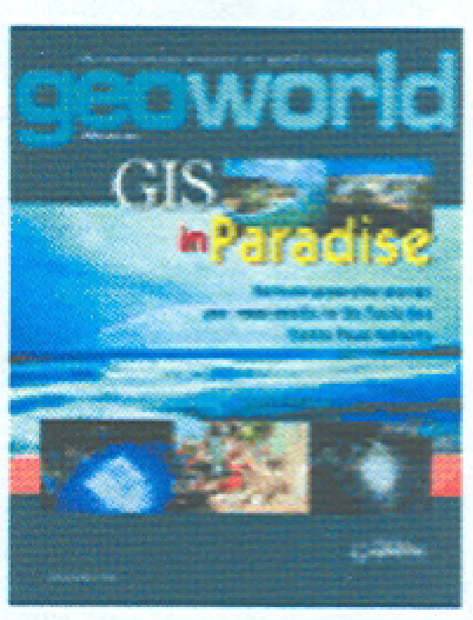
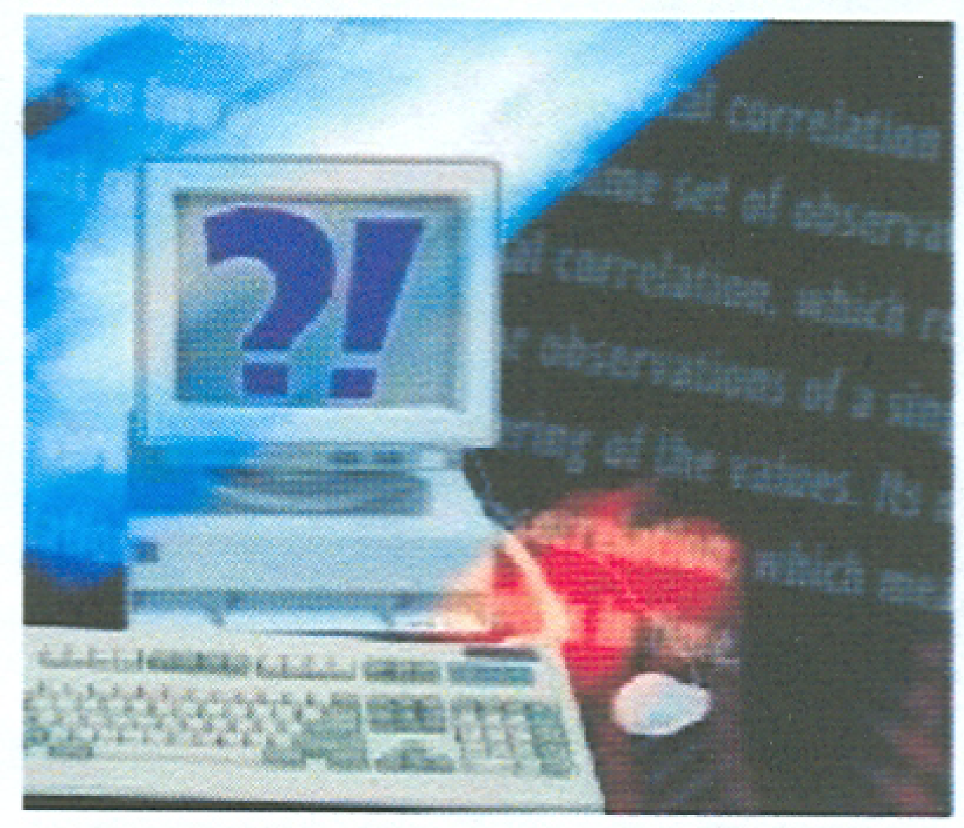


geoworld

Spatial Statistics

Opposites Don't Attract in Spatial Autocorrelation



By Anthony C. Lea and Daniel A. Griffith.Lea is managing director, research and development, MapInfo Canada/Compusearch Group; e-mail: tony_lea@mapinfo.com. Griffith is a professor of Geography at Syracuse University; e-mail:griffith@maxwell.syr.edu.

The most common measure of statistical correlation is between two variables for the same set of observations. A variant of this concept is serial correlation, which references the correlation among values for observations of a single variable, according to some ordering of the values. Its most common version is temporal autocorrelation, which measures the correlation among adjacent points in time.

The geographic extension of this concept is spatial autocorrelation, which measures the correlation among values at one point in space and adjacent or nearby values of the same variable. Spatial autocorrelation is cumbersome to compute, because it's difficult to identify adjacent or nearby locations.

Positive spatial autocorrelation indicates that geographically nearby values tend to be similar. Most social science and marketing variables tend to be positively spatially autocorrelated because of the way people, households and businesses organize themselves geographically.

Unfortunately, many researchers and analysts don't take spatial autocorrelation into account, because they're not aware of its importance. This has serious consequences for the accuracy and efficiency of conventional statistical measures. In most geographical analyses, and especially geodemographic analyses, failure to take spatial autocorrelation into account, through the use of adjusted formulae and procedures, can result in analyses that are inefficient, biased or simply wrong.

Significance of Spatial Autocorrelation

Household income is a good example of a variable exhibiting positive spatial autocorrelation. Neighborhoods next to higher-income neighborhoods tend to have higher incomes, whereas neighborhoods next to lower-income neighborhoods tend to have lower incomes. Households with high incomes prefer to live next to others with high incomes, bidding up land values, demanding larger houses and causing high-income neighborhoods to be unaffordable for low-income households. As a result, low-income households are constrained to live in lower-rent areas.

Why is this important? Put simply, if this wasn't the case, much geographical analysis in general would be uninteresting, and most of the discipline of business geographics wouldn't exist. Statistical clustering of households would yield groups lacking a geographic expression. We wouldn't be able to target likely Mercedes Benz buyers simply by targeting high-income areas, because such areas wouldn't exist. Ethnic communities would be absent from urban landscapes, as would the characteristic renter areas, areas of families with kids, university student populations, gay communities, etc.

Much of what's practiced as direct marketing to targeted areas, and the large part of retail location theory that focuses on demographics and lifestyle, would be futile. In a society characterized by no spatial autocorrelation, people and households would be arbitrarily and thoroughly mixed with respect to the variables we use most for marketing purposes--they would be "located randomly."

Correlation Coefficients

Understanding spatial autocorrelation requires an understanding of conventional correlation. The Moran Coefficient is a popular measure of spatial autocorrelation; it behaves much like a conventional correlation coefficient, and roughly ranges between -1 and 1, with a near-zero value denoting the absence of geographic correlation.

To illustrate spatial autocorrelation, we selected some common socioeconomic and demographic variables for two areas in the United States and undertook our analysis using census tracts as observations. Most of the variables are 2000 estimates of 1990 census variables. The value of the Moran Coefficient for these variables is reported in Table 1 for Syracuse, N.Y., and in Table 2 for Houston.

Table 1.

Moran Coefficient Values for Judiciously Selected Attributes: 1990 and 2000, Syracuse, N.Y., Census Tract Geography (n = 208)

Attribute	Moran Coefficient	Percent of variance accounted for by nearby values
Percent Male	0.28137	17.7
Population Density	0.70312	72.4
Percent Black/White Ratio	0.64495	56.5
Median Household Income	0.63207	55.5
Percent Widowed	0.27647	15.3
Percent University Degree	0.46436	42.8
Percent Chinese	0.42170	33.5

Table 2. Moran Coefficient Values for Judiciously Selected Attributes: 1990 and 2000, Houston, Census Tract Geography (n = 696)

Attribute	Moran Coefficient	Percent of variance accounted for by nearby values
Percent Male	0.35563	30.9
Population Density	0.52903	51.2
Percent Black/White Ratio	0.69310	62.6
Median Household Income	0.57342	52.7
Percent Widowed	0.54800	48.8
Percent University Degree	0.69693	67.9
Percent Chinese	0.52261	50.2

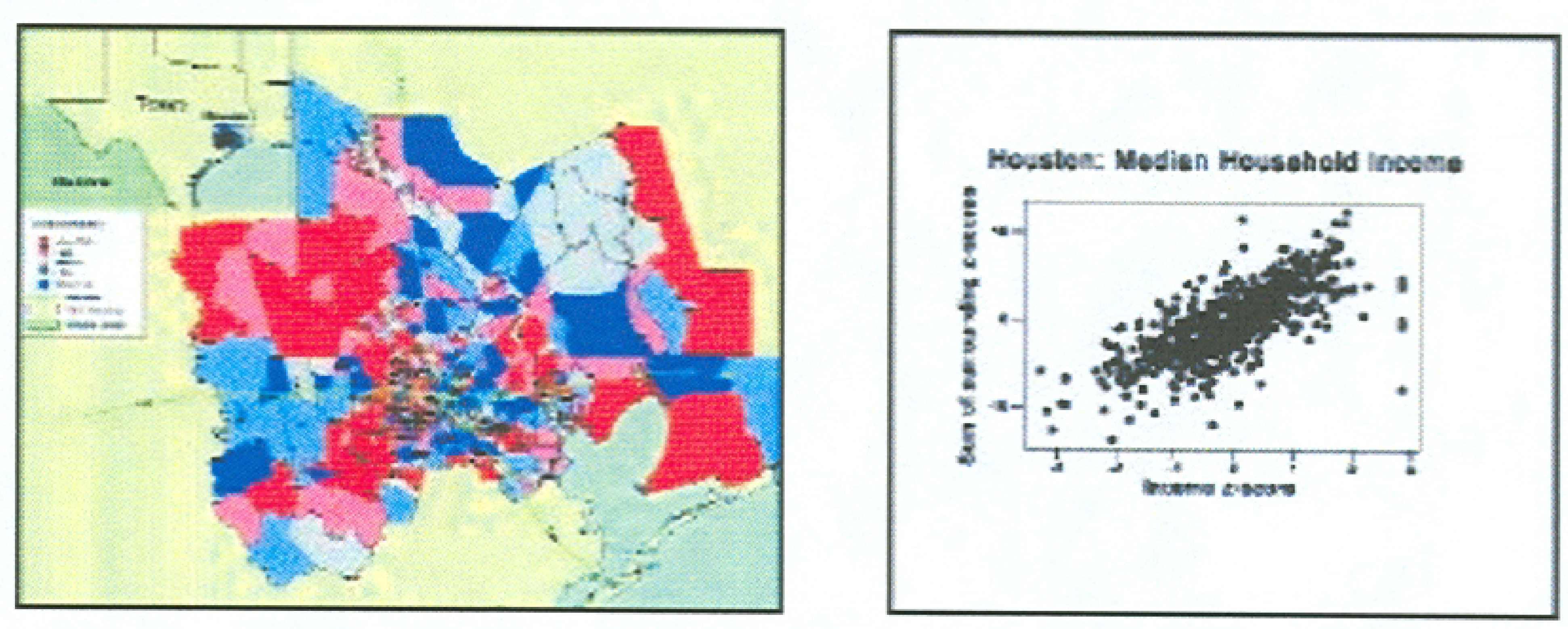
The percent of the population that's male was selected with the expectation that it would be randomly distributed across a city--its Moran Coefficient would exhibit weak spatial autocorrelation (nearby male percentages would account for little of the geographic variation in this variable). The remaining attributes were selected with the expectation that each would exhibit strong spatial autocorrelation. We expected that the black/white ratio and median household income would exhibit especially strong spatial autocorrelation. Most of the variables were transformed in some way (e.g., logarithm or power functions), so they would be approximately normally distributed before autocorrelation was computed.

We suspect that most analysts will be surprised at the extent of positive spatial autocorrelation in common socioeconomic and demographic variables. For most variables, the spatial autocorrelation is stronger in Houston. Syracuse has a higher autocorrelation value in population density, because it's a monocentric city.

The Autocorrelation Challenge

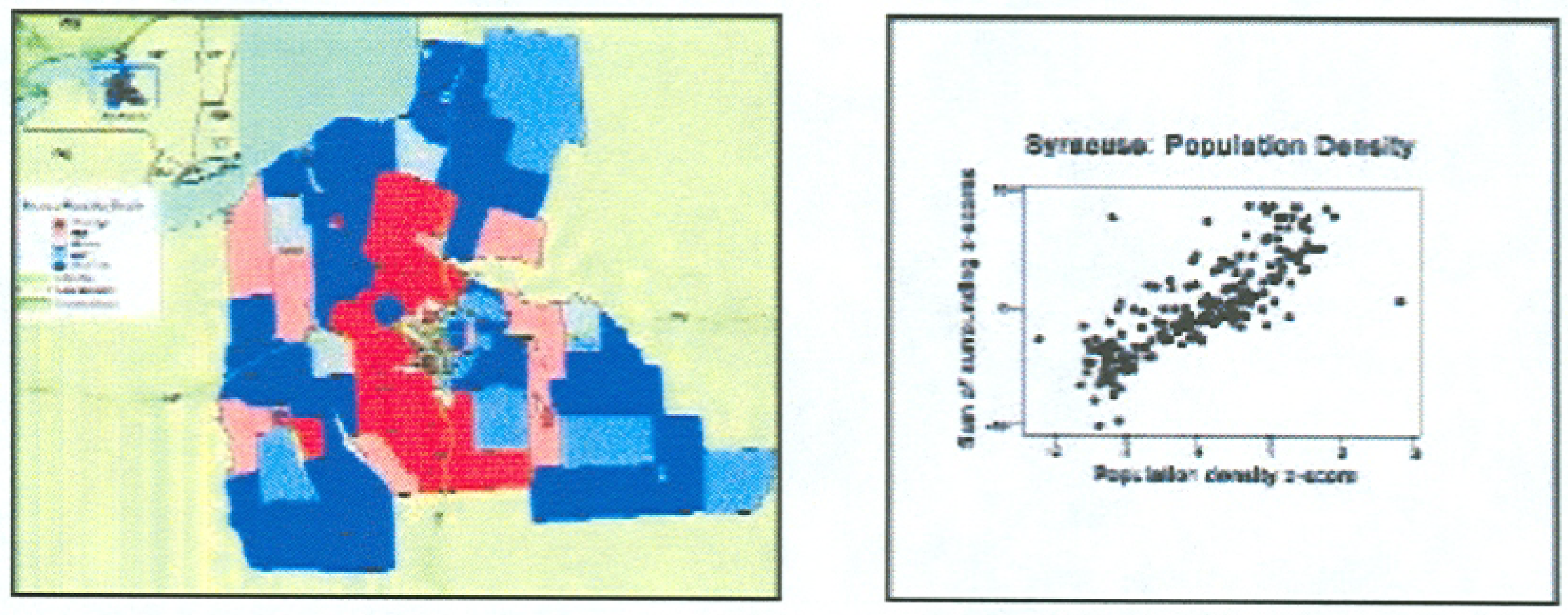
Spatial autocorrelation complicates statistical analysis by altering the variance of variables. Positive spatial autocorrelation causes the variance estimate to be inflated, changing the probabilities that statisticians commonly attach to making incorrect statistical decisions.

In a hypothesis-testing framework, there's an increased tendency to reject the null hypothesis when it's true. Another way of viewing the complication is that spatial autocorrelation captures the presence of redundant information in data. In the absence of spatial autocorrelation (i.e., the independent observations assumed in classical statistics), knowing attributes of a household wouldn't help one guess attributes of nearby households. As positive spatial autocorrelation of a household attribute increases, the ability to correctly guess the attribute's value for nearby households improves.



A Houston map shows median household income (left), and a diagram shows a Houston median household income Moran scatterplot (right).

If perfect positive spatial autocorrelation characterizes a given household attribute, then after the attribute is observed for any household, it's known for all households. The presence of this redundant information affects the contribution each georeferenced observation makes to statistics calculated with a database. The value of the first observation entering a calculation furnishes all new data.



A Syracuse map shows population density (left), and a diagram shows a Syracuse population density Moran scatterplot (right).

When subsequent observed values are for nearby observations, only part of the additional data obtained with them is truly new, because part already has been obtained with previous, spatially autocorrelated observations. Consequently, more spatially autocorrelated than independent observations need to be included in a calculation to attain an equally informative statistic.

New Products

There now is a well-developed body of theory, with revised formulae and computational processes, about how one should take spatial autocorrelation into account. Unfortunately, the necessary computations tend to involve a lot of additional work. Decades ago, those who wanted to deal with spatial autocorrelation had to write their own computer code. But times are changing. There now are several software packages on the market with spatial statistics modules that make short work of the extra computational burden involved and report correct statistics for georeferenced data. The most common are listed in Table 3.

Table 3. Common Software Packages with Spatial Statistic Capabilities

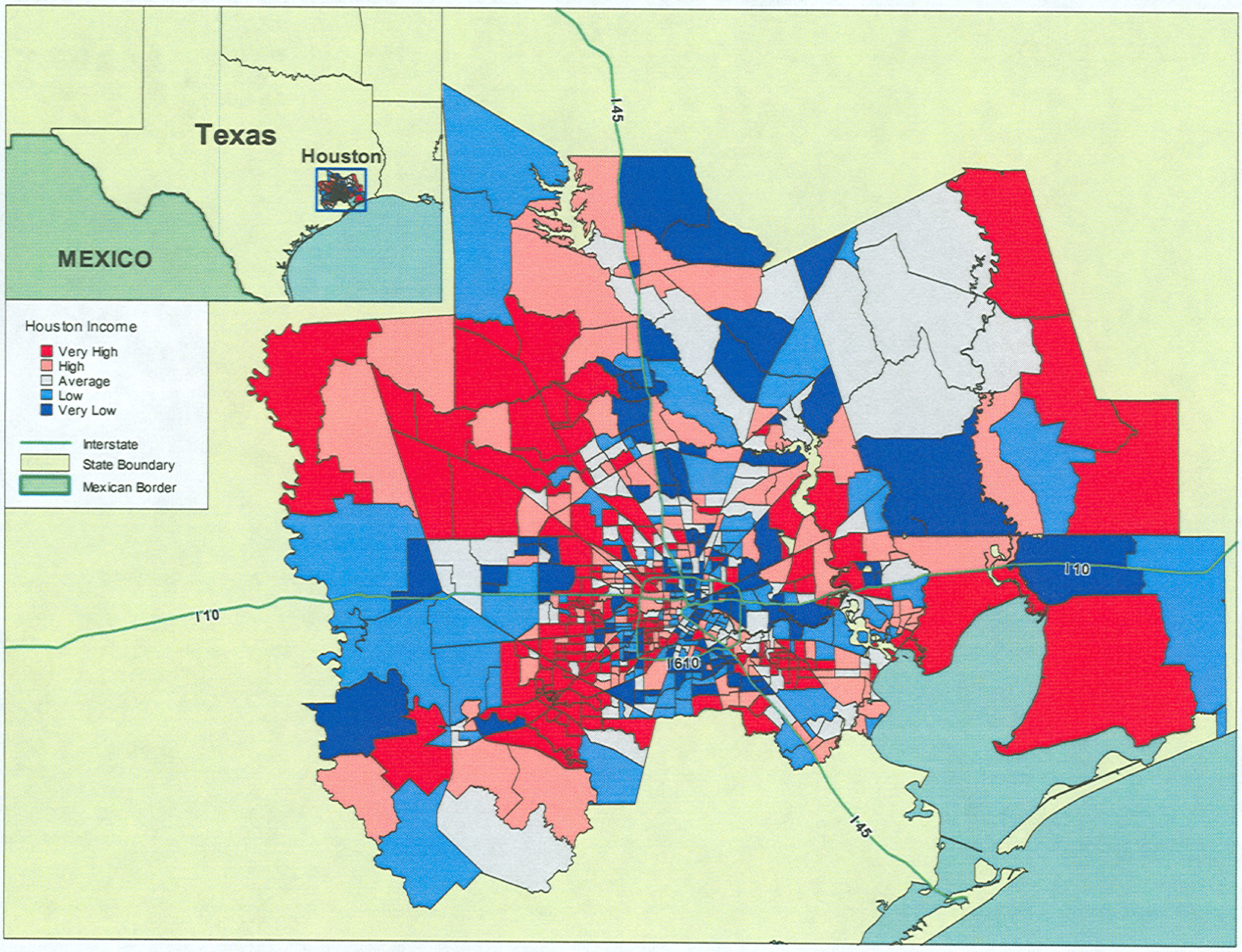
Software	Selected spatial autocorrelation capabilities
SAS	Code in Griffith and Layne (1999)
S+	Spatial stats module
SPSS	Code in Griffith and Layne (1999)
MINITAB	Macros by Griffith
SpaceStat	Autoregression
Idrisi	Moran Coefficient
ArcView	Scripts by Zhang and Griffith (1997, 2000)
ArcInfo	Moran Coefficient, Geary Ratio

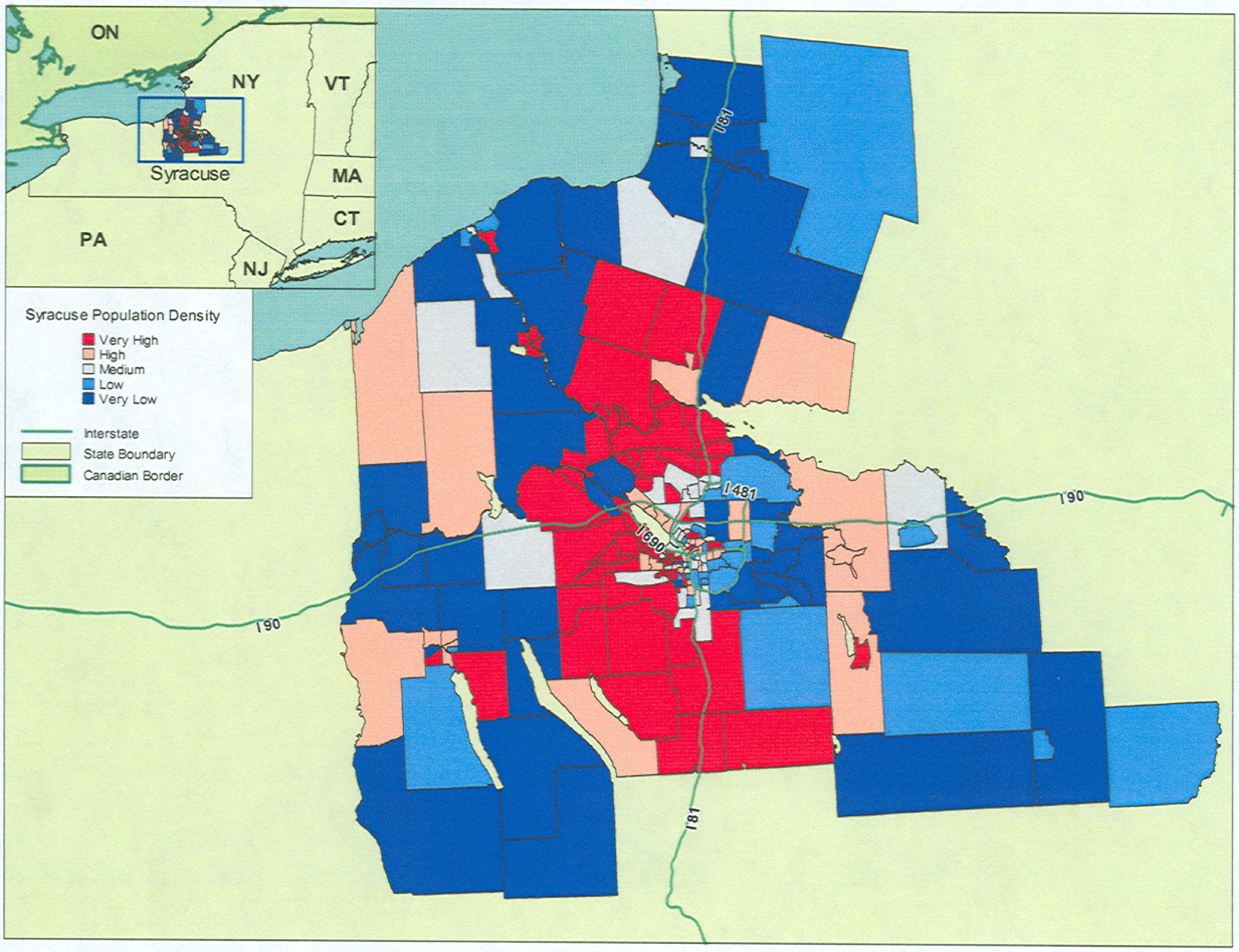
The adjusted results are achieved by using spatial autoregressive techniques to convert the total information contained in a dataset into n smaller, independent pieces from which spatial autocorrelation effects have been removed. The increase in statistical modeling complexity is required to have a sound statistical basis for making decisions.

Failure to take into account the positive spatial autocorrelation that characterizes almost all data can result in wrong inferences and decisions. A further motivation for using spatial autocorrelation is furnished by the increased percentage of variance explained for the dependent variable of a predictive model in business geographics--exploiting spatial autocorrelation typically increases the R-squared value by about 5 percent (Griffith and Layne, 1999). Obtaining 5 percent additional explanatory power in this way is much easier and more available than from collecting and cleaning additional data or using different statistical methods.

References

Griffith, D., and Layne, L. 1999. *A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Datasets*, Oxford University Press, New York.


Texas
Houston
MEXICO
Houston Income
Very High
High
Average
Low
Very Low
Interstate
State Boundary
Mexican Border
I 45
I 10
I 10
I 610
I 45


ON
NY
VT
MA
CT
PA
NJ
Syracuse
Syracuse Population Density
Very High
High
Medium
Low
Very Low
Interstate
State Boundary
Canadian Border
I 81
I 481
I 690
I 90
I 90
I 81