow for localized instability in spatial patterns, Aneslin (1995) proposed derivation of LISA statistics from global indicators:

$$\sum_i L_i = \gamma\Lambda. \tag{3.1}$$

where Λ is a global indicator of spatial autocorrelation (e.g., Moran's I) and γ is a scalar parameter, such that the sum of individual LISAs, L_i, is proportional to Λ.

This chapter focuses on the LISA most widely used in applied research, the local Moran's I_i statistic (denoted I_i to distinguish from the global indicator, I). For more technical details see Anselin (1995) and Ord and Getis (1995). As with the global Moran's I, a positive I_i value indicates that an individual unit is surrounded by similar units and thus a potentially significant cluster amid the global study sample, that is, high values surrounded by high values, low values surrounded by low values. A negative I_i value, alternatively, points to an individual that is surrounded by unlike units, that is, high values amid a neighborhood of low values and vice versa. More detailed steps for computation of I_i are outlined below.

Computation. When analyzing spatial data, researchers must first identify units to be investigated. In higher education, those will typically be point-based units like student addresses, non-contiguous locations on a single campus, or locations across multiple institutions. Alternatively, researchers might also be working with grouped aggregates of individual

points or polygons like counties, Census tracts, or other administrative borders. The next stage involves referencing those point-based or grouped units with geographic coordinates, a process called *geocoding*. Some commonly used data sources include geographic information, such as the National Center for Education Statistics' Integrated Postsecondary Education Data System (IPEDS). There, researchers can find longitude and latitude coordinates for most higher education institutions (coordinates will represent a central location on campus, such as the main administration building). Geocoding other addresses can be accomplished through a number of free sources such as the Census Bureau for locations in the United States (https://www.census.gov/geo/maps-data/data/geocoder.html) or through GIS and statistical programs. Table 3.1 shows an example of geocoded institution-level data obtained from IPEDS. Once in possession of geocoded data, researchers can move forward with identification of global and local spatial dependence.

Exact steps for processing geocoded data for analysis will vary based on software. Commonly used GIS software includes ArcGIS, GeoDa, and QGIS in addition to the considerable spatial functionality of *R* and Python, all of which have a variety of resources for identification of global and local spatial association and requisite tools for creating spatial weights matrices from geocoded data. Once ready for analysis, an initial global Moran's *I* test allows for broad measurement of spatial dependence in a given data set prior to identification of hot spots, cold spots, and outliers and/or estimation of spatial regression models. Software capable of handling geocoded data will have functions that test the null hypothesis of spatial independence across the study area by calculating a Moran's *I* statistic for a provided variable and spatial weights matrix. For example, the "moran.test" and "moran.mc" functions in the "spdep" package in *R* calculate Moran's *I* indices and test significance under the assumption of normality or by comparison against a distribution of simulated *I* statistics, respectively. The distinction is important as researchers consider distributional assumptions about the local or global spatial dependence under investigation, such that a pseudo *p*-value for indicators can still be gleaned from Monte Carlo simulations independent of restrictive mathematical proofs (Anselin, 1995). A statistically significant global Moran's *I* test makes identification of hotspots a particularly useful descriptive statistic, as the presence of spatial autocorrelation in dependent variables will require incorporating spatial weights in predictive models to obtain accurate estimates.

As noted previously, the local Moran's I_i index is a positive or negative value calculated for each unit that, if significant, represents particularly potent sources of spatial autocorrelation. Resulting indices fall into four categories: high-high (high values surrounded by high values), low-low (low values surrounded by low values, i.e., cold spots), and outliers, high-low and low-high. Figures 3.2 and 3.3 in the example below contain examples of each. Like the global Moran's *I*, different software programs

Table 3.1. Sample Geocoded Data from IPEDS

Unit ID	*Institution Name*	*Street or P.O. Box*	*City*	*State*	*ZIP Code*	*Longitude*	*Latitude*
110608	California State University-Northridge	18111 Nordhoff St	Northridge	California	91330	−118.526817	34.23671
110617	California State University-Sacramento	6000 J St	Sacramento	California	95819-2694	−121.42194	38.55745
110635	University of California-Berkeley	200 California Hall	Berkeley	California	94720	−122.260463	37.871918

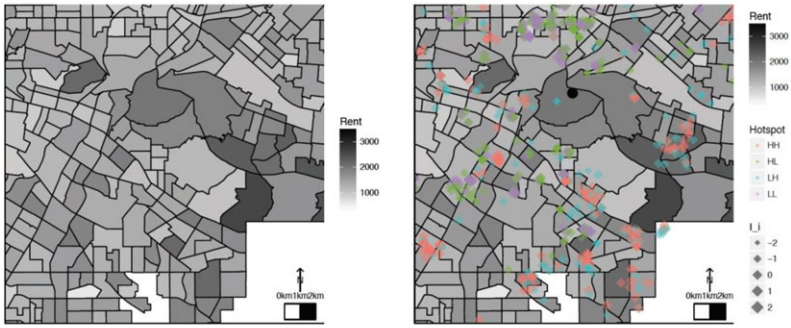
Figure 3.2. LISA hot spots near college

have different functions for calculating *p*-values and pseudo *p*-values for I_i statistics under the assumption of normality or based on permutations. The latter is the default method for calculating pseudo *p*-values in ArcGIS and GeoDa, while the "localmoran" function in the "spdep" package in *R* returns standardized I_i statistics and *p*-values (Bivand & Wong, 2018). As consensus among most spatial researchers is to shy away from distributional assumptions about I_i, Caulley et al. (2017) implemented a Monte Carlo simulation-based function for calculating local Moran's I_i in *R*. The next section in this chapter puts to practice methodological details discussed so far using administrative data from a college in southern California.

Empirical Application

The following section illustrates how LISAs can be used in practice by institutional researchers to (1) identify hot spots, cold spots, and outliers of a given variable within a student population and (2) show how those resuls can be combined with other sources of data. The latter is significant as much of the value of using spatial analysis techniques comes from their ability to convey information about social processes that

Figure 3.3. Median rent and LISA hot spots



(a) Median Rent by Census Tract (b) With LISA Hotspots

intuitively connects with administrators and policymakers, that is, most people can generally interpret maps and on a surface level understand the relationship between education variables and their geographic context. Below, I identify spatial dependence in GPAs among a group of community college students in California and estimate LISAs. I then map LISAs—clusters of high GPAs, low GPAs, and outliers—with neighborhood-level rental prices to show one way in which researchers can contextualize spatial dependence.

Data and Variables. Data come from student-level records for academic year 2017–2018 from a community college in Los Angeles County. Each observation contains personal information including home address, as well as academic information like GPA. Geocoding student addresses was carried out using a web-based geocoder from UCLA (gis.ucla.edu/geocoder).

A sample of 7,935 unique student-address observations were culled from a larger sample of 8,969 students enrolled in the fall of 2017, with a number of observations dropped based on administrative input issues for addresses and/or other missing data. The dependent variable used is student total GPA, given noted effects of neighborhoods on various academic outcomes (Chetty, Hendren, & Katz, 2016). Additional neighborhood data comes from the Census Bureau's American Community Survey for year 2016.

Methodology. I first tested for spatial autocorrelation in total GPA variables using the global Moran's *I* statistic. The spatial weights matrix used is a near-neighbor matrix with *k-NN*, the number of *j* neighbors, equal to 9. An iterative process, I selected spatial weights based on maximization of the *I* statistic, though a number of qualitative and quantitative distinctions could be made here (Getis, 2009). For $k\text{-}NN = 9$, a global Moran's *I* index of 0.01 rejected the null hypothesis of spatial independence in total GPA variables at $\alpha = 0.05$ based on 5,000 permutations (pseudo-$p = 0.023$). Though

Table 3.2. Hotspots—Spatial Clusters and Outliers (*k-NN* = 9)

Not Significant ($p < 0.05$)	*High-High*	*Low-Low*	*High-Low*	*Low-High*
6,906	275	236	317	201

such a small I value might not warrant consideration of spatial autocorrelation in further estimation of regression models (i.e., global Moran's I can range from −1 to 1), I nonetheless proceeded with calculating local Moran's I_i statistics from GPA variables using the Monte Carlo function outlined by Caulley et al. (2017).

Local Moran's I_i Results. Table 3.2 shows local Moran's I_i results with hot spots—significant clusters and outliers—identified from 500 permutations. A total of 1,029 locations (student addresses) were GPA hot spots, cold spots, and outliers (non-corrected pseudo-$p < 0.05$) and were evenly distributed between hot and cold clusters ($n = 511$) and outliers ($n = 517$). It follows that 6,906 observations were not significant. Figure 3.2 shows a selection of hot spots, cold spots, and outliers in a researcher-defined geographic area around the college, with noticeable high-high clustering on the southern portion of the map and low-low clustering in the northern portion. The black circle in Figure 3.2 is the location of the college. A number of outliers are scattered throughout, which is to be expected with student-level units and relatively weak spatial autocorrelation across the entire sample. Nevertheless, the LISA analysis suggests that geography may be a factor impacting student GPAs. As a practical note, researchers maintain the discretion to define a region that best represents the relationship being described in LISA values when visually mapping hot spots. In the present case, displayed is a region surrounding the college that is densely populated with students. In an institution-level analysis with a national sample of colleges, alternatively, researchers might highlight hot spots in an entire state or even a full map of the United States. Next, I discuss how to use neighborhood-level variables to aid in LISA interpretation and offer practical recommendations for campus leaders.

Discussion. As there is reasonable evidence for the presence of spatial effects confounding student total GPA variables in the example above, the question of what factors are driving those local spatial effects arises. Local Moran's I_i values showed a number of hot spots, both clusters of students with high GPAs relative to the sample mean and students with lower GPAs, as well as outliers (high GPAs with low GPA neighbors and vice versa). Contextualizing LISA results can be accomplished by inclusion of geographic independent variables in descriptive and/or regression-based analyses. For example, socioeconomic status (SES) has consistently been shown to impact academic achievement holding other variables constant (Sirin, 2005) and as such, *spatializing* predictors of college GPA might incorporate neighborhood-level economic variables. Using median rent as

a geographic-based proxy for SES, I overlay LISA results for student GPAs in the preceding analysis on census tracts in the user-defined area shown in Figure 3.2. Panel a in Figure 3.3 shows median rent prices by census tract (five-year estimates, 2012–2016) while Panel B shows student GPA LISAs overlaid on the median rent map. A visual overview of the maps indicates that there may be a relationship between SES and GPA using rent as a space-based proxy for SES.

Additional spatial regression-based analyses would be required to make any definitive conclusions, but those methods are computationally intensive and out of the scope of this chapter. Nonetheless, this exercise illustrates the ease with which LISAs can be visualized and incorporated with other spatial data sources to respond to common questions asked by institutional researchers. Indeed, a primary benefit of *spatializing* interactions between measures of student and/or institutional variables and individual and contextual factors comes from the ability to leverage rich secondary data (e.g., student addresses, U.S. Census) in order to pinpoint specific populations or colleges that might otherwise go unnoticed in aggregated analyses. Campus leaders can then use that information to target certain areas for expanded student services like transportation and tutoring or to adjust other campus policies in response to neighboring institutions.

Limitations of LISAs

The most significant limitation of LISAs is their geographic context. That is to say, the nature of LISAs requires research questions with a specific spatial relationship in mind prior to carrying out analyses and interpretation of results thus requires connecting statistical inputs, outputs, and their relationships within a broader analysis of space, place, and geography. Be it lack of geocoded data for a given research question or spatially uncorrelated variables, use of LISAs to describe a particular process affecting students or a campus may not be as universal as analyzing descriptive trends or regression estimates.

Conclusion

While LISAs incorporate large amounts of data in a specific spatial context, they present institutional researchers with an important exploratory tool to better understand *localized* social-spatial dynamics impacting students and campuses. As shown in the empirical example here, visualization of LISAs is an accessible method of spatial analysis that offers an alternative to predictive models that sometimes require difficult to justify assumptions. Together with other methods and techniques outlined in this volume, LISAs complement a broader spatial analysis toolkit that can advance understanding of often overlooked processes impacting students, administrators, and other higher education stakeholders.

References

Anselin, L. (1995). Local indicators of spatial association–LISA. *Geographical Analysis*, *27*(2), 93–115.

Anselin, L., & Rey, S. (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis*, *23*(2), 112–131.

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association *TEST*, *27*(3), 716–748.

Caulley, L., Sawada, M., Hinther, K., Ko, Y. T. I., Crowther, J. A., & Kontorinis, G. (2017). Geographic distribution of vestibular schwannomas in West Scotland between 2000–2015. *PloS One*, *12*(5), e0175489.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, *106*(4), 855–902.

Cliff, A., & Ord, J.K. (1981). *Spatial processes: Models and applications*. London: Pion.

Getis, A. (2009). Spatial weights matrices. *Geographical Analysis*, *41*(4), 404–410.

González Canché, M. S. (2014). Localized competition in the non-resident student market. *Economics of Education Review*, *43*, 21–35.

LeSage, J. P., & Pace, R. K. (2010). Spatial econometric models. In Manfred M. Fischer, Arthur Getis (Eds.), *Handbook of applied spatial analysis* (pp. 355–376). Berlin: Springer.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*(4), 286–306.

Rios-Aguilar, C. (2015). Using big (and critical) data to unmask inequities in community colleges. *New Directions for Institutional Research*, *2014*(163), 43–57.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

Sokal, R. R., & Oden, N. L. (1978). Spatial autocorrelation in biology: 1. Methodology. *Biological Journal of the Linnean Society*, *10*(2), 199–228.

Titus, M. A., Vamosiu, A., & McClure, K. R. (2017). Are public master's institutions cost efficient? A stochastic frontier and spatial analysis. *Research in Higher Education*, *58*(5), 469–496.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(Supplemental Proceedings), 234–240.

AUSTIN LYKE is a PhD student in the Higher Education and Organizational Change program at the Graduate School of Education and Information Studies at UCLA. His research focuses on organizational and spatial contexts of higher education in the United States.