Extreme coefficients in Geographically Weighted Regression and their effects on mapping

Seong-Hoon Cho
Assistant Professor, University of Tennessee, 2621 Morgan Circle, 314-A Morgan Hall, Knoxville, TN 37996-4518, Phone: (865) 974-7408, Fax: (865) 974-9492, email: scho9@utk.edu

Dayton M. Lambert
Assistant Professor, University of Tennessee, 2621 Morgan Circle, 321-A Morgan Hall, Knoxville, TN 37996-4518, Phone: (865) 974-7472, email: dmlambert@utk.edu

Seung Gyu Kim
Graduate Research Assistant, University of Tennessee, 2621 Morgan Circle, 401 Morgan Hall, Knoxville, TN 37996-4518, Phone: (865) 974-7658, email: sgkim@utk.edu

Su Hyun Jung
Graduate Research Assistant, University of Tennessee, 2621 Morgan Circle, 401 Morgan Hall, Knoxville, TN 37996-4518, Phone: (865) 974-7658, email: suhyunj@utk.edu

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association 2009 AAEA & ACCI Joint Annual Meeting, Milwaukee, Wisconsin, July 26-29, 2009

Extreme Coefficients in Geographically Weighted Regression and Their Effects on Mapping

Abstract: This study deals with the issue of extreme coefficients in geographically weighted regression (GWR) and their effects on mapping coefficients using three datasets with different spatial resolutions. We found that although GWR yields extreme coefficients regardless of the resolution of the dataset or types of kernel function, 1) the GWR tends to generate extreme coefficients for less spatially dense datasets, 2) coefficient maps based on polygon data representing aggregated areal units are more sensitive to extreme coefficients, and 3) coefficient maps using bandwidths generated by a fixed calibration procedure are more vulnerable to the extreme coefficients than adaptive calibration.

Keywords: extreme coefficient; fixed and adaptive calibrations; geographically weighted regression; Mapping

Extreme Coefficients in Geographically Weighted Regression and Their Effects on Mapping

INTRODUCTION

Cleveland and Devlin (1988) originally introduced local spatial regression techniques. A subclass of these regressions was renamed 'geographically weighted regression' (GWR) (Brunsdon et al., 1996, 1999, 2001). GWR has recently been applied intensively to test the assumption that regression parameter are globally stationary (McMillen 1996, Brunsdon et al., 1996, 1999, 2001; Fotheringham and Brunsdon, 1999; Fotheringham et al., 1998, 2002; Leung et al., 2000a, 2000b; Huang and Leung, 2002; Yu and Wu, 2004; Laffan et al., 2005; Lambert et al., 2006; Cho et al., 2006, 2008a; Yu, 2006, 2007; Deller and Lledo, 2007; Partridge and Rickman, 2007; Lo, 2008).

The main appeal of GWR is its ability to generate parameter estimates for every regression point by using observations in a given neighborhood. Typically, the parameter estimates are mapped to highlight spatial variation (Mennis, 2006). Resulting maps are thought to be didactic aids for policymakers, and for summarizing the large amount of data generated by the procedure. The Google Scholar and Web of Science identified 1,210 web links, with word searches using 'GWR' and 'spatial' (April 10, 2008). An examination of the first 100 papers that used GWR revealed that 94 papers mapped the parameter estimates in ex post discussion of the results. Despite the merit and increasing popularity of GWR, there are potentially serious problems associated with the approach as noted in the literature: 1) spatial error dependence (Leung et al., 2000b; Fotheringham et al., 2002), 2) potential multicollinearity among local regression coefficients (Wheeler and Tiefelsdorf, 2005), and 3) extreme coefficients including sign reversals (Farber and Páez, 2007).

GWR results that have policy implications typically use mapping to present results to policymakers. The issue of extreme coefficients in GWR is particularly important in these instances because it directly affects the visual pattern generated by the parameter distribution. Of the 94 papers that mapped GWR parameter estimates, 51 papers used some form of cluster analysis (i.e., local indicator of spatial association (LISA, Anselin, 1995), Gi-statistics) or interpolated data between the GWR regression parameters. Extreme coefficients can severely affect the determination of spatial clusters (Castro and Singer, 2001) which makes inferences based on cluster maps generated by GWR tenuous. Ex post interpolation of GWR coefficients may also introduce 'phantom' trends or patterns across a geographic area that would not otherwise exist in the actual data used to generate those patterns (Anselin, 2001); patterns that may contribute to biased interpretation and misinformed policies (Lambert et al., 2007).

This paper examines the effects of extreme GWR coefficients generated by a series of hedonic housing price models on mapping of spatial clusters and their visual interpretation. Three data sets are compared, each recorded at different spatial resolutions. The first data set is point data of home sales transaction in Knox County, Tennessee. The second and third data sets are polygon data. The second data is census-block level information about median home value in the Southern Appalachian of the U.S. The third data set is at the county-level, including the change in median home value during the 1990's for the U.S.

Because the locations of specific census-block groups and counties are proxied by their centroids coordinates in establishing the neighborhood effect in GWR (i.e., the weight matrix for the neighborhood size), centroids of larger census-block groups and counties represent larger areas.[1] As the spatial scale of data changes, spatial processes exhibit new interactions and

[1] Mean areas for census-block groups of the Southern Appalachian data and counties of the U.S. data are 9.56 and 979.11 square miles, respectively.

relationship (Nelson et al., 2007). The larger the area represented by the centroid, the wider and larger the area represented by the optimal neighborhood size if the trace of the weight matrix is allowed to expand and contract at each location, commonly referred to ‘adaptive’ function. Likewise, for the case of parcel level data, less dense data exist in more rural areas because parcels are more sparsely distributed, thus resulting in larger spatial neighborhoods. On the other hand, if expansion and contraction of the weight matrix is not allowed and there is a cutoff threshold (Fotheringham et al., 2002, p.57), the standard errors of the coefficients are higher because the number of data points used is small. In this case, the kernel is commonly referred to ‘fixed’ function.

Thus, one might expect that extreme values in GWR coefficients are likely to be found in areas where data points are relatively sparse using both fixed and adaptive weight matrices. Because the size of a census-block group is larger in a more rural area, the county size is larger in the Western U.S., and a parcel and a census-block group is more sparsely distributed in edge areas, the hypothesis is equivalent to testing the existence of more extreme coefficients 1) in more rural areas for the parcel and census-block group data, 2) in the West U.S. region for the county data, and 3) in edge areas for the parcels and census-block group data. Using the diagonal elements of a projection matrix estimated using the set of GWR coefficients, extreme coefficients are detected. Extreme coefficients will inevitably influence any projection of a map and its interpretation.

DATA

For the example, three hedonic models were estimated, each with different data sets. One data set and model was used by Cho et al. (2008b) in a study of open space amenity valuation in Knox

County, Tennessee. It consisted of 2,889 observations of single-family house sales during 2000 and land cover information derived from Landsat 7 imagery for 2001. The amenity value of open space area within a buffer of 1.0 mile drawn around each house sale transaction was estimated in the study. Cho et al. (2008c) used the second dataset in a hedonic study of the value of spatial configuration of forests within the Southern Appalachian Highlands. This dataset consisted of 4,915 observations of median housing values at the census-block group level in 2000. The amenity values of mean patch size, patch density, and edge density, and composition of forest species for the deciduous, *evergreen*, and mixed forest classifications were estimated in the study. Kim (2007) used the third dataset in a hedonic air quality study of the continental U.S. The dataset consisted of 3,102 observations of median housing values at the county level. The amenity value of a decrease in total suspended particulates (TSP) was estimated in the study. Table 1 summarizes the three models.

EMPIRICAL MODEL

Geographically Weighted Regression

The hedonic housing price model is:

$$\ln p_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i\,,\ \ i = 1, \dots, n,\ k = 1, \dots, \mathrm{m}, \qquad (1)$$

where $\ln p_i$ is the natural log of the housing price (or value) of *i*th observation; x_{ik} is the *i*th observation of the k^{th} of *m* variables; ε_i is a random error; (u_i, v_i) denotes the location coordinates of *i*th observation; and $\beta_0(u_i, v_i)$ and $\beta_k(u_i, v_i)$ are local parameters for *i*th observation. The p_i for the model using Knox, Southern Appalachian, and U.S. data are housing sale price, median home value, and change in median home value, respectively. The complete lists of 43, 29, and 20 explanatory variables used in each model are presented in the Table 1. A natural log transformation for the distance, dollar, and area-related variables was used and it is symbolized as $\mathrm{Ln}(\cdot)$ in the Table 1.

The GWR estimator is:

$$\hat{\beta}(u_i, v_i) = \left[(X'W(u_i, v_i)X)\right]^{-1} X'W(u_i, v_i)P \tag{2}$$

where $\hat{\beta}$ represents an $n \times m$ matrix with elements $\hat{\beta}_k(u_i, v_i)$; X is an $n \times m$ matrix containing a vector of the x_{ik} variables; P is a vector of $\ln p_i$; and $W(u_i, v_i)$ is an $n \times n$ weight matrix in which the diagonal elements w_{ij} are geographical weights for each of the n observations for regression point i.

Different kernel functions $K(d_{ij}/b)$ determine the diagonal elements of the weight matrix, with d_{ij} the distance between point *i* and *j*, *b* a value that minimizes the residual sum of squares of predicted values (e.g., a cross-validation (CV) procedure). Fotheringham et al. (2002) suggest using a fixed Gaussian kernel, with $K(d_{ij}/b) = \exp\left[-(d_{ij}/b)^2/2\right]$; or an adaptive bi-square function, with $K(d_{ij}/d_{\max}) = \left[1-(d_{ij}/d_{\max})^2\right]^2$ if *j* is one of the *N*th nearest neighbors of *i* and $K(d_{ij}/d_{\max}) = 0$ otherwise. For the adaptive kernel, d_{max} is the maximum distance between observation *i* and its optimal number of neighbors. For the fixed kernel, *b* is distance that is also used as a cutoff value for all observations. GWR 3.0 software was used for fitting the model (Fotheringham et al., 2002).

Screening Extreme Coefficients

Extreme values in GWR coefficients are detected using a construct that follows the basic theory behind the 'hat matrix' commonly used in econometric studies to determine influential data points. The diagonal elements of the hat matrix are $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$. To identify influential or 'extreme' observations, Belsley et al. (1980) propose a cutoff of $2k/n$, where n is the number of observations used to fit the model, and k is the number of parameters in the model. Values of h_i exceeding this cut-off exert significant leverage in the coefficient design space, and typically warrant further investigation. In our approach, we consider the n by k matrix of GWR coefficients as a surrogate hat matrix; $h_i^{\beta} = \boldsymbol{\beta}_i(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_i'$. Observations with h_i^{β} values above the cutoff of $2k/n$ are considered as extreme GWR coefficients. We use this metric to isolate extreme GWR coefficients that may influence patterns observed in ex post map making activities.

Coefficient Mapping With and Without Extreme GWR Estimates

GWR parameter estimates are often mapped to facilitate interpretation. This application focuses on cluster mapping using the GWR coefficients generated from three different hedonic models in three distinct data sets. The first set of GWR coefficients we map are the coefficients associated with an open space measure corresponding with the Knox County housing data set. The second set of GWR coefficients we map correspond with a variable measuring patch density of evergreen forest found in the census-block group of Southern Appalachia home value data set. The last set of GWR coefficients we map correspond with an air quality improvement measure reflected in change in Total Suspended Particulates (TSP) emission density found in the US county-level home value data set.

Cluster patterns of parameter estimates from the same regression are mapped with and without the extreme coefficients to investigate the leverage effects of these extraordinary points.

The GWR regressions are based on two runs to compare the sensitivity of the results. The first run applies an adaptive bi-square function. A fixed Gaussian kernel is used in the second set of GWR regressions. Maps are generated from each regression with each data set. In an ex post analysis, Local Indices of Spatial Association (LISA, Anselin, 1995) are estimated to compare the extent to which extreme GWR coefficients influence cluster formation.[2]

EMPIRICAL RESULTS

Table 2 shows that distribution of coefficients are quite different, whereas coefficients of determination (R^2) and sums of squared errors (SSE) are fairly close among the GWR estimates based on the fixed and adaptive calibrations for all three databases. The distance associated with the fixed bandwidth parameter is shorter in the parcel-level point data (i.e., Knox database), while bandwidth (e.g., the optimal number of neighbors) estimated by the adaptive kernel is larger in the same data.

Diagonal elements of the GWR coefficient hat matrices generated from each data set for the fixed and adaptive estimates are mapped in Figure 1. The extreme coefficients identified by the $2k/n$ cutoff across the three datasets indicate that extreme coefficients are located more frequently near the edge of the study area. The extreme coefficients appear more pronounced for the estimates generated under the fixed calibration assumption as opposed to those estimated under the adaptive bandwidth assumption.

The Figure 1 shows there are more extreme coefficients identified in rural areas with lower densities of observations than in urban areas at the Knox and the Southern Appalachian datasets. For example, there are no extreme coefficients identified in the city center for the estimates generated by the fixed calibration procedure in the Knox dataset. Average sizes of

[2] An inverse distance weight matrix was used to estimate the LISA.

census-block groups with extreme coefficients in the Southern Appalachian dataset are 5.88% and 6.24% larger than those with non-extreme coefficients for the fixed and adaptive estimates, respectively. Note that the greater size of the census-block groups is constructed for areas that are more rural. Figure 1 also shows that significantly more extreme coefficients are found in the Western regions than the rest of the country for the coefficients estimated using the fixed bandwidth in the U.S. dataset. Notably smaller numbers of extreme coefficients are generated for the coefficients estimated using the adaptive bi-square function than for the estimates generated by the fixed Gaussian kernel in the U.S. dataset.

We used a Kolmogorov-Smirnov test (KS-test) to determine if the distributions of h_i^{β} with and without extreme coefficients were significantly different. The null hypothesis that the distributions were the same was rejected at the level of 1% in each scenario calibrated using the fixed kernel. With the adaptive calibration, the null hypothesis of similar distributions was rejected for the Knox and Southern Appalachian datasets, but not for the U.S. dataset at the 1% level. This shows, in general, the distribution are statistically different after removing the extreme coefficients.

More extreme coefficients identified near the edge and in rural areas may be explained by the smaller number of observations in the bandwidth for the regression near the edge and rural area than inner points and urban area, respectively. Consequently, in regions where data are scarce, the standard errors of the coefficients when fixed kernels are used are higher because the number of data points used is small if cutoff threshold is enforced (Fotheringham et al., 2002). On the other hand, the extreme coefficients for adaptive estimates may be explained by Farber and Páez (2007)'s justification of more heterogeneous local neighborhoods resulting from larger bandwidths in terms of distance. To find the same number of nearest neighbors for the adaptive calibration, a location near the edge or in rural area needs to cover a longer distance than one in the centre or in urban area, respectively.

Smaller numbers of extreme coefficients are generated by the adaptive calibration relative to fixed calibration in all three datasets, reflecting greater propensity of the fixed calibration procedure to generate extreme GWR coefficients. This implies that the number of extreme coefficients generated by the fixed kernel is greater than the number generated by adaptive calibration.

Figures 2, 3, and 4 show the LISA cluster maps for the high-high and low-low coefficients for the fixed and adaptive estimates before and after removing the extreme coefficients using the three datasets. The spatial patterns of clusters using the Knox dataset in Figure 2 show apparently little effect of extreme coefficients for both fixed and adaptive estimates, as the clusters with and without the extreme coefficients are comparable for both estimates.

The spatial clusters of coefficients associated with patch density of evergreen forest from the Southern Appalachian dataset shown in Figure 3 are more visibly affected by the extreme coefficients than the spatial patterns of clusters using the Knox dataset for both cases in Figure 2. Although the overall pictures of the clusters (i.e., majority of high-high and low-low clustered areas), are not affected by the extreme coefficients, visible differences due to extreme coefficients exist in Southern Appalachian estimates. The spatial patterns of clusters using the U.S. dataset shown in Figure 4 are affected visibly significantly by the extreme coefficients for both fixed and adaptive estimates. As the extreme coefficients are removed from the LISA cluster map, most of the clusters in the West disappear in the fixed estimates. It is not clear whether this result is an artifact of the large geographic unit or it simply indicates that the spatially varying coefficients are showing that change in median house value in these regions are actually more responsive to variation in air quality, or alternatively, there is a form of model misspecification related to geography that this coefficient is capturing.

Simple statistics of clusters re-confirm the visible pattern of higher effect of extreme coefficients on cluster mapping in greater size of data points (see Table 2). For example, the % of observations identified as high-high clusters changed by -1.76%, +4.55%, and +11.27% in Knox, Southern Appalachian, and U.S. datasets, respectively, by removing the extreme coefficients for the fixed case. For the adaptive calibration case, the relatively larger effects exist in the U.S. dataset where the % of observation in the high-high clusters changed by -3.35%, whereas the changes are -2.82% and +1.07% for the Knox and Southern Appalachian datasets, respectively. This implies that extreme coefficients can alter the cluster mappings considerably for both fixed and adaptive estimates. Particularly the effects are greater using greater size of data points, i.e., county level data such as U.S. dataset in our example.

Removing the extreme coefficient changes distribution of spatial clusters. For example, changes of cluster pattern appeared in the Southern Appalachian data appears greater than changes in the Knox data although the % change of observation in the high-high clusters following the removal of extreme coefficients is greater in the Knox data than in the Southern Appalachian data. This may be due to visible distinction in the resolution of data (parcels as opposed to polygons).

CONCLUSIONS

Despite the current popularity of GWR, more and more applied researchers have identified potentially serious problems with the approach. This study addresses with the issue of extreme coefficients generated by GWR, and their effects on mapping coefficients based on the analysis of three datasets. We found that although GWR yields extreme coefficients regardless of the resolution of the dataset or types of kernel function, 1) the GWR tends to generate extreme coefficients for less spatially dense datasets, 2) coefficient maps based on polygon data representing aggregated areal units are more sensitive to extreme coefficients, and 3) coefficient maps using bandwidths generated by a fixed calibration procedure are more vulnerable to the extreme coefficients than adaptive calibration.

The causes of the extreme coefficients may be artifacts of the resolution and distribution of the dataset, spatially varying coefficients, and/or misspecification related to geography captured by the coefficients. This study suggests that researchers and those who advise policymakers should be more cautious in interpreting maps as extreme coefficients are likely to affect maps of cluster patterns and the extreme coefficients may be caused by factors other than spatially varying coefficients. This leaves us a future research need to develop a model that can filter extreme coefficients caused by the spatially varying coefficients from all the other sources.

References

Anselin, L., 1995, "Local Indicators of Spatial Association-LISA," *Geographical Analysis*, 27: 93-115.

Anselin, L., 2001, "Spatial Effects in Econometric Practice in Environmental and Resource Economics," *American Journal of Agricultural Economics* 83(3): 705-710.

Belsley, D. A., Kuh, E., and R. E. Welsch, 1980, *Regression Diagnostics,* New York: Wiley.

Brunsdon, C., Fotheringham, A. S. and M. E. Charlton, 1996, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," *Geographical Analysis,* 28: 281-298.

Brunsdon, C., Fotheringham, A. S. and M. E. Charlton, 1999, "Some Notes on Parametric Significance Tests for Geographically Weighted Regression," *Journal of Regional Science,* 39(3):497-524.

Brunsdon, C., McClatchey, J. and D. R. Unwin, 2001, "Spatial Variations in the Average Rainfall/Altitude Relationship in Great Britain: An Approach Using Geographically Weighted Regression," *International Journal of Climatology*, 21:455-466.

Castro, M. C. and B. Singer, 2001, *Malaria Foci and Colonization Processes on the Amazon Frontier: New Evidence from a Spatial Analysis and GIS Approach,* Princeton University, NJ: Office of Population Research.

Cho, S., Bowker, J. M. and W. M. Park, 2006, "Measuring the Contribution of Water and Green Space to Housing Values: An Application and Comparison of Spatially-Weighted Hedonic Models," *Journal of Agricultural and Resource Economics,* 31(3): 485-507.

Cho, S., Clark, C. D. Park, W. M. and S. Kim, 2008a, "Spatial and Temporal Variation in the Housing Market Values of Lot Size and Open Space," *Land Economics,* 85: 51-73.

Cho, S., Lambert, D. and S. Kim, 2008b, “The Value of Open Space: Implications for Preservation Decisions,” Working Paper, Department of Agricultural Economics, University of Tennessee.

Cho, S., Jung, S. and S. Kim, 2008c, “Valuations of Forests: Spatial and Species Configurations in the Southern Appalachian Highlands,” Available online: http://www.springerlink.com/content/c265t175530l6082, *In press, Environmental Management.*

Cleveland, W. S., and S. J. Devlin, 1988, “Locally Weighted Regression: An Approach to Regression Analysis by Local Ritting,” *Journal of the American Statistical Association* 83(403): 596-610.

Deller, S.C. and V. Lledo, 2007, “Amenities and Rural Appalachian Growth,” *Agricultural and Resource Economics Review,* 36(1): 107-132.

Farber, S. and A. Páez, 2007, “A Systematic Investigation of Cross-Validation in GWR Model Estimation: Empirical Analysis and Monte Carlo Simulations,” *Journal of Geographical Systems,* 9(4): 371-396.

Fotheringham, A.S. and C. Brunsdon, 1999, “Local Forms of Spatial Analysis. *Geographical Analysis,”* 31(4): 340-358.

Fotheringham, A. S., Brunsdon, C. and M. E. Charlton, 1998, “Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis,” *Environment and Planning A*, 30(11): 1905-1927.

Fotheringham, A. S., Brunsdon, C. and M. E. Charlton, 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationship*, West Sussex: John Wiley & Sons Ltd.

Huang, Y. and Y. Leung, 2002, “Analyzing Regional Industrialization in Jiangsu Province Using Geographically Weighted Regression,” *Journal of Geographical Systems*, 4: 233-249.

Kim, S., 2007, Measuring Value of Air Quality: Application of Spatial-Hedonic Model. Thesis (MS). Agricultural Economics, University of Tennessee.

Laffan, S. W, Nielsen, O. M., Silcock, H. and M. Hegland, 2005, “Sparse Grids: A New Predictive Modeling Method for the Analysis of Geographic Data,” *International Journal of Geographical Information Science*, 19(3): 267-292.

Lambert, D. M., McNamara, T. K., and M. I. Garrett, 2006, “An Application of Spatial Poisson Models to Manufacturing Investment Location Analysis,” *Journal of Agricultural and Applied Economic,* 38(1): 105–121.

Lambert, D.M., Schaible, G., Johansson, R. and U. Vasavada, 2007, “The Value of Integrated CEAP-ARMS Survey Data in Conservation Program Analysis,” *Journal of Soil and Water Conservation* 62(1): 1-10.

Leung, Y., Mei, C. and W. Zhang, 2000a, “Statistical Tests for Spatial Nonstationary Based on the Geographically Weighted Regression Model,” *Environment and Planning A*, 32(1): 9-32.

Leung, Y., Mei, C. and W. Zhang, 2000b, “Testing for Spatial Autocorrelation among the Residuals of the Geographically Weighted Regression,” *Environment and Planning A,* 32(5): 871-890.

Lo, C. P., 2008, “Population Estimation Using Geographically Weighted Regression,” GIScience & Remote Sensing, 45(2): 131-148.

McMillen, D.P., 1996, “One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach,” *Journal of Urban Economics,* 40(1): 100-124.

Mennis, J., 2006, "Mapping the Results of Geographically Weighted Regression," *The Cartographic Journal*, 43(2): 171-179.

Nelson, A. Oberthür, T. and S. Cook, 2007, "Multi-Scale Correlations Between Topography and Vegetation in a Hillside Catchment of Honduras," *International Journal of Geographical Information Science*, 21(2): 145-174.

Partridge, M. and D. Rickman, 2007, "Persistent Pockets of Extreme American Poverty and Job Growth: Is There a Place-based Policy Role?" *Journal of Agriculture and Natural Resource Economics* 32(1): 201-224.

Wheeler, D. and M. Tiefelsdorf, 2005, "Multicollinearity and Correlation Among Local Regression Coefficients in Geographically Weighted Regression," *Journal of Geographical Systems*, 7: 161-187.

Yu, D. and C. Wu, 2004, "Understanding Population Segregation from Landsat ETM+ Imagery: a Geographically Weighted Regression Approach," *GIScience and Remote Sensing*, 41: 145-164.

Yu, D., 2006, "Spatially Varying Development Mechanisms in the Greater Beijing Area: A Geographically Weighted Regression Investigation," *Annals of Regional Science*, 40: 173-190.

Yu, D., 2007, "Modeling Owner-Occupied Single-Family House Values in the City of Milwaukee: A Geographically Weighted Regression Approach," GIScience & Remote Sensing, 44(3): 267-282.

Table 1. Summary of the three models: Variables used in the model and the data types and sources.

Knox County Model	Southern Appalachian Model	U.S. Model
Dependent Variable		
Ln(Housing sale price)	Ln(Median home value)	Ln(Change in median home value)
Explanatory Variables		
Variable Used for Mapping		
Ln(Area of open space)	Patch density of evergreen Forest	Change in TSP emission density
Structural Variables	***Spatial Configuration and Composition of Forest Species***	***Structural Variables***
Ln(Lot size)	Mean patch size of deciduous forest	Change in % of houses built in last 10 years
Age of house	Edge density of deciduous forest	Change in % of houses built 10-20 years ago
Brick siding (1, if brick)	Patch density of deciduous forest	Change in % of houses built before 1939
Swimming pool (1, if pool)	Mean patch size of evergreen forest	Change in % of houses without plumbing
Garage (1, if garage)	Edge density of evergreen forest	***Census-Block Group Variables***
Number of bedrooms	Mean patch size of mixed forest	Change in population density
Number of stories	Edge density of mixed forest	Change in % of white
Number of fireplaces	Patch density of mixed forest	Change in % of age above 65
Quality of construction (1, if excellent or good)	***Structure Variables***	Change in % of persons with high school graduate
Condition of structure (1, if excellent or good)	% of house with 3 or more rooms	Change in % of persons with college graduate
Ln(Total finished area of house)	% of house with kitchen	Change in % of urban population
Census Block-Group Variables	% of house with plumbing	Change in % of persons in poverty
Vacancy rate	Age of house	Ln(Change in household income)
Housing density	***Census-Block Group Variables***	Change in unemployment rate
Unemployment rate	Urban (1, if 100% in urban)	Change in % of manufacturing employment
Travel time to work	Interface (1, if in mixed urban-and-rural)	Change in % of vacant house
Ln(Median household income)	Ln(Per capita income)	Change in % of owner-occupied house
Distance Variables	Housing density	***Tax and neighborhood***
Ln(Distance to central business district)	Travel time to work	
Ln(Distance to greenway)	Vacancy rate	
Ln(Distance to railroad)	Unemployment rate	
Ln(Distance to sidewalk)	Stability	
Ln(Distance to park)	% of people with college	
Ln(Distance to golf course)		
Ln(Distance to water body)		
Ln(Park size)		
Ln(Water body size)		
High School Dummy		

Variables	degree	***variables***
Doyle	% of people over 65 years old	Change in per capita taxes
Bearden	***Distance Variables***	***Environmental Variables***
Carter	Ln(Distance to major city)	Natural amenity scale
Central	Ln(Distance to major road)	Rural urban continuum code
Fulton	Ln(Distance to National Park or National Forest)	Elevation
Gibbs	Ln(Distance to lake or reservoir)	
Halls	***Environmental Variables***	
Karns	Elevation	
Powell	Emission of Nitrogen Oxides	
Farragut		
Austin		
Other Spatial Dummy Variables		
Knoxville		
Flood		
Interface		
Urban growth boundary (1, if in urban growth boundary)		
Planned growth area (1, if in planned growth area)		
Real Estate Market Variable		
Season of sale (1, if April through September)		
Data Type		
Parcel level point data	Census-block group level boundary data	County level boundary data
Data Sources		
2001 National Land Cover Data, 2004 Environmental Systems Research Institute Data & Maps, GeoLytics Census CD, The Knoxville, Knox County, KUB Geographic Information System (Cho et al. 2008a)	2001 National Land Cover Data, 2004 Environmental Systems Research Institute Data & Maps, GeoLytics Census CD (Cho, Jung, and Kim 2008)	U.S. EPA 2007, County and City Data Books 2003, GeoLytics Census CD, USDA Economic Research Service, U.S. Census Bureau 2007 (Kim 2008)

Table 2. Goodness of fit, bandwidth, and summary of parameter estimates and clusters with and without extreme GWR coefficients.

Dataset		Knox		Southern Appalachian		U.S.	
Kernel Function		Fixed	Adaptive	Fixed	Adaptive	Fixed	Adaptive
Coefficient at Lower Quartile		-0.020	-0.044	-0.464	-0.622	-0.001	-0.002
Coefficient at Median		0.027	-0.004	0.043	-0.008	-0.000	-0.000
Coefficient at Upper Quartile		0.049	0.036	0.895	1.026	0.001	0.000
Adjusted R-square		0.74	0.77	0.78	0.77	0.71	0.71
SSE		187.2	188.0	146.0	155.3	26.1	25.5
Bandwidth[a]		3.16	1,240	22.15	896	195.24[b]	389
Total Number of Obs.		2,889	2,889	4,915	4,915	3,102	3,102
Number of Observed Extreme Coefficients		201 (144)	181 (144)	322 (246)	284 (246)	357 (155)	37 (155)
Number of Obs. in High-high Cluster	With Extreme Coefficients	1,476	1,273	1,493	1,380	1,029	624
	Without Extreme Coefficients	1,326	1,117	1,604	1,350	1,220	514
Number of Obs. in Low-low Cluster	With Extreme Coefficients	1,165	1,198	1,729	1,752	829	154
	Without Extreme Coefficients	998	1,133	1,749	1,708	713	418

[a] Distance in mile for fixed kernel and number of observation for adaptive kernel.
[b] Converted from 2.87 decimal degree by estimating the distance in Arcmap that 1.47 decimal degree is approximately 100 mile.

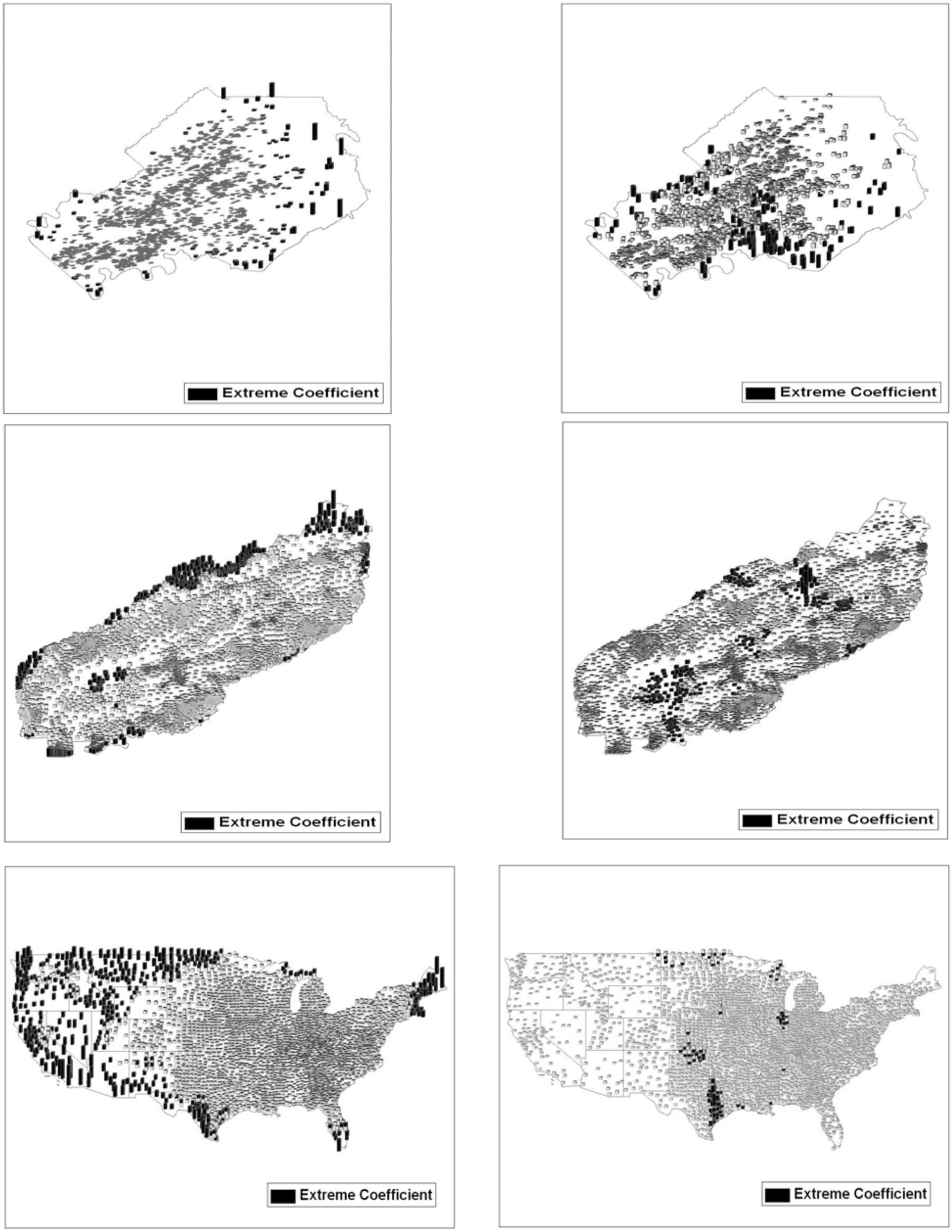
Extreme Coefficient
Extreme Coefficient
Extreme Coefficient
Extreme Coefficient
Extreme Coefficient
Extreme Coefficient

Fig. 1.

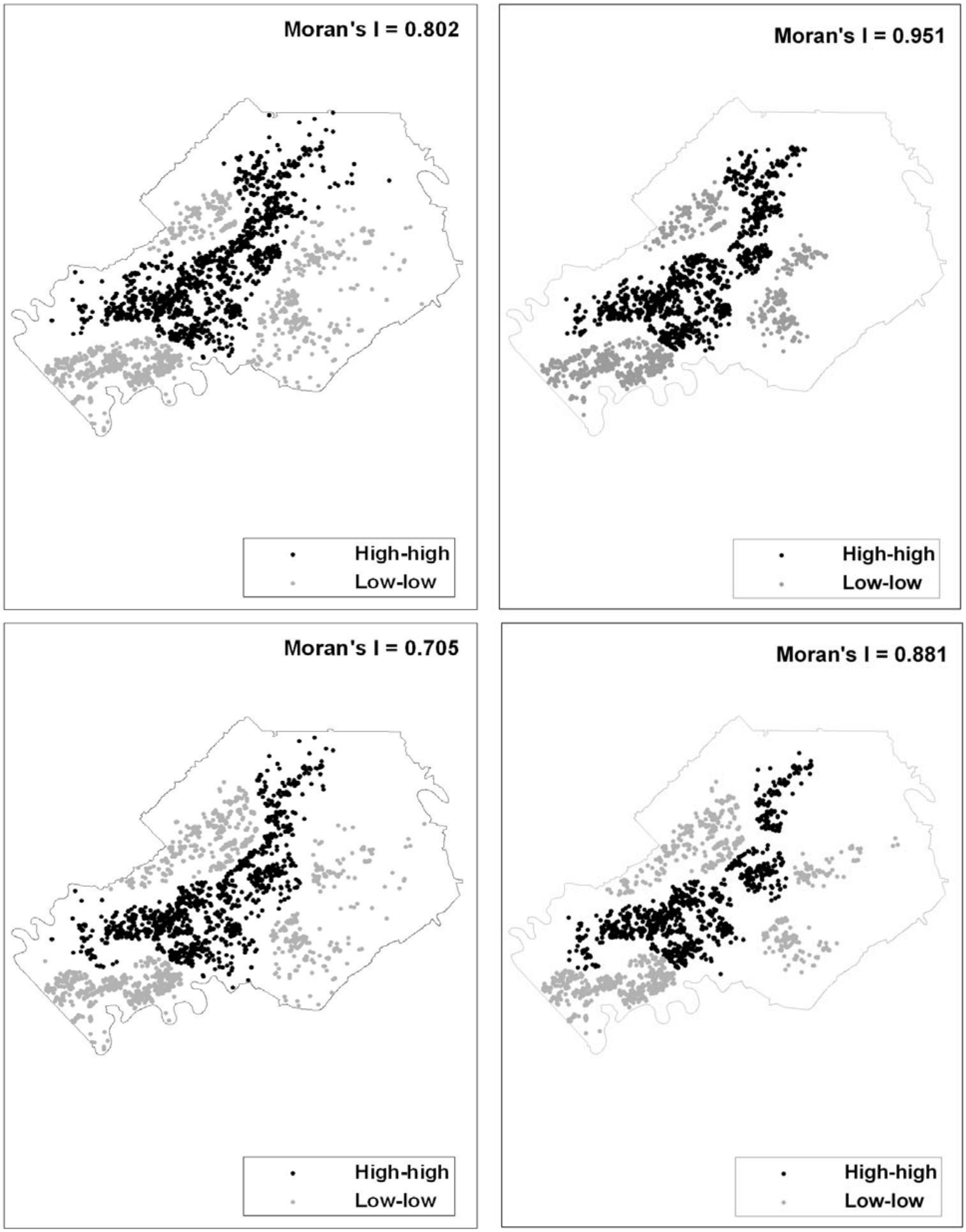
Moran's I = 0.802
High-high
Low-low
Moran's I = 0.951
High-high
Low-low
Moran's I = 0.705
High-high
Low-low
Moran's I = 0.881
High-high
Low-low

Fig. 2.

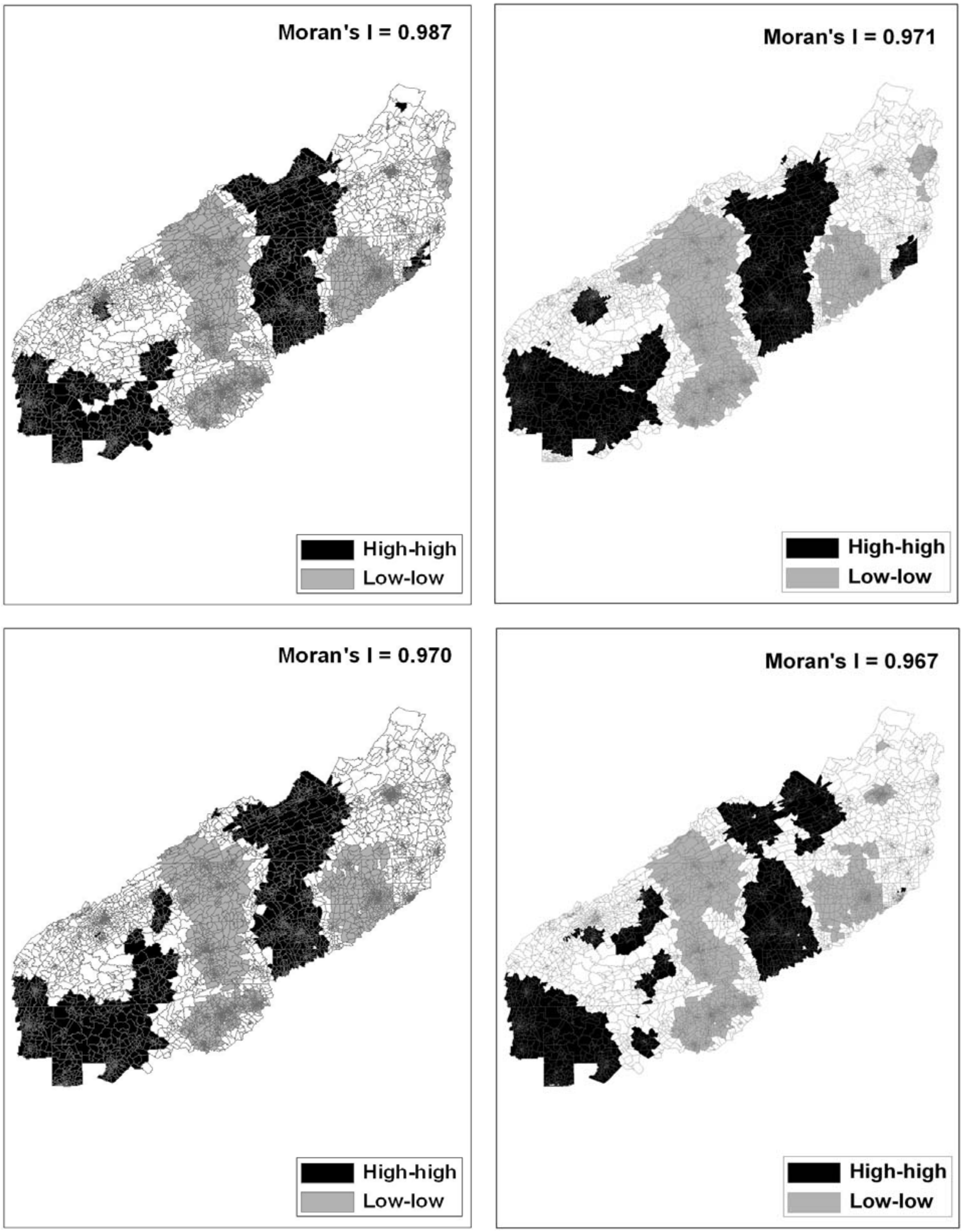
Moran's I = 0.987
High-high
Low-low
Moran's I = 0.971
High-high
Low-low
Moran's I = 0.970
High-high
Low-low
Moran's I = 0.967
High-high
Low-low

Fig. 3.

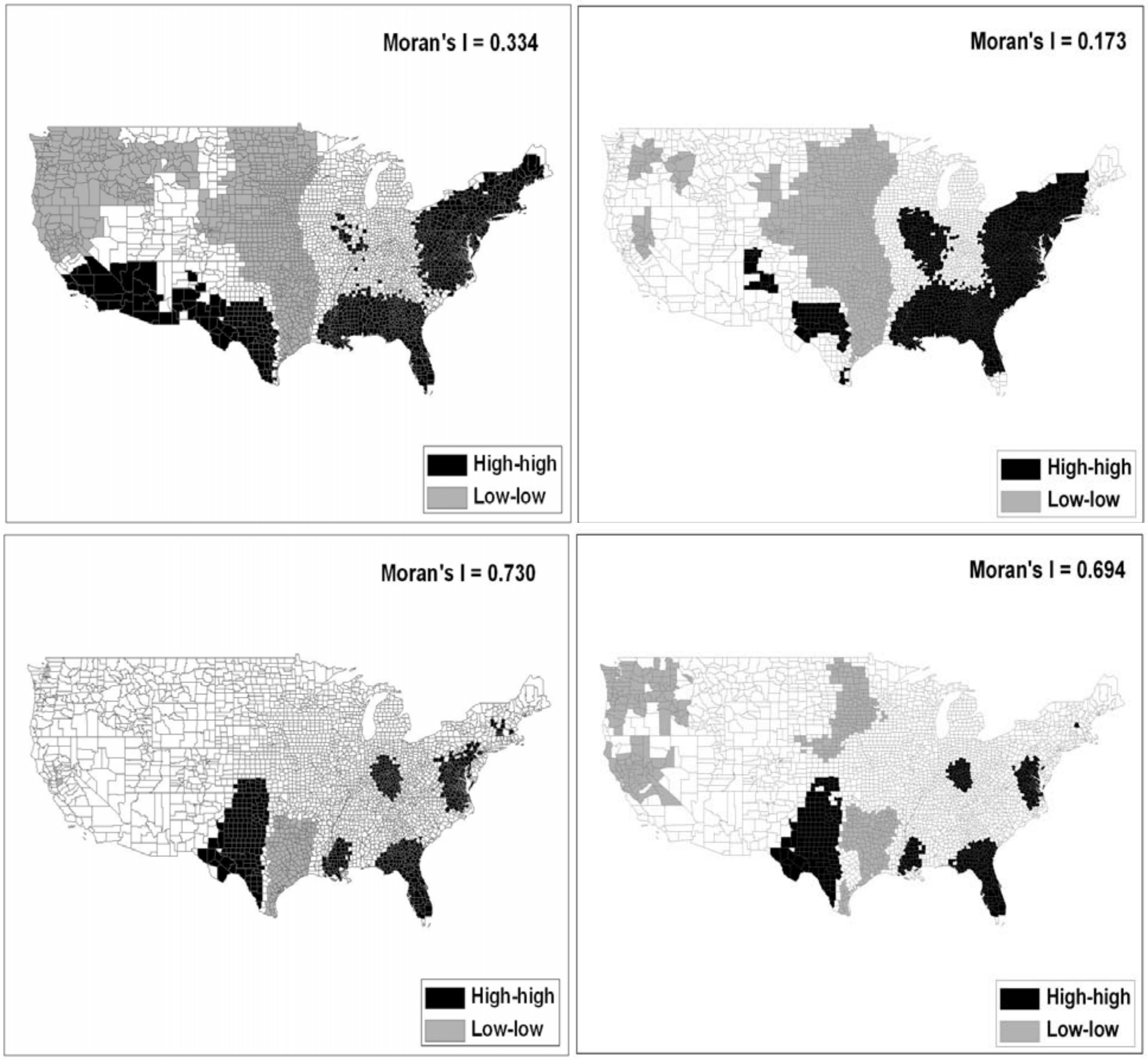
Moran's I = 0.334
High-high
Low-low
Moran's I = 0.173
High-high
Low-low
Moran's I = 0.730
High-high
Low-low
Moran's I = 0.694
High-high
Low-low

Fig. 4.

Fig. 1. The extreme coefficients identified by the cutoff of *2k/n* for fixed (left) and adaptive (right) estimates using Knox (top), Southern Appalachian (middle) and the US (bottom) datasets.

Fig. 2. The LISA cluster maps for spatial clustering of high-high and low-low coefficients at the 5% level for fixed (top) and adaptive (bottom) estimates before (left) and after (right) removing the extreme coefficients using Knox dataset

Fig. 3. The LISA cluster maps for spatial clustering of high-high and low-low coefficients at the 5% level for fixed (top) and adaptive (bottom) estimates before (left) and after (right) removing the extreme coefficients using Southern Appalachian dataset.

Fig. 4. The LISA cluster maps for spatial clustering of high-high and low-low coefficients at the 5% level for the fixed (top) and adaptive (bottom) estimates before (left) and after (right) removing the extreme coefficients using US dataset