[AM-03-023] Local Measures of Spatial Association

Abstract

Local measures of spatial association are statistics used to detect variations of a variable of interest across space when the spatial relationship of the variable is not constant across the study region, known as spatial non-stationarity or spatial heterogeneity. Unlike global measures that summarize the overall spatial autocorrelation of the study area in one single value, local measures of spatial association identify local clusters (observations nearby have similar attribute values) or spatial outliers (observations nearby have different attribute values). Like global measures, local indicators of spatial association (LISA), including local Moran's I and local Geary's C, incorporate both spatial proximity and attribute similarity. Getis-Ord Gi*, another popular local statistic, identifies spatial clusters at various significance levels, known as hot spots (unusually high values) and cold spots (unusually low values). This so-called "hot spot analysis" has been extended to examine spatiotemporal trends in data. Bivariate local Moran's I describes the statistical relationship between one variable at a location and a spatially lagged second variable at neighboring locations, and geographically weighted regression (GWR) allows regression coefficients to vary at each observation location. Visualization of local measures of spatial association is critical, allowing researchers of various disciplines to easily identify local pockets of interest for future examination.

Keywords: geographically weighted regression, hot spot analysis, local indicators of spatial association (LISA), local spatial statistics, spatial heterogeneity, spatial non-stationarity, spatial statistics

Author & citation

Livings, M. and Wu, A-M. (2020). Local Measures of Spatial Association. The Geographic Information Science & Technology Body of Knowledge (3rd Quarter 2020 Edition), John P. Wilson (Ed.). DOI: 10.22224/gistbok/2020.3.10

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Local measures of spatial association. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Introduction
2. Spatial Relationships
3. Local Indicators of Spatial Association
4. Getis-Ord Gi* Statistic
5. Bivariate Local Moran's I
6. Geographically Weighted Regression
7. Applications of Local Measures of Spatial Association

1. Introduction

Spatial association describes how values of observations or samples are related in space. This spatial relationship is based upon Tobler's First Law of Geography – "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236) – and can be measured globally or locally. Global measures of spatial association focus on detecting global spatial dependence and quantifying the overall similarity between neighbor observations by providing one single statistic for the entire study area. One can use global measures to test whether values of observations nearby are similar (i.e., positive spatial autocorrelation) or dissimilar (i.e., negative spatial autocorrelation). Refer to Global Measures of Spatial Association for more details.

As an example, let's say we want to determine whether there is spatial autocorrelation of poverty rates between counties in Georgia. Global Moran's I provides a summarized value that indicates a positive spatial autocorrelation of poverty rates in the state of Georgia (e.g., using a first-order queen contiguity-based spatial weights matrix, I = 0.479). To examine spatial relationships that are not constant across space (i.e., spatial non-stationarity or spatial heterogeneity), and to identify where there are clusters of unusually high (or low) poverty rates in Georgia, we would need to employ local measures.

Local measures of spatial association describe spatial variations in different regions of the study area under the assumption of spatial heterogeneity. Local measures allow users to identify local "pockets" of spatial autocorrelation that may not be recognized when using global measures.In this entry, we briefly discuss common spatial relationships used for constructing the spatial weights matrix before introducing local indicators of spatial association (LISA), namely local Moran's I and local Geary's C, and the bivariate extension of local Moran's I. We look into another popular local statistic, Getis-Ord Gi*, and discuss the basic models of geographically weighted regression (GWR), a spatial regression technique that handles spatial heterogeneity. The principle of GWR is included in this entry for its capability to vary relationships between multiple variables geographically (hence considered 'local'); extensive methodology development and variations of GWR are not covered here and should be referred to in a separate entry (Geographically Weighted Regression). We close the entry by discussing applications of local measures, specifically discipline uses and available software environments.

2. Spatial Relationships

When calculating local measures of spatial association, we use spatial relationships between the observations in local neighborhoods to construct the spatial weights matrix. Common spatial relationships include weighting based on distance (e.g., inverse distance or inverse distance squared; fixed radius; k nearest neighbors) for point data, or spatial contiguity (e.g., Rook's case; Queen's case) for polygon data. Inverse distance or inverse distance squared relationships involve nearer observations having a stronger influence and that influence decaying with increasing distance; this relationship is commonly applied to continuous data, such as temperatures or air pressures. Fixed radius relationships apply the same influence to all points within a fixed radius of neighborhood area, and k nearest neighbors ensures that every observation has a specified number of neighbors, "k," a

relationship often applied when the density of observations varies widely across the study region.

Spatial contiguity is most commonly used to represent polygon features with connected neighborhoods, such as census tracts or counties, using the Rook's case to define neighbors as sharing a common edge boundary or the Queen's case to define neighbors as sharing both edges and corner vertices. Other spatial relationships such as travel time may be used when addressing specific goals in a spatial analysis. As local measure results are highly sensitive to how the "neighborhood" is defined, understanding spatial relationships in the data and consequently constructing a representative spatial weights matrix is a critical step before computing any local measures of spatial association. (See Global Measures of Spatial Association for more discussion on the spatial weights matrix).

3. Local Indicators of Spatial Association

Local indicators of spatial association (LISA) were proposed by Luc Anselin in order to decompose global indicators, namely Moran's I and Geary's C, into the contributions of individual observations to identify local cluster patterns or spatial outliers (Anselin, 1995). Statistics classified as LISAs must satisfy two requirements. First, the LISA for each observation represents the degree of spatial clustering of similar values around that observation. Second, the sum of LISAs for all observations is proportional to the associated global indicator. Here we introduce the two LISA statistics, local Moran's I and local Geary's C.

3.1 Local Moran's I

Local Moran's I is the most widely used LISA statistic that describes spatial clustering of observations in high or low values. For each observation i and neighboring observations j, the equation for local Moran's I incorporates deviations from the mean ($x_i - \bar{x}\ and\ x_j - \bar{x}$, respectively) to measure variable similarity, quantifying how similar or different the variable values for each observation and their neighbors are when compared to the global average. Local Moran's I equation is shown in Eq. 1:

$$I_i = \left[\frac{x_i - \bar{x}}{S_i^2}\right] * \sum^{N}{}_{j=1} w_{ij}(d)(x_j - \bar{x}) \qquad \text{(Eq. 1)}$$

where $S_i^2 = \frac{\sum^{N}{}_{j=1} w_{ij}(d)}{N-1}$, $(x_i - \bar{x})$ and $(x_j - \bar{x})$ represent the covariance of i and j, wij(d) is the spatial weights matrix, and j≠i. Similar to global Moran's I, local Moran's I incorporates the spatial weights matrix $\sum^{n}{}_{j} w_{ij}(d)$ to represent neighborhood structure, but only within a defined distance d. Unlike global Moran's I, which provides one single statistic, local Moran's I is calculated for each individual observation.

Statistical significance of local Moran's I can be determined using several approaches. A basic approach involves calculating z scores of local Moran's I at each observation i. A large

positive z score (e.g. $z > +1.96$) indicates that the observation and its neighbors have significantly similar values (spatial clusters consisting of either high or low values). A large negative z score (e.g., $z < -1.96$) indicates that the value at observation i is significantly different than its neighbors (i.e., spatial outliers).

Alternatively, a permutation approach can be used by randomly rearranging all values over all sampled locations a large number of times (e.g., 999 iterations) and computing local Moran's I for each permuted data set. The resulting empirical distribution represents the extremeness (or lack thereof) of the observed LISA statistic, relative to the randomly permuted values.

Local Moran's I can be visualized in a Moran scatterplot and on a map. These two visualizations complement each other. A Moran scatterplot displays the relationship between the variable value at observation i and the average variable value in the neighborhood, organized into four quadrants: High-High, High-Low, Low-Low, Low-High. The Moran scatterplot computation is detailed in Global Measures of Spatial Association. Local Moran's I outcomes can then be color-coded, based on scatterplot quadrant, onto a map illustrating spatial clusters (High-High, Low-Low) and spatial outliers (High-Low, Low-High). Figure 1 illustrates the Moran scatterplot and corresponding map using county-level poverty rates in Georgia as an example using a first-order queen contiguity-based spatial weights matrix. The counties with "High-High" values (i.e., high-poverty counties surrounded by high-poverty neighbors) are those falling in Quadrant I of the Moran scatterplot (Figure 1, bottom left) and shaded red on the map (Figure 1, right).

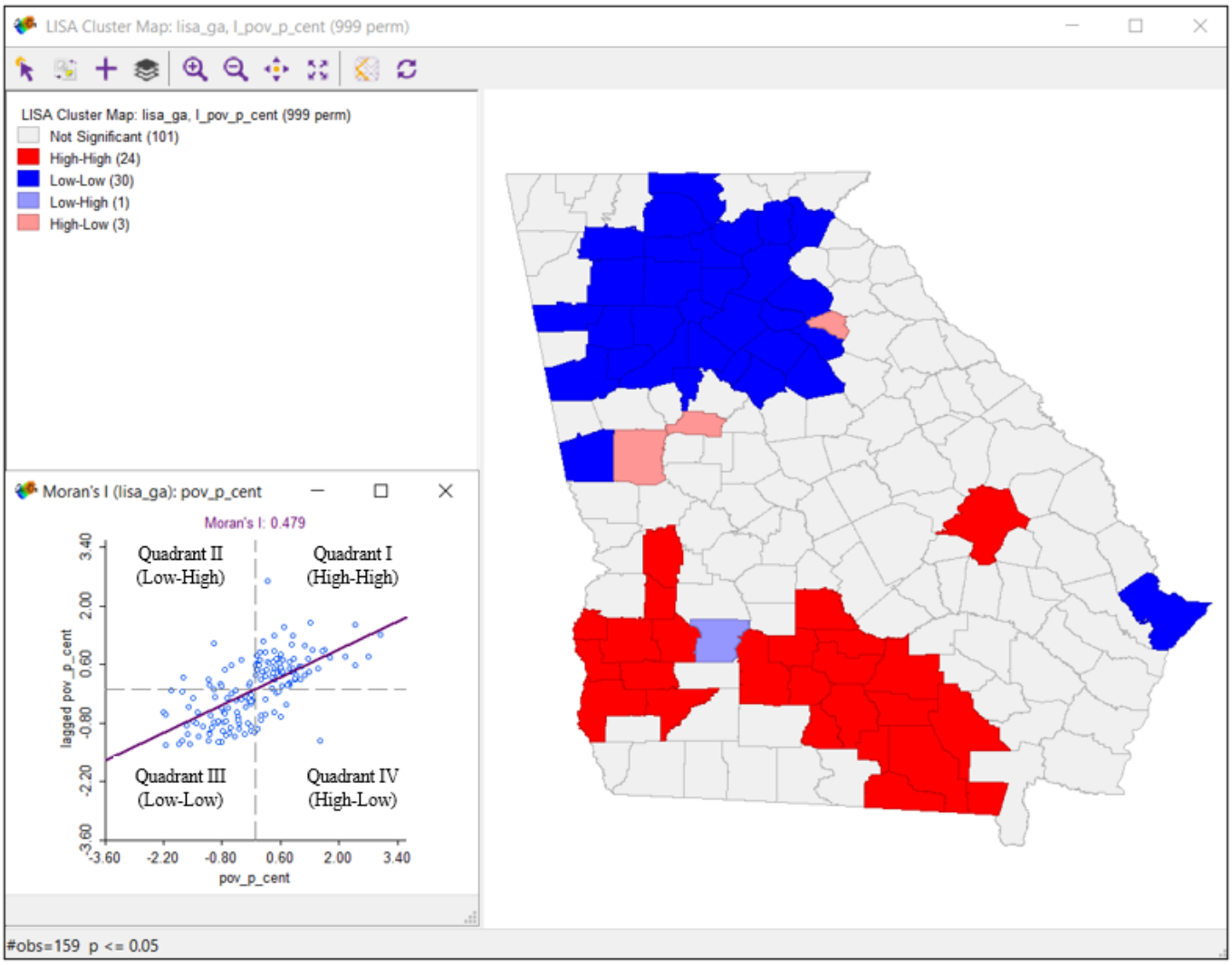


Figure 1. A Moran scatterplot (lower left) and corresponding map (right) showing counties that are spatial clusters (High-High, Low-Low) or spatial outliers (High-Low, Low-High) of poverty rates in the state of Georgia. Source: authors. Note: an earlier version of this figure contained a labeling error; a revised version was published January 4, 2021.

3.2 Local Geary's C

Similar to local Moran's I, local Geary's C incorporates the spatial weights matrix $\sum^{n}{}_{j} w_{ij}(d)$. Unlike local Moran's I, which is based on covariance, local Geary's C incorporates the squared difference between each point and its neighbors, as shown in Eq. 2, emphasizing the differences between locations:

$$c_i = \sum^{n}{}_{j} w_{ij}(d)[(x_i - \bar{x}) - (x_j - \bar{x})]^2 \qquad \text{(Eq. 2)}$$

where $[(x_i - \bar{x}) - (x_j - \bar{x})]^2$ represents the squared difference between i and j, and

$\sum_{j}^{n} w_{ij}(d)$ is the spatial weights matrix.

Similar to local Moran's I, the sum of local Geary's C is proportional to global Geary's C. Geary's C is not applied as often as its counterpart, Moran's I.

4. Getis-Ord G Statistic

The family of G statistics, introduced by Getis and Ord (1992), examines spatial clustering patterns that may not be identified by other global measures. The General G statistic may be used as a global indicator that measures overall concentration of similar or dissimilar values located within a specified distance of one another. This statistic incorporates the spatial weights matrix $\sum_{i}^{n} \sum_{j}^{n} w_{ij}(d)$ and involves multiplying all pairs of values (xi, xj) in which xi and xj are within distance d of each other (Eq. 3):

$$G(d) = \frac{\sum^{n}{}_{i} \sum^{n}{}_{j} w_{ij}(d) x_i x_j}{\sum_{i}^{n} \sum^{n}{}_{j} x_i x_j}, i \neq j \qquad \text{(Eq. 3)}$$

The most widely used G statistic is Gi*, which identifies local spatial clustering patterns, namely "hot spots" (spatial clusters of high values) and "cold spots" (spatial clusters of low values). Computation of Getis-Ord Gi* incorporates the spatial weights matrix with all neighboring values (xj) of any observation i within a specified distance (d) including itself (i.e. j=i) in the numerator (Eq. 4). The denominator is the sum of all neighboring values (xj), again including observations for which j=i.

$$G_i^*(d) = \frac{\sum^{n}{}_{j} w_{ij}(d) x_j}{\sum_{j}^{n} x_j} \qquad \text{(Eq. 4)}$$

Getis-Ord Gi* can be applied to all vector types (i.e., points, lines, polygons), although polygons (areal units) are most commonly seen.

Statistical significance of G statistics may be determined via hypothesis testing using standardized z scores calculated for each observation (e.g., county polygons) and corresponding p values. The z scores for each observation and corresponding levels of significance are consequently used in recognizing hot spots, clusters of observations with variable values significantly larger than expected, and cold spots, clusters of observations

with variable values significantly smaller than expected.

The most common method of visualizing Gi* statistic outcomes is a hot spot map. Hot spots are commonly color-coded using shades of red, with the darkest red corresponding to z scores larger than +2.58 (99% confidence), and medium and light shades of red corresponding to z scores between +1.96 and +2.58 (95% confidence) and z scores between +1.65 and +1.96 (90% confidence), respectively. In contrast, cold spots are commonly color-coded using shades of blue. The ranges for negative z scores correspond to those for positive z scores, with the darkest blue representing 99% confidence, medium blue representing 95% confidence, and light blue representing 90% confidence. Areal units that are not part of statistically significant clusters (i.e., z scores between -1.65 and +1.65) are shown as white or gray. Figure 2 shows an example visualization of a Getis-Ord Gi* result, again using data representing county-level poverty rates in Georgia based on a first-order queen contiguity-based spatial weights matrix:

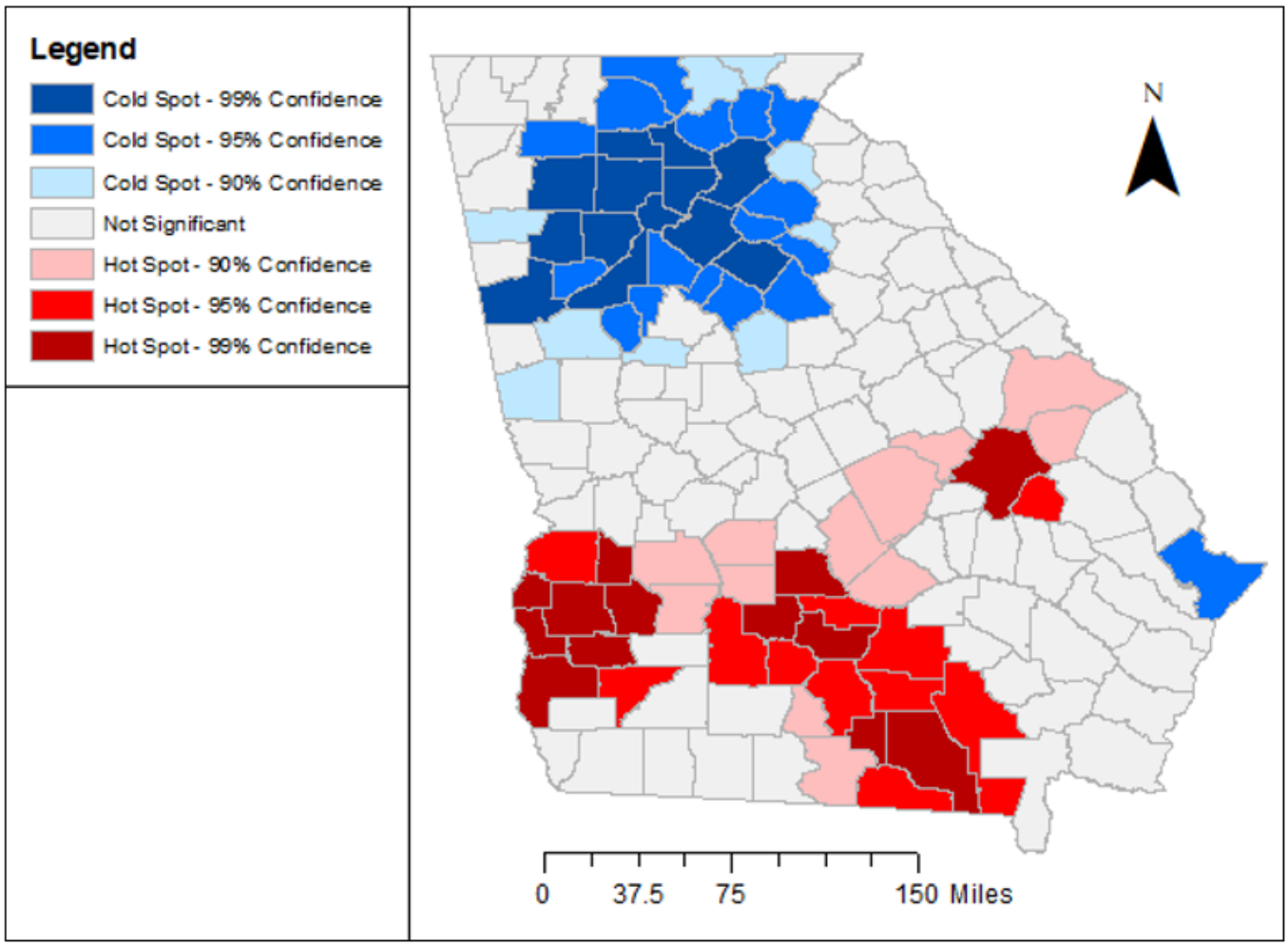


Figure 2. A hot spot map of county-level poverty rate in the state of Georgia generated using Getis-Ord Gi*. Source: authors.

Local Moran's I and Getis-Ord Gi* both measure local spatial association, with slightly different applications. Local Moran's I is used to describe the relationship between each

individual observation and its neighbors, identifying both clusters (High-High, Low-Low) and spatial outliers (High-Low, Low-High), whereas Getis-Ord Gi* is used exclusively for cluster identification, highlighting clusters of high values (hot spots) and clusters of low values (cold spots). Computation of Getis-Ord Gi* yields both a Gi* statistic for each observation and corresponding statistical significance (i.e., 90%, 95% or 99% confidence). In contrast, computation of local Moran's I identifies clusters and spatial outliers only at the α = 0.05 significance level.

5. Bivariate Local Moran's I

Bivariate spatial correlation, in the form of the bivariate local Moran's I statistic (shown in Eq. 5), describes the statistical relationship between a variable at a location and a spatially lagged second variable at neighboring locations (Anselin, 2019):

$$I_B = \frac{\sum_i (\sum_j w_{ij}(d) x_i \times y_j)}{\sum_i x^2{}_i} \qquad \text{(Eq. 5)}$$

where xi x yj is the cross product of the first variable at location i and the second variable at each neighboring location j, and wij(d) is the spatial weights matrix representing neighborhood structure within a specified distance d. Similar to the univariate local Moran's I, bivariate local Moran's I can be visualized in a bivariate Moran scatterplot and on a map. The bivariate Moran scatterplot displays the relationship between the first variable value at observation i and the spatially lagged second variable value at observation j, organized into four quadrants. Figure 3 illustrates the bivariate Moran scatterplot and its corresponding map using Georgia county-level poverty rates as x and spatially lagged county-level unemployment rates as y, using a first-order queen contiguity-based spatial weights matrix. The counties with "High-High" values are high-poverty counties surrounded by high-unemployment neighbors and are shown in Quadrant I of the bivariate Moran scatterplot (Figure 3 bottom left) and shaded red on the map (Figure 3, right).

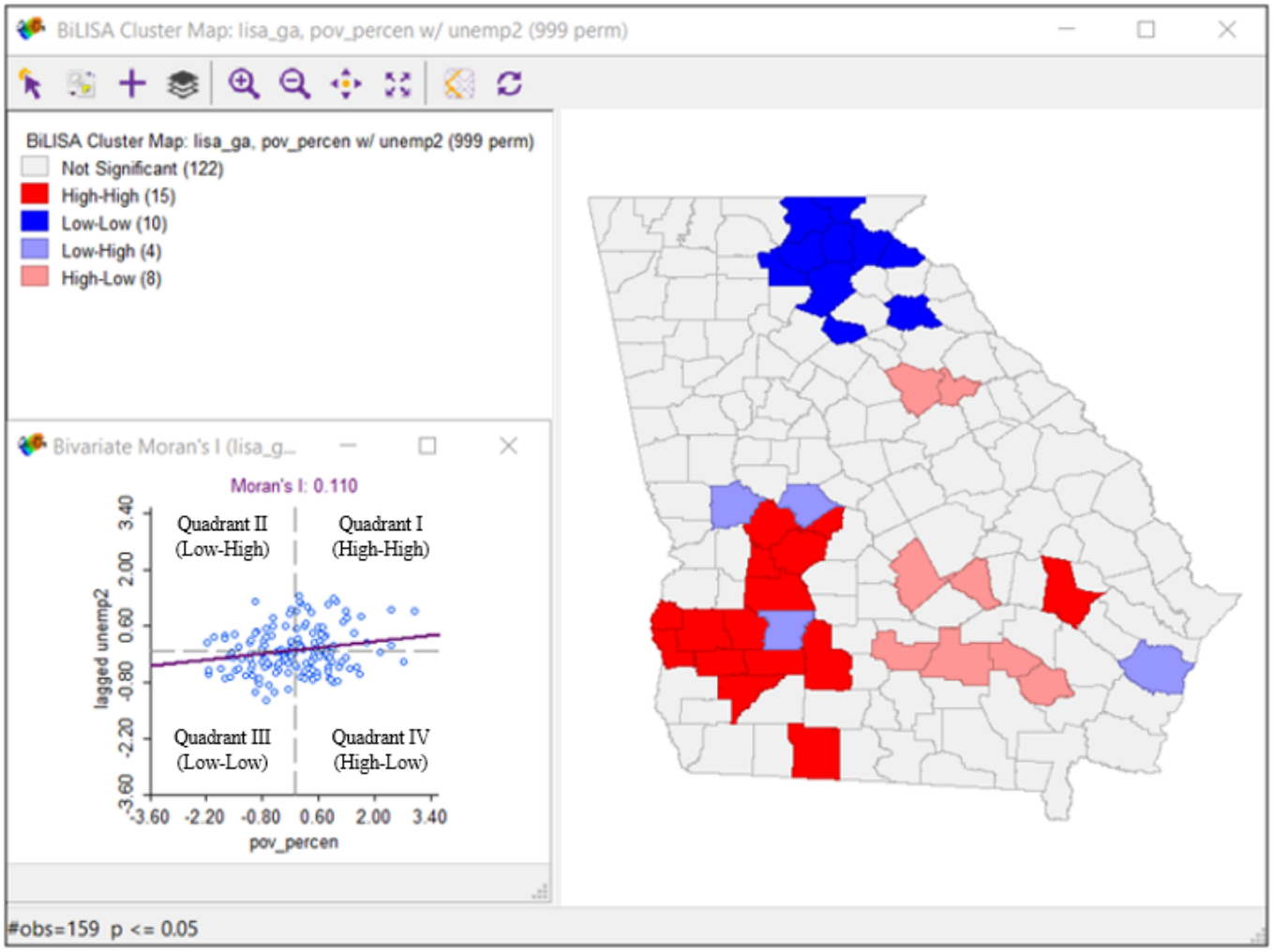


Figure 3. A bivariate Moran scatterplot (lower left) and corresponding map (right) showing counties that are bivariate spatial clusters (High-High, Low-Low) or bivariate spatial outliers (High-Low, Low-High) of poverty rates and unemployment rates in the state of Georgia. Source: authors. Note: an earlier version of this figure contained a labeling error; a revised version was published January 4, 2021.

6. Geographically Weighted Regression

Geographically weighted regression (GWR) is a spatial regression modeling technique that handles spatial heterogeneity by allowing regression coefficients ("parameter estimates") to vary spatially. Introduced by Brunsdon et al. (1996), GWR fits ordinary least square (OLS) regression models at individual observations (usually areal units, e.g., country polygons) using variable values from neighboring observations. A basic OLS regression model consisting of n observations is shown in Eq. 6:

$$y = \beta_0 + \sum_{k}^{n} \beta_k x_k + \epsilon$$

(Eq. 6)

where y is the dependent variable and xk is the corresponding kth independent variable; β0 is the intercept and βk is the coefficient of the kth independent variable; and ε is the error term. OLS estimates coefficients (βk) of the independent variables (xk) by minimizing the sum of the squares of the differences between the observed dependent variable (y) and its predicted counterpart ($\hat{y}$).

Just like Moran's I, it is necessary to fit a global model (i.e., OLS) before running local models such as GWR. The global OLS regression helps model specifications (i.e., selecting xik) for GWR. A basic GWR equation looks similar to OLS, but consists of a model for each location i (Eq. 7):

$$y_i = \beta_{i0} + \sum_{k}^{n} \beta_{ik} x_{ik} + \epsilon_i \qquad \text{(Eq. 7)}$$

Each local model is spatially weighted, with higher weights given to nearer observations (kernels), within a fixed or spatially adaptive distance. The spatial weights matrix is used to provide weights in parameter estimates (Eq. 8):

$$\hat{\beta}_i = (X^T W_i X)^{-1} W_i X^T y \qquad \text{(Eq. 8)}$$

where Wi is the spatial weights matrix of each observation i calculated with a kernel function and $\hat{\beta}_i$ is the vector of (n+1) parameter estimates in local regression models. The kernel function, including form (typically Gaussian or bisquare) and bandwidth, works the same as that in kernel density estimation (see Kernels and Density Estimation) and can be impactful to GWR outcomes. Usually setting with a fixed distance (but can also be spatially adaptive using a fixed local sample size) of each kernel, the bandwidth can be optimized by minimizing a cross-validation score (for prediction accuracy) or Akaike Information Criterion (AIC; for model parsimony).

Essentially, GWR identifies whether or not spatial relationships vary around observation locations. The outcome of GWR is usually compared with the global OLS regression model using a measure such as AIC. The AIC provides an estimate of out-of-sample prediction error, essentially estimating the quality and fit of each model relative to other candidate models; the model with the smallest AIC value provides the best fit to the data. If there is little difference between the AIC values for a global regression model and a GWR model, the spatial components of the GWR do not greatly improve the fit of the model and GWR is likely unnecessary.

Because the parameter coefficients and level of significance (t-values, p-values) are calculated for each observation, results from a GWR analysis may be displayed as a series of maps or a composite map. Again, we use the example of county-level poverty rates in

Georgia as the dependent variable, with county-level unemployment rates as the independent variable. A composite map in Figure 4 displays the GWR parameter estimates for unemployment rates in counties that are statistically significant ($p < 0.05$); this example was computed using a fixed kernel and the bandwidth was optimized by minimizing the AIC. By allowing parameter estimates to vary at each observation, GWR minimizes the potential bias caused by using one 'averaged' parameter coefficient for each independent variable across space.

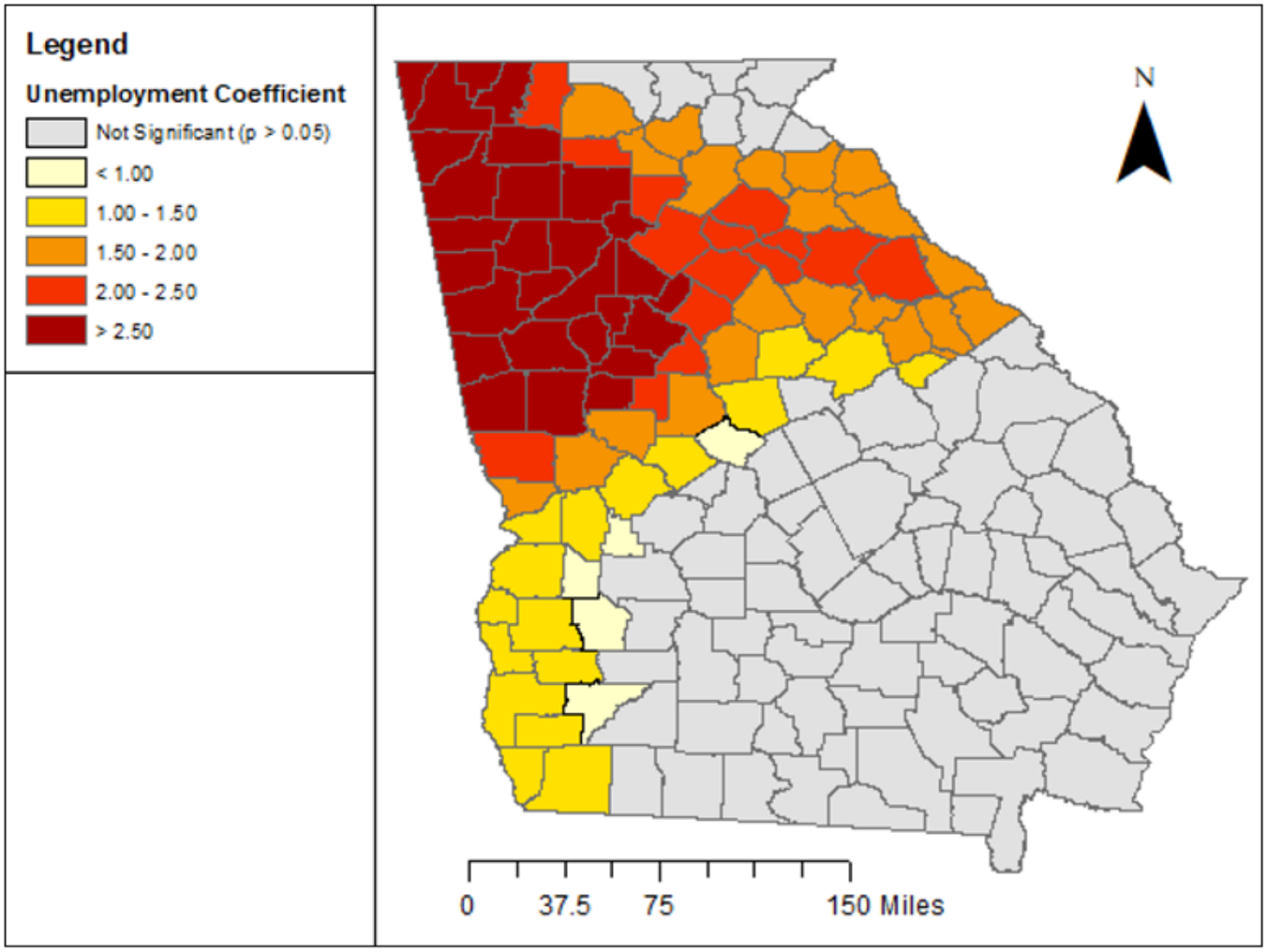


Figure 4. An outcome map of GWR showing parameter coefficients of unemployment rates (independent variable) in relation to poverty rates (dependent variable) in Georgia counties. Source: authors.

Besides a classic GWR using normally distributed (Gaussian) data, GWR has also been successfully applied to Poisson (count) and binary data. Further, GWR has been given much attention in various research fields, and there has been continuous improvement and development of GWR (e.g., Geographically Weighted Principal Component Analysis, Geographically Weighted Discriminant Analysis, Multiscale GWR) in the past two decades. These variations of GWR are not covered in this entry and should be referred to in a separate entry (Geographically Weighted Regression).

While GWR is an essential approach to examine relationships between variables across the study area, the parameters (i.e., independent variables) are often not independent of one another across locations (e.g., air temperature and elevation). This potentially leads to multicollinearity, in which two or more independent variables are highly correlated. If a GWR model displays multicollinearity, one or more independent variables that are highly correlated should be removed from the model. A “multicollinearity condition number” or “variance inflation factor” can be used to determine whether multicollinearity is serious and should be dealt with before finalizing the best-fitting GWR model (Kim, 2019).

7. Applications of Local Measures of Spatial Association

Local measures of spatial association have been widely used in various disciplines to provide critical information about ‘where’ local variation occurs. Measures of spatial outliers or clustering like LISA and Getis-Ord Gi* help researchers identify locations of interest in public health, environmental sciences, and social sciences. Recent applications include identifying disproportionate burdens of disease transmission in areas with certain sociodemographic characteristics (Chang et al., 2017), discovering local areas most prone to crime (Malleson & Andresen, 2015), and recognizing soil contamination hotspots of rare earth elements or heavy metals (Yuan et al., 2018; Wu & Johnston, 2019). With considerations of spatial heterogeneity in multiple regression, GWR has also been applied in various fields of study, ranging from environmental conditions (Luo et al., 2017) to regional development (Yu, 2014) to public health (Matthews & Yang, 2012) and sociodemographic equity (Gilbert & Chakraborty, 2011; Nicholson et al., 2019).

Due to the advance of computing power for data handling and processing, local measures of spatial association have been developed to explore variability not only across space but also over time (Diawara et al., 2019; Yu, 2014; Harris et al., 2017). Spatiotemporal analysis allows examinations of changes both spatially and temporally, therefore helping to identify both spatial and temporal trends in data. Common computing environments for local spatial association measures include GIS software such as ArcGIS, GeoDa, ENVI, MGWR, and QGIS, as well as spatial functions in the programming languages R (e.g., lctools, spdep, spgwr, GWmodel) and Python.

References

Anselin, L. (1995). Local Indicators of Spatial Association -- LISA. Geographical Analysis, 27(2):93-115.

Anselin, L. (2019). Global spatial autocorrelation (2): Bivariate, Differential and EB Rate Moran Scatter Plot.

Brunsdon, C., Fotheringham, A.S., & Charlton, M.E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281-298.

Chang, B.A., Pearson, W.S., & Owusu-Edusei, K., Jr. (2017). Correlates of county-level nonviral sexually transmitted infection hot spots in the US: application of hot spot analysis and spatial logistic regression. Annals of Epidemiology, 27(4), 231-237.

Diawara, N., Waller, L., King, R., & Lorio, J. (2019). Simulations of Local Moran's Index in a Spatio-temporal Setting. Communications in Statistics-Simulation and Computation, 48(6), 1849-1859.

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189-206.

Gilbert, A., & Chakraborty, J. (2011). Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. Social Science Research, 40(1), 273-286.

Harris, N.L., Goldman, E., Gabris, C., Nordling, J., Minnemeyer, S., Ansari, S., Lippmann, M., Bennett, L., Raad, M., Hansen, M., & Potapov, P. (2017). Using Spatial Statistics to Identify Emerging Hot Spots of Forest Loss. Environmental Research Letters, 12(2), 024012.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology, 72(6), 558-569.

Luo, J., Du, P., Samat, A., Xia, J., Che, M., & Xue, Z. (2017). Spatiotemporal Pattern of PM2.5 Concentrations in Mainland China and Analysis of Its Influencing Factors using Geographically Weighted Regression. Scientific Reports, 7(1), 1-14.

Malleson, N., & Andresen, M.A. (2015). The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. Cartography and Geographic Information Science, 42(2), 112-121.

Matthews, S.A., & Yang, T.C. (2012). Mapping the results of local statistics: Using geographically weighted regression. Demographic Research, 26(6), 151-166.

Nicholson, D., Vanli, O.A., Jung, S., & Ozguven, E.E. (2019). A spatial regression and clustering method for developing place-specific social vulnerability indices using census and social media data. International Journal of Disaster Risk Reduction, 38, 101224.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46: 234–240.

Wu, A.-M., & Johnston, J. (2019). Assessing Spatial Characteristics of Soil Lead Contamination in the Residential Neighborhoods Near the Exide Battery Smelter. Case Studies in the Environment, 3, 1-9.

Yu, D. (2014). Understanding regional development mechanisms in Greater Beijing Area, China, 1995–2001, from a spatial–temporal perspective. GeoJournal, 79(2), 195-207.

Yuan, Y., Cave, M., & Zhang, C. (2018). Using Local Moran's I to identify contamination hotspots of rare earth elements in urban soils of London. Applied Geochemistry, 88, 167-178.